UNIVERSITY OF
LIVERPOOL

Functional and Comparative Genomics Of *Enterococcus faecium* Isolated From Animals

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by Ashwag Shami

Date: 18-12-2014

General Abstract

Enterococci are Gram-positive bacteria that inhabit the gastrointestinal tract of humans and animals as commensal flora. In recent years two species, *Enterococcus faecalis* and *Enterococcus faecium*, have become an increasing medical concern by virtue of their ability to gain and spread antibiotic resistance.

In this study, genomes of vancomycin-resistant isolates of *E. faecium* from pig, chicken and calf were sequenced using 454 and PacBio platforms. The assembled genomes were annotated and compared with human *E. faecium* isolates to identify their repertoire of genes potentially associated with colonising each host. Phylogenomics of *E. faecium* was used to investigate the relationship between animal and human strains. The genomes of the chicken, pig and calf isolates differed in size (2.5 Mb to 3.3 Mb) with the size difference due to horizontally-acquired elements (mostly phage, transposons and insertion sequences); the chicken isolate genome contained five prophages.

A mega-plasmid present in each of the sequenced *E. faecium* was revealed to be integrated into the genome of the chicken isolate. Comparison of the three genomes identified putative niche adaptation genes with a variety of proposed functions, particularly carbohydrate utilisation. Possible factors that explain *E. faecium* sub-populations, including clinical, commensal and animal isolate clades were examined. Use of the PhenoLink relationship tool to examine the *E. faecium* sub-populations identified that putative niche specific genes include carbohydrate utilisation genes and mobile genetic elements.

Temperate bacteriophages are known to be important drivers of genome plasticity in *E. faecium* species. The diversity of prophages and their relationship between was investigated after locating 56 prophage elements containing integrase and lysin genes encoded in the 139 publicly available *E. faecium* genomes. Comparative analysis of these prophages identified eight sequence types, which differed in size and gene content. The prophage genomes comprised between 17 to 72 ORFs and their size ranged from 13.9 to 55.1 kb with 35% to 37.9% average G+C content. Based on alignment analyses of the major functional proteins encoded in the prophage genomes (integrase, terminase large subunit, tail protein and holin) each was assigned a sequence type. All of the prophage integrases were identified to be tyrosine (XerC) recombinases and many of their respective *attP/attR* sequences were identified. The mosaic nature of *E. faecium* prophage genome sequence types supports previous hypotheses that extensive genetic recombination drives chimeric phage types.

Acknowledgements

Working on this Ph.D. has been a wonderful and often great experience. It is hard to say whether it has been grappling with the topic itself which has been the real educational experience, or grappling with how to give talks, working in a group and staying focused ... In any case, I am thankful to many people for making my Ph.D.an unforgettable experience.

First and foremost, praises and thanks to the God, for giving me blessings throughout my Ph.D. to complete the research successfully.

I wish to acknowledge the Royal Saudi Government and Ministry of Higher Education for providing me with Masters and Ph.D. scholarships to study in the UK.

I would like to express my deep and honest appreciation to my research supervisors Dr. Malcolm Horsburgh and Dr. Alistair Darby who have been a wealth of knowledge and support throughout my study. Their understanding, interest and kindness were infinite. As supervisors of a student originally from a non-English speaking country, they consumed extensive time and effort in improving my English communication skills, which will indeed have great influence on my future research career. Sincerely my thanks to Dr. Malcolm, it was a great pleasure and honour to work and study under his supervision. I am extremely grateful for what he has offered me. I would also like to thank him for his friendship, empathy, and great sense of humour. I am extending my heartfelt thanks to his wife and family for their acceptance and patience during thesis preparation.

I would particularly like to thank Paul Loughnane and Pisut Pongchaikul whose excellent expertise, patience and kindness helped me to get through all the difficult stages of my project in the lab and in bioinformatics.

I owe particular thanks to my husband Moiad Smeer Almadani. He always supports me and is a source of pride for me over many years. His faith, sustenance, understanding and companionship were the sources of my strength to pursue this dream.

I am extremely grateful to my parents for their love, prayers, caring and sacrifices for educating and preparing me for my future. I am very much thankful to my daughters Maiar and Yara for their love and patience. Also I express my thanks to my sisters-in-law Suha Khan and Shahd Nassar and my brothers Abdul Aziz and Ammar for their support and valuable prayers.

They always stood beside me and they supported me to get rid of everything that annoys me.

Finally, I extend my gratitude to my friends who gave me all the emotional support throughout my years of study who helped me to open up my heart in ways I wouldn’t have thought were possible.

Declaration

The work in this dissertation was carried out in accordance with the Regulations of the University of Liverpool. This work is my own original research, except where acknowledged in the text. No part of this thesis has been submitted for any other degree. The dissertation has not been submitted to any other University.

Signed

Ashwag Shami

Date

18-12-2014

Table of Contents

List of Tables

List of Figures

List of Abbreviations

(v/v)	Volume/volume
(w/v)	Weight/Volume Concentration
Att	Prophage attachment site
Bp	Base pairs
CAT	Chloramphenicol acetyl transferase
CC17	Clonal complex-17
COG	Clusters of Orthologous Group
Esp	Enterococcal surface proteins
GC	Guanosine-cytosine content
GEIs	Genomic islands
GRE	Glycopeptide-resistant enterococci
HGT	Horizontal gene transfer
IS	Insertion sequences
Kb	kilobase pairs, (thousand of base pairs)
LB	Luria Bertani Broth
LCBs	Locally collinear blocks
Mb	Megabase pairs, (millions of base pairs)
MGEs	Mobile genetic elements
MLSB	Macrolides, Lincosamides and Streptogramin B
MLST	Locus Sequence Typing
MSCRAMM	Microbial surface components identifying adhesive matrix molecules
NCBI	National Center for Biotechnology Information
OD	Optical density
ORF	Open reading frame
PacBio	The Pacific Biosciences
PAI	Pathogenicity-associated island
PBPs	Penicillin binding proteins
PCR	The polymerase chain reaction
PFGE	Pulsed Field Gel Electrophoresis
PFU	Phage counting Plaque forming unit
PTS	Phosphotransferase system
PYR	L-pyrrolidonyl-B-naphthylamide
RT	Room temperature
SSRs	Short sequence repeats
ST	Sequence types
STRs	Short tandemly repeated sequence
THB	Todd Hewitt Broth
Tn	Transposon
UV	Ultraviolet
Van A	Vancomycin resistance type A
Van B	Vancomycin resistance type B
VRE	Vancomycin-resistant *Enterococcus faecium*
β	Beta

Chapter One :Introduction.

1.1 History of the Genus *Enterococcus*

Enterococcus was historically termed as a taxonomically diverse genus identified as being 'faecal streptococci', associated with the gastrointestinal tract of human (Giraffa 2002). Thiercelin in 1899 first coined the term 'enterocoque' to describe a newly found Gram-positive diplococcus species. Andrews and Horder in 1906, isolated the same organism from an endocarditis patient and named it *'Streptococcus faecalis'* (Murray 1990).

Based on antigens identified as being group-specific, enterococci were placed in *Streptococcus* group D, while pyogenic streptococci belong to groups A, B, C, E, F or G using antisera. Enterococci were thus classified as group D streptococci because of their morphology and Lancefield antigenicity. The antigenicity of the carbohydrate moiety of the cell wall is distinguished according to a system devised by Lancefield in the 1930s (Smith, Niven *et al.* 1938). The established lancefield antigen of *Streptococcus* is a virulence determinant. For example, in group A streptococci it plays a significant role in resistance to platelet-derived antimicrobials in serum, neutrophil killing and the cathelicidin antimicrobial peptide LL-37 (van Sorge, Cole *et al.* 2014).

Many efforts were made to classify these organisms into better taxonomic groups due to their great diversity. A new classification pattern was proposed by Sherman in 1937 that classified streptococci into four main groups namely pyogenic, viridans, lactic streptococci and enterococci. In 1984 research carried out using nucleic acid hybridization revealed the latter

group showed only weak association to streptococci (Sherman, Mauer *et al.* 1937). Subsequently based on nucleic acid techniques, DNA hybridisation, DNA: rRNA hybridisation and 16S rRNA sequencing revealed that *S. faecalis* and *S. faecium* were only distantly related to other streptococci. The new genus named *Enterococcus* was proposed and *S. faecalis* and *S. faecium* were removed from the genus *Streptococcus* and renamed as *Enterococcus faecalis* and *Enterococcus faecium*, respectively (Schleifer, Kilpper-Balz *et al.* 1984, Ludwig, Seewaldt *et al.* 1985). The classification of enterococci has always been challenging because it is a heterogeneous group of Gram-positive cocci which is more closely related to the genera *Carnobacterium*, *Lactococcus* and *Vagococcus*, yet has many characteristics of the genus *Streptococcus* (Leclerc, Devriese *et al.* 1996).

The genus of *Enterococcus* is composed of more than forty species (The National Center for Biotechnology Information, NCBI), classified on the basis of pigment production, motility and ability to generate acids from a range of carbohydrates (Fischetti, Novick *et al*. 2006)**.** Based on the chemotaxonomic and phylogenetic studies, the establishment of 16S rRNA sequences led to the description of seven clonal complexes within the genus namely (i) *E. faecalis*, *E. haemoperoxidus* and *E. moraviensis*; (ii) *E. faecium*, *E. durans*, *E. hirae*, *E. mundtii*, *E. pocinus*, and *E. villorum*; (iii) *E. avium*, *E. pseudoavium*, *E. malodoratus*, and *E. raffinosus*; (iv) *E. casseliflavus*, *E. gallinarum* and *E. flavescens*; (v) *E. cecorum* and *E. columbae*; (vi) *E. dispar* and *E. asini*; (vii) *E. saccharolyticus* and E. *sulfureus.* Other species are *E. gilvus*, *E. pallens* and *E. ratti* (Klein 2003).

While there are multiple species in the genus *Enterococcus*, two are associated with the majority of human infections, *E. faecalis* and *E. faecium* (Magi, Capretti *et al.* 2003).

1.2 General Characteristics

Species of the genus *Enterococcus* are facultative anaerobic cocci which grow as short to medium length chains or as pairs in liquid culture. They are catalase negative and have a fermentative metabolism (Hollenbeck and Rice 2012). The optimum growth temperature of enterococci is 37 ºC although they are capable of growing over a temperature range of 10 to 45 ºC. They have an ability to survive at 60 ºC for 30 minutes, survive at a high pH, hydrolyse bile-esculin and L-pyrrolidonyl-B-naphthylamide (PYR) and grow in the presence of 6.5% sodium chloride (Hollenbeck and Rice 2012). Since *Entercoccus* species are resistant to harsh environmental conditions they are sensitive indicators of faecal contamination (Franz, Stiles *et al.* 2003).

1.3 Habitat and Distribution

Enterococci are generally considered to be commensal flora in the gastrointestinal tract of humans and warm-blooded animals (Kuhn, Iversen *et al.* 2005, Santagati, Campanile *et al.* 2012). However, they are not restricted to these niches and enterococci are resilient species of insects and reptiles. They can be isolated from many plants and it has been proposed that enterococci are spread between plants by insects (Mundt 1961).

Different species of enterococci exhibit some host specificity. Most frequently, *E. faecalis* and *E. faecium* are found in humans and farmed livestock. *E. faecium* is the predominant species isolated from chicken and pig. *E. durans* is found both in humans and poultry. *E. avium* and *E. gallinarum* are restricted to poultry (Nowlan and Deibel 1967), *E. columbus* is specific to pigeons (Devriese, Ceyssens *et al.* 1990) and *E. asini* is specific to donkeys (de Vaux, Laguerre *et al.* 1998). The distribution of enterococcal species varies across age groups. *E. faecalis* is principally present in the intestinal microflora of young poultry, while, *E. faecium* and *E. caecorum* dominate in chickens around 12 weeks (Devriese, Hommez *et al.* 1991).

1.4 *Enterococcus* as a commensal

Commensalism is the relationship between two organisms in which one or both organisms gets benefits and the other organism is not harmed. In the colon of nearly all humans and most animals enterococci are minor residents, present at ~10^8 colony forming units per g of faeces (Gilmore 2002). Enterococci have effectively evolved various genetic traits which helps maintain their stable colonisation. Commensal isolates of *E. faecium* and *E. faecalis* are genetically distinct compared to infection isolates. The differences may be unclear, however, since immunocompromised patients are more susceptible to infection even with commensal strains (Jett, Huycke *et al.* 1994, Huycke, Sahm *et al.* 1998).

1.5 Enterococcal infections

Over recent decades enterococci have been identified as an important opportunist pathogen causing nosocomial infections such as bacteremia, infective endocarditis, urinary tract infections, intra-abdominal, pelvic and soft tissue infections as well as surgical wound infections. The identification of different species of enterococci causing these infections provided information for epidemiological surveillance (Huycke, Sahm *et al.* 1998, Lester, Sandvang *et al.* 2008). Fisher *et al* (2009) demonstrated that the majority of *Enterococcus* infection can be considered endogenous, by translocation of the bacteria within epithelial cells of the intestine, which later cause infection through lymph nodes and consequently extend to other cells inside the body.

1.6.1 Pathogenesis of enterococcal disease and virulence factors

To cause disease enterococci must colonise host tissues, defend against host immune mechanisms and express factors that enable persistence. Multiple factors are known that regulate the virulence of *Enterococcus* species, for example ability to colonise the gastrointestinal tract, ability to adhere to a variety of extracellular matrix components, including vitronectin, thrombospondin and lactoferrin, and ability to adhere to oral cavity epithelia, urinary tract epithelia and human embryo kidney cells (Fisher and Phillips 2009). Pathogenicity of enterococci has been related to several key virulence traits associated with adhesion, translocation and immune evasion (Johnson 1994).

1.5.2 Adhesins

The first important step for the bacteria in infection is to adhere to the host tissues. The most significant adhesion factors are extracellular surface protein (Esp), aggregation substance (Asa), *Enterococcus faecalis* antigen A (EfaA), and endocarditis and biofilm-associated pili (Ebp) (Fisher and Phillips 2009). Surface proteins called adhesins play a crucial part in binding to their eukaryotic receptors on the surface of epithelial cells, endothelial cells, leukocytes and the extracellular matrix. Adhesins also have many different roles in enhancing phagocytosis, acting as toxins and initiating or decreasing host inflammatory responses (Jett, Huycke *et al.* 1994).

1.5.2.1 Enterococcal surface proteins (Esp)

Extracellular surface protein (Esp) was described in *Enterococcus* species by Shankar *et al* (1999). These proteins were first identified in *E. faecalis* and are highly conserved in *E.faecium* sub-populations (Willems, Homan *et al.* 2001). *Esp* encodes a cell-wall-associated protein frequently associated with clinical isolates. This protein has a significant role in promoting adhesion, colonisation, immune avoidance, and has a role in antibiotic resistance (Foulquie Moreno, Sarantinopoulos *et al.* 2006).

Esp is associated with enterococcal biofilm formation, which might lead to adhesion to eukaryotic cells, such as those of the urinary tract, and increases resistance to environmental stresses (Borgmann, Niklas *et al.* 2004). Comparison of the incidence of virulence and antibiotic resistance between

E. faecium strains of dairy, animal and clinical origin was performed by Mannu *et al* (2003) and they suggested that the *esp* gene may correlate with pathogenicity, since *esp* was absent in dairy isolates, comparing with 21 of 28 clinical strains that had the gene. Conjugation rates and resistance to ampicillin, ciprofloxacin and imipenem were also higher in *E. faecium* strains with *esp* than strains without it.

1.5.2.2 Aggregation Substances Agg

Agg is a pheromone-inducible surface glycoprotein that facilitates aggregate formation through conjugation, enhances adhesion to a range of eukaryotic surfaces and plasmid to transfer (Koch, Hufnagel *et al.* 2004). The existence of Agg raises the hydrophobicity of the enterococcal cell surface promoting localisation of cholesterol to phagosomes and many interrupt or inhibit fusion with lysosomal vesicles (Eaton and Gasson 2002).

Pulsed-field gel electrophoresis analysis performed by Billstrom *et al* (2008) indicated that the gene encoding Agg exists in clinical isolates of *E. faecalis* but not *E. faecium*. Adhesion to collagen of *E. faecalis* (Ace) or *E. faecium* (Acm) is another cell-surface protein belonging to the microbial surface components identifying adhesive matrix molecules (MSCRAMM) family (Fisher and Phillips 2009).

Sex pheromones were recognised in *E. faecalis* by identifying a clumping reaction that occurs through conjugative transfer of plasmids (Wirth 1994). The pheromones are chromosomally encoded small peptides composed of

seven to eight amino acids encouraging a mating response in cells with corresponding conjugative plasmids. Sex pheromones trigger chemoattraction of neutrophils causing granule enzyme secretion and respiratory burst (Ember and Hugli 1989).

1.5.3 Biofilm

Singh *et al* (2007) suggested that the capability of enterococci to generate biofilms is essential in producing endodontic, endocarditis and urinary tract infections. The formation of pili is required for biofilm formation. The endocarditis- and biofilm-associated pili gene cluster (ebp) contributes to the production of biofilm in enterococci. The *ebp* operon contains *ebpA*, *ebpB ebpC* and encoding pilus subunits *srtC* encoding sortase C that catalyses their covalent attachment to peptidoglycan and are found on the surface of *E. faecalis* and *E. faecium* (Nallapareddy, Singh *et al.* 2006, Sillanpaa, Prakash *et al.* 2009). Enterococcal pili are heterotrimeric and the pilus shaft contains two minor pilins

1.5.4 Secreted virulence factors

1.5.4.1 Cytolysin

Cytolysin (also called haemolysin) is a bacterial toxin that has β-haemolytic properties and is bactericidal against other Gram-positive bacteria (Koch, Hufnagel *et al.* 2004, Billstrom, Lund *et al.* 2008). Cytolysin was found in several *E. faecalis* and *E. faecium* isolates and its haemolytic and bactericidal activity has higher occurrence in clinical isolates compared to

food isolates. It is regulated by a quorum-sensing mechanism via a two-component system (Fisher and Phillips 2009). Clewell (1990) indicated that cytolysins are generally encoded by highly conserved conjugative plasmids like pAD1, although they can be encoded chromosomally.

1.5.5 Hydrolytic enzymes

1.5.5 .1 Gelatinase and serine protease

The fundamental role of both gelatinase and serine protease in enterococcal pathogenesis is assumed to be in generating nutrients for the bacteria by degrading host tissue; these proteases also have functions in biofilm formation (Mohamed and Huang 2007). Gelatinase (GelE) is an extracellular zinc metallo-endopeptidase that is able to hydrolyse haemoglobin, gelatin and casein, and other bioactive peptides. The gene (*gelE*) is chromosomally located and is expressed in a cell-density dependent manner. The gene *sprE* is located directly downstream it is co-transcribed with *gelE* and encodes a serine protease. Gelatinase is secreted by *E. faecalis* strains (Koch, Hufnagel *et al.* 2004, Fisher and Phillips 2009).

1.5.5.2 Hyaluronidase

Hyaluronidase is a cell surface-associated enzyme. In *Enterococcus*, hyaluronidase may act as a virulence factor by hydrolysis of hyaluronic acid and is associated with tissue damage (Jett, Huycke *et al.* 1994). The mucopolysaccharide moiety of connective tissue is effectively depolymerised by hyaluronidase enabling the spread of enterococci as well

as their toxins across host tissue (Kayaoglu and Orstavik 2004). The gene encoding hyaluronidase (*hyl*) is located on the chromosome in both *E. faecalis* and *E. faecium* (Vankerckhoven, Van Autgaerden *et al.* 2004).

1.5.6 Lipoteichoic acid

Membrane-associated lipoteichoic acids are amphipathic polymers comprised of a hydrophilic polyglycerolphosphate backbone connected through an ester bond to a hydrophobic glycolipid tail. Lipoteichoic acids are common among prokaryotic organisms. For enterococci these surface molecules have been shown to be identical to the group D antigen (Wicken, Elliott *et al.* 1963, Jett, Huycke *et al.* 1994, Ginsburg 2002). Surface molecules like D-alanine lipoteichoic acid (LTA) present several roles in Gram-positive bacteria, for example modulation of autolysin and cation homeostasis. Alanine esters of enterococcal lipoteichoic acid play a significant role in biofilm formation and resistance to antimicrobial peptides (Fabretti, Theilacker *et al.* 2006).

1.6 Enterococcal epidemiology

Studies of ecology and epidemiology of *Enterococcus* have stated *E. faecium* and *E. faecalis* are commonly isolated from sausages, cheese, minced beef, fish and pork. Foods that originate from animals are often connected with infectivity by *Enterococcus* species, as they are capable of surviving in the heating process. Mainly it is the contamination with *Enterococcus* species that is the reason for the association of these species

with foods from animal origin (Klein 2003, Foulquie Moreno, Sarantinopoulos *et al.* 2006).

Kuhn *et al* (2003) indicated that the prevalence of *Enterococcus* species differs across Europe. *E. faecalis* and *E. faecium* are the most commonly isolated species from environmental and clinical sources in UK and Spain. *E. faecium* has lower a incidence in Sweden with *E. hirae* having a higher isolation rate, while *E. hirae* in Denmark is the most common species and is isolated mostly from slaughtered animals.

Resistance to glycopeptide antimicrobials, teicoplanin and vancomycin was first reported in 1986 in Europe, followed by related reports in the USA in 1987 and in Singapore in 1994 (Leclercq, Dutka-Malen *et al.* 1992, Chlebicki and Kurup 2008). From 1989 to 1993, the proportion of vancomycin resistant isolates in the USA increased from 0.3 % to 7.9 % (Centers for Disease and Prevention 1993). About 28 % of enterococcal isolates were resistant to vancomycin in 2003 and the incidence of this resistance has been rising steadily over the years (National Nosocomial Infections Surveillance 2004).

In the USA, vancomycin-resistant *Enterococcus faecium* (VRE) established mostly in patients exposed to healthcare settings and studies showed no link between VRE and farm animals in 1990. In Europe and Asia the situation was different because of the use of avoparcin glycopeptide as a growth promoter in animal husbandry which consequently directed a high rate of

VRE colonisation in animals. The VRE transfer to human subsequently occurred by direct contact with animals or by eating contaminated products (Leclercq, Dutka-Malen *et al.* 1992, Chlebicki and Kurup 2008).

The epidemiology of VRE distribution has been diverse in US with VRE endemic in hospitals for many years or linked to foreign travel and consumption of imported food, but colonisation absent in healthy people. In contrast, in Europe outbreaks of VRE seldom arise in hospitals but have been isolated from healthy individuals and farm livestock and food (Coque, Tomayko *et al.* 1996, Wegener, Madsen *et al.* 1997, Bonten, Willems *et al.* 2001).

Pulsed Field Gel Electrophoresis (PFGE) studies discovered similar PFGE-patterns in humans and animal isolates, not only from the same geographic region but also from very distinct epidemiological environments (Stobberingh, van den Bogaard *et al.* 1999, Hammerum, Fussing *et al.* 2000, van den Bogaard, Willems *et al.* 2002). Amplified Fragment Length Polymorphism (AFLP) study of VRE populations performed by Willems *et al.* (2000) reported 11% of the human clinical isolates associated to clusters also present in poultry and pig and also found specificity in host colonisation. Further analyses of gene clusters responsible for vancomycin resistance and Tn*1546* in *E. faecium*, reported that humans and animal isolates have identical Tn*1546* types, suggesting that horizontal gene transfer occurs between human and animal *E. faecium* (Stobberingh, van den Bogaard *et al.* 1999, van den Bogaard, Willems *et al.* 2002).

Using antimicrobials as growth promoters is an efficient approach of enhancing productivity and animal health in livestock production. Avoparcin, which is a glycopeptide, produces cross-resistance to vancomycin, is an example of a growth promoter that has been used in agricultural systems in Europe but not USA, particularly in the pig and poultry industries (van den Bogaard and Stobberingh 1999). Avoparcin has been proposed as a significant effect in the emergence and spread of resistance to vancomycin in enterococcal populations (Bager, Aarestrup *et al.* 1999) For this purpose, the use of avoparcin was excluded in Denmark in 1995 followed by the rest of the EU in 1997. In addition, Virginomycin is used as an additive to animal food in agriculture industry and the overuse of virginamycin could have led to the acquired resistance to steptogramins. The use of virginamycin was excluded in Denmark in 1998 and through the EU in 1999 (Aarestrup 2000).

VRE have been associated globally with hospital outbreaks and the vancomycin resistance gene (*vanA*) has transferred to methicillin- resistant *Staphylococcus aureus*. Evolutionary genetics, population structure, and geographic distribution of VRE isolated from nonhuman and human sources and community and hospital reservoirs recognised a genetic lineage of *E. faecium* (clonal complex-17) that has spread worldwide. The CC17 lineage is associated with ampicillin resistance, a pathogenicity island, and is linked with hospital outbreaks. CC17 is a model of accumulative evolutionary developments that enhanced the relative fitness of bacteria in hospital environments (Willems, Top *et al.* 2005).

1.7 Antimicrobial Resistance

Enterococcus *faecalis* and *E. faecium* have succeeded as nosocomical pathogens because of their ability to gain and spread antibiotic resistance to commonly used antibiotics (Leclercq 1997). Two types of antimicrobial resistance are associated with *Enterococcus* species, namely intrinsic and acquired resistance. Intrinsic resistance is chromosomally encoded within the core genome of all members of the species and occurs naturally, whereas horizontal transfer of genetic material or sporadic mutations account for acquired resistance (Hollenbeck and Rice 2012, Gilmore MS, Clewell DB *et al.* 2014).

1.7.1 Intrinsic resistance

Enterococcus species are naturally resistant to the most commonly used antimicrobial classes, for example β-lactams and aminoglycosides, which are typically affective for the treatment of Gram-positive infections. Low-level intrinsic resistance is found in *Enterococcus* species in respect of resistance to cephalosporins, trimethoprim-sulfamethoxazole and lincosamide (Leclercq, Dutka-Malen *et al.* 1992). In addition, *Enterococcus* species are frequently resistant to tetracycline, rifampicin, quinolones, macrolides, chloramphenicol and fosfomycin, and these antibiotics are rarely used to treat enterococcal infections (Hollenbeck and Rice 2012).

The typical treatment for enterococcal infections is a bactericidal and synergistic mixture of a cell wall synthesis inhibitor such as a β-lactam antibiotic (benzylpenicillin or ampicillin) or glycopeptide, with an

aminoglycoside (streptomycin or gentamicin). However, the efficacy of this combination has been compromised by the emergence of enterococci with high-level aminoglycoside resistance (Leclercq, Dutka-Malen *et al.* 1992).

1.7.1.1 β-lactams

The β-lactam ring is part of the main structure of numerous antibiotic families for example, cephalosporins, penicillins, carbapenems, and monobactams. Nearly all of these antibiotics work by inhibiting bacterial cell wall biosynthesis and they are extremely efficient versus Gram-positive and Gram-negative bacteria (Thomson and Bonomo 2005).

The reason for the intrinsic resistance to β-lactam agents in *Enterococcus* is low affinity of penicillin binding proteins (PBPs) for β-lactams. β-lactams bind to the PBPs, enzymes associated with the cross linking of pentapeptide molecules in the peptidoglgcan layer of the bacterial cell wall. The association of a β-lactam with PBPs disrupts the growth of the bacteria by weakening the cell wall and resulting in programmed cell death (Klare, Badstubner *et al.* 1999, Hollenbeck and Rice 2012).

1.7.1.2 Aminoglycoside

Low-level intrinsic resistance to aminoglycosides is exhibited by all enterococci including to gentamicin, which is the most common aminoglycoside used with enterococcal infections. The aminoglycosides target enterococci by binding to the 16S rRNA of the 30S ribosomal subunit and thereby inhibit protein synthesis. Enterococci that possess the gene

aac(6')-Ie-aph(2')-Ia are resistant to almost all aminoglycosides including gentamycin, amikacin, tobramycin, kanamycin and netilmycin, but remain sensitive to streptomycin (Chow 2000). The enzyme encoded by *aac(6')-Ie-aph(2')-Ia* modifies the antibiotic by phosphorylating and simultaneously acetylating it at two different positions. This impairs the binding of the aminoglycoside to the 30S ribosomal subunit (Leclercq 1997, Chow 2000, Hollenbeck and Rice 2012).

1.7.1.3 Streptogramins

Streptogramins target ribosomes at the level of inhibition of translation through binding to the bacterial ribosome and interfere with protein synthesis. Resistance to streptogramins appears via a number of mechanisms involving target modification, efflux, and enzyme catalyzed antibiotic modification (Johnston, Mukhtar *et al.* 2002).

Streptogramins show bactericidal activity when they act synergistically as two components, streptogramin A and B. these components alone show a weak bacteriostatic activity whereas the combination can act bactericidally (Kehoe, Snidwongse *et al.* 2003). Simjee *et al* (2002) stated that resistance to both A and B components of streptogramin was found in enterococcal species, including *E. faecium*, *E. gallinarum* and *E. hirae*.

1.7.1.4 Glycopeptides

Glycopeptides are rigid, large molecules that inhibit cell wall peptidoglycan synthesis in a late stage of bacterial growth (Reynolds 1989). Glycopeptides

disrupt enterococci by interfering with cell wall synthesis by attaching to the terminal acyl-D-alanyl-D-alanine (D-Ala-D-Ala) of the precursors used in peptidoglycan synthesis (see 1.5.9.2.5). Several motile species of enterococci such as *E. flavescens*, *E. gallinarum* and *E. casseliflavus* express low levels of intrinsic resistance to glycopeptides (Gholizadeh and Courvalin 2000).

1.7.2 Acquired resistance

Enterococci acquire resistance to many antibiotics via the acquisition of genetic material or via sporadic mutations. Horizontal gene transfer among enterococci appears most frequently by the movement of transposons and the transfer of pheromone-sensitive broad host range plasmids (Hollenbeck and Rice 2012) and to an unknown extent by phage transduction (Yasmin, Kenny *et al.* 2010).

1.7.2.1 β-lactams

Enterococci can express high-level resistance to β-lactams or other penicillin drugs when there is overproduction of low affinity penicillin binding proteins (PBPs). Resistance also occurs through acquisition of β-lactamases or mutations in PBP4/5 targets, which results in poor, or no binding to these targets (Fontana, Ligozzi *et al.* 1996). Synthesis of β-lactamase of high levels may result in resistance to β-lactam antibiotics. This secreted enzyme is overproduced when the operon repressor protein is absent and this occurs most often in *E. faecalis*, rather than *E faecium* strains (Murray 1992). Recently, over 80% of clinical *E. faecium* isolated

from all over the world are ampicillin (Zhang, Paganelli *et al.* 2012).

Plasmid-mediated genes encoding β-lactamases (*bla)* were first defined in *E. faecalis* in 1983. Since then the production β-lactamase in enterococci has been rare and described mainly in *E. faecalis*. Genes encoding β-lactamases in *Enterococcus* and *S. aureus* are identical and are frequently encoded by staphylococcal β-lactamase transposon Tn*552*. Hollenbeck *et al* (2012) suggested that high-level penicillin resistance in *E. faecium* is generally related to accumulation of point mutations in the penicillin-binding region of PBP5. While these point mutations are expected to originate *de novo* in single bacteria due to selective pressure from antibiotics, chromosome-to chromosome mobilisation of low affinity *pbp5* genes has been recognised *in vitro* and is expected to explain the distribution of high-level penicillin resistance in *E. faecium* (Rice, Carias *et al.* 2005).

1.7.2.2 Aminoglycosides

Most commonly, high-level resistance to aminoglycosides occurs due to the production of aminoglycoside modifying enzymes which are plasmid-mediated. These enzymes also nullify the synergistic killing effect of aminoglycoside in combination with cell wall-active agents (Chow 2000, Kotra, Haddad *et al.* 2000). Aminoglycoside resistance due to mutation of the ribosomal targ*et al*so occurs as does reduced antibiotic transport. These mechanisms are chromosome-mediated (Kotra, Haddad *et al.* 2000).

1.7.2.3 Macrolides, Lincosamides and Streptogramin B (MLS_B)

The term macrolide defines drugs with a macrocyclic lactone ring of 12 or more elements (Kanoh and Rubin 2010). Macrolide antibiotic include erythromycin, clarithromycin, and azithromycin (Alvarez-Elcoro and Enzler 1999).

Lincosamide antimicrobials including lincomycin and clindamycin act by inhibiting peptidyltransferase activity of the 50S ribosomal subunit. This ultimately interferes with protein synthesis. The gene *Inu(B) (linB)*, responsible for resistance to lincosamide, encodes the enzyme nucleotidyltransferase which adenylates a hydroxyl group on the lincosamide (Tenson, Lovmar *et al.* 2003).

Type B streptogramins and macrolides act on the 50S ribosomal subunit in a similar fashion and cause interference in the same binding site. Regularly, resistance to both classes of antibiotics occurs through a common mechanisms, for instance, resistance against macrolides, lincosamides and streptogramins B (MLS_B) via enzymatic methylation or mutation of adenine 2058 (Pernodet, Boccard *et al.* 1988, Vannuffel and Cocito 1996). The *ermB* gene borne on conjugative plasmids encodes for resistance to MLS_B by methylating the adenosine residue in 23S rRNA of the 50S ribosomal subunit (Jensen, Frimodt-Moller *et al.* 1999).

1.7.2.4 Streptogramin A

Three main mechanisms are involved in the acquired resistance for streptogramins (i) acetylation of the antibiotic, (ii) efflux of the antibiotic and (iii) dimethylation of the 23S rRNA target site. As of now 12 genes in *Enterococcus* species have been reported for streptogramin resistance (Hollenbeck and Rice 2012) .

1.7.2.5 Glycopeptide

For the past 25 years, the acquisition of glycopeptide resistance by enterococci has been an epidemiological and antimicrobial challenge. In 1988, glycopeptide-resistant enterococci (GRE) were first described. *E. faecium* is the species that exhibits greatest resistance to glycopeptides compared with *E. faecalis* and other *Enterococcus* species (Farrell, Mendes *et al.* 2011).

The cell wall of *Enterococcus* is composed mostly of peptidoglycan, with teichoic acid and polysaccharide. Teichoic acid is found only in Gram-positive bacteria and not in Gram-negative bacteria (Cheng, McCleskey *et al.* 1997). The carbohydrate moiety is cross-linked with peptide side chains in the peptidoglycan layers. The glycans and the peptides are connected through amide linkages, which link the carboxyl group of the muramyl residues and the terminal amino group of the peptides. D-Ala:D-Ala ligase and MurF enzymes catalyse the addition of D-alanyl-D-alanine to UDP-MurNAc pentapeptide precursor for peptidoglycan biosynthesis (Neuhaus and Struve 1965). Cell wall synthesis is inhibited by the antibiotic

vancomycin, which is a commonly prescribed glycopeptide. Unlike penicillins which bind to the enzyme, vancomycin binds to (acyl-D-alanyl-D-alanine) via 5 hydogen bonds. Transglycosylation and transpeptidation is thereby inhibited and the peptide precursors lose the ability to cross-link. Therefore cell wall integrity is lost and cell death occurs (Figure1.1) (Arthur, Molinas *et al.* 1993, Walsh, Fisher *et al.* 1996).

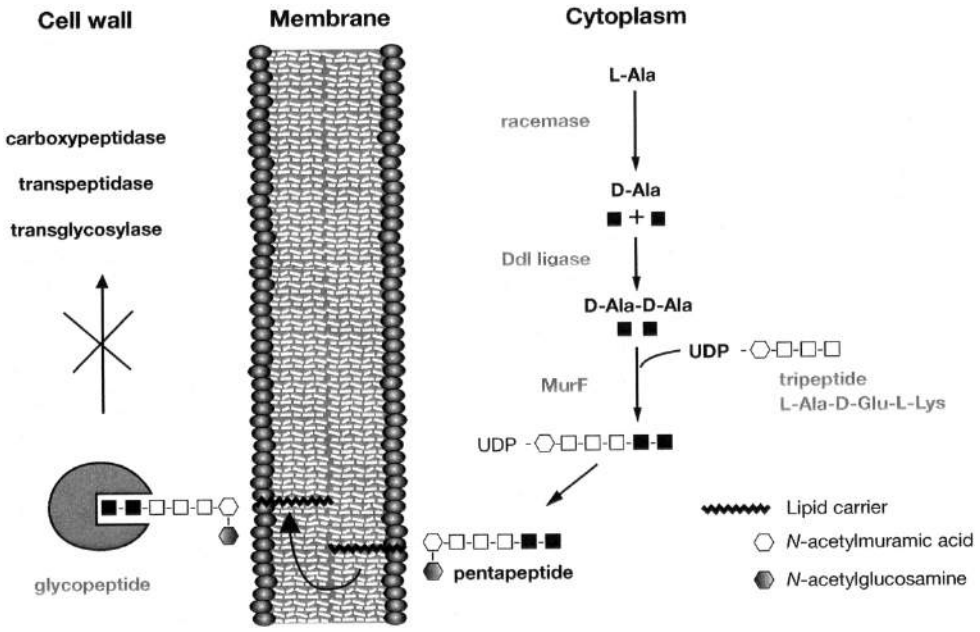


Figure1.1: Peptidoglycan biosynthesis and the mechanism of vancomycin. Association of the antibiotic to the C-terminal d-Ala–d-Ala of late peptidoglycan precursors stops catalysed reactions by transpeptidases, transglycosylases, and carboxypeptidases reproduced from Courvalin 2006.

1.7.2.5.1 Vancomycin resistance

The first vancomycin-resistant enterococci were reported in 1988 in Europe. Since then vancomycin resistance has spread rapidly. There are six different types of vancomycin resistance in *Enterococcus*. VanA, B, D, E, and G types relate to acquired resistance; VanC is an intrinsic resistance of *E. gallinarum*, *E. casseliflavus and E. flavescens* (Arthur, Reynolds *et al.* 1996).

The MIC of vancomycin and teicoplanin due to different gene types overlaps, consequently differentiation of glycopeptide resistance is presently established by sequencing of the structural genes for the resistance ligases (Courvalin 2006). VanA type isolates show high levels of inducible resistance to both vancomycin and teicoplanin, while VanB type isolates show flexible levels of inducible resistance to vancomycin only (Arthur, Reynolds *et al.* 1996). VanD type isolates are considered by constitutive resistance to sensible levels of vancomycin and teicoplanin (Depardieu, Reynolds *et al.* 2003). VanC, VanE and VanG type isolates are resistant to low levels of vancomycin however stay susceptible to teicoplanin (Reynolds and Courvalin 2005).

Though the six types of resistance include correlated enzymatic functions, they can be discriminated via the position of the corresponding genes and via the kind of regulation of gene expression. The *vanA* and *vanB* operons have been found on plasmids or in the chromosome, while the *vanD, vanC, vanE* and *vanG* operons have been found in the chromosome only

(Courvalin 2006). Willems *et al.* (1999) demonstrated that type A resistance is the most prevalent type in enterococci producing a high level of inducible resistance to vancomycin and teicoplanin.

Resistance to vancomycin emerged as a result of the presence of operons that encode enzymes for synthesis of low-affinity precursors. Mechanistically in which the C-terminal d-Ala residue is replaced by d-lactate (d-Lac) or d-serine (d-Ser), therefore adjusting the vancomyin-binding target; removal of the high-affinity precursors that are typically formed by the host consequently eliminates the vancomycin-binding target (Arthur, Reynolds *et al.* 1996).

The *vanA* operon responsible for the resistance phenotype is present on a mobile element, the non-conjugative Tn*1546* transposon (Figure 1.2), as part of a self-transferable plasmid. Tn*1546* can also be found integrated on the bacterial chromosome (Arthur, Molinas *et al.* 1993). VanR and VanS are involved with regulation and VanS recognises vancomycin whereby VanR controls induction of other Tn*1546*-encoded genes. The *vanH* encoded dehydrogenase produces D-lactate that is associated with D-alanine by the *vanA* encoded ligase. Van H, Van A and Van X are required for glycopeptide resistance by inhibiting vancomycin binding and restoring cell wall synthesis. Finally, the accessory proteins Van Y and Van Z are not necessary for resistance but are frequently colocalised (Courvalin 2006) (Figure 1.2).

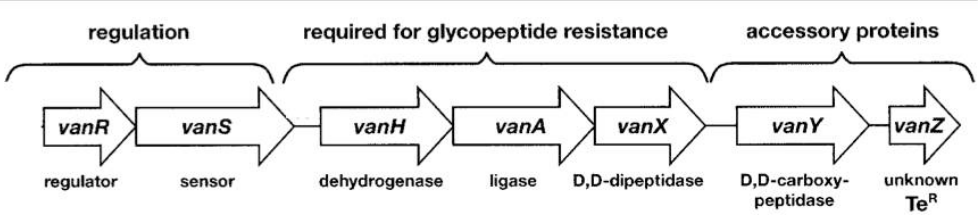


Figure 1.2: Organisation of VanA-type glycopeptide resistance operon. The arrows show regulatory and resistance and the accessory coding sequences reproduced from Courvalin 2006.

1.7.2.5.1.1 Target modification

The VanH dehydrogenase encoded by the transposon (Tn*1546)* converts pyruvate to d-Lac and the ligase (VanA) catalyses an ester bond between d-Ala and d-Lac. The subsequent d-Ala-d-Lac depsipeptide switches with the d-Ala-d-Ala dipeptide in peptidoglycan synthesis, a substitution that lowers affinity for glycopeptides significantly (Bugg, Wright *et al.* 1991, Arthur, Reynolds *et al.* 1996).

1.7.2.5.1.2 Removal of the susceptible target

Attachment of glycopeptides to peptidoglycan precursors that comprise d-Ala-d-Ala prevents peptidoglycan synthesis. The association between vancomycin and its target is inhibited via the elimination of the susceptible precursors that terminate in d-Ala. The VanX D,D-dipeptidase and The VanY D,D-carboxypeptidase enzymes elaborate this outcome (Figure 1.3); VanX enhances the host d-Ala:d-Ala ligase (Ddl) to hydrolyse the d-Ala-d-Ala dipeptide that is synthesised and when removal of d-Ala-d-Ala by

VanX is incomplete then VanY eliminates the C-terminal d-Ala residue of late peptidoglycan precursors.

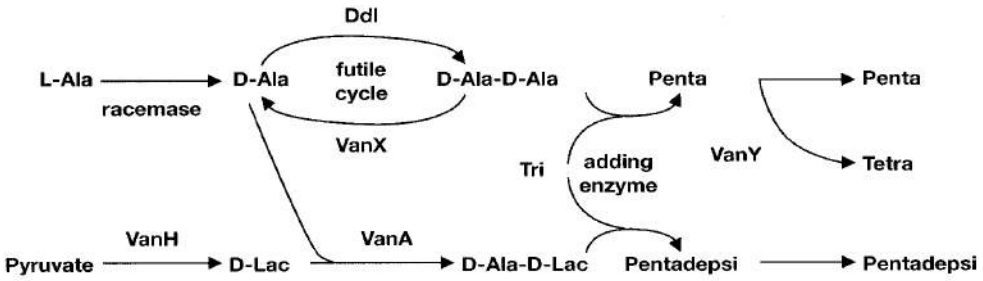


Figure 1.3: VanA-type glycopeptide resistance. Synthesis of peptidoglycan precursors in a VanA-type resistant strain reproduced from Courvalin 2006.

1.7.2.6 Chloramphenicol

Chloramphenicol inhibits protein synthesis by binding to a receptor site on the 50S subunit of the bacterial ribosome, inhibiting peptidyltransferase. Chloramphenicol acetyl transferase (CAT) is the enzyme responsible for chloamphenicol resistance in enterococcal species. The chloramphenicol resistance gene, *cat*, exists across the streptococci, staphylococci and enterococci from horizontal transfer of genetic material between these organisms. Resistance can be plasmid-encoded or present on the chromosome. Chloramphenicol resistance has been observed in *E. faecalis* and *E. faecium* isolates (Pepper, Le Bouguenec *et al.* 1986, Pepper, Horaud *et al.* 1987, Klare, Konstabel *et al.* 2003).

1.7.2.7 Tetracycline

Protein synthesis is inhibited via tetracycline antibiotics by a reduction of the affinity within regions of bacterial 30S rRNA for aminoacyl-tRNA. Resistance to tetracycline occurs through two main mechanisms. (i) The genes *tetK* and *tetL* encode transporters for active efflux of the antibiotic across the cell membrane. The *tetL* gene, which is the most common of these two efflux genes in enterococci, is located on conjugative plasmids or the chromosome (Roberts and Hillier 1990, Bentorcha, De Cespedes *et al.* 1991). (ii) The genes *tetM*, *tetO* and *tetS* encode for proteins which bind to the ribosome and prevent tetracycline binding. The *tetM* gene is the most frequent tetracycline resistance gene present in enterococci and is located on the chromosome. This gene is commonly associated with a family of genetic elements, which drive their own transfer from donor to recipient bacteria through conjugative plasmids, such as Tn*916* (Rudy, Taylor *et al.* 1997, Rice 1998).

1.8 Genome sequencing

In the 1970s, the Lambda bacteriophage (50,000 nucleotides) was the first genome that was sequenced by Sanger *et al.* (Sanger, Nicklen *et al.* 1977). Since that the DNA sequencing was accomplished at that time for small genomes of organelles and viruses. Complete sequencing of a bacterial genome was not yet possible because of economic and technical restrictions. Subsequent development in sequencing technologies was required to enable whole genome sequencing of bacteria. *Haemophilus influenzae* was the first bacterial genome to be sequenced using a shotgun method developed by

Sanger *et al.* (Sanger, Nicklen *et al.* 1977, Fleischmann, Adams *et al.* 1995).

The shotgun method of sequencing using cloned fragments has limitations. The technique uses randomly sampling and the generation of 500–700 nucleotide reads, which are then assembling to reconstruct the DNA sequence. The assembly process is built by determining regions that overlap, whereby more than 1 million bases of sequence reads are essential to sequence a 1 Mb genome (Fraser and Fleischmann 1997).

Since the mid-1990s, next generation sequencing technologies have arisen, which are high-throughput y*et al*so relatively cheap. Four next generations sequencing platforms, including 454 sequencing platform, Miseq, Hiseq 2000 and GAIIx were used until recently. In Miseq, Hiseq 2000 and GAIIx methods, the construction of clonal DNA colonies (DNA clusters) is prepared by the attachment of DNA molecules and primers on a slide which are then amplified with DNA polymerase. Four types of fluorescently labeled reversible-terminator nucleotides are added to evaluate the DNA sequence, and the combined nucleotides are imaged. The next cycle is started when the fluorescent dye with the terminal 3' blocker is chemically removed from the DNA (Rothberg and Leamon 2008, Rothberg, Hinz *et al.* 2011). Contrastingly, the first step in 454 method is preparation of the sample which involves the following; DNA fragmentation, end repair, capture of the fragments on beads, polymerase chain reaction (PCR), clonal amplification of the captured fragments in aqueous-oil emulsion

microreactors, breaking of the microreactors and enrichment of beads with amplified DNA (Rothberg and Leamon 2008).

More recently in 2011, two major sequencing platforms were released, namely the Ion Torrent Personal Genome Machine (ITPGM) and the Pacific Biosciences (PacBio) RS. Single molecules are sequenced in real time without amplification by using PacBio. In this method, a conjugate of DNA template and DNA polymerase are attached to 50 nm-wide wells. Using nucleotide fluorescently labeled with γ-phosphate, second strand DNA synthesis is carried out by the DNA polymerase. Combinations of bases during DNA synthesis are identified by incomes of a different pulse of fluorescence. ITPGM differs from previous next generation sequencing methods since polymerisation events are distinguished by pH variations instead of light. A bead with DNA fragments carrying specific adapter sequences are connected together and then clonally amplified by emulsion PCR. A chip that contains the template beads has proton-sensing wells that are applied on a silicon wafer, and sequencing is primed from a prearranged location in the adapter sequence. As bases are combined during the sequencing progression, protons are discharged and a signal is revealed relative to the number of bases combined (Donkor 2013) .

1.9 Enterococcal genomes and genome based studies

The genome is the entire coding and non-coding genetic element present in an organism. Deciphering genome sequences has provided a wealth of information about different aspects of the virulence of microorganisms. *E.*

faecium and *E. faecalis* are the two species are responsible for most enterococcal infections and are frequently compared. Comparative genome hybridization (CGH) studies indicated that *E. faecium* and *E. faecalis* have a substantial amount of inter species genomic diversity. This is due to variation in their accessory genomes including wide variety of plasmids, phages, conjugative elements and pathogenicity islands (van Schaik, Top *et al.* 2010). Our understanding of *E. faecium* fundamental biology and virulence-associated traits has been limited due to fewer genome sequences, with *E. faecalis* strains being more widely sequenced and studied.

The genome of *E. faecalis* V583 was determined in the late 1990s and completed in 2002 (Paulsen, Banerjei *et al.* 2003). The first partially assembled, draft genome sequence for *E. faecium* strain TX0016 (formerly *E. faecium* DO strain) isolated in 1992 from a case of endocarditis, was published in 2000. The VanB vancomycin-resistant *E. faecium* strain (Aus0004) was isolated from the bloodstream in 1998. Both of these genome sequences were completed in 2012 (van Schaik, Top *et al.* 2010, Lam, Seemann *et al.* 2012, Qin, Galloway-Pena *et al.* 2012).

From 2002 until 2014 only four other *E. faecalis* genome sequences were published (OG1RF, EF62, D32 and Symbioflor1), and the publicly available genome sequence is not completely annotated. Furthermore, seven *E. faecium* strains, isolated from different ecological niches were reported by van Schaik *et al* (2010), using pyrosequencing technology. Their conclusions can be highlighted in three significant points: (i) hospital-

associated isolates acquire genomic differences associated with colonisation genes and antibiotic resistance; (ii) strains related to the same clonal complex, such as CC17, are related in the core genome, nevertheless the gene content still has large difference; and (iii) the pan-genome of *E. faecium* specified that the gene pool in this species is open most likely as it is subject to multiple ecological niches that this species can colonize. The gain and /or loss of mobile genetic elements is the most significant driving force in enterococci.

In addition to this study the draft genome sequences for 28 enterococcal strains of different origin, including the species *E. faecalis*, *E. faecium, E. casseliflavus*, and *E. gallinarum*, were also published in 2010 (Palmer and Gilmore 2010).

More recently, a report of 51 strains of *E. faecium* isolated from various ecological environments including hospital-isolated, commensal-isolated and animal-isolated was published. The study has contributed to our understanding of genomic diversity of *E. faecium* species. The conclusions of Lebreton *et al* (2013), have confirmed the significant points previously stated by van Schaik *et al* (2010) that (i) The epidemic hospital lineage of *E. faecium* is quickly developing and emerged approximately 75 years ago, associated with the presence of antibiotics, from a population that comprises the majority of animal strains, and not from human commensal lines. (ii) The lineage that comprised most animal strains separated from the human commensal line around 3,000 years ago, a time that matches the

urbanisation of humans, increase of hygienic practices, and domestication of animals, (iii) The acquisition of new metabolic capabilities, colonisation traits, and gain and or loss of mobile elements and function were playing an important role on each bifurcation.

1.10 *E. faecium* genome

Prokaryotic genome sizes differ over more than a twentyfold range. In the prokaryote group, distinct phyla cover approximately overlapping size ranges. Large-scale diversity is observed within species; more than 1,000,000 bp variations have been shown in the genomes of *Streptomyces coelicolor*, *Prochlorococcus marinus* and *Escherichia coli*. Horizontal acquisition, gene duplication and lineage-specific gene loss are genetic events, which can affect bacterial genome size (Bentley and Parkhill 2004, van Schaik, Top *et al.* 2010).

Palmer *et al* (2012) suggested that the size of bacterial genomes correspond with the number of genes in the genome (coding capacity) and consequently the complexity of its encoded activities. The genome of *E. faecium* strains vary in size from 2.50 to 3.14 Mb while the number of ORFs range from 2587 to 3118. The first strains with fully assembled genome sequences of *E. faecium* TX0016 (DO strain) and Aus0004 have genome sizes of 2.69 Mb and 2.95 Mb, respectively each has three circular plasmids (Lam, Seemann *et al.* 2012, Qin, Galloway-Pena *et al.* 2012).

1.10.1 *E. faecium* Sub-populations

There are two subpopulations of *E. faecium*, commensal or community-associated strains (CA clade) and hospital-associated strains (HA clade). Almost all hospital-associated strains encode pathogenicity islands, mobile genetic elements such as IS, plasmids, phage or genes coding for antibiotic resistance, colonisation and virulence (Top, Willems *et al.* 2008, Galloway-Pena, Roh *et al.* 2012).

Molecular epidemiology studies using MLST and eBURST analysis have shown that most of the HA clade of *E. faecium* are associated with a lineage called CC17 (clonal complex17) (Willems, Top *et al.* 2005, Top, Willems *et al.* 2008). The HA clade of *E. faecium* belonging to the lineage CC17 has particular characteristics such as ampicillin and quinolone resistance, in addition, CC17 strains contain the *esp* gene that carried by pathogenicity islands. These factors could be a reason for their improved survival in the hospital environment (Bonten, Willems *et al.* 2001, Willems, Homan *et al.* 2001, Top, Willems *et al.* 2008).

1.11 Mobile genetic elements

Segments of DNA that encode enzymes and proteins that facilitate the movement of DNA inside genomes (intracellular mobility) or among bacterial cells (intercellular mobility) are called mobile genetic elements (MGEs). Intercellular movement of DNA proceeds by three forms in prokaryotes: transformation, conjugation and transduction (Frost, Leplae *et al.* 2005). MGEs play an important role in the evolution of a wide range of

bacteria and are involved in the distribution of variable genes, such as virulence and antibiotic resistance genes causing innovation of 'hospital superbugs', in addition to the formation of new metabolic pathways by catabolic genes (Juhas, van der Meer *et al.* 2009).

1.11.1 Insertion sequences elements and transposons

Intracellular DNA movement is facilitated by transposons and insertion sequences. Insertion sequences (IS) are the simplest transposable elements and are widely distributed in bacteria (Kusumoto, Ooka *et al.* 2011). IS elements are usually less than 2.5 kb in size and commonly defined as carrying only the genetic information associated with their transposition and its regulation. Transposons are larger and carry genes encoding other functions such as antibiotic resistance (Schneider and Lenski 2004).

IS elements are found more often in clinical *E. faecium* strains than community-associated strains. Previously IS16 was considered to be molecular marker for the identification of pathogenicity in clinical *E. faecium* strains (Leavis, Willems *et al.* 2007, Werner, Fleige *et al.* 2011). However, these IS elements are not found in all clinical *E. faecium* strains. A total of 180 IS elements and transposase related genes were located in the complete genome of *E. faecium* TX0016 and almost half of these elements were present on plasmids. Some IS elements are present on the chromosome and plasmids in several copies at definite loci (Qin, Galloway-Pena *et al.* 2012).

IS elements have a significant role in the exchange of the genetic material in *E. faecium.* The element EfaB5 is present at the 3' end of the virulence gene *esp* in *E. faecium*. EfaB5 belongs to the family of conjugative and integrative elements of Gram-positive bacteria, which gives evidence for the horizontal gene transfer in *E. faecium* (van Schaik, Top *et al.* 2010). Tn*1545* have been shown to transfer at a low frequency from *E. faecium* to *Listeria monocytogenes* in the intestinal tract of gnotobiotic mice. Tn*916* is the well-characterized conjugative transposon (Jett, Huycke *et al.* 1994). It was shown using an *E. faecalis* donor that when two distinguished derivatives of Tn*916* are present the conjugative transfer of one transposon is accompanied by the other. The frequency rate of transfer was up to 50% (Hammerum, Flannagan *et al.* 2001).

1.11.2 Plasmids

A plasmid is a group of functional genetic segments that are structured into a steady, self-replicating replicon, which is smaller than the cellular chromosome and which typically does not comprise genes required for vital cellular functions. Plasmids can be circular double-stranded DNA molecules or linear double-stranded DNA (Hinnebusch and Tilly 1993).

Several plasmids have been reported in *Enterococcus* that contain antimicrobial and heavy metal resistance genes and play a significant role in virulence and DNA repair mechanisms (Arias, Panesso *et al.* 2009, Garcia-Migura, Hasman *et al.* 2009). Most of the antibiotic resistance genes are existent on the plasmids, which can be confirmed from the occurrence of

plasmid replicating genes and toxin/antitoxin genes in the same contig as that of antibiotic resistance genes (van Schaik, Top *et al.* 2010). Some genes present on plasmids and transposons encode for traits such as antibiotic resistance, virulence and bacteriocin activity and utilisation of unusual substrates. These traits help the organism to survive in challenging environments. In enterococci, the virulence genes are present on conjugative plasmids, which are horizontally transferred to other strains (Jett, Huycke *et al.* 1994).

1.11.3 Bacteriophages

The viruses that infect bacteria are named bacteriophages (phage). Phages must attach to the host to initiate their life cycle and it is not able to propagate in the absence of a host. Phages are associated with almost all identified bacteria and are consequently discovered in distinct environments ranging from oceans and soil to deserts. Phages can be found as free virions in the environment or associated with their bacterial hosts. Phages are discovered in almost all places where their bacterial hosts occur (Wommack and Colwell 2000, Pedulla, Ford *et al.* 2003, Prigent, Leroy *et al.* 2005, Prestel, Salamitou *et al.* 2008, Srinivasiah, Bhavsar *et al.* 2008).

In recent years, many phage genome sequences have become accessible. It is noticeable from phage genome sequences that phage genomics are extremely different. This variety in genetic makeup results from the particular replication of phage particles through infection of their hosts. During these infections phages can exchange their DNA with host genomes

by recombination and this generates diversity in the phage genome (Hendrix, Smith *et al.* 1999).

The order Caudovirales represent tailed phages with dsDNA and an isometric capsid and contains the vast majority of phages. Caudovirales include three phylogenetically-related families distinguished by tail morphology: *Myoviridae* (long contractile tails), *Siphoviridae* (long non-contractile tails), and *Podoviridae* (short tails) (Ackermann 2007, Krupovic, Prangishvili *et al.* 2011). Phages that infected *Escherichia coli* are the most well-studied tailed phages included T4 (*Myoviridae*), coliphages ʎ, (*Siphoviridae*), and T7 (*Podoviridae*) (Ptashne, Jeffrey *et al.* 1980, Johnson, Poteete *et al.* 1981, Tabor and Richardson 1985, Miller, Kutter *et al.* 2003). Non-tailed phages have many families with different morphologies, comprising polyhedral (vesicular and envelope like), filamentous (long filaments to short rods), and pleomorphic (including lemon, droplet and ampule shaped) (Ackermann 2007).

Phages can enhance the environmental fitness and virulence of the bacterium by lysogenic conversion (van Schaik, Top *et al.* 2010). Temperate phages can carry genes coding for virulence factors which gets integrated into the bacterial genome and can be expressed by the pathogen (Bensing, Siboo *et al.* 2001, Chibani-Chennoufi, Dillmann *et al.* 2004). Once the genome of temperate phages becomes integrated into the host chromosome specific genes are expressed for maintenance of lysogeny and for repression of the lytic life cycle. Antibiotics like norfloxacin and

mitomycin C or physical stress such as, UV radiation can be used to induce the lytic cycle (Duerkop, Palmer *et al.* 2014).

The family of phages that infect *E. faecium* and *E. faecalis* are *Siphoviridae.* These phages have an isometric head about 40nm and a non-contractile tail, which is long ranging from 70 nm to 220 nm (van Schaik, Top *et al*. 2010, Yasmin, Kenny *et al*. 2010).

1.11.4 Genomic islands

Genomic islands or GEIs mediate a considerable part of genetic recombination in bacteria. They play an important part in bacterial evolution by spreading of antibiotic resistance and virulence genes and by producing new clinical strains. GEIs are distinct DNA regions which are mobile or non-mobile and vary between strains. They have the ability integrate into the host and excise themselves and transfer to new bacteria by transformation, conjugation or transduction (Juhas, van der Meer *et al.* 2009). Genomic islands are usually 10 to 200kb in size and carry not only virulence genes but also other genes for symbiosis, aromatic compound metabolism, resistance to mercury or siderophore synthesis (Hacker and Kaper 2000, Sullivan, Trzebiatowski *et al.* 2002, Juhas, van der Meer *et al.* 2009).

The *esp* gene located on genomic island in *E. faecium* and *E. faecalis* can transfer between the two species. The genomic island in *E. faecalis* consists of phage related integration, excision proteins, homologs of plasmid

conjugation functions and terminal direct repeat. This suggests that the genetic transfer of genomic island or associated genes may occur as an integrative conjugative element although this has not been proved (Manson, Hancock *et al.* 2010).

Aims of the study

At the start of this PhD project in 2010 there were 72 sequenced genomes of *Enterococcus* sequenced including 23 of *E. faecium* and none of these genomes were closed. None of the *E. faecium* genomes were from animal isolates. The lack of animal *E. faecium* isolate genomes inspired the research aims.

Multi-drug resistant enterococci, particularly those that are vancomycin resistant, are a major cause of concern for the medical community; it has been shown that the genes responsible for this resistance have the potential to be transferred to other Gram-positive pathogens such as *Staphylococcus aureus*. Antimicrobials used as growth promoters to enhance productivity and animal health has produced cross-resistance and has led to the emergence and spread of resistance to vancomycin in enterococcal populations. A greater genome-based insight is needed to integrate the relationship between *E. faecium* from animals and humans and the study presented here has sought to achieve this.

General aims

The primary aim of this research is to answer two key questions that are:

(i) Are strains from animals discrete from human isolates and have they acquired genes specific for colonising an animal host?

(ii) Which mobile genetic determinants are carried by animal strains of *E. faecium* and are these common to or distinct from human isolates?

Specific aims

(i) Complete and annotate the genome sequence of the vancomycin-resistant isolates from animals; E429 (chicken), E172 (calf), and E142 (pig).

(ii) These genomes sequences will be compared with each other and to the reported human and animal isolates.

(iii) Analyse phylogenetic relationships within *E. faecium* strains by investigating the molecular evolutionary connections between animal strains that will be represented through phylogenomic trees.

(iv) Compare *E. faecium* prophage genomes to identify the differences between the phage types resident in this species.

Chapter Two: Materials and methods.

2.1 Media, Strains and Antibiotics

2.1.1 Growth Media

All broth media were prepared according to the manufacturer's instructions unless specified. Media was prepared in distilled water and sterilised by autoclaving for 15 min at 120°C at 15 psi, unless otherwise stated.

Todd Hewitt Broth (THB)

36.4 g l^{-1} of THB powder was prepared and sterilised by autoclaving. 1.5% (w/v) Agar –Agar (Merck) was added to the broth prior to sterilisation to obtain THB agar media.

Luria Bertani (LB) Media

25 g l^{-1} of LB broth powder was prepared and sterilised by autoclaving. 37 g l^{-1} of LB agar powder was prepared and sterilised by autoclaving.

Bottom Agar

A solution containing 18.2 g l^{-1} of THB powder and 7.5 g l^{-1} of High clarity agar was prepared in 500 ml of distilled water and sterilised by autoclaving.

Top soft Agar

7.28 g l^{-1} of THB powder and 0.8 g l^{-1} of High clarity agar were added to 200 ml of distilled water and sterilised by autoclaving. 2 ml of 1 M $CaCl_2$ was then added prior to use.

2.1.2 Strains and culture conditions

The bacteria used in this study are listed in Table 2.1. Cultures were stored at -80 °C in THB containing 15% (v/v) glycerol. Cultures were maintained on THB agar and stored at 4 °C.

Standard culture conditions for *E. faecium* in this study were 10 ml THB in a universal tube with shaking at 250 rpm overnight at 37 °C. For larger scale cultures, a ratio of 1: 10 media to conical flask volume was maintained. Over growth, absorbance at 600 nm was monitored against sterile THB blank.

2.1.3 Antibiotics

Antibiotics used in this study are listed at selective concentrations in Table 2.2. Stock solutions of antibiotics were prepared in ethanol or distilled water followed by filter sterilisation and stored at -20 °C.

Table 2.1: List of bacterial strains used in this study for experimental and bioinformatics analyses

Strain	Source	Information
EnGen0009-E1573	Bison	Rumen, Belgium
E172	Calf	ST1, VanA resistance strain, Netherland
EnGen0028-E1604	Cheese	Norway
E429	Chicken	ST8, VanA resistance strain, Netherland
EnGen0001-E1575	Chicken	Belgium
LIV294	Chicken	Chicken faeces
EnGen0005-E0045	Chicken	Faeces, UK
EnGen0048-E4215	Chicken	Sweden
EnGen0043- E2134	Chicken	Netherland
LIV302	Dog	Dog faeces
E4452	Dog	Faeces, Netherland
E4453	Dog	Faeces, CC17, Netherland
EnGen0020-E1574	Dog	Belgium
EnGen0057-E4389	Dog	Faeces, Denmark
EnGen0029-E1613	Fish burger	Norway
EnGen0042-E1861	Hospitalized patient	Faeces, Spain
EnGen0047-E3548	Hospitalized patient	Blood, Netherland
1_141_733	Hospitalized patient	Wound, USA
1_230_933	Hospitalized patient	Blood, CC17, USA
1_231_408	Hospitalized patient	Blood, CC17, USA
1_231_410	Hospitalized patient	Skin and soft tissue, CC17, USA
1_231_501	Hospitalized patient	Blood, USA
1_231_502	Hospitalized patient	Blood, CC17, USA
Aus0004	Hospitalized patient	Blood, CC17, Australia
DO	Hospitalized patient	Blood, CC17, USA
E1071	Hospitalized patient	Faeces, Netherland
E1162	Hospitalized patient	Blood, CC17, France
E1636	Hospitalized patient	Blood, Netherland
E1679	Hospitalized patient	Vascular catheter tip, Brazil
EnGen0002-E1133	Hospitalized patient	Faeces, CC17, USA
EnGen0004- E1258	Hospitalized patient	Blood, Spain
EnGen0011-E1185	Hospitalized patient	Blood, France
EnGen0012_E0120	Hospitalized patient	Ascites, Netherland
EnGen0013-E0333	Hospitalized patient	Blood, CC17, Israel
EnGen0016-E1392	Hospitalized patient	CC17,UK
EnGen0021-E1552	Hospitalized patient	Faeces, Netherland
EnGen0024-E1904	Hospitalized patient	Urine, Netherland
EnGen0025-E1626	Hospitalized patient	Stomach, Netherland
EnGen0026-E2039	Hospitalized patient	Germany

EnGen0030-E2883	Hospitalized patient	Blood, Netherland
EnGen0031-E1623	Hospitalized patient	Pus, Netherland
EnGen0033-E1972	Hospitalized patient	Blood, Germany
EnGen0034-E2297	Hospitalized patient	Urine, CC17, USA
EnGen0035-E1627	Hospitalized patient	Gut, Netherland
EnGen0036-E1731	Hospitalized patient	Blood, CC17, Tanzania
EnGen0038-E2620	Hospitalized patient	Blood, Netherland
EnGen0040-E1634	Hospitalized patient	Netherland
EnGen0045-E6012	Hospitalized patient	CC17, Latvia
EnGen0046-E2560	Hospitalized patient	Blood, CC17, Netherland
EnGen0049-E6045	Hospitalized patient	CC17, Portugal
EnGen0050-E2369	Hospitalized patient	Wound, CC17, Hungary
EnGen0051-E1644	Hospitalized patient	CC17, Germany
EnGen0052-E3346	Hospitalized patient	Blood, Netherland
EnGen0054-E1321	Hospitalized patient	Catheter, CC17, Italy
EnGen0056-E3083	Hospitalized patient	Blood, Netherland
TX0082	Hospitalized patient	Blood, USA
TX0133A	Hospitalized patient	Blood, USA
TX0133B	Hospitalized patient	Blood, USA
TX0133C	Hospitalized patient	Blood, USA
TX0133a.01	Hospitalized patient	Blood, USA
TX0133a.04	Hospitalized patient	Blood, USA
U0317	Hospitalized patient	Urine, CC17, Netherland
LIV66	Human	TX0016, Endocarditis isolate
LIV153	Human	VanA resistance strain
ERV26	Human	Airways
P1139	Human	Urinogenital tract
V689	Human	Skin
C1904	Human	Blood
C309	Human	China
C497	Human	Blood
ERV161	Human	Blood
ERV165	Human	Gastrointestinal tract
ERV168	Human	Skin
LCT-EF128	Human	Bronchoalveolar lavage, China
P1123	Human	Blood
P1137	Human	Skin
P1986	Human	Blood
S447	Human	Urinogenital tract
ERV102	Human	Oral cavity
503	Human	Unpublished

504	Human	Unpublished
505	Human	Unpublished
506	Human	Unpublished
509	Human	Unpublished
510	Human	Unpublished
511	Human	Unpublished
513	Human	Unpublished
515	Human	Unpublished
ERV38	Human	Unpublished
ERV69	Human	Unpublished
ERV99	Human	Unpublished
R446	Human	Unpublished
R494	Human	Unpublished
R496	Human	Unpublished
R497	Human	Unpublished
R499	Human	Unpublished
R501	Human	Unpublished
TC_6	Human	Derived from the ampicillin resistant, Tn916-containing strain D344R
ERV1	Human	Airways
LIV296	Jaguar	Jaguar faeces- Chester zoo
D344SRF	Lab strain	Lab strain, USA
EnGen0032-E1622	Mouse	Netherland
Com12	Non-hospitalized individual	Faeces, USA
Com15	Non-hospitalized individual	Faeces, USA
E1039	Non-hospitalized individual	Faeces, Netherland
E980	Non-hospitalized individual	Faeces, Netherland
EnGen0015-E1007	Non-hospitalized individual	Faeces, Netherland
EnGen0017-E1050	Non-hospitalized individual	Faeces, Netherland
TX1330	Non-hospitalized individual	USA
EnGen0018-E1576	Ostrich	Caecum, South Africa
LIV297	Otter	Mouth swab
LIV298	Otter	Mouth swab
LIV303	Otter	Mouth swab
EnGen0007- E1578	Pig	Faeces, Germany
E142	Pig	ST6, VanA resistance strain, Netherland
EnGen0008-E0688	Pig	Spain
EnGen0014-E0679	Pig	Belgium

EnGen0019-E0680	Pig	Germany
EnGen0044-E2071	Poultry	Denmark
EnGen0039-E1630	River water	Netherland
LIV299	Rodent	Irish rodent faeces
EnGen0022-E0269	Turkey	Faeces, Netherland
EnGen0010-E0164	Turkey	Faeces, Netherland
LCT-EF20	Unpublished	Culture of *Enterococcus faecium* that spent 17 days in space aboard the Shenzhou 8 spacecraft, China
LCT-EF258	Unpublished	Culture of *Enterococcus faecium* that spent 17 days in space aboard the Shenzhou 8 spacecraft, China
LCT-EF90	Unpublished	China
TX1337RF	Unpublished	Gastrointestinal tract
NRRL	Milk and dairy utensils	Unpublished
Aus0085	Unpublished	Unpublished
C621	Unpublished	Unpublished
E1590	Unpublished	Unpublished
E1620	Unpublished	Unpublished
P1140	Unpublished	Unpublished
E417	Unpublished	Unpublished

Table 2.2: List of antibiotics used in this study.

Antibiotics*	Concentration (µg ml^{-1})
Tetracycline	5
Ampicillin	50
Chloramphenicol	5
Spectinomycin	50
Erythromycin	10
Gentamycin	500
Vancomycin	10

2.2 Reagents

2.2.1 General Reagents and Buffers

Stocks solutions of buffers were prepared with the ingredients listed below. The components were dissolved in 1 L of water, sterilised by autoclaving and stored at RT. Diluting with water as required made a working solution of each buffer. The working solutions were also sterilised by autoclaving before use and stored at RT.

Phosphate Buffered Saline (PBS)

1 x PBS

NaCl 8 g l^{-1}

KCl 0.2 g l^{-1}

Na_2HPO_4 1.4 g l^{-1}

KH_2PO_4 0.24 g l^{-1}

Tris-HCl Buffer

Tris 2 M

Tris-HCl buffer was prepared from stock by diluting in water, with pH adjusted using conc. HCl. Tris-HCl Buffer was then sterilised by autoclaving and stored at RT.

Phage buffer (SM), pH 7.8

Tris/HCl pH 7.8 50 mM

NaCl 10 mM

$MgSO_4$ 1 mM

$CaCL_2$ 4mM

Gelatin 1% (w/v)

Enzymatic Lysisbuffer, pH 8.0

Tris/HCl 20 mM

EDTA 2 mM

Triton X-100 1.2% (v/v)

TAE 50X, pH 8.0

Tris 2 M

EDTA 50 mM

Acetic acid 1 M

TE buffer, pH 7.5

Tris/HCl, pH 8.0 10 mM

EDTA, pH 8.0 1 mM

DNA loading buffer

0.25% (w/v) bromophenol blue

30% (v/v) glycerol in water

2.3 Enzymes

Enzyme	Enzyme Source
Lysozyme	Sigma
Proteinase K	Sigma
Ribonuclease A	Sigma
ExoSAP-IT	Usb.Affymetrix,Inc
pfx polymerase	Invitrogen
BioMix Red	Bio Line
Taq polymerase	Thermo

2.4 Kits

Kits	Manufacturer
ISOLATE PCR and Gel Kit	BioLine
QIAprep Miniprep kit	Qiagen
QIAGEN DNeasy Blood & Tissue kit	Qiagen
ISOLATE Plasmid DNA mini Kit	BioLine

2.5 Methods

2.5.1 DNA purification

DNA was isolated and purified using a QIAGEN DNeasy Blood & Tissue kit according to the manufacturer's instructions, as follows. Two colonies of strain E429 were used to inoculate 10 ml of THB and grown overnight at 37°C. The cells were pelleted (7,500 g; 10 min) and resuspended in 180 µl of enzymatic lysis buffer. After careful vortexing, 20 µl of lysosyme was added and incubated for at least 30 min at 37°C.

After that, 25 µl of proteinase K and 200 µl of Buffer AL were added and incubated at 56°C for 30 min. 200 µl of 100% (v/v) ethanol was added to the sample and mixed by vortexing. Then, the mixture was transferred to a DNeasy Mini spin column and centrifuged (8000 rpm; 1 min). 500 µl Buffer AW1 was added and centrifuged (8000 rpm; 1 min) followed by 500 µl of Buffer AW2 and centrifuged (14,000 rpm; 3 min) to dry the membrane of the DNeasy Mini spin column. The genomic DNA was eluted by the addition of 200 µl Buffer AE and centrifuged (8000 rpm; 1 min).

Finally, the genomic DNA was aliquotted and stored at -20 °C until use. DNA was quantified using NanoDrop and Qubit Fluorometer (Invitrogen).

2.5.2 Plasmid purification

Miniprep plasmid isolation

Plasmids were isolated and purified using QIAprep® Miniprep kit and ISOLATE Plasmid DNA mini Kit according to the manufacturer's instructions, as follows. Two colonies of strains were used to inoculate 10 ml of THB and grown overnight at 37 °C. The overnight culture was centrifuged and bacterial pellet was resuspended tin 250 µl Buffer P1 and transferred to a microcentrifuge tube. 250 µl Buffer P2 was added and mixed gently by inverting the tube. Then, 350 µl Buffer N3 was added and the tube was inverted immediately to mix the solution gently. The solution was centrifuged (13,000 rpm; 10 min).

The supernatant was applied to a QIAprep spin column and centrifuged for 30–60 s and the flow-through was discarded. Finally, 0.75 ml Buffer PE was added to wash the QIAprep spin column and centrifuged for 30–60 s and the flow-through was discarded and QIAprep spin column was centrifuged for 1 min to remove residual wash buffer. Place the QIAprep column in a clean 1.5 ml microcentrifuge tube. The DNA was eluted by add 50 µl Buffer EB (10 mM Tris·Cl, pH 8.5) or water to the center of each QIAprep spin column and the column was standed for 1 min, and centrifuged for 1 min.

Chloramphenicol amplification

A colony of each strain was inoculated into 10 ml of THB then incubated overnight with shaking at 37°C. 0.1 ml of the overnight culture was transferred to 50 ml of fresh THB and incubated with shaking at 37°C for 4 h. 25 ml of cells were added to 500 ml of THB and incubated with shaking at 37°C for exactly 2 h. Chloramphenicol was added to the culture to a final concentration of 170 µg ml^{-1} and incubated contrived at 37°C overnight. Plasmids were purified using QIAGEN Plasmid Mini prep as per manufacturer guidelines but with several modifications. 100 ml of the overnight culture was used to extract the plasmid. Lysozyme was added to a final concentration of 100 µg ml^{-1} and the cells were incubated for 10 min at 37°C prior to plasmid purification. For cell lysis, buffer P2 was added and cells were incubated for 5 min at 37°C. Finally, the resulting plasmid DNA extract was electrophoresed on 1% (w/v) agarose gel and electrophoresed for 2 h at 90V.

Whole cell mini lysate

2 ml of overnight culture was resuspended in 5 ml of lysis buffer. 5 μg ml^{-1} lysosyme was added and incubated for 45 min. The culture was pumped up and down using 1.5 ml syringe with 0.5 X 16mm needle. 25% glycerol was added and samples were separated on 5% (w/v) agarose gel and electrophoresed for 1 h at 80V.

2.6 Genetic Manipulations by Polymerase Chain Reaction (PCR)

2.6.1 Primer design and synthesis

Primers were designed using the online Primer3-plus software. The following guidelines were used for designing primers:

a) Primers should be 18 - 27 bases in length

b) 50% of GC content

c) Melting temperature (Tm) ideally over 60°C

d) Primers with a terminal T should be avoided

e) Primers with 3' complementary ends should be avoided, as they can result in primer dimerisation.

Primers were synthesised by Eurofins genomic (http://www.eurofinsgenomics.eu/) or Sigma-Aldrich (http://www.sigmaaldrich.com) supplied at a concentration of 100 μMol. All primers used in this study are listed in Table 2.3.

Table 2.3: Genome coordinates and sequence of primers used for closing animal *E. faecium* gaps strain E429 isolated from chicken.

Genome coordinates of primers	Primer sequence	Gap location
Gap 1_F Gap 1_R	5'-tgctttggcttcagttccta-3' 5'-cgttgttagtggtccgttca-3'	96941 - 97767
Gap 2_F Gap 2_R	5'-aatgaaacttccaacatggga-3' 5'-tgcaaatgcaactattttcaataaa-3'	474153 - 474812
Gap 3_F Gap 3_R	5'-ccaatcattaacagtgtttggaa-3' 5'-tgaagcgcattttggatctg-3'	693583 - 694463
Gap 4_F Gap 4_R	5'-ccaacgagtaaggagtcacca-3' 5'-ggtttgaaaaaccaagttatggtc-3'	747122 - 747857
Gap 5_F Gap 5_R	5'-tggatatgatcgaaaaatatcaagg-3' 5'-ttcaaaaagaaaaataggctgaa-3'	911130 - 911859
Gap 6_F Gap 6_R	5'-agtagggcaccgaagaaatg-3' 5'-ccaagaatcgacttcttggatga-3'	1221329 - 1222141
Gap 7_F Gap 7_R	5'-caagtagggcaccgaagaaa-3' 5'-tcgctttagtcaattttggtca-3'	1344280- 1345140
Gap 8_F Gap 8_R	5'-tgtgaatttcaactccttctaaattg-3' 5'-tggtataattttcttatcggtaagtgg-3'	1364480 -1365335
Gap 9_F Gap 9_R	5'-catgaacgtgcagggaagta-3' 5'-gatgaaatattcacaaagctaacca-3'	1400219 -1400897
Gap 10_F Gap 10_R	5'-ttttatgatgctccagaagtgaa-3' 5'-tgattcgatccccttttgtta-3'	1477585- 1478432
Gap 11_F Gap 11_R	5'-gatcgcgatcggtcaatttg-3' 5'-acgttgtttcccaatgccta-3'	1641260 – 1642088
Gap 12_F Gap 12_R	5'-tgccatgtcctgtcgttctc-3' 5'-tatggacatggaccgttcac-3'	1854861 -1855469
Gap 13_F Gap 13_R	5'-atcaagtaaaattgtctgcagga-3' 5'-aagtggaaatggatgggaca-3'	1977394- 1978118
Gap 14_F Gap 14_R	5'-aacggagttaacggctttcc-3' 5'-gcggaatggaacggtattta-3'	1985357 -1986083
Gap 15_F Gap 15_R	5'-tcgaaacgtttaggccatag-3' 5'-tttgcggtacaggggggtta-3'	2019530- 2020164
Gap 16_F Gap 16_R	5'-tccaattgcttccttccatc-3' 5'-cagttgagtcgtggaaaacg-3'	2209198- 2209989
Gap 17_F Gap 17_R	5'-tcatcccctaactgcagaaga-3' 5'-aagtgaattctgcaccagca-3'	2302574 -2303221
Gap 18_F Gap 18_R	5'-tcataagcgccgtaccttcc-3' 5'-acgaactcatgcagtccaca-3'	2363691- 2364340
Gap 19_F Gap 19_R	5'-tcagcaacttttctattctctttg-3' 5'-gacgttaaccattgaaaacatcc-3'	2562483- 2563373
Gap 20_F Gap 20_R	5'-aaattgagtggttttgaccttga-3' 5'-tattcccaaaaatttcgtgac-3'	2607921- 2608662

Genome coordinates of primers	Primer sequence	Gap location
Gap 21_F Gap 21_R	5'-tgcaaaattggagaacgaaa-3' 5'-gcggtcaagtttgttttgaa-3'	2674655- 2675490
Gap 22_F Gap 22_R	5'-gttttttgaaagcataattgcaataa-3' 5'-aggcccccaacattaaaatc-3'	2678784- 2679615
Gap 23_F Gap 23_R	5'-attttggggagcgtcaataa-3' 5'-caaaggaagtattgagctatgcg-3'	2687147- 2688025
Gap 24_F Gap 24_R	5'-ccatttttggataactggttttcc-3' 5'-ctacggactgaattaacggc-3'	2829762- 2830574
Gap 25_F Gap 25_R	5'-ttcagaatgcaattgattaaacg-3' 5'-ttggcaaaagatagcgaagg-3'	2836122 -2836824
Gap26_F Gap26_R	5'-attggctgaccaagcaaaag-3' 5'-tcgtcttgtagtatagttgaaaaatcc-3'	392623-393893
Gap28_F Gap28_R	5'-gcaatttcctaatgaagaatctctg-3' 5'-tcgtattcttccagcgaatg-3'	413432-414602
Gap29_F Gap29_R	5'-accatagacgaactgacaatga-3' 5'-acctaagccgaaagctccag-3'	419900-421179
Gap30_F Gap30_R	5'-cagccatccacaagtaaacatta-3' 5'-ttatgggtgcgagtcaaaga-3'	2394372-2395598
Gap31_F Gap31_R	5'-atattgcaattcccgattcc-3' 5'-gctgtacgctccaatcatca-3'	631078-632119
Gap32_F Gap32_R	5'-catgtgtatgtctaaacccatga-3' 5'-taaaagctgcgaaagccgta-3'	693582-694464
Gap33_F Gap33_R	5'-gaaatcctcgacagatgaatac-3' 5'-ggaaattgagttaaatcaccaaca-3'	706741-707895
Gap34_F Gap34_R	5'-ttcgacgaaatcatgttctagaaag-3' 5'-aagctctctagtaattgttgattaagg-3'	730239-731380
Gap35_F Gap35_R	5'-cagctagtatttatggatggcagc-3' 5'-cgaccgttccttatctaaacg-3'	818407-819420
Gap36_F Gap36_R	5'-tgtttttccttccatcagca-3' 5'-aaatggcattcaaaatggca-3'	1096043-1097041
Gap37_F Gap37_R	5'-aatggcgaagaaaggagtga-3' 5'-tttcattcgaagcttggctg-3'	1123082-1124080
Gap38_F Gap38_R	5'-caaacaattttgtaagttcatcataag-3' 5'-gatctcgtcctgcggtttg-3'	1149869-1150900
Gap39_F Gap39_R	5'-atatcgaacagacggtaacc-3' 5'-tgcaggattagaaggaagctg-3'	1167768-1168737
Gap40_F Gap40_R	5'-tgcagaaatccaagaattatca-3' 5'-gtgtttaacaaaggatcgattgac-3'	1264364-1265676
Gap41_F Gap41_R	5'-gacctttgagggtgcagttg-3' 5'-ttaacgctttgggcattttc-3'	1377756-1378716
Gap42_F Gap42_R	5'-agtccatcactgttattcaaatca-3' 5'-cctgttacgttgtagttggatctg-3'	1433512-1434668
Gap43_F Gap43_R	5'-ttcggcttatttccgaagaa-3' 5'-aagggctgtgacaaatgtacc-3'	1509550-1510784

Genome coordinates of primers	Primer sequence	Gap location
Gap45_F Gap45_R	5'-atttcaggtggtttctggac-3' 5'-cactagaaggcgcatgtgag-3'	1752915-1753930
Gap46_F Gap46_R	5'-ttcttactagtccgaatgtatccaa-3' 5'-ttaccttctgcttgctctaaactg-3'	1951699-1952864
Gap47_F Gap47_R	5'-ctcacacaaagtgtaactaattttgac-3' 5'-cttgtgagtggcttattgatcc-3'	2089221-2090244
Gap48_F Gap48_R	5'-tcctgctcaaaacaaaaagatg-3' 5'-tttgcggagttgtaataggttataatg-3'	2160136-2161168
Gap49_F Gap49_R	5'-aattgttccattgcggtttc-3' 5'-tccgctttcataaatctcgaa-3'	2221256-2222385
Gap50_F Gap50_R	5'-ggacgcttgtctatttcatgg-3' 5'-acgtttttcagagccatttc-3'	2227163-2228772
Gap51_F Gap51_R	5'-ggaagattttaactgtttgctatagat-3' 5'-ttccataaaaatcccgaatcc-3'	2689338-2692879
Gap52_F Gap52_R	5'-tcttggtctcggacaaactct-3' 5'-ccagagattcttcattaggaaattg-3'	2695011-2700691
Gap53_F Gap53_R	5'-ccactatcacctttattcctggt-3' 5'-cattaaaacgaaatatgtatgattctg-3'	2702122-2705334
Gap54_F Gap54_R	5'-gaattgattctgtagtgacccc-3' 5'-aagatagaaatgttgccccc-3'	2707245-2708776
Gap55_F Gap55_R	5'-gggaaaacacacggacattc-3' 5'-tgaccccgtaactcacacttc-3'	2716482-2717794
Gap56_F Gap56_R	5'-ccctctacagaagaatcgctatc-3' 5'-ttaaaaacttctaatggtgtttggt-3'	2719389-2721404
Gap57_F Gap57_R	5'-cgtccatataaatagcggcata-3' 5'-tcatgaacaatatgatgtgatcg-3'	2722748-2725013
Gap58_F Gap58_R	5'-tttaccgaatgaacagatagc-3' 5'-tgaatcattttaaaggcaaacaa-3'	2731923-2732472
Gap59_F Gap59_R	5'-ttttctttaactagagcgctttttatg-3' 5'-actcatgatacgcagctcca-3'	2734766-2735707
Gap60_F Gap60_R	5'-ttgcttttgcaactctaagtgaa-3' 5'-gggagcatttatacccacca-3'	2762959-2764253
Gap61_F Gap61_R	5'-tgcggacattattgatctagc-3' 5'-tcggaggaatattatgtgagtaca-3'	2765573-2766660
Gap62_F Gap62_R	5'-cttttccaagccatactcca-3' 5'-tcattggtctgctcgatgac-3'	2776481-2779588
Gap63_F Gap63_R	5'-gaaacgctctgtaacgcttct-3' 5'-tgctacagtacttgttgatgtggtt-3'	2781077-2784542
Gap64_F Gap64_R	5'-ttggtcagaatgaagaataacagc-3' 5'-aatcagataataacccctatacaacg-3'	2786496-2790062
Gap65_F Gap65_R	5'-aggagacgatgagtttgaaca-3' 5'-gccgtgggatactatcttcg-3'	2894371-2810045

Genome coordinates of primers	Primer sequence	Gap location
Gap66_F Gap66_R	5'-ttgcaattcctaatgggagtg-3' 5'-tggccttcgtattctcaaca-3'	2812453-2816178
Gap67_F Gap67_R	5'-cttttgttgcacaccctgag-3' 5'-tgaaaggtcggaaagaacaaa-3'	2821217-2824460
Gap68_F Gap68_R	5'-gaggagtcgaacccctaacc-3' 5'-aagcattttcttattgactttcaca-3'	2829762-2830574
Gap69_F Gap69_R	5'-ctagaatggggagtggcaaa-3' 5'-tttctatcaattattaaaacggtgga-3'	2836122-2836824
Gap70_F Gap70_R	5'-tcgtggatatgctgcttttt-3' 5'-tgcaacttgatgcaaacaca-3'	2844043-2844142
Gap71_F Gap71_R	5'-aaaagcatcggtgcagtgtt-3' 5'-acgacgactgctcccagtaa-3'	2858252-2861585

Table 2.4: Antibiotic resistance gene primers used in this study.

Antibiotic gene primers	Primer sequence
TetM_F TetM_R	5'-ttttgggcttttgaatggag-3' 5'-tctatccgactatttggac-3'
pbp1_F pbp1_R	5'-gcaagaatggcaaatgaac-3' 5'-cagcttggtacatgattt-3'
Van_F Van_R	5'-catccccgttttatttgg-3' 5'-accagttacatacgtcggg-3'

Table 2.5: Phage integrase primers used in this study.

Phage integrase primers	Primer sequence
E429_phage_int_F E429_phage_int_R	5'-ggcgaaaaatattggggatt-3' 5'-cgaagcaccactttcaaaca-3'
E429_DOphage_int_F E429_DOphage_int_R	5'-caaagatgggcgattcaagt-3' 5'-tttgaaaatcggtcacctg-3'

Table 2.6: Housekeeping gene primers used in this study.

Housekeeping genes primers	Primer sequence
Adk F Adk R	5'-tatgaacctcattttaatggg-3' 5'-gttgactgccaaacgatttt-3'
Adk2 F Adk2 R	5'-gaacctcattttaatgggg-3' 5'-tgatgttgatagccagacg-3'

2.6.2 PCR conditions and reactions

A 25-µl PCR mixture was used to generate PCR products for sequencing and contained 12.5 µl of BioMix Red (Bio Line), 0.75 µl of each (10 mM) primer, 0.5 µl of (10 ng) E. *faecium* DNA and 10.5 µl of autoclaved distilled water. The PCR mixtures were subjected to thermal cycling (2 min at 94°C and then 30 cycles of 30 s at 94°C, 30 s at 55°C and 90 s at 68°C with a 7-min final extension at 68°C).

Alternatively, PCRs were performed using pfx polymerase (Invitrogen) in the following standard protocol. A 50 µl PCR mixture contained 3 µl of 10 mM primer mix (Table 2.3), 1 µl (10 ng) of *E. faecium* DNA, 1.5 µl of 10 mM deoxyribonucleoside triphosphate mixture (Bio Line), 1 µl of 50 mM $MgSO_4$, 5 µl of pfx Amplification Buffer, 38.1 µl of autoclaved distilled water. Finally, 0.4 µl platinum pfx DNA polymerase was added to the PCR mixtures. The PCR mixtures were subjected to the same thermal cycling condition as above.

Alternatively PCR conditions were used as follows:

1- Temperature gradient PCR (10°C): this procedure was used for the primers that did not work with the first two reactions conditions:

50 µl PCR mixtures were used to generate PCR products for sequencing using BioMix Red as above with temperature gradient 10°C for 30 s (50°C, 55°C, 60°C).

2- Magnesium chloride with two different thermal cycles: this step was used for the primers that had not previously worked:

The 50 µl PCR mixtures were used as above except 5 µl of 50 mM $MgCl_2$ was added and the amount of the autoclaved distilled water was changed to be 16µl. The first thermal cycle condition was (5 min at 94°C and then 30 cycles of 30 s at 94°C, 30 s at 55°C and 1.30 min at 68°C with a 7-min final extension at 68°C) and the second thermal cycle condition was (5 min at 94°C and then 35 cycles of 30 s at 94°C, 30 s at 53°C and 2 min at 68°C with a 7-min final extension at 68°C).

3- Taq polymerase enzyme: this step was used for the primers that had not previously worked. PCRs were performed using: PCR buffer (45 mM Tris-HCL, pH 8.8; 11 mM $(NH_4)_2$ $S0_4$; 4.5 mM $MgCl_2$; 6.7 mM 2-mercaptoethanol; 4.4 µM EDTA; 113 µg/ml BSA; 1 mM of each of 4 deoxyribonucleotide triphosphates), 1 µM of each primer and 0.5 µl of Taq polymerase (Thermo) per 10 µl reaction. The thermal cycling conditions were 5 min at 94°C and then 35 cycles of 30 s at 94°C, 30 s at 53°C and 2 min at 68°C with a 7-min final extension at 68°C.

2.7 Agarose gel electrophoresis

DNA was analysed by electrophoresis on 0.5-1% agarose gels depending on the size of the DNA being loaded at. Agarose was added to 1X TAE buffer and melted in a microwave. Once the agarose had cooled to about 50°C, ethidium bromide was added to a final volume of 1 µg ml^{-1} and the gel was poured. The gel was allowed to set for at least 30 minutes, then transferred to a horizontal electrophoresis tank containing TAE buffer, with the gel submerged to a depth of 2-5 mm. The sample DNA was mixed with DNA loading buffer and then added onto the gel. DNA electrophoresis was usually performed at 110 V for 30 min. Positive PCR products resulted in a single clear band in the agarose gels under UV light with no band in the negative control that did not include the template DNA.

2.8 PCR purification

PCR products were purified directly using the ISOLATE PCR and Gel Kit (BIOLINE) for removal of the remaining enzyme and primers following the manufacturer's instructions. Gel extraction was used where multiple bands were visualised by UV.

The required DNA band was excised with a clean scalpel and purified from the gel using ISOLATE PCR and Gel Kit (BIOLINE) according to the manufacturer's instructions, with excision of up to 300 mg agarose gel fragment. The gel slice was transferred to a 2 ml tube. The gel slice was dissolved by incubating it for 10 min at 50 °C with vortexing. 50 µl of Binding Optimize solution was added and vortexed. Then, 750 µl of the sample was transferred to a spin column and centrifuged at 10,000 g for 1

min and filtrate was discarded. This step was repeated by reusing the collection tube. 700 µl of Wash Buffer A was added and centrifuged at 10,000 g for 1 min and filtrate was discarded. This step was repeated by reusing the collection tube, which was centrifuged at maximum speed for 2 min. The column was placed in a 1.5 ml Elution tube and 50 µl of Elution buffer was added directly to the spin column membrane. The column was incubated for 1 min at RT and then centrifuged at 6000 g for 1 min to elute the DNA. For the isolation of PCR products, 100 µl of PCR mixture was added to spin column after addition of 500 µl of Binding buffer.

The solution was mixed well by carefully pipetting and then centrifuged (10,000 g; 2 min). The collection tube was discarded and the column placed in a 1.5 ml Elution tube. 20 µl of Elution buffer was added directly to the spin column membrane and incubated at RT for 1 min, then centrifuged (6000 g;1 min) to elute the PCR product.

2.9 Sequencing of PCR products

PCR products were treated using ExoSAP-IT (Usb.Affymetrix, Inc). ExoSAP-IT mixture was prepared by mixing 0.5 µl of Exonuclease I, 5.0 µl of SAP and 194.5 µl of distilled water. To treat the PCR product, 25 µl of the product was mixed with 10 µl of ExoSAP-IT mixture. The reaction was then performed at 37°C for 30 min then at 95°C for 5 min to inactivate ExoSAP-IT. After treating the PCR products with ExoSAP-IT, the products were sent to GATC BIOTECH http://www.gatc-biotech.com for sequencing.

2.10 Bioinformatics analysis of PCR products.

The PCR product sequences were analysed using Codon Code Aligner software http://www.codoncode.com/aligner/new.htm and the sequences were assembled using Geneious 5.0.4.

2.11 Induction of bacteriophages

To determine whether prophage could be induced to enter the lytic cycle, thereby releasing free virus, the strains were induced using chemical (norfloxacin, Mitomycin C) and physical (UV) agents. The host range of released phage was tested using 15 different indicator animal isolates of *E. faecium*.

2.11.1 Norfloxacin induction

Bacteria cultured on THB broth were diluted 10–fold in 10 ml of fresh broth and grown to an optical density of 0.6 to 0.7 at 600 nm. Norfloxacin was supplemented to broth at 1 μg ml^{-1} and incubated for 1 h at 37°C. 1 ml of the bacteria was then sub-cultured in 10 ml fresh broth supplemented with 0.01 M $CaCl_2$ and incubated for 2 h at 37°C. Finally, the phage lysate was filtered through 0.2-μm membrane.

2.11.2 UV induction

Bacteria cultured on THB were diluted 10–fold in 10 ml fresh broth, grown to an optical density of 0.4 to 0.5 at 600 nm. The cultures were centrifuged at 10844 g for 10 min and resuspended in 1 mM $CaCl_2$. The cultures then

were exposed to UV radiation (366 nm) for 40-60 s. 1 ml of the treated bacteria was then added to 10 ml of fresh broth supplemented with 1 mM $CaCl_2$ and incubated for 2 h at 37°C. Supernatants were filtered through 0.2-µm membrane. 100 µl of the host cells were mixed with 5 ml of top soft agar and poured on bottom agar. 10 µl of the filtrate was pipetted on the top agar and incubated overnight at 37°C proceeding to plaque observation.

2.11.3 Mitomycin C induction

Bacteria cultured on THB broth were diluted 10–fold in 25 -50 ml of fresh THB broth and incubated at 37°C with shaking for 3 h (OD 600 around 0.2-0.4). Mitomycin C (Sigma) was supplemented to broth at 4 µg ml^{-1} and incubated for 4 h at 37°C. Finally, the phage lysate was filtered through 0.2-µm membrane.

2.12 Phage propagation

Host bacteria were cultured overnight at 37°C in THB. 0.1 ml of phage stock solution and 0.1 ml of overnight bacterial culture were added into 3 ml of pre warmed soft agar and poured as overlay agar onto bottom agar plates. Agar was allowed to set then incubated at 37°C overnight. The plates were then observed for the plaques. For the plating method, dilutions of phage stock solutions were added to 3 ml of molten soft agar inoculated with 100 µl of log-phase culture. The mixture was poured onto bottom agar plates and incubated overnight at 30°C.

2.13 Phage lysate

Strains that contained antibiotic markers were cultured in LB broth at 37°C overnight. Lysates of the donor strain were generated by mixing 5 ml of cells (OD 600 ~ 0.5) with 5 ml of phage buffer and 50 µl (10^9 pfu ml^{-1}) of stock lysate. The mixture was incubated at 30°C until complete lysis was observed (2 - 4 hours), then it was filter sterilised and stored at 4°C.

2.14 Phage counting Plaque forming unit (PFU)

A sensitive strain was cultured in THB broth until log phase. Phage lysate was diluted in phage buffer to 10^{-7}. 100 µl of diluted phage was mixed with 200 µl of bacterial culture and 50 µl of 1 M $CaCl_2$ was added. 5 ml of phage top agar was added to phage mixture and overlaid on phage bottom agar plate. Once the phage top agar was set plates were incubated at 37°C overnight. The number of plaques were counted and pfu ml^{-1} was calculated using the formula pfu ml^{-1} = number of plaques x 10^8.

2.15 Phage Transduction

0.5 ml of recipient cells culturing overnight in LB containing of 10 mM $CaCl_2$ was added to 100 µl of phage lysate and 1 ml LB containing 10 mM $CaCl_2$. The mixture was incubated stationary at 37°C for 25 min, followed by 15 min on an orbital shaker 250rpm at 37°C. The mixture was centrifuged (13,000 rpm; 10 min) and all of the supernatant was removed. The cells were resuspended in 1 ml of 0.02 M sodium citrate and incubated on ice for 20 min. 100 µl aliquots were spread on to LB plates containing 0.05 % (w/v) sodium citrate and selective antibiotic. Plates were incubated

at 37°C for 90 min and overlaid with 5ml of LB Top agar containing selective antibiotics. Plates were incubated for 24-48 hours at 37°C. PCR amplification for antibiotic genes was performed to confirm transduction (Table 2.5).

2.16 Preparation of bacteriophage DNA; PEG precipitation/ purification

Phage DNA was purified from the free phage after grow the bacteria in 50 to 100 ml THB for overnight at 37°C and the free phage lysate was filtered through 0.2-µm membrane and stored at 4 °C. phage DNA was purified after 45 min of adding the induction agent and then the phage lysate was filtered through 0.2-µm membrane and stored at 4°C. Phage DNA was purified after 4 of adding induction agent and phage lysate was filtered and store at 4°C. PEG precipitation was carried out on the phage stock to isolate the phage DNA. 30 µl chloroform was added to each 10 ml of the phage stock to lyse any remaining bacteria. 5 µg ml^{-1} DNase and 1 µgml^{-1} RNase were added and incubated at 37°C for 4h. Bacteriophages were precipitated by incubation with 33 % (w/v) Polyethylene glycol (PEG) on ice for 30 min. Precipitated bacteriophage were then harvested by a 10 min centrifugation at 10000 rcf.

Supernatant was discarded and the pellet was resuspended in 1 ml of SM buffer. 5 µg ml^{-1} DNase and 1 µg ml^{-1} RNase were added and incubated at 37°C overnight. DNA was purified by the addition of an equal volume of equilibrated phenol:chloroform:isoamyl alcohol (25:24:1, pH 8). The mixture was centrifuged at 14500 rcf for 5 min and the resulting aqueous

phase was transferred to a new tube, this step was repeated twice. The DNA was precipitated by the addition of 0.6 volume isopropanol. Finally, the mixture was centrifuged at 14500 rcf for 30 min. The supernatant was discarded and the pellet was resuspended with 70 % (v/v) ethanol prior to resuspension in 100 µl distilled water. DNA was quantified using a Qubit fluorometer (Invitrogen). PCR amplification was used to identify the presence of the phage DNA by using phage integrase primers (Table 2.5) and the housekeeping gene *adk* and *adk2* (Table 2.6) and to determine purity of the phage DNA relative to genomic DNA.

2.17 Bacteriocin induction

Bacteria were cultured in THB at 37 °C and harvested at four different time points 2, 4, 8 h and overnight. The cultures were filtered through a 0.2 µm membrane. Host bacteria were grown overnight at 37°C in THB. 0.1 ml of overnight bacterial culture was added into 3 ml of pre-warmed soft agar and poured as overlay agar onto agar plates and allowed to set. 10 µl of cell filtrates stocks were spotted on the plates and incubated at 37°C overnight. The plates were observed for zones of growth inhibition.

For the plating method, dilutions of stock solutions were added to 3 ml of molten soft agar inoculated with 100 µl of log-phase culture. The mixture was poured onto bottom agar plates and incubated over night at 30°C. Size excision columns (Centricon plus-20) were used to discriminate between phage and bacteriocins. The stock was centrifuged at 4000 rpm for 1 h. The solution that was passed through the occlusion membrane was spotted onto

plates containing host bacteria to assay growth inhibition zones or plaques, compared with unfiltered material.

2.18 Bioinformatics tools

The following section provides a description of bioinformatics tools and resources evaluated/used during this study.

2.18.1 Sequence Analysis Tools

Basic Local Alignment search tool BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

This tool was established to discover (local) homology between two sequences. Protein and nucleotide sequence databases can be used for a given sequence of interest. This program calculates the statistical significance of an alignment (Altschul, Gish *et al.* 1990).

The BLAST algorithm has many variation; BLASTN, BLASTP, BLASTX, TBLASTN, mega BLAST and psi-BLAST. These different algorithms use are according to the query input (nucleotide, protein or translated sequences) with searches against a vast number of organism sequences.

MUMmer (http://www.tigr.org/software/mummer)

MUMmer 3.0 is open-source software that enables genome sequence comparison of large genomes. MUMmer can align incomplete genomes from a shotgun sequencing project using the NUCmer program included

with the system. The graphical viewing tools afford different ways to analyse genome alignments (Kurtz, Phillippy *et al.* 2004).

Artemis (http://www.sanger.ac.uk/Software/Artemis/v8/)

Artemis is a DNA sequence viewer and annotation tool that allows visualisation of sequence features and the results of analyses within the context of next generation data.

CLUSTALW (http://www.ebi.ac.uk/Tools/msa/clustalw2/)

CLUSTALW (1.83) is one of the most powerful programs used to achieve multiple sequence alignments. This program allows the presentation of multiple nucleotide and protein sequence alignments (Larkin, Blackshields *et al.* 2007).

MUSCLE (http://www.drive5.com/muscle/)

MUSCLE (v3.6) is a computer program most widely used in biology to create multiple sequence alignments of proteins. MUSCLE uses different algorithms including fast distance estimation and progressive alignment. The accuracy and speed of the program is better than CLUSTALW, since hundreds of sequences can be aligned in seconds (Edgar 2004).

FigTree (http://tree.bio.ed.ac.uk/software/figtree/)

FigTree (v1.3.1) is a program for graphical viewing of phylogenetic trees. The program was designed to show summarized and annotated trees formed by BEAST.

FastTree (http://www.microbesonline.org/fasttree/)

FastTree (v2.1.7) is an open-source software construct, which can infer maximum likelihood phylogenetic trees from alignments of nucleotide or protein sequences. Millions of alignments can be done in a reasonable amount of time and memory (Price, Dehal *et al.* 2010).

Geneious (http://www.geneious.com/)

By using Geneious (v7.1.3) software, one can analyse integrated protein and DNA sequences, perform BLAST and get access to public databases. The most powerful analysis that can be done using this software is the sequence alignments manageability for both pair-wise and multiple sequence alignments and visualization of the sequence alignments. The alignment results can be viewed as phylogenetic trees.

OrthoMCL (http://www.orthomcl.org/orthomcl/?rm=orthomcl)

OrthoMCL (v1.4) is one of the most commonly used programs to perform identification of orthologous groups. In addition, access to these groups is extremely important for study gene/protein evolution and comparative genomics and genome annotation.

All against All BLASTP between species and within species with Markov Cluster algorithm methods can be performed to find all orthologous groups with any recent paralogs. Ortholog analysis by using OrthoMCL can be applied with two genomes or it can be extensive to cluster orthologs from multiple species in order to constructing orthologous groups (Li, Stoeckert *et al.* 2003).

Mauve (http://gel.ahabs.wisc.edu/mauve)

Mauve (v2.3.1) software is a powerful package applied to determine the presence of rearrangements and horizontal transfer in a genome. It is used for the identification and alignment of conserved genomic DNA (Darling, Mau *et al.* 2004). Mauve alignments were used in this study to draw comparison between whole genomes as well as examine the reasons of rearrangements within genomes of *E. faecium*.

BRIG (http://sourceforge.net/projects/brig/)

The BLAST Ring Image Generator BRIG (v1.0) is a desktop application written in Java 1.6. This application was used in genome comparisons and generates a circular image for the genome. The comparison in this application depends on the Basic Local Alignment Search Tool (BLAST) and CGView for image rendering. For generating genomes maps in BRIG in this study DNA or protein files were used.

MeV (http://www.tm4.org/mev.html)

Multi experiment Viewer MeV (v10.2) is a beneficial microarray data analysis tool, including high-level algorithms for statistical analysis, classification, clustering, visualization, and biological argument discovery (Chu, Gottardo *et al.* 2008). MeV was used in this study for clustering orthologous groups and for cladogram analysis.

Unipro UGENE (http://ugene.unipro.ru)

UGENE (v1.11.5) is open-source software that can be used as a multiplatform software. It offers visualization of annotated genome sequences, multiple sequence alignments and phylogenetic trees (Okonechnikov, Golosova *et al.* 2012). In this study UGENE software was used to identify and map the repetitive units in the genomes.

Phenolink (http://bamics2.cmbi.ru.nl/websoftware/phenolink/)

Phenolink is a web-tool to identify genetic links between phenotypes. It uses ~omics technologies that connect phenotypes with high-throughput molecular biology information. The purpose is to see through cellular mechanisms underlying an organism's phenotype (Bayjanov, Molenaar *et al.* 2012). A default parameter was used to identified *E. faecium* phenotypes.

C**RISPRs Finder (http://crispr.u-psud.fr/Server/CRISPRfinder.php)**

CRISPRFinder is a free access web service. CRISPRs stands for Clustered regularly interspaced short palindromic repeats. Five tools are available in CRISPRs Finder, which can be used for:

1. Detecting very short CRISPRs that consist of one or two motifs.
2. Identifying highly conserved regions (DR) and extracting similarly sized unique sequences, which lie between the DRs called spacers.
3. Obtaining the AT-rich leader sequence, which flanks the CRISPR cluster on one side.
4. To do BLAST searches to look for spacers in the Genbank database.

5. To identify the highly conserved regions (DR) are present in other prokaryotic sequenced genomes (Grissa, Vergnaud *et al.* 2007).

Island Viewer (http://www.pathogenomics.sfu.ca/islandviewer)

IslandViewer is a freely accessible web service that provides detection of gene clusters likely to be of horizontal origin, called Genomic islands (GIs). These clusters contain genes such as virulence, antibiotic resistance or other important adaptation genes. IslandViewer uses a graphical interface that allows easy viewing and the island data of both the chromosome and the gene level can be downloaded. The server uses three methods to identify the GI regions. IslandPick; comparative genomic GI prediction method to advance stringent data sets of GIs and non-GIs, SIGI-HMM; This method measures codon usage to identify possible GIs by using Hidden Markov Model (HMM). Finally, IslandPath-DIMOB; this method visualises several common characteristics of GIs such as abnormal sequence composition or the occurrence of genes that are functionally related to mobile elements (Langille and Brinkman 2009).

PHAST (http://phast.wishartlab.com)

PHAST is a fast web server used to distinguish, annotate and graphically present prophage sequences and prophage features within bacterial genomes or plasmids (Zhou, Liang *et al.* 2011).

IS Finder (https://www-is.biotoul.fr//)

IS Finder is a database provides a list of insertion sequences elements isolated from Eubacteria and *Archaea*. The IS elements in this database are defined in individual files which contains their general features such as name, size and family plus their DNA and protein sequences. In addition, for the comparison an on-line BLAST search is available.

2.18.2 Databases and Genome Resources

NCBI (http://www.ncbi.nlm.nih.gov/)

The NCBI server provides a wide range of bioinformatics tools. Inside the molecular databases there are nucleotide, protein, structure, taxonomy, genome, expression and chemical databases. In addition, NCBI offers a literature database, which includes research articles (e.g. PubMed) and pools of reference overviews. BLAST, genome map viewer and ORF finder are the tools available in NCBI.

EBI-EMBL (http://www.ebi.ac.uk/)

The European Bioinformatics Institute (EBI) research centre and bioinformatics service provides and hosts literature, pathway, sequence, networks, microarray and ontology databases. In addition, it offers some of the most recognized EBI tools such as UniProt, Ensembl, ArrayExpress, Biomart and InterPro.

Antibiotics Resistance Database

Resfinder (http://cge.cbs.dtu.dk/services/ResFinder/)

ResFinder is a database used to identify the antimicrobial resistance genes. BLAST for identification of acquired antimicrobial resistance genes in whole genome data is the main method that is used in this database (Zankari, Hasman *et al.* 2012).

CARD: (http://arpcard.mcmaster.ca)

The comprehensive Antibiotics Resistance Database (CARD) is a tool used to analyse the genetics and genomics of antibiotic resistance and to identify antibiotic resistance genes in new unannotated genome sequences (McArthur, Waglechner *et al.* 2013).

Virulence factors database

VFDB (http://www.mgc.ac.cn/VFs/)

The virulence factors database (VFDB) is an integrated and comprehensive resource of virulence factors for bacterial pathogens. Two different tools regular BLAST and PSI/PHI BLAST, can be used to identify: offensive virulence factors with roles for adherence, invasion and toxins; defensive virulence factors such as secretion systems type III, IV, VI and VII and autotransporter type V; nonspecific virulence factors such as iron uptake systems, magnesium transport and exoenzymes; and finally, the regulation of virulence-associated genes.

2.19 Genome sequencing

The genome of three vancomycin-resistant animal *E. faecium* isolates has been sequenced by whole genome shotgun using 454 pyrosequencing. The pyrosequencing were performed by generating standard fragment template 8 Kb DNA libraries, which were multiplex identifier (MID) tagged to allow multiple samples to be run in a single plate region, using the GS-FLX 454 Life Sciences through The Center for Genomic Research (CGR) in Liverpool University.

The genome *E. faecium* isolated from calf has been sequenced by Pacific Biosciences PacBio RS. A total of 56 µg μl^{-1} of DNA was sent The Center for Genomic Research (CGR) in Liverpool University, where a single 10 kb SMRT-bell sequencing library (Pacific Biosciences) was constructed. The SMRT-bell library was sequenced using 2 SMRT cells (Pacific Biosciences).

2.19 Structural and functional annotation

Genome annotations were managed using RAST server (http://rast.nmpdr.org) and IMG/ER (Integrated Microbial genomics) (https://img.jgi.doe.gov). Gene structure was assigned by the automated gene-calling algorithm, Prokka (version1.8) (http://www.vicbioinformatics.com/software.prokka.shtml) using default parameters. To validate the prokka gene prediction, the open reading frames (ORF) were compared to published sequences using BLASTn. After the gene-finding progression, different types of investigation were made in order to predict the function of the encoded proteins.

BLAST search algorithm was used to examine the homology of the putative ORFs (DNA and protein). Functional classification of ORFs was based on homology search against COGs. Protein function annotation was constructed based on the homology search against NCBI protein database.

2.20 Genome map

The BLAST Ring Image Generator BRIG version 0.95 (http://sourceforge.net/projects/brig/) was used to create circular plots for visualising *E. faecium* genomes.

Gene bank file was used. The map had the information of gene name and the start and end positions within the genome. The program was performed using BLASTp with an upper identity threshold of 70 % and lower identity threshold of 50 %.

2.21 Ortholog analysis

Orthogroups are genes that probably have the same function and possibly some paralogs. Paralog is a duplication of gene that has acquired new functions. The occurrence of orthogroups across all of the genomes were determined using OrthoMCL, with a threshold BLAST e-value of 10^{-5}.

2.22 Phylogenetic construction

A phylogenetic tree of all *E. faecium* genomes was calculated using a distance method based on pairwise protein sequence alignments using

Geneious software. Rapid bootstrapping option for nucleotide sequences, using 1000 bootstrap replicates was used.

In order to maximise resolution on the tree, we used all single-copy core orthogroups in our *E. faecium* genomes. Both protein and nucleotide sequence trees were established for both core genes and accessory genes to compare the relationship within the branches in the phylogenetic tree. This was done to check and compare the two trees and make sure of the primary reason for that drive the clade. The phylogenetic trees were inferred by both neighbor-Joining, and split decomposition analysis. Phylogenetic trees were edited with Fig Tree , which is a graphical viewer of the phylogenetic tree.

Core genome phylogeny, firstly, OrthoMCL was performed. Then, core genes were defined by selecting genes, which were present only one gene in each strain. The sequence alignments of those genes were conducted using MUSCLE, and then they were trimmed and concatenated. A core phylogenetic tree was constructed using fasttree, with bootstrapping supports obtained from seqboot.

2.23 Pan genome analysis

Compute pan-genome and core-genome sizes and their evolutions for a genome set were determined using the R project for statistical computing using (gplots package). In addition, the common and variable genome proportion for each group of *E. faecium* genome was detected. The pan-genome analysis is computed using the OrthMcl results. If an orthologue is associated with every compared genome, this orthologue is a part of the core-genome. If an orthologue is associated with 1 > n of the compared

genomes, it is a part of the variable-genome. If an orthologue is not clustered with any compared genomes, it is a singleton and is a part of the variable-genome. The size of the core and pan-genomes was estimated by fitting an exponential curve through medians.

2.24 Phage identification

Prophage genomes were obtained from the sequence of their hosts that were available from the NCBI database and were predicted from these genomes using the PHAST algorithm. One complete prophage of *E. faecium* IME-EFm1 was reported previously (Wang, Wang *et al.* 2014). To predict phage-related genes in each genome, Artemis and BLAST were used to compare genes against the PHAST database.

2.24.1 Sequence clustering and phylogenetics

Mauve progressive alignments to determine conserved sequence segments most likely to be conserved in recombinational events were determined using the Mauve algorithm. Alignments of specific genes were done using Geneious. The phylogenetic trees of several selected genes were constructed with Geneious using the Neighbor-Joining algorithm. Trees were bootstrapped for 1000 times. Tree was visualized using FigTree.

2.24.2 Putative prophage attachment sites

In the lysogenic isolate the prophage is expected to be bordered by short directly repeated sequence (the attL, attR of the prophages).

Consequently, to detect the putative attachment sites, genomic sequences of the lysogenic *E. faecium* strains were analysed for the presence of directly repeated sequences flanking the prophages using Unipro UGENE.

Chapter Three: Genome sequencing of three animal isolates of *Enterococcus faecium.*

3.1 Introduction

Bacterial diseases represent a major source of morbidity and mortality amongst humans and animals. Pathogenic bacteria comprise a diverse range of species, which have discrete virulence mechanisms. A good knowledge and understanding of these mechanisms is necessary to design successful new therapies against bacterial diseases and manage the emergence of novel isolates. The design of therapies is limited due to the extent of information about the pathogenesis of some diseases being limited or non-existent (Donkor 2013).

Genome sequencing, combined with interpretation using bioinformatic analyses of genome data, has dramatically extended our understanding of bacterial pathogens, particularly with respect to their ecology, evolution, and pathogenesis (Tang and Holden 1999, Donkor 2013). Doolittle (1999) states that the ability to exploit complete genome sequences of microbes offers many opportunities for medicine and delivers an abundance of knowledge for interrogating evolutionary networks. Greater than 1,800 bacterial genomes, including the majority of bacterial pathogens, have now been completely sequenced (Ribeiro, Przybylski *et al* . 2012). The resource of sequenced genomes and the direct access to genome data have advanced studies in biology and has given birth to a new science called genome-based biology (Garcia-Vallve, Romeu *et al* . 2000).

The typical bacterial genome consists of a single circular chromosome, however there are exceptions, with several medically significant bacteria

having two or more chromosomes, including *Burkholderia*, *Brucella*, *Vibrio*, and *Leptospira* species; several species have linear chromosomes, for example *Borrelia burgdorferi* (Guzman, Romeu *et al* . 2008). Allen *et al* (2006) indicated that the majority of bacterial genomes are smaller than 5 Mb in size, although species have been described with genomes up to 30 Mb, for example *Bacillus megaterium* (Allen, Price *et al* . 2006).

Guzman *et al* (2008) establish that the difference in bacterial genome size appears related to lifestyles, whereby obligate pathogen species have smaller genomes than parasitic species, which in turn have smaller genomes than free-living species. The nucleotide composition in bacterial genomes varies across bacteria. The GC (guanosine-cytosine) content may differ locally within a genome, but it is relatively constant within a bacterial genus and species, varying from ~25% GC in *Mycoplasma spp* to ~75% in *Micrococcus* species. The variation in GC content within a single genome was used to determine the acquisition of genomic portions by horizontal gene transfer, classically pathogenicity islands, since these frequently have a different GC ratio (Walk, Alm *et al* . 2007).

On average, a bacterial genome comprises around 2,500 genes. The genome encodes all of the biochemical functions that are necessary for survival of an individual species, and additionally those functions necessary for virulence within the genome of pathogenic bacteria. Bacterial genomes contain few non-coding regions (Jacob and Monod 1961, Allen, Price *et al* . 2006).

Between closely related organisms (based on phylogenetic distances) the gene content and gene order are well-conserved, however, among more distantly related organisms it becomes less conserved (Guzman, Romeu *et al* . 2008). An evolutionary tree of microorganisms can be constructed from comparative analyses of the nucleotide sequences of genes encoding ribosomal RNAs or core genome proteins, such as, CTP synthetase and the cell adhesion protein FtsY(Pennisi 1998).

Specific Aims

The aims were to sequence the genomes of three vancomycin-resistant isolates of *E. faecium* from chicken, calf and pig using next generation pyrosequencing on the Roche 454 titanium platform. These genomes were selected specifically to investigate host adaptation in mammalian hosts. A further aim was to attempt closure of gaps in one of these genomes to produce a closed *E. faecium* from animals. This would enable comparative genomics.

3.2 Results

3.2.1 Genome sequencing and assembly

The genome sequences of *E. faecium* strain E429, isolated from chicken, strain E172, isolated from calf and strain E142, isolated from pig, were determined using the GS-FLX sequencing platform (454 Life Sciences), as described in section 2.19.2. The insert library representing each genome was sequenced extensively to provide reads for each *E. faecium* isolate of 849,986, 366,122 and 335,440, respectively for E429 (chicken), E172 (calf)

and E142 (pig) (Table 3.1). For each respective strain, these reads were assembled into 922, 786 and 136 contigs respectively. The longest scaffold gives the best approximation for the size of the three genomes although the number of scaffolds obtained for strain E429 (chicken), E172 (calf) and E142 (pig) were 19, 18 and 3. The chromosome of the animal strains of *E. faecium* varies in size, therefore, from approximately 3.38 Mb in the chicken strain to 2.94 Mb in the calf strain and 2.52 Mb in the pig strain, with a GC-content of 38.75%, 38.67% and 38.13%, respectively (Table 3.1). Associated with each genome assembly are 62, 67 and 55 tRNAs respectively for strain E429 (chicken), E172 (calf) and E142 (pig) and markedly different numbers of ribosomal genes with 11 rRNAs (1 x 5S, 3 x 16S and 7x 23S), 14 rRNAs (2 x 5S, 4 x 16S and 8 x 23S) and 3 rRNAs (1 of 5S, 1 of 16S and 1 of 23S), respectively for chicken, pig and calf.

3.2.2 Annotation of the *E. faecium* genome animal strains

The genomes of the chicken, calf and pig strains were annotated using IMG-ER (Integrated Microbial Genomes Expert Review). The initial annotation analysis identified 3,574, 2,892 and 2,641 protein coding genes in chicken, calf and pig, respectively. Approximately 2%, 2.72% and 2.15% of the genes in the animal strain genomes, respectively, determine structural RNAs. The remaining 98% of predicted ORFs in strain E429 (chicken), 97.28% in strain E172 (calf) and 97.85% in strain E142 (pig) were studied using homology analyses with sequence databases, which identified that 74% (2,708), 78% (2,325) and 79% (2,147) of the predicted ORFs,

respectively, were likely to be functional proteins. Nearly 10% of the genomes are non-AGCT bases.

Table 3.1: Structural features associated with the sequenced genomes of *E. faecium* strains E429, E172 and E142.

Genomic features	E429 (Chicken)	E172 (Calf)	E142 (Pig)
Estimated genome size	3.4 MB	2.9 MB	2.5 MB
Number of scaffolds	19	18	3
Shortest scaffold (bp)	2063	2053	3397
Largest scaffold (bp)	2868347	2291364	2454786
N50 scaffold size	2868347	2291364	2454786
Number of contigs scaffolded	179	204	85
Number of contigs scaffold bases	2984142	2678841	2500548
Scaffold G + C content	38.1%	38.2%	38.1%
Non-ACGT bases	321618	398256	106125
Number of contigs	922	786	136
Shortest contig (bp)	101	102	101
Largest contig (bp)	96987	186193	190330
Total number of assembled bases	3383541	2948249	2525775
N50 large contig size	33454	28565	47222

3.2.3 General genome features of the three animal strains of *E. faecium*

The general genome features of each animal strain from 454 sequence data analysis are described in the following table.

Table 3.2: Genome composition features of strains E429, E172 and E142.

Feature	E429 (chicken)		E172 (calf)		E142 (pig)	
	Number	%	Number	%	Number	%
DNA, total number of bases	3383541	100	2948249	100	2525775	100
DNA coding number of bases	2700854	79.8	2310519	78.3	2156842	85.3
DNA G+C number of bases	1311102	38.7	1140083	38.6	963197	38.1
Genes total number	3647	100	2973	100	2699	100
Protein coding genes	3574	98	2892	97.2	2641	97.8
Protein coding genes with function prediction	2708	74.2	2325	78.2	2147	79.5
Protein coding genes without function prediction	866	23.7	567	19.0	494	18.3
Protein encoding enzymes	666	18.2	639	21.4	608	22.5
Protein coding genes connected to KEGG pathways	735	20.1	720	24.2	676	25.0
Protein coding genes connected to KEGG Orthology (KO)	1332	36.5	1280	43.0	1214	44.9
Protein coding genes with COGs	2437	66.8	2186	73.5	2056	76.1

3.2.4 Ribosomal genes

Within the genome isolated from a chicken (E429), two copies of 23S rRNA are identical and five copies differ by 6 to 11 nucleotide bases and the three copies of 16S rRNA differ by 4 to 5 nucleotide bases.

Four genes of 23S rRNA are identical in the calf genome (E172) and three are different by 7 nucleotide bases while one copy differs by 440 nucleotide bases. The four copies of 16S rRNA range in size from 1332 to

1561bp.BLAST against RNA genes found in *E. faecium* genomes showed that most of the 23S rRNA found in animal *E. faecium* genomes are unique. One RNA operon was found in each animal genome of animal *E. faecium*, comprising 23S, 16S, 5S and at least one tRNA, while most of the rRNA genes were found at the end of the genome assemblies and surrounded by phage genes, transposase and insertion elements which suggested that the rRNA genes were not assembled correctly (Figure 3.1).

Table 3.3 **A**: Comparative genome features of *Entercoccus* species retrieved from the Integrated Microbial Genomes database. The table displays the variation in copy number of rRNAs genes among a selection of *Entercoccus* species genomes.

Genome Name	Size (Mb)	Protein coding genes	rRNA Genes	5S	16S	23S
Enterococcus sp. 7L76	3.09	2348	3	1	1	1
E. faecalis V583	3.35	3390	12	4	4	4
E. faecalis Symbioflor 1	2.81	2808	12	4	4	4
E. faecalis 62	3.10	3094	12	4	4	4
E. casseliflavus EC20	3.42	3189	15	5	5	5
E. hirae ATCC 9790	2.85	2845	18	6	6	6
E. faecium DO	2.83	3148	16	4	6	6
E. faecium Aus0004	2.96	2934	18	6	6	6
E. faecium NRRL	2.84	2772	18	6	6	6

Table 3.3 **B**: Comparative genome features of *E. faecium* strains retrieved from the Integrated Microbial Genomes database. *refers to closed genomes. The table displays the variation in copy number of rRNAs genes among a selection of *E. faecium* isolates from humans (clinical and commensal strains) compared with animal strains.

Genome Name	Size (Mb)	Protein coding genes	GC%	rRNA Genes	5S	rRNA 16S	23S	tRNA Genes	Source
E429	3.38	3647	39	11	1	3	7	62	Chicken
E172	2.9	2973	39	14	2	4	8	67	Calf
E142	2.52	2699	38	3	1	1	1	55	Pig
1,141,733	2.92	2829	38	3	1	1	1	43	Human clinical
Com15	2.80	2786	38	3	1	1	1	59	Human commensal
***DO**	2.83	3148	38	16	4	6	6	31	Human clinical
***Aus0004**	2.96	2934	38	18	6	6	6	47	Human clinical
***NRRL**	23.4	2304	38	3	1	1	1	52	Milk

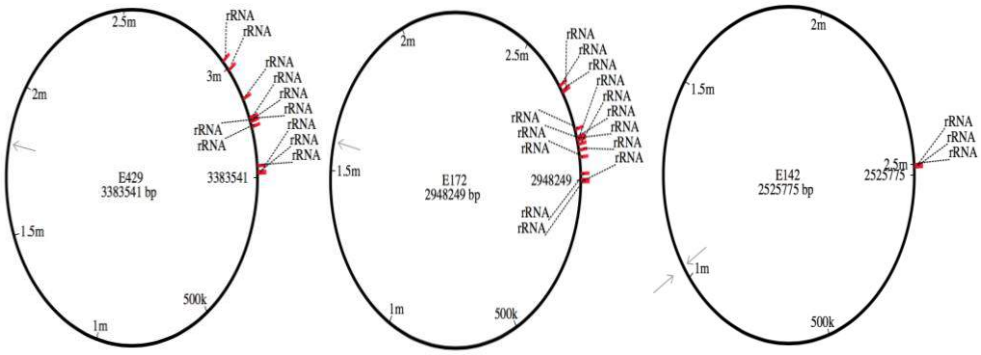


Figure 3.1: Syntenic ribosomal rRNA gene organisation in the genomes of chicken (E429), calf (E172) and pig (E142) strains**.**

3.2. 5 GC- content

The GC-content of most sequenced isolates of *E. faecium* is 38%. The *E. faecium* genomes isolated from chicken (E429) and calf (E172) have a slightly higher G+C content of 39% (Table 3.3B). Across the genus the G+C content varies from 35-43% (Table 3.4).

Table 3.4: Genome features of *Enterococcus* species retrieved from Integrated Microbial Genomes database https://img.jgi.doe.gov/cgi-bin/er/main.cgi?logout=1.

Genome Name	GC %
***E .durans* ATCC 6056**	38
***E .faecalis* 02-MB-P-10**	37
***E. faecium* Aus0085**	38
***E .flavescens* ATCC 49996**	42
***E .gilvus* ATCC BAA-350**	41
***E .haemoperoxidus* ATCC BAA-382**	36
***E .hirae* ATCC 9790**	37
***E .italicus* DSM 15952**	39
***E .malodoratus* ATCC 43197**	40
***E .moraviensis* ATCC BAA-383**	36
***E .pallens* ATCC BAA-351**	40
***E .phoeniculicola* ATCC BAA-412**	36
***E .raffinosus* ATCC 49464**	39
***E .saccharolyticus* 30_1**	41
***E .sulfureus* ATCC 49903**	38
***E .villorum* ATCC 700913**	35
***E. avium* ATCC 14025**	39
***E. caccae* ATCC BAA-1240**	36
***E. casseliflavus* 14-MB-W-14**	43
***E. cecorum* DSM 20682**	36
***E. gallinarum* EG2**	41
***E. mundtii* ATCC 882**	38

3.2.6 Genome synteny

The evolutionary relationships between organisms and a prediction of gene function can be examined by a comparison of gene order between genomes. Multi-gene regions with conserved DNA sequence and gene order are described as having genome synteny (Bentley and Parkhill 2004). A comparison of gross organisation of the genomes using the software package Mauve, which is a multiple-genome alignment program and visualiser, identifies locally collinear blocks of DNA (LCBs). These blocks correspond to regions of the chromosome devoid from genome

rearrangements. The blocks reveal that the genome isolates from a chicken (E429) has gene clusters that are organised in the reverse complement in the calf (E172) and pig (E142) genomes (Figure 3.2). The majority of homologous genes in the calf genome (E172) and pig genome (E142) are located as collinear clusters. One explanation for the extent of inversion present in the chicken strain could be that repetitive sequences in the genomes were a driver for recombination events.

Species of *Enterococcus* show varying degrees of synteny based on their overall protein sequences and gene order comparing different species of *Enterococcus* with *E. faecium* this synteny varies from extensive (*E. hirae*, *E. durans*, *E. mundtii* and *E. villorum*) to minimal (*E. caccae*, *E. haemoperoxidus*, *E. gallinarum*, *E. casseliflavus*) (Figure 3.3). This comparison was performed using the complete Aus0004 *E. faecium* genome and some of the genomes used in these comparisons are fragmented and this may affect the apparent synteny of the compared genomes.

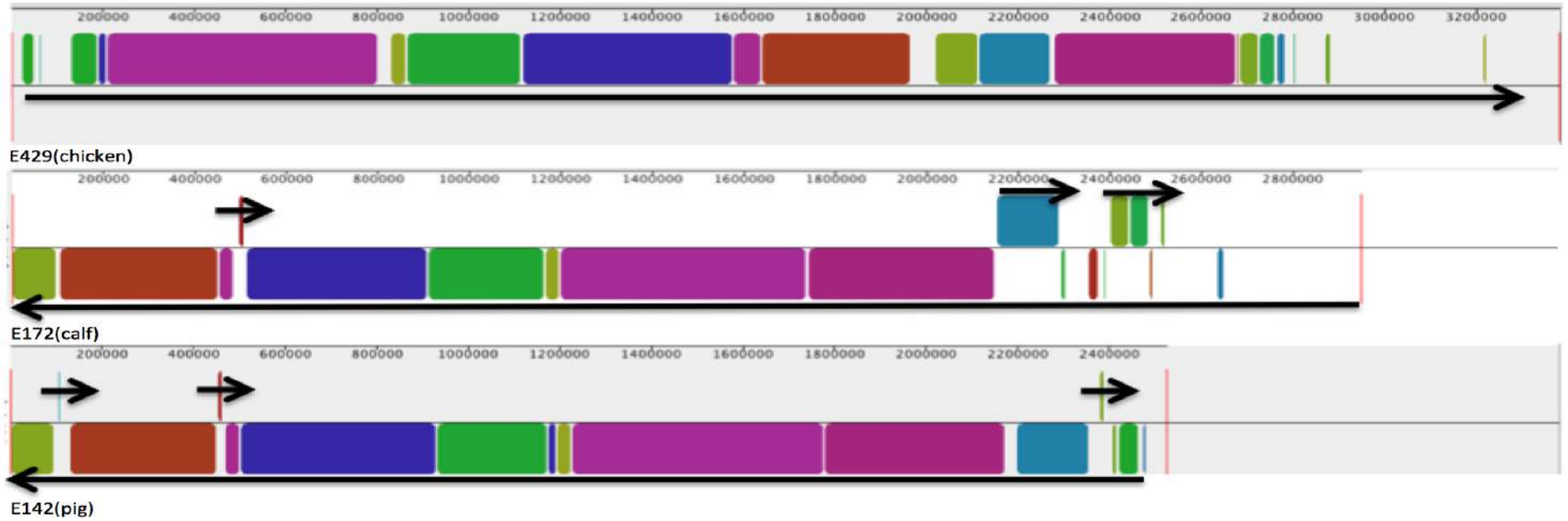
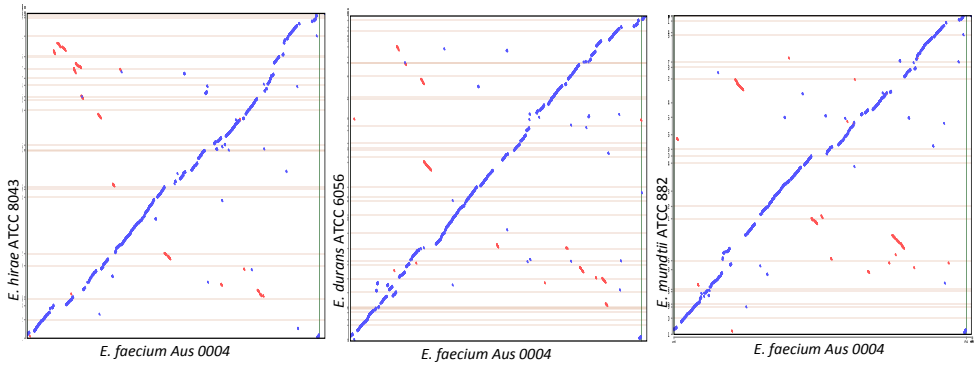
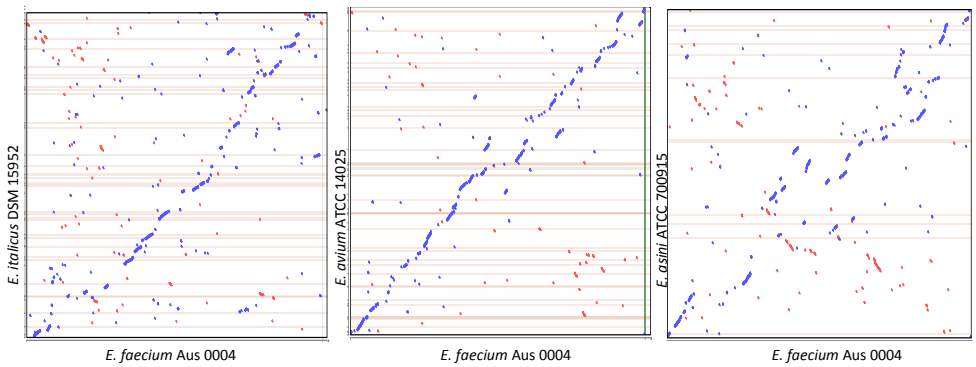


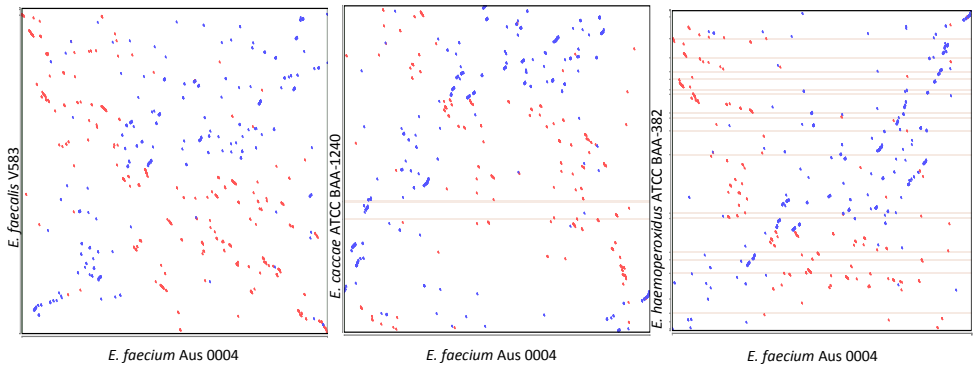
Figure 3.2: Locally Collinear Blocks (LCBs) identified in a comparison of *E. faecium* animal genomes. Each contiguously coloured region is a locally collinear block of homologous backbone sequence. LCBs below the centreline are in the reverse complement orientation relative to the reference genome (E429). The black arrows show the orientation in the LCBs compared to the reference genome.



A.



B.



C.

Figure 3.3: Genome synteny between *E. faecium* Aus0004 and other *Enterococcus* species. **A.** Mummer plot identifies a high degree of relatedness based on the overall protein sequence homology and gene order between the complete genome of *E. faecium* Aus0004 and the genomes of *E. hirae* ATCC 8043, *E. durans* ATCC 6056 and *E. mundtii* ATCC 882.

B. Mummer plot identifies a lesser degree of relatedness based on their overall protein sequence homology and gene order between the complete genome of *E. faecium* Aus0004 and the genomes of *E. italicus* DSM 15952, *E. avium* ATCC 14025 and *E. asini* ATCC 700915. **C.** Mummer plot identifies a low degree of relatedness based on their overall protein sequence homology and gene order between the complete genome of *E. faecium* Aus0004 and the genomes of *E. faecalis* V583, *E. caccae* ATCC BAA-1240 and *E. haemoperoxidus* ATCC BAA-382. The blue dashed line represents the homology between the two strains. The red dashed lines represent inverted regions between the two strains. X-axis shows Aus0004 genome. Y-axis shows the *Enterococcus* species genomes.

3.2.6 Genome inversion in *E. faecium* genomes

Possible explanations for the large inversions in the human strains, Aus0004 and DO, relative to the animal strains E429 (chicken), E172 (calf) and E142 (pig) were examined. Using the software package Mauve, several IS elements were located at the boundary of each collinear block. There is an apparent inversion in the regions in both Aus0004 and DO genomes respectively, when the circular chromosome is taken into account.

Blocks 1 and 3 are bordered by an integrase gene in the genome of Aus0004, to each side (positions 722300-723217 and 2211763-2212716) and these genes are 917 bp and 953 bp in size respectively (red arrows) these could explain the region 1 and 3 inversion in the genome. In addition, several integrase and IS elements were also spread adjacent to boundaries of

the blocks: ISEfa7 (position 604818-606263) and 1.4 kb in size; integrase (position 303078-304031) and 953 bp in size; ISEfm1 (positions 297341 - 298102 and 286975 – 287883) and 761 bp and 908 bp in size, respectively, plus IS1251 in position 44706 - 45896 and 1.1 kb in size. The inversion of the chromosome in block 2 of the Aus0004 genome could have occurred due to adjacent prophages (presented as orange arrows) (Figure 3.4).

An explanation for the inversion observed in the animal genomes E429 (chicken), E172 (calf) and E142 (pig) compared with the DO strain could be recombination due to due to transposases at the boundaries of the inverted section of chromosome. Both transposases are identical in size, (953 bp) and are located immediately adjacent (719177 - 720130 and 2110894 – 2111847) (blue arrows) (Figure 3.4). A further copy of this transposase is located at 291446 - 292399.

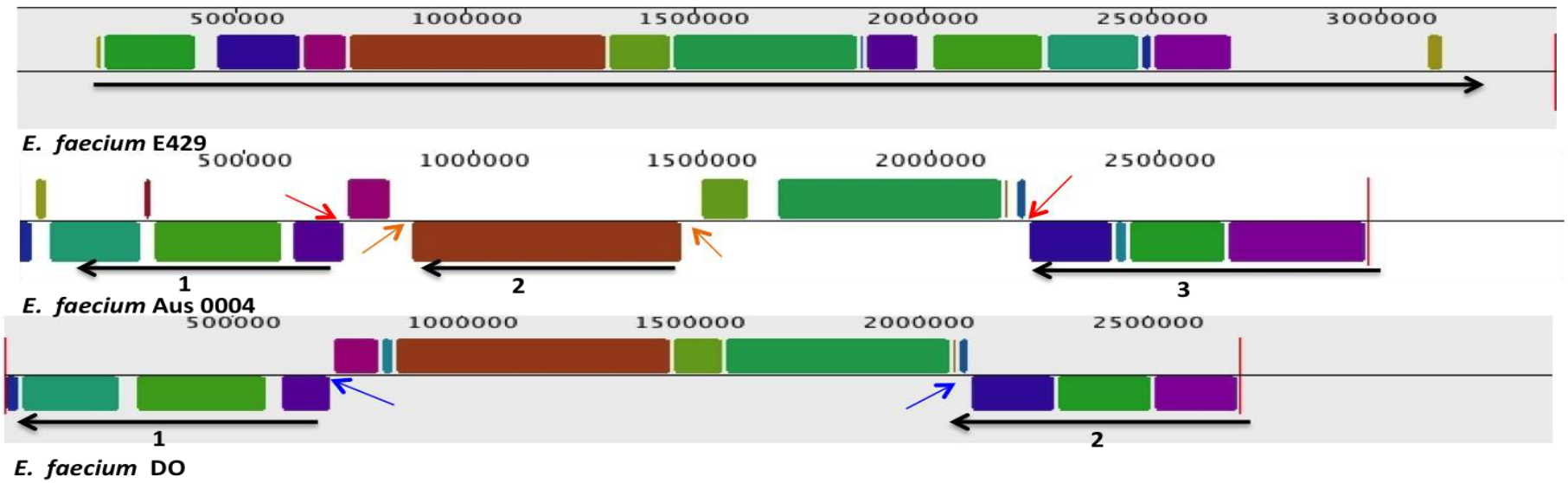
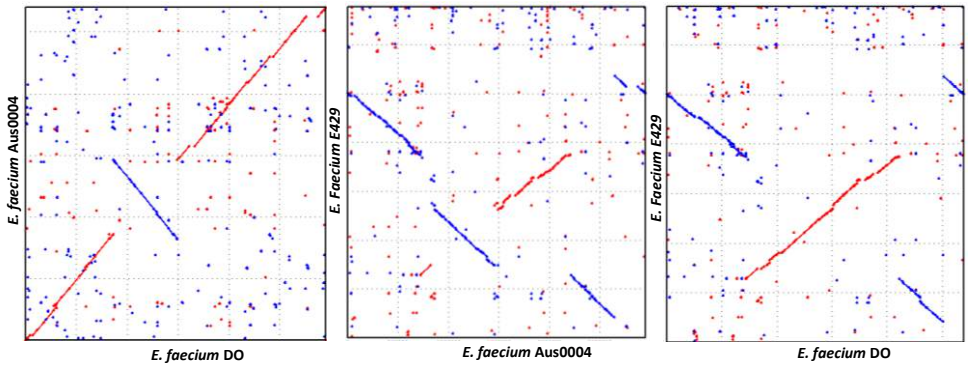


Figure 3.4: Locally Collinear Blocks (LCBs) identified among the *E. faecium* genome isolates from a chicken and the complete genomes Aus0004 and DO. Each contiguously coloured region is a locally collinear block of homologous backbone sequence. LCBs below a genome's centreline are in the reverse complement orientation relative to the reference genome (E429). The black arrows show the orientation of the LCBs compared to the reference genome. Red arrows show the location of the integrase in the genome of Aus0004. Orange arrows show the presence of prophages in the genome of Aus0004. Blue arrows show the transposons located in the genome of DO strain.



A. **B.** **C.**

Figure 3.5**:** Genome synteny of *E. faecium.* Mummer plot shows the existence of a large inversion within *E. faecium* strains. **A.** Mummer plot shows the existence of the inversion within the two complete genomes Aus0004 and DO strain. X-axis shows DO genome. Y-axis shows the Aus0004 genome. **B.** Mummer plot shows the existence of inversion within the complete genome Aus0004 and chicken strain (E429). X-axis shows the Aus0004 genome. Y-axis shows E429 genome. **C.** Mummer plot shows inversion exists within the complete genome DO and the chicken strain (E429). X-axis shows DO genome. Y-axis shows the E429 genome. The plots present the homology between the two strains.

3.2.7 Repetitive sequence elements in the sequenced *E. faecium* genomes

Many bacterial genomes have been described to contain repetitive DNA. These repeat sequences are typically 400 bp in size (Delihas 2011). Analysis of the genomes of the animal *E. faecium* strains using the software package Unipor UGENE determined that there were 1885, 1758 and 1422 short tandemly repeated sequences (STRs) in the chicken, calf and pig strains,

respectively (Figure 3.6). These STRs have a repeat length of 3 bp and tandem size from 9-10 bp. In addition, 750, 550 and 285 short sequence repeats (SSRs) were found in strains E429, E172 and E142, respectively; with a minimum repeat length of 15 bp and a distance between the repeats of 2 bp to 2000kb (Figure 3.7).

UNIPOR-UGENE displays approximate repeat sequences found in the DNA sequence. The repetitive sequence elements in the animal *E. faecium* genome sequences have a high sequence identity and high copy number. The observed genome inversions could be derived from these repeats.

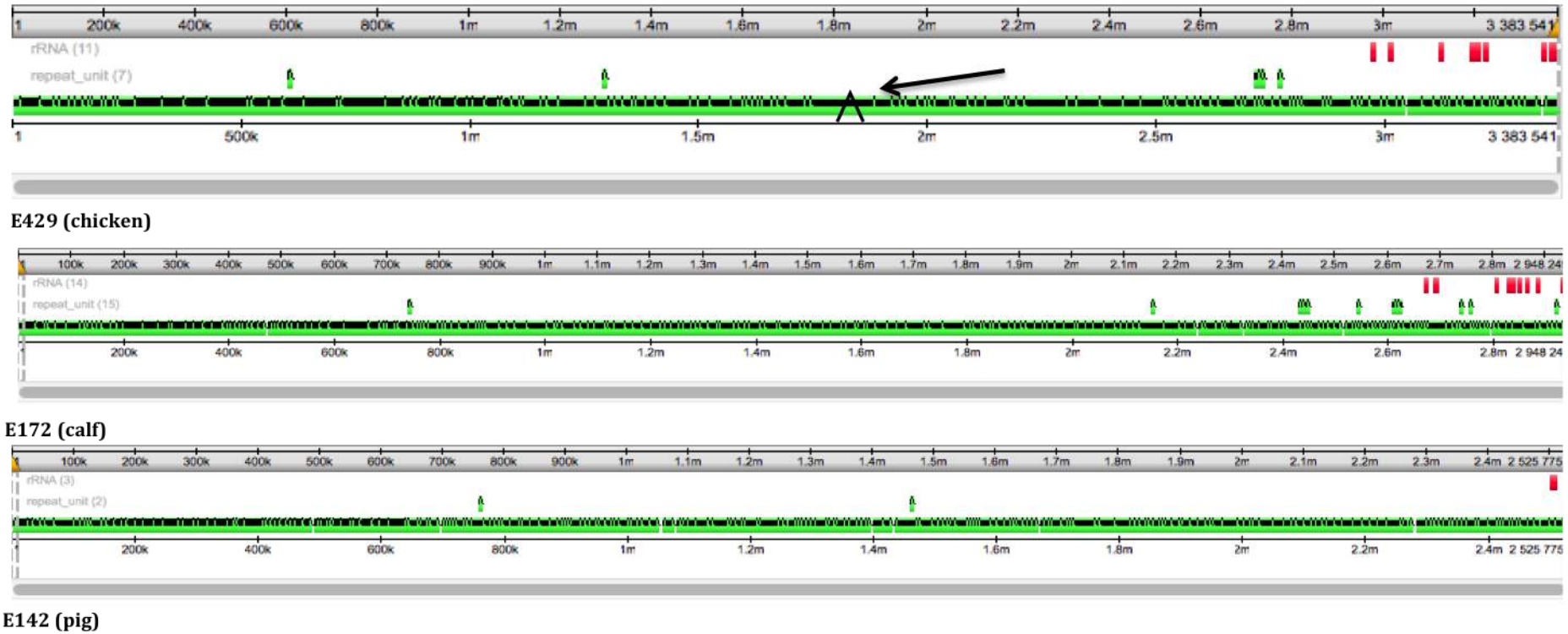


Figure 3.6: Short tandemly repeated sequence (STRs) in animal *E. faecium* strains. STRs covering almost the whole genome of chicken, calf and pig. STRs annotations are located side by side in green (black arrows) and red verticals show rRNA operons.

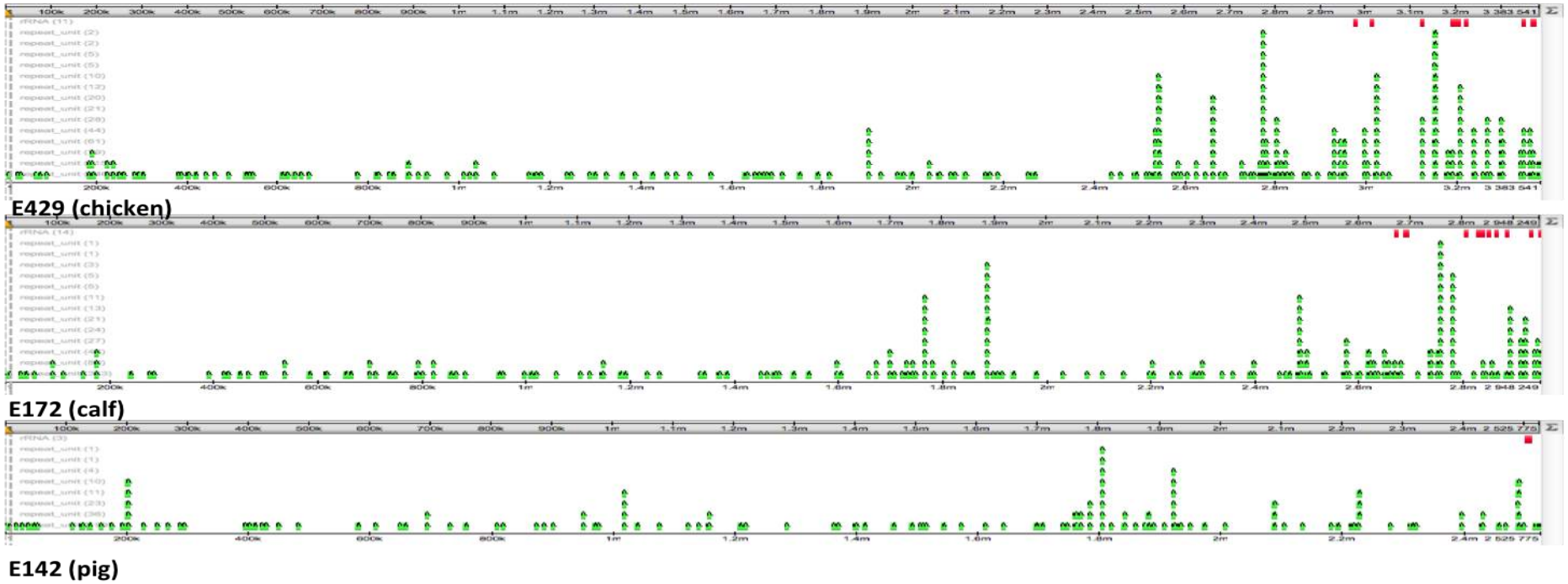
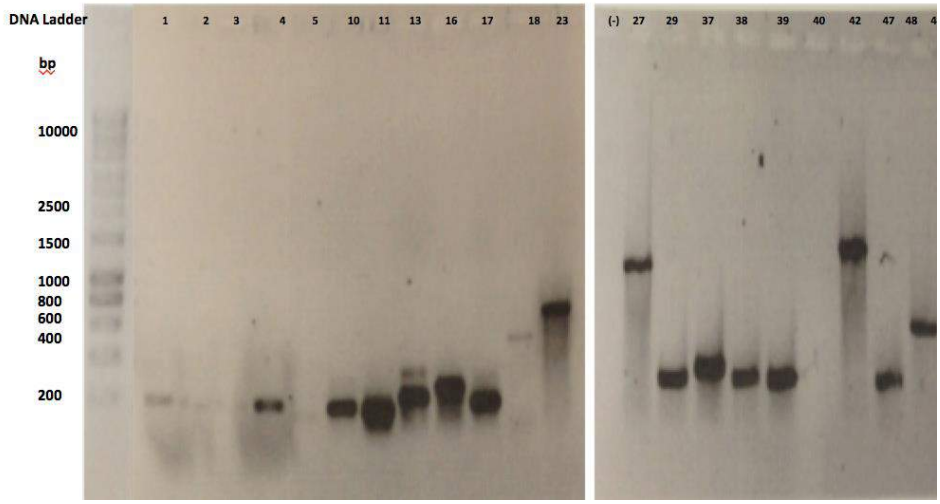


Figure 3.7: Short sequence repeats (SSRs) in animal *E. faecium* strains. SSRs covering the animal *E. faecium* genomes. SSRs annotations are located side by side in green and red blocks show rRNA operons.

3.2.4 Genome gap closure

3.2.4.1 Gap closure

A major starting aim of the study was to sequence the genomes of *E. faecium* isolates from animals, and since there was not a closed animal *E. faecium* genome so generate one to enhance comparative studies. The genome closure stage consisted of PCR amplification that bridged gaps in the sequence assembly. PCR amplifications were performed using primers designed from sequence approximately 100 bp from the 5' and 3' edges. The purified amplicons were sequenced using the PCR primers. Seventy-one pairs of PCR primers were designed to yield 1-6 kb amplicons spanning each gap (Table 3.5). The sequencing results obtained were assembled into the E429 genome. The gap sizes range between 400 bp to about 6 kb and the successfully amplified PCR products ranged between 200-2500 bp (Figure 3.8).



A. **B.**

Figure 3.8: PCR amplifications of the *E.faecium* E429 genome gaps. The size of the PCR products varied between 200-2500 bp. Positive PCR products resulted in a single clear band in the agarose gels, with no band in the negative result. (A) amplicons covering gaps 1, 2, 3, 4, 5, 10, 11, 13, 16, 17, 18 and 23. (B) amplicons covering gaps 27, 29, 37, 38, 39, 40, 42, 47, 48 and 49. (-) indicates the negative control.

Table 3.5: PCR amplification result for *E. faecium* E429 gaps. +++ Indicates very strong band, ++ shows strong band, + weak band and - is negative result.

Primers name	Expected product size	Product	Gap closed
Gap 1F.Gap 1R	1052	+++	**Yes**
Gap 2F,Gap 2R	871	+++	**Yes**
Gap 3F,Gap 3R	1041	+++	**Yes**
Gap 4F,Gap 4R	942	+++	**No**
Gap 5F,Gap 5R	871	++	**Yes**
Gap 6F,Gap 6R	1001	-	**No**
Gap 7F,Gap 7R	123998	++	**No**
Gap 8F,Gap 8R	1008	++	**Yes**
Gap 9F,Gap 9R	874	++	**No**
Gap10F, Gap 10R	1045	+++	**Yes**
Gap11F, Gap 11R	1037	++	**No**
Gap 12F,Gap 12R	917	-	**No**
Gap 13F,Gap 13R	951	+++	**Yes**
Gap 14F,Gap 14R	1095	+++	**No**
Gap 15F,Gap 15R	817	++	**No**
Gap 16F,Gap 16R	1064	+++	**Yes**
Gap 17F,Gap 17R	851	+++	**Yes**
Gap 18F,Gap 18R	811	++	**No**
Gap 19F,Gap 19R	1057	-	**No**
Gap 20F,Gap20R	968	++	**Yes**
Gap 21F,Gap 21R	1048	++	**No**
Gap 22F,Gap 22R	1014	-	**No**
Gap 23F,Gap 23R	995	+++	**Yes**
Gap 24F,Gap 24R	1052	++	**Yes**
Gap 25F,Gap 25R	1024	++	**No**
Gap 26F,Gap 26R	1672	++	**No**
Gap 28F,Gap 28R	1503	-	**No**
Gap 29F,Gap 29R	1602	++	**Yes**
Gap 30F,Gap 30R	1599	+++	**No**
Gap 31F,Gap 31R	1408	++	**No**
Gap 32F,Gap 32R	1195	-	**No**
Gap 33F,Gap 33R	1569	++	**No**
Gap 34F,Gap 34R	1550	-	**No**
Gap 35F,Gap 35R	1367	++	**No**
Gap 36F,Gap 36R	34583	++	**No**
Gap 37F,Gap 37R	259820	++	**No**
Gap 38F,Gap 38R	1340	+++	**Yes**
Gap 39F,Gap 39R	1320	+++	**Yes**
Gap 40F,Gap 40R	1586	+++	**No**
Gap 41F,Gap 41R	1276	+	**Yes**
Gap 42F,Gap 42R	1448	+++	**Yes**
Gap 43F,Gap 43R	1562	++	**No**
Gap 45F,Gap 45R	1346	+++	**Yes**
Gap 46F,Gap 46R	1487	-	**No**
Gap 47F,Gap 47R	1367	+++	**Yes**
Gap 48F,Gap 48R	1344	+++	**No**
Gap 49F,Gap 49R	1307	+++	**Yes**
Gap 50F,Gap 50R	1913	+++	**Yes**
Gap 51F.Gap 51R	3542	-	**No**
Gap 52F,Gap 52R	5681	+++	**No**
Gap 53F,Gap 53R	3213	-	**No**
Gap 54F,Gap 54R	1532	++	**No**
Gap55F, Gap 55R	1313	-	**No**
Gap 56F,Gap 56R	2024	+++	**No**

Gap 57F,Gap 57R	2266	+++	**No**
Gap 58F,Gap 58R	550	-	**No**
Gap 59F,Gap 59R	940	-	**No**
Gap 60F,Gap 60R	1295	+++	**Yes**
Gap 61F,Gap 61R	1088	+++	**No**
Gap 62F,Gap 62R	3108	+++	**No**
Gap 63F,Gap 63R	3475	-	**No**
Gap 64F,Gap 64R	3567	+++	**No**
Gap 65F,Gap 65R	5675	++	**No**
Gap 66F,Gap 66R	3726	+++	**No**
Gap 67F,Gap 67R	3244	-	**No**
Gap 68F,Gap 68R	4868	+++	**No**
Gap 69F,Gap 69R	410	++	**Yes**
Gap 70F,Gap 70R	1234	+++	**Yes**
Gap 71F,Gap 71R	908	+++	**Yes**

Fifty-three gap PCRs out of seventy-one were successfully amplified. Eighteen expected products were never successfully amplified despite extensively optimising PCR conditions, meaning that there were gaps that would not be closed.

After sequencing, followed by attempts to incorporate the sequence data, twenty-five regions (7, 8, 9, 12, 15, 17, 18, 21, 25, 26, 31, 33, 35, 36, 41, 45, 54, 56, 57, 58, 61, 62, 63, 66 and 67) remained as unassembled sequence gaps. Seven gaps (4, 11, 14, 30, 37, 40, 48) were not closed since the PCR product sequence did not close the gap between two contigs (see Table 3.6). However, a small number of gaps that were sequenced closed the entire gap between two contigs, such as gaps number 10 and 13 (Figure 3.9).

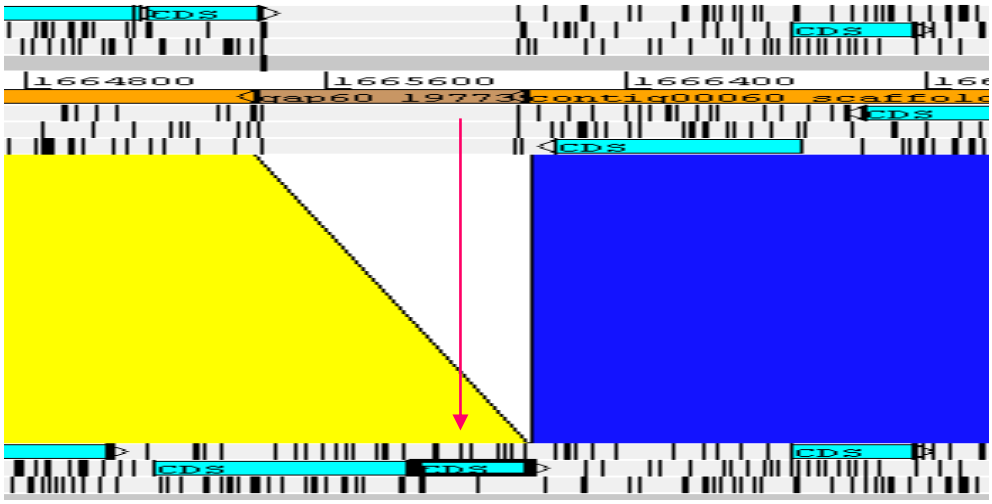


Figure 3.9: Gap closure of chicken *E. faecium* genome. Gap number 13 located between contig00059 (blue) and contig00060 (yellow), which was successfully closed. The top genome represents the genome with gaps and the bottom genome represents the genome after gap closure.

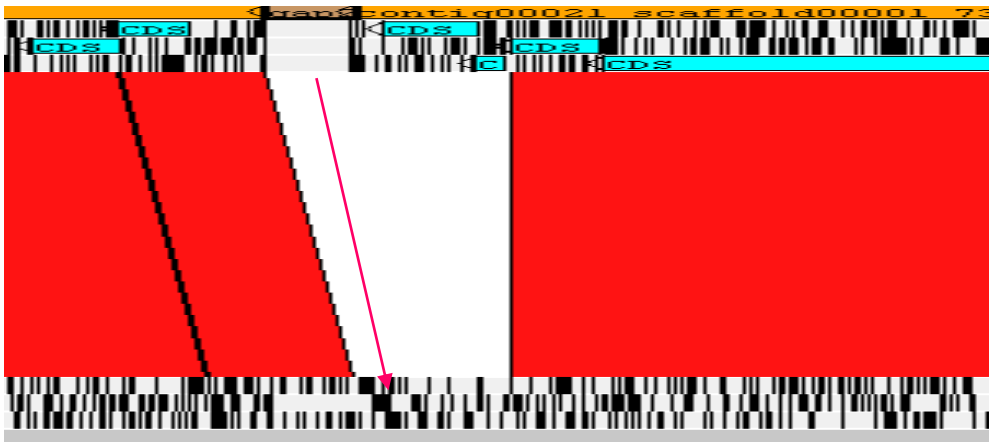


Figure 3.10: Gap closure of chicken *E. faecium* genome. Gap number 4 located between contig00021 and contig00022, which was not closed completely. The red arrow shows the location of the remaining gap. The top genome represents the genome with gaps and the bottom genome represents the genome after gap closure.

Analysis of the coding potential of the sequenced gap regions using RAST identified three potential virulence or colonisation genes. The first of these genes contains a potential adhesin gene encoding a protein annotated as '*Streptococcus pyogenes* recombinatorial zone', this gene subsystem has homology with a group A streptococcal genomic region that is highly recombinatorial among closely related strains, this adhesin has been proposed to play an important role in pilus-production and adhesion to human tissues (Bessen and Kalia 2002). A second identified gene encodes a protein with homology to cobalt-zinc-cadmium resistance protein CzcD. The third gene potentially encodes an iron scavenging mechanism, within whereby hemin uptake and utilisation systems in Gram positives bacteria its role is Sortase A that catalyses the covalent attachment of LPXTG proteins to peptidoglycan. The remainder of the gaps contain mostly mobile element genes encoding transposases, plasmid and phage proteins plus various metabolism and cell wall and capsule genes (Table 3.6).

Table 3.6: Gap sequence information of *E. faecium* E429. Gap location and the BLAST results for the PCR reactions. *indicates the gap that is not completely closed.

Gap number	Gap location (contig)	Gap result
1	contig00001 -contig00002	Zinc-containing alcohol dehydrogenase
2	contig00012 -contig00013	Hydrolase NUDIX family
3	contig00016-contig00017	Plasmid pVEF4
***4**	contig00021-contig00022	Response regulator
5	contig00024-contig00025	Integral-membrane protein
10	contig00049-contig00050	Lysis protein
***11**	contig00052-contig00053	Glucose uptake protein
13	contig00060-contig00061	Hypothetical protein
14	contig00061-contig00062	ISEf1, transposase
16	contig00071-contig00072	Hypothetical protein
20	contig00087-contig00088	Hypothetical protein
23	contig00091-contig00092	Transposon IS elements IS905
24	contig00109-contig00110	Transposon IS elements IS905
***29**	contig00010-contig00011	pVEF3 plasmid
***30**	contig00081-contig00082	Hypothetical protein
***37**	contig00033-contig00034	Putative tail or base plate protein gp19 [Bacteriophage A118]
38	contig00034-contig00035	4-carboxymuconolactone decarboxylase
39	contig00036-contig00037	Hypothetical protein
***40**	contig00040-contig00041	Cell division trigger factor
42	contig00047-contig00048	Helix-turn-helix domain-containing protein hypothetic
***43**	contig00022-contig00023	Transposon Tn*1546* insertion sequence ISEfa10 transposase genes
60	contig00101-contig00102	Hypothetical protein
69	contig00003-contig00004	Hypothetical protein
70	contig00004-contig00005	Extracellular protein
71	contig00005-contig00006	Murein_hydrolase_LrgA

In addition, the NCBI web server (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was used to investigate gaps that use of the RAST server (http://rast.nmpdr.org/) failed to annotate. BLASTp was utilised to examine the coding potential of these gaps. This identified plasmid-derived regions p*VEF4* and p*VEF3* in gaps 3 and 29, respectively. BLASTp comparisons also identified insertion sequence and transposon sequences in several gaps, including ISEf1 in gap 14, IS*905* in gap 23 and 24 and Tn*1546* was found in gap 43 (Table 3.6).

3.2.4.2 A fully sequenced *E. faecium* genome

The assembly of complex genomes using short sequence reads remains a challenge, mostly because of the occurrence of repeats, which cannot be assembled unambiguously. The repeat sequences in the strains studied here added additional complexity due to their high copy number and high sequence identity. These repeats also lead to extra complexity due to genomic rearrangements. Consequently, the 454 sequencing platform with de novo assembly approaches could not resolve to completion the assembly of the animal *E. faecium* genomes.

To circumvent theses issues the Pacific Biosciences RS (PacBio) platform was applied to fully sequence *E. faecium* strain E172 isolated from calf. The PacBio long-read sequencing platform provides increased read length and equitable genome coverage making it possible to construct assembled genome sequence data comprising few or no gaps by generating longer contigs (Ferrarini, Moretto *et al* . 2013).

A total of 65,958 PacBio RS reads were recovered with a mean read length of 7,505 bp totalling 495,063,288 nucleotides and representing an average depth of coverage of 115.87 of the *E. faecium* E172 genome. The dataset covered the entire *E. faecium* genome strain E172 in ten contigs (100% coverage). Genome annotation using Prokka identified 2,900 *E. faecium* genes most of which matched the 454 sequence data. Additional genes filled the gaps which matched those identified by 454 sequencing (Figure 3.11).

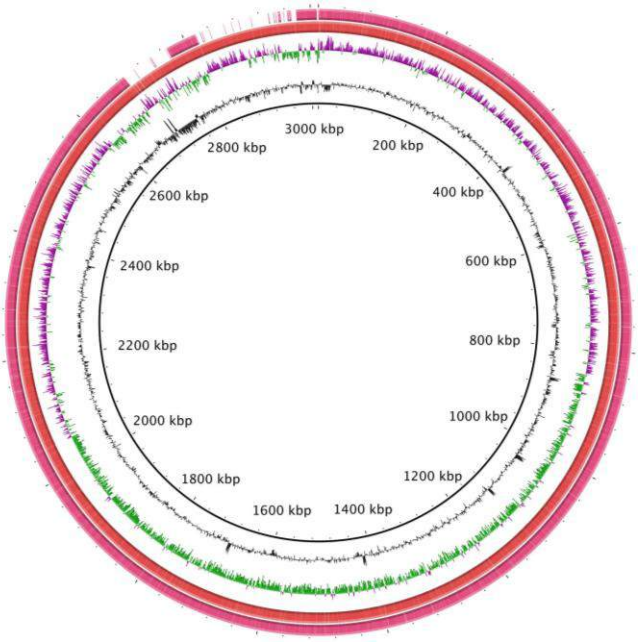


Figure 3.11: Genome map of the complete *E. faecium* strain E172. The black ring represents the complete genome of E172 (calf) using long reading platform (PacBio). The ring represents the draft genome of E172 using short read platform (454).

The chromosome of E172 has a high gene density with 2,823 predicted ORFs with a coding area of 96.22%. Annotation of the genome using IMG-ER revealed 2,099 (72.5%) of the predicted ORFs were orthologous to clustered ORFs of published genomes and in total 2,071 (70.95%) and 2,428 (82.75%) ORFs had homology with ORFs in COGs and Pfam databases, respectively (Table 3.7). The majority of genes (81.29%) could be assigned a function, however, only 684 (23.31%) of these genes were assigned to enzymes, and 770 (26.24%) were present in the KEGG database. Of the 2,823 predicted ORFs, 78 encoded proteins contain signal peptides. Of these

secreted proteins 30 have unknown function, 18 are predicted to be cell wall-associated proteins and 30 are predicted to be carbohydrate-binding and associated with an ABC transporter. Associated with the E172 PacBio genome assembly there are 70 tRNAs and 18 rRNAs (6 x 5S, 6 x 16S and 6x 23S). Nearly 10% of the genomes are non-AGCT bases in both 454 and PacBio, which may reflect the high number of the repetitive sequence in animal *E. faecium* genomes.

Table 3.7: Structural features associated with the sequenced genomes of *E. faecium* strains E172 using the 454 sequencing and PacBio platforms.

Genomic features	E172 (PacBio)	E172 (454)
Estimated genome size	3.0 MB	2.9 MB
Non-ACGT bases	321618	398256
Number of contigs	10	786
Shortest contig (bp)	100	102
Largest contig (bp)	2505612	186193

Table 3.8: Genome composition features of strains E172 using 454 sequencing and PacBio platforms.

Feature	E172 (PacBio)		E172 (454)	
	Number	%	Number	%
Genes total	2934	100	2973	100
Protein coding genes	2823	96.22	2892	97.2
Protein coding genes with function prediction	2385	81.29	2325	78.2
Protein coding genes without function prediction	438	14.93	567	19.0
Protein encoding enzymes	684	23.3	639	21.4
Protein coding genes connected to KEGG pathways	770	26.2	720	24.2
Protein coding genes connected to KEGG Orthology (KO)	1377	46.9	1280	43.0
Protein coding genes with COGs	2071	70.59	2186	73.5

3.3 Discussion

3.3.1 Genome analysis

Qin *et al* (2012) demonstrate that the genome size of *E. faecium* isolated from humans ranges from 2.50 Mb (strain E1039) to 3.14 Mb (strain 1,230,933). The numbers of protein-coding genes range from 2,587 (E1039) to 3,118 (strain TX0133A). By comparing the size of human *E. faecium* sequenced strains with animal *E. faecium* sequenced strains in this study, it is clear that the calf strain has the largest genome among all *E. faecium* strains in the database.

The large size of the genome could reflect a capacity of the bacterium to compete and survive in a nutritionally complex niche. The nutritional and physiochemical environment of the gastrointestinal tract might demand increased capability and versatility of this species relative to human isolates. When compared with other *Enterococcus* species including *E. faecalis, E. gallinarum* and *E. casseliflavus, E. faecium* isolates were found to have an intermediate genome size. *E. gallinarum* and *E. casseliflavus* have the largest genome size range from 3.4 to 3.6 kb (IMG- Integrated Microbial Genomes, Palmer, Godfrey *et al* . 2012, Qin, Galloway-Pena *et al* . 2012). van Schaik *et al* (2010) explained that this variation in genome size across *Enterococcus* species was proposed to occur due to expansion within species due to duplication and horizontal gene transfer.

The mean genome size of the majority of human infection isolates and epidemic isolates, including the clonal complex 17 (CC17) genogroup, is significantly larger (2.84 to 2.98 Mb) than that of isolates from faeces of non-hospitalised humans (2.71 to 2.84 Mb) or animal isolates and sporadic human infection isolates (2.59 to 2.75 Mb). This difference could represent the effect of cycles of infection and survival in the hospital being correlated with the acquisition of new genes (Lebreton, van Schaik *et al* . 2013).

3.3.2 Genome synteny

Genetic maps of bacteria reveal that only certain gene clusters are syntenic and homologous genes are maintained at the same relative position (Tamames 2001).

High to very low synteny was found when comparing *Enterococcus* species. Some of the genomes used in this comparison are fragmented and this can have effects in the appearance synteny. The genome backbones of the *Enterococcus* species were clearly related but were distinct, and large inversions were revealed within *E. faecium* strains.

Comparing the gene order within a selection of strains of *Enterococcus* species showed that *E. faecium* (Aus0004) and the genomes of *E. hirae* (ATCC 8043), *E. durans* (ATCC 6056) and *E. mundtii* (ATCC 882) shared a very conserved DNA sequence and gene order.

An intermediate degree of relatedness was found between *E. faecium* (Aus0004) and the genomes of *E. italicus* (DSM 15952), *E. avium* (ATCC 14025) and *E. asini* (ATCC 700915). At the other extreme the genomes of *E. faecium (*Aus0004) and *E. faecalis* (V583), *E. caccae* (ATCC BAA-1240) and *E. haemoperoxidus (*ATCC BAA-382) possess very different gene orders (Figure 3.3). A phylogenetic tree of *Enterococcus* species previously described by Carvalho Mda *et al* (2004) using 16S rDNA seequnces identified that synteny correlated with the species evolution relationships (Figure 3.12). The species that share high synteny with *E. faecium* are branched close to *E. faecium* in the phylogenetic tree, while the species that share low synteny are branched far from *E. faecium*

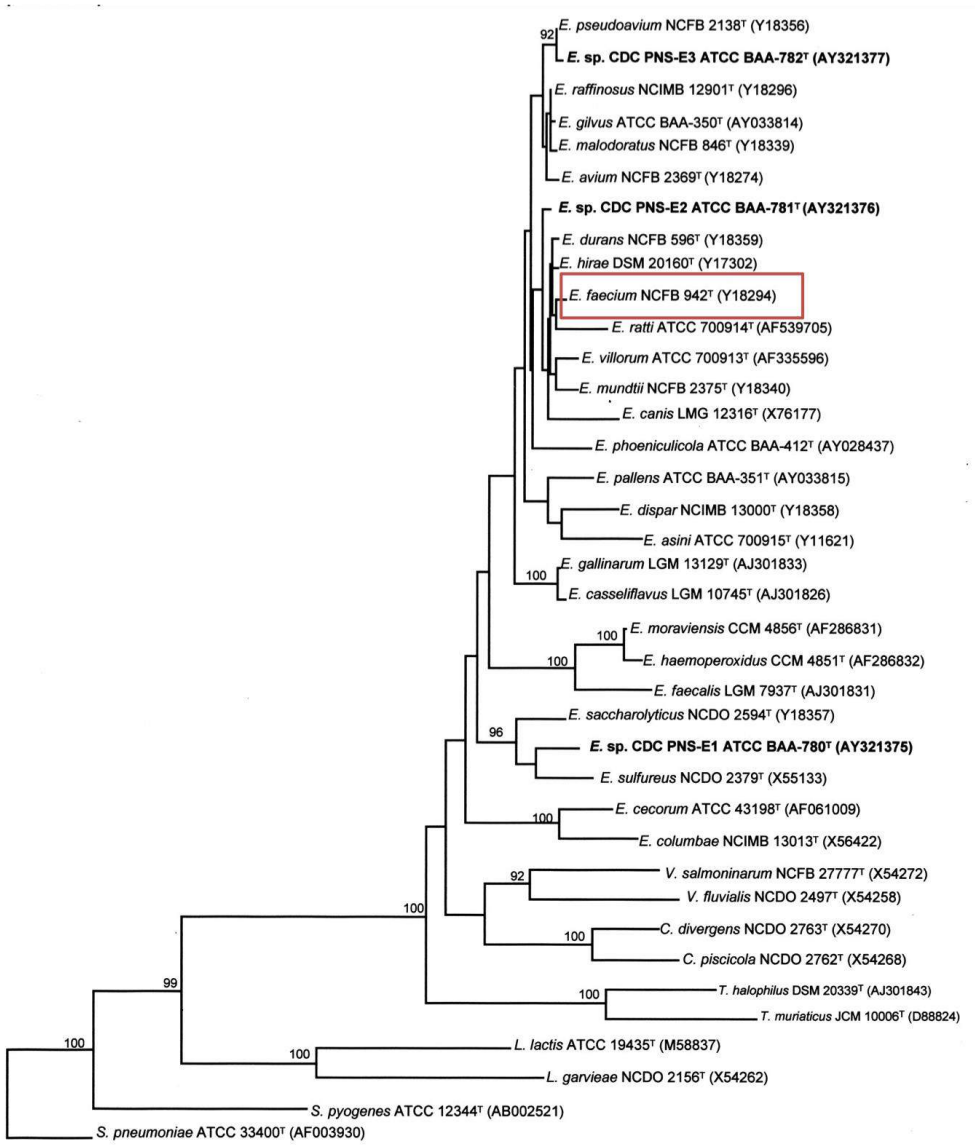


Figure 3.12: Phylogenetic tree of enterococci constructed by (Carvalho Mda, Steigerwalt *et al* . 2004) and based on comparative analysis of 16S rDNA sequences.

3.3.2.1 Genome inversions in animal strains of *E. faecium*

The complete genomes of *E. faecium* Aus0004 and DO reveal a large (683 kb) inversion, relative to each other. Lam *et al* (2012), stated that prophages

present at the boundary of the inversion could be the factor for facilitating the chromosomal inversion across the replication terminus.

Further reasons were found that might explain the chromosomal inversion across the replication terminus of Aus0004 and DO genomes relative to the animal *E. faecium* genomes (Figure 3.4). The genomes of *E. faecium* Aus0004 and animal *E. faecium* have three inversions. Prophages are present of the boundary of the inversion in block 2 and could have facilitated this chromosomal inversion given that these prophages exhibit a high degree of similarity. Blocks 2 and 3 of the animal genomes may have inverted by recombination due to the presence of integrases. Equally, the high numbers of IS elements may have played an important role in facilitating the chromosomal inversion since there are multiple IS elements at the boundary between the inverted regions. For example IS*1251* and ISEfm1 are present at the boundary of the inverted regions in Aus0004, relative to the three animal *E. faecium* genomes.

3.3.3 Gap closure

Gap spanning PCR products were amplified by using a range of different reaction conditions. Sequences obtained from closure of genome gaps revealed that the majority were transposon and plasmid sequences, which are known to contain repetitive sequences. For example, transposons found in gaps matched Tn*6085* and Tn*1546* and identified *Enterococcus* plasmids matched pVEF4, pVEF3 and p5753.

Twenty-five gaps were successfully completed to leave around 215, and since the rate of closure slowed it made this aim of the research unfeasible. While some of this failure was clearly due to operator errors the assembly of the genome at junction regions, combined with the repetitive sequences in the sections being amplified and potentially sequence errors due to the 454 technology all conspired against successfully completing the *E. faecium* E429 genome.

Currently there are high numbers of bacterial genomes sequenced to high-quality draft stand by using short read sequence data combined with whole genome assembly techniques. However, the high quality genome drafts almost always contain gaps. There are known limitations with the input data and the techniques used to construct draft assemblies. Factors such as repetitive genomic features, genomic polymorphism and sequencing biases complicate assembly of some regions (English, Richards *et al* . 2012).

Recently, an automated approach using long-reads from the Pacific Biosciences RS (PacBio) platform has enabled the completion of entire bacterial genomes. The software tool (PBJelly) uses PacBio reads to close gaps and preserve annotations. The arrival of (PacBio) sequencing has brought further advances in genome sequencing by increasing throughput and decreasing cost and the time taken to complete a genome (English, Richards *et al* . 2012).

The PacBio RS sequencing data of *E. faecium* E172 generally improved scaffolding, gap filling and genome sequence finishing comparing with the 454 sequencing platform. Assemblies using the 454 data include multiple gaps that leading to a large number of contigs and scaffolds even in a smaller sized genome such as animal *E. faecium* isolated from pig. The E172 genome data using 454 comprise 786 contigs with more than hundred gaps comparing with only 10 contigs using PacBio. In addition, large numbers of nucleotides of the genome for example the rRNA genes were not assembled in 454 sequence data and thus contigs must be recovered from the genome assembly.

Chapter Four: Comparative genomics of *Enterococcus faecium*, isolated from animals.

4.1 Introduction

The genome data publicly available for bacterial species and their closely related isolates have greatly expanded our understanding of bacterial specialisation. Geographic separation or habitat specialisation can potentially account for the genetic diversity observed within a bacterial species (Ellegaard, Klasson *et al*. 2013). It remains unclear whether bacteria that are not isolated by geographic or physical barriers branch into distinct groups, however, studies of bacteria such as *Bacillus, Vibrio* and *Synechococcus,* which are free-living, identified clustering sequences that correlate with ecological specialisation. Moreover, recombination and horizontal gene transfer between species could effect speciation in bacteria (Gogarten, Doolittle *et al*. 2002, Connor, Sikorski *et al*. 2010, Ellegaard, Klasson *et al*. 2013).

The origin of a DNA sequence, together with its phenotypic and ecological effects can determine whether individual bacteria belong to distinct clusters (Cohan 2001). For instance, on the basis of metabolic and other phenotypic characteristics, 315 isolates of *Neisseriaceae,* which is a family containing pathogens that cause the diseases gonorrhea and meningitis, were spread into 31 different clusters (Barrett and Sneath 1994). Phenotypic clustering (based mostly on metabolic characters) has long been proposed as a mechanism for bacterial speciation. Genotypic clustering has largely substituted phenotypic clustering as a primary principle for defining bacterial species. For many years, clustering was derived from whole-genome DNA hybridisation between pairs of strains, which aided the

differentiation of species. Currently, 16S rRNA and protein-coding gene sequence clusters are used for species differentiation (Ellegaard, Klasson *et al*. 2013).

The *Enterococuus* genus presently consists of 37 species that inhabit a wide range of niches that includes the gastrointestinal microbiota of almost every animal phylum. Intrinsic resistance to harsh conditions and metabolic versatility are proposed to explain the ability of this genus to colonise broadly (Ramsey, Hartke *et al*. 2014). A comparative genome analysis performed by van Schaik *et al* (2010) indicated that there are differences in the carbohydrate metabolic pathways, oxidative stress defence mechanisms and particular protein families between *Enterococcus* species. For example, *E. faecium* has the ability to utilise carbon sources from plant polysaccharides (arabinose), while *E. faecalis* does not. *E. faecalis* has the ability to use ethanolamine as a carbon source in the presence of cobalamin while this is absent from *E. faecium* (Del Papa and Perego 2008). Van Schaik *et al* (2010) indicated that a potential defence mechanism to oxidative stress is delivered by glutathione (g -GluCysGly; GSH), which can be synthesised by *E. faecium* and *E. faecalis*. However, *E. faecium* has a glutathione peroxidase enzyme, which may play a more prominent role in the oxidative stress response while E. *faecalis* lacks this particular enzyme.

Carbohydrate fermentation allows enterococci to succeed in distinct environments. Each *Enterococcus* species is known to utilise at least 13 sugars and over 30 additional sugars are utilised by several species. The ability to utilise a broad range of carbohydrates appears to result from the

capability of *Enterococcus* to share carbon utilisation mechanisms among strains and species, frequently on mobile elements (Ramsey, Hartke *et al*. 2014).

Population biology-based studies have indicated that there are specific lineages of human and animals. Isolates of *E. faecium* from animal have also the ability to act as a reservoir of antibiotic resistance genes (Bonten, Willems *et al*. 2001, Willems, Top *et al*. 2005). Comparative analyses between *Enterococcus* species identified genes, such as *esp*, that are horizontally transferred by conjugation, transformation and transduction between animal and human isolates, as well as from *E. faecium* to *E. faecalis* (van Schaik, Top *et al*. 2010).

Molecular epidemiological studies of *E. faecium* that were based on Multi-Locus Sequence Typing (MLST) indicated that commensal strains of *E. faecium* are distinct from clinical infections strains. The clinical infections subpopulation commonly has IS16, pathogenicity island(s), and plasmids or genes associated with antibiotic resistance, colonisation, and/or virulence. 3–10% sequence difference was found in four genes between clinical clade and commensal clade, including 5% difference between pbp5-R (ampicillin-resistant) from clinical isolates and pbp5-S (ampicillin-sensitive) from commensal isolates (Galloway-Pena, Roh *et al*. 2012). Lam *et al* (2012) suggested that the genomic plasticity detected in *E. faecium* isolates is could be responsible for the diverse properties shown by commensal and clinical isolates.

A lack of information about animal strains of *E. faecium* means that the degree of variation among human commensal, hospital and animal isolates was not clear. The original aim of this study was to genome sequence several animal strains of *E. faecium* to compare with human strains. While this study was ongoing, in 2013 Lebreton *et al* published the sequences of animal and human commensal and hospital isolates of *E. faecium* (Lebreton, van Schaik *et al*. 2013). In addition, the first complete human clinical isolate genomes of *E. faecium* Aus0004 and DO were published.

Specific aims

This chapter will expand genome comparisons to include animal, clinical and commensal *E. faecium* isolates from different niches to consider the reason for demarcation in the *E. faecium* species. The three animal strains were isolated from chicken, calf and pig will be compared with each other and with all other isolates from animals and humans. The comparison will determine whether these strains differ from human and other animal isolates and whether they have acquired genes specific for colonising their animal host.

4.2 Results

4.2.1 Comparative genomics of *Enterococcus faecium*

The numbers of *E. faecium* sequenced strains used in this study are 129; which include 42 clinical, 8 commensal and 21 animal isolates (Table 2.1).

4.2.1.1 Core and pan-genome of *E. faecium*

In this study, an effort was made to define a conserved core genome of the 129 *E. faecium* strains (Table 2.1), and suggest those genes likely to be essential for cell function, in contrast to the variable genes that are not conserved and are subject to horizontal gene transfer in the *E. faecium* genomes. The core and pan-genome of *E. faecium* were identified using OrthoMCL and were analysed using R statistical software (Section 2.23). As a result, 1,467 orthologous clusters that were found in the 129 strains of *E. faecium* were allocated as core genome and 11,669 orthologous clusters were allocated as pan- genome (Figure 4.1 and Figure 4.2). The pan-genome of the *E. faecium* confirmed that the genome of *E. faecium* is open. Moreover, the ratio of the horizontal gene transfer in the genome is high. A pan-genome can be considered to be essentially unlimited in size when each new genome is added the size of the pan-genome increases (“ open”) or in contrast to have a finite size genome (“ closed”).

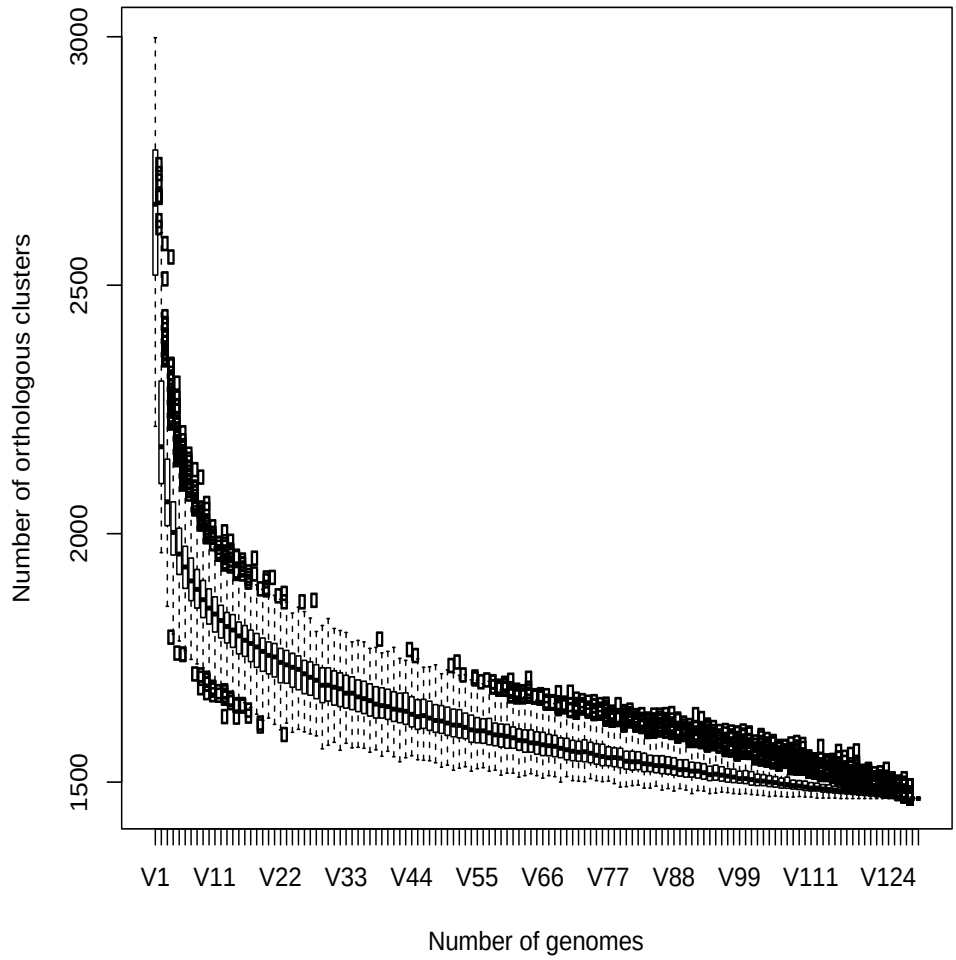


Figure 4.1: Genome structure of *E. faecium*. The core genome of the 129 strains of *E. faecium*. Circles represent the number of core genes when each genome is added. Black bars indicate median values.

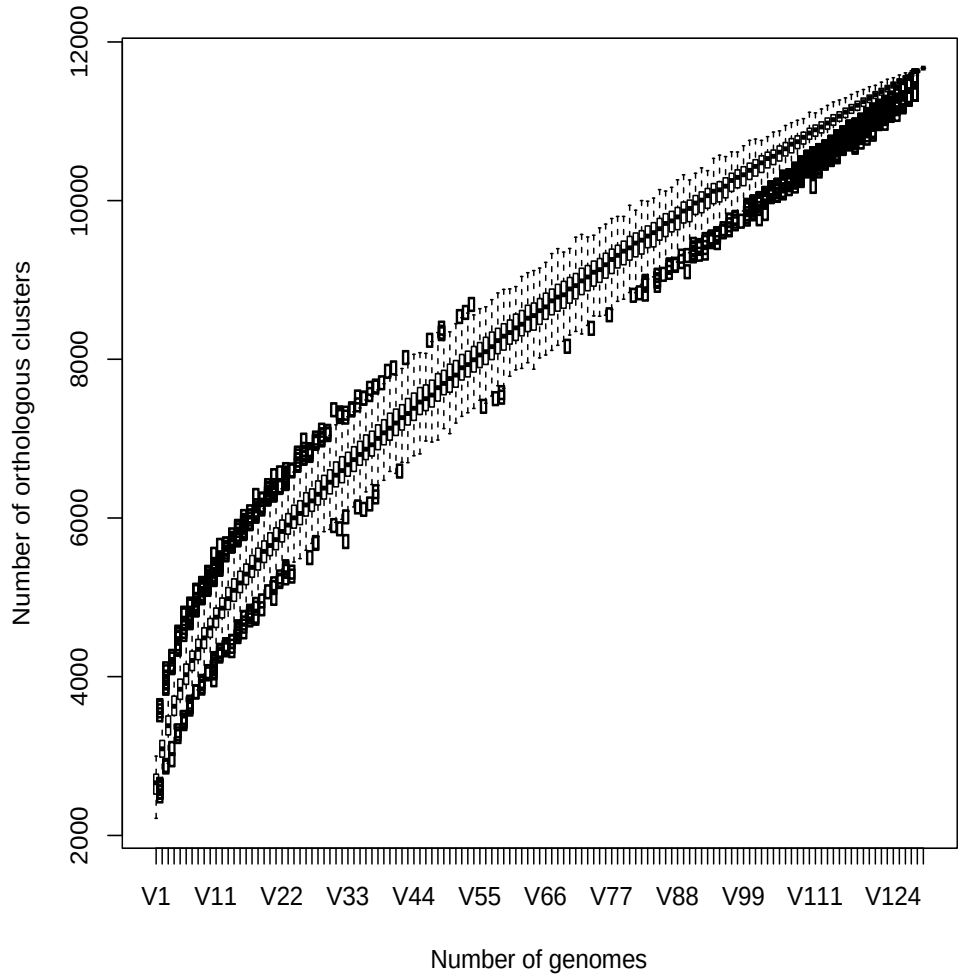


Figure 4.2: Genome structure of *E. faecium*. Pan-genome determined from 129 strains of *E. faecium*. The pan-genome is indicated for increasing numbers of sequenced *E. faecium* genomes. Circles represent the number of new genes when a genome is added. Black bars indicate median values.

The core gene set identified in the 129 strains (Table 2.1) of *E. faecium* encodes fundamental functions, such as information storage and processing, cellular processes and metabolism. Approximately 14.8% of the core orthologues belong to unknown function categories. The core functional

classes of *E. faecium* are those involved in intracellular trafficking, secretion (18.4%), carbohydrate transport and metabolism (9.54%) translation (9.3%), transcription (7.8%), amino acid transport and metabolism (7.1%), replication, recombination and repair (6.4%), cell wall/membrane biogenesis (5.6%), energy production and conversion (4.2%) and nucleotide transport and metabolism (4.2%). In addition, 7% of the core orthologues do not match any functional categories in the COGs database (Table 4.1). Mobile genetic elements including phage, plasmid and IS elements genes appear at a low frequency in the core genome of *E. faecium*, compared with the pan-genome of *E. faecium.*

About 43% of orthologues present in the pan-genome of *E. faecium* have no defined function in the COGs database and 21.8% are unknown function. The remaining functional ortholog classes in the pan-genome include replication, recombination and repair (14%), carbohydrate transport and metabolism (14.6%), transcription (8.7%), cell wall/membrane biogenesis (7.8%), defence mechanisms (4.3%) and amino acid transport and metabolism (3.8%). Within the pan-genome of *E. faecium* the frequency of replication, recombination and repair functions are twice that of the same function in the core genome (Table 4.2).

E. faecium species appear capable of utilising multiple sugars, such as aldose, mannitol, ribulose, arabinose, lactose, xylose, maltose, glucitol, sorbitol and mannose, the genes for which are located in the core genome.

Fructose, galactitol, glucose, rhamnose, sorbose and sucrose utilisation genes are found in the pan genome of *E. faecium* (Supplemental File, S1).

Table 4.1: Core clusters of Orthologous Groups (COGs) of *E. faecium*. Table shows the numbers of COGs in the core genome of *E. faecium* and the percentage of each functional category relative to total COGs in the core genome.

COG	COG Definition	Total	%
	Information storage and processing		
(J)	Translation	126	9.32
(K)	Transcription	107	7.92
(KL)	Atp-Dependent DNA helicase	2	0.14
(KT)	hydrolase_RelA	1	0.07
(L)	Replication,_recombination,_repair	87	6.43
(LU)	Protein involved in DNA mediated transformation	1	0.07
	Cellular processes		
(D)	Cell_cycle_control,_mitosis,_meiosis	17	1.25
(M)	Cell_wall/membrane_biogenesis	76	5.62
(MNOU)	Flagellum-specific muramidase which hydrolyses the peptidoglycan layer to assemble the rod structure in the periplasmic space	1	0.07
(NOT)	Adaptor protein; enables recognition and targeting of proteins for proteolysis, involved in negative regulation of competence	1	0.07
(O)	Posttranslational_modification,_protein_turnover,_chaperones	49	3.62
(T)	Signal_transduction_mechanisms	37	2.73
(U)	Intracellular_trafficking,_secretion	14	18.42
(V)	Defence _mechanisms	28	2.07
	Metabolism		
(C)	Energy_production,_conversion	57	4.21
(CP)	ABC transporter (permease)	1	0.07
(E)	Amino_acid_transport,_metabolism	97	7.17
(EGP)	Major facilitator superfamily protein	1	0.07
(EQ)	Hydantoinase/Oxoprolinase	1	0.07
(F)	Nucleotide_transport,_metabolism	58	4.29
(FG)	Histidine triad (HIT) protein	1	0.07
(FJ)	Deaminase	1	0.07
(G)	Carbohydrate_transport,_metabolism	113	8.36
(GK)	ROK family protein	1	0.07
(GM)	NAD-dependent epimerase/dehydratase	29	2.14
(H)	Coenzyme_transport,_metabolism	34	2.51
(I)	Lipid_transport,_metabolism	3	0.22
(IQ)	Short-chain dehydrogenase/reductase	70	5.18
(P)	Inorganic_ion_transport,_metabolism	7	0.51
	Poorly characterised		
(R)	General_function_prediction	129	9.54
(S)	Function_unknown	200	14.8
Grand Total		1351	92.02
Not in eggNOG db		117	7.97

Fructose uptake systems (PTS) (*EIIABC-Fru*) were mainly found in human *E. faecium.* However, two animal strains isolated from pig (E0680) and chicken (E429) have two distinct fructose PTS systems a duplication found only in two clinical strains isolated from blood (E1636 and E1185) and a commensal strain (E1039) (ORTHOMCL4064 and ORTHOMCL4184). The pig strain E1578 has novel fructose PTS systems (ORTHOMCL4968) (Supplemental File, S1). Galactitol uptake systems (PTS) were found broadly across 115 strains of *E. faecium* including clinical, commensal and animal (ORTHOMCL2001). In contrast, glucose uptake systems (PTS) (ORTHOMCL2654) were exclusively found in the clinical strains and only one dog (E4389) strain. The other identifiable carbohydrate uptake systems were variably present across the strain groups (Supplemental File, S1).

Table 4.2: Clusters of Orthologous Groups (COGs) of *E. faecium*. Table shows the numbers of COGs in the pan-genome of *E. faecium* and the percentage of each functional category relative to total COGs in the core genome.

COG	COG Definition	Total	%
	Information storage and processing		
(J)	Translation	27	1.05
(K)	Transcription	225	8.75
(KL)	Atp-Dependent DNA helicase	2	0.07
(KT)	hydrolase_RelA	1	0.03
(KOT)	Accessory gene regulator protein	1	0.03
(L)	Replication,_recombination,_repair	362	14.08
	Cellular processes		
(D)	Cell_cycle_control,_mitosis,_meiosis	28	1.08
(M)	Cell_wall/membrane_biogenesis	201	7.82
(N)	Cell_motility	3	0.11
(NOU)	Cleaves type-4 fimbrial leader sequence and methylates the N-terminal (generally Phe) residue protein	1	0.03
(NU)	Mannosyl-Glycoprotein endo-beta-N	1	0.03
(O)	Posttranslational_modification,_protein_turnover,_chaperones	45	1.75
(T)	Signal_transduction_mechanisms	66	2.56
(U)	Intracellular_trafficking,_secretion	16	0.62
(V)	Defence _mechanisms	112	4.35
	Metabolism		
(C)	Energy_production,_conversion	61	2.37
(CT)	Adenylate/Guanylate	1	0.03
(E)	Amino_acid_transport,_metabolism	100	3.89
(EH)	D-Isomer specific 2-hydroxyacid dehydrogenase	2	0.07
(EQ)	Hydantoinase/Oxoprolinase	3	0.11
(ET)	-	1	0.03
(F)	Nucleotide_transport,_metabolism	23	0.89
(FG)	Histidine triad (HIT) protein	1	0.03
(G)	Carbohydrate_transport,_metabolism	376	14.6
(GK)	ROK family protein	3	0.11
(GKT)	Sugar:Hydrogen symporter protein	6	0.23
(GM)	NAD-dependent epimerase/dehydratase	14	0.54
(H)	Coenzyme_transport,_metabolism	36	1.40
(HI)	Citrate lyase	1	0.03
(I)	Lipid_transport,_metabolism	20	0.77
(IQ)	Short-chain dehydrogenase/reductase	70	5.18
(P)	Inorganic_ion_transport,_metabolism	82	3.19
(Q)	Secondary_metabolites	16	0.62
	Poorly characterised		
(R)	General_function_prediction	161	6.26
(RM)	Phosphatase	1	0.03
(S)	Function_unknown	562	21.8
Grand Total		2570	56.24
Not in eggNOG db		1999	43.75

4.2.1.2 Phylogenetic tree

As a means to further investigate the relationship between the panel of *E. faecium* genomes a phylogenetic study was approached. The phylogenetic analysis (Section 2.22) was performed based on the distinction of 1,467 shared, single copy orthologous groups and delivers a complete vision of the evolutionary descent of the 129 sequenced *E. faecium*, comprising the human infection isolates, including clonal complex CC17, non-hospitalized human isolates and animal isolates. This phylogenetic tree of the complete set of *E. faecium* isolates in the database was expected to enhance our understanding of the evolution of *E. faecium* (Figure 4.3). The generated tree (neighbour-joining tree) from the core orthologues revealed clustering of strains in clades associated largely with their source.

Based on the distinction of 1,467 shared, single-copy orthologous groups (core genome), *E. faecium* strains separate into three distinct clades A, B and C. Within the branch forming clade A, the majority of human infection contains sequence types (STs) from the clonal complex 17 (CC17) genotype (sequence type 17 [ST17], ST117, and ST78) are grouped together. Moreover, nearly all isolates belonging to the CC17 group cluster together forming clade A1. The remainder the sporadic human infection isolates are mixed together with the animal isolates to forming clade B.

Unexpectedly, one of the strains in clade A1 has an animal origin (dog), which potentially reveals links between hospital strains and household pets. Strain 1231408, consisting of a background genome of clade A was

unexpectedly found as sister group with clade C, which contains most of the commensal isolates.

Animal isolates form the major group within clade B. Most of *E. faecium* isolated from birds are grouped together in one branch in clade B, which also includes a calf strain (E172). A chicken strain (E429) is associated with a different subgroup of clade B, that contains most of the pig isolates.

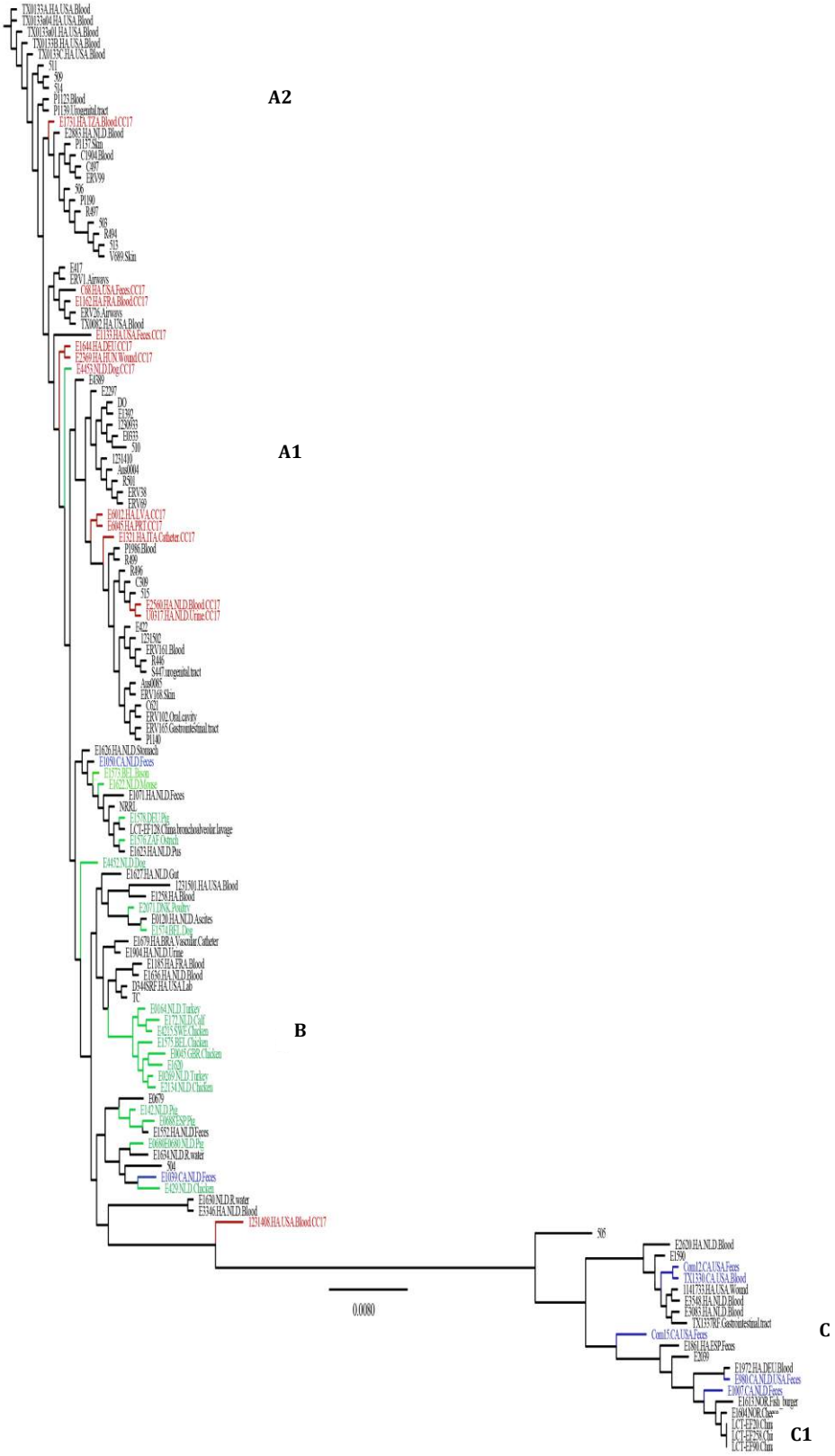

A2
A1
B
C
C1
0.0080

Figure 4.3: Neighbour-joining tree of *E. faecium.* The tree is based on the concatenated alignments of 1,467 single-copy shared core genes in 129 *E. faecium* genomes. Bootstrapping was performed with 1,000 replicates. The origins of the strains are indicated. Green indicates animal origin, blue is commensal origin, red is CC17 origin and black indicates sporadic human infection strains. Clade C indicates most of commensal strains; clade B indicates a mix of animal strains and other hospital strains. Clade A indicates most of the hospital strain, with A1 representing strains that belong to CC17; clade A2 contains most of the sporadic human infection strains.

4.2.1.1 Heat map analyses

A heat map of the genetic correlations between the 129 *E. faecium* strains was generated using the R programme for statistical computing. The presence/absence of 11669 gene clusters (pan-genome) was used to construct a heat map. The number of clusters is related to the cluster variation between all strains. The generated heat map is composed of three main groupings labelled A, B and C. Group A comprises clinical strains mainly related to clonal complex 17 (CC17); group B consists of animal strains and clinical strains that do not belong to CC17 and group C contains most of the commensal strains (Figure 4.4).

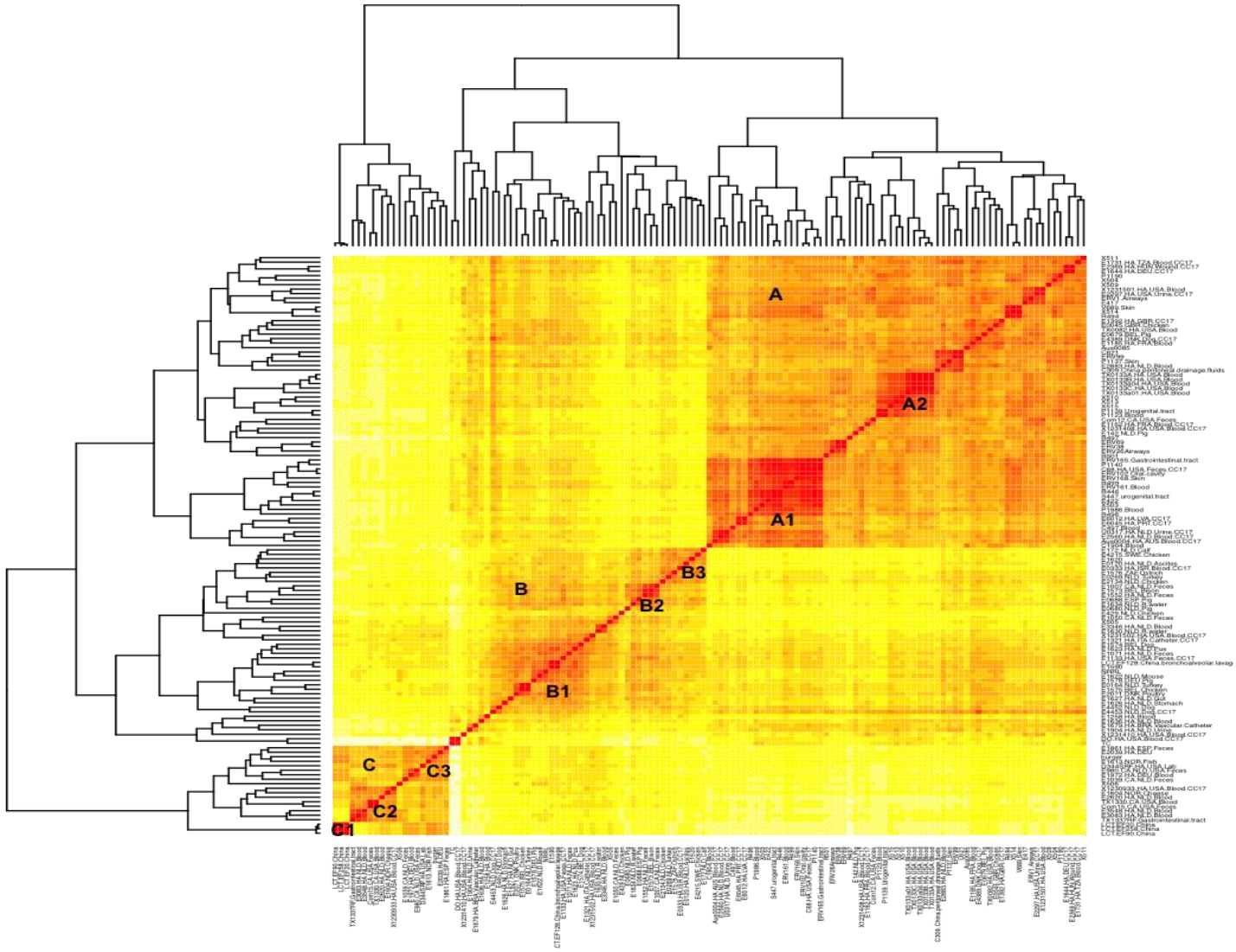

A
A2
A1
B
B3
B2
B1
C
C3
C2
C1

Figure 4.4: Heat map of the genetic correlations between the 129 *E. faecium* strains. Group A, B and C are of mixed strain origin. Group A represents hospital-associated strains, mostly of CC17 origin; group B comprises animal-associated strains and group C consists of mixed sources including commensal strains. The identical set of trees is represented on the x-axis and y-axis, the correspondence between colour scale and genetic correlation levels are presented on the right-hand side of the heat map (Red shows absent clusters, yellow shows present clusters). Also, since the same set of trees is symbolised on the x and y axis the color values along the heatmap are bright red. This is because a tree matched to itself will not have any branching differences.

The presence and absence of the 11669 accessory orthologous groups in the 129 strains of *E. faecium* also revealed smaller subgroups. CC17 strains were grouped together in group A in the heat map and this group also contains the highly-related hospital-associated strains, from Texas TX0133a01, TX0133C, TX0133a04, TX0133B and P1123, P1139, X515, X513 and X510 (group A2) (Figure 4.4).

The majority of the strains in group A are blood isolates. Unexpectedly, strain Com 12, which is a commensal strain, and animal strains E0045 (chicken), E0679 and E142 (pig) and E4389 (dog) are also located in this group. Most of the strains in clade (A1) are not associated with metadata to describe their source.

The majority of the animal *E. faecium* isolates are grouped in clade B (Figure 4.4), which comprises three small subgroups, B1, B2 and B3. The largest subgroup (B1) contains hospital isolates and three isolates belonging to CC17 group (E4453 (dog), E1133 and E1321). Most of the strains in this group are from the same geographic region (The Netherlands). Clade B2 contains two pig strains (E0680 and E0688) a bison strain E1573, a chicken strain (E429), a strain isolated from river water (E1634) and one clinical isolate (E1552). Most of the strains in this subgroup including the river water strain were isolated from The Netherlands. Subgroup B3 includes most of the bird isolates, a calf strain and one CC17 strain (E0333), most of the strains in this subgroup are isolated from The Netherlands (Figure 4.4).
The majority of commensal strains are located in group C, however, the group also includes clinical isolates. C1 contains strains isolated from the same geographic region (China); one CC17 genotype strain (1,230,933) and a food strain isolated from cheese (E1604) form C2 and clade C3 contains a mixture of clinical, commensal and a food strain isolated from fish burger (Figure 4.4).

Both analyses, phylogenetic tree and presence/absence tree, indicate that clinically-isolated, commensal strains and animal strains are similarly clustered together in specific clades (Figure 4.3 and Figure 4.4). Commensal strains of *E. faecium* cluster together as a group in both analyses, however, the clinical strains are split into two distinct groups. In addition, the clinical strains of type CC17 are grouped together in a branch distant from other clinical strains. The general genetic background of animal strains of *E.*

faecium suggests that they are part of the pathogenic *E. faecium* groups, but from a different origin (Figure 4.3 and Figure 4.4). However, several strains; 1230933; 506; E1039 and D344SRF, have different core genomes according to their placement in the phylogenetic tree (Figure 4.3) but they group together according to their pan genome to form clade C in the heat map in Figure 4.4.

4.2.2 Comparative genomics of animal *Enterococcus faecium*

At the start of this project, none of the *E. faecium* genomes that were sequenced were isolated from animals while at the time of writing this thesis 18 *E. faecium* strains isolated from animals had been sequenced and partially assembled. These animal strains include four isolated from chicken, four from dog, four from pig, two from turkey, one from bison, one from mouse, one from poultry and one from ostrich (Table 2.1). Despite these numbers of animal *E. faecium* genomes that have now been sequenced none has yet been closed.

4.2.2.1 Core and pan-genome of animal *E. faecium*

In this study, an attempt was made to define conserved core orthologues in animal *E. faecium* genomes. The aim was that it would identify those that are essential for colonisation of animal hosts and distinct from variable of genes that are not conserved and which are likely to be subject to horizontal gene transfer (HGT) in animal *E. faecium.* The core and pan-genomes of animal *E. faecium* were identified using OrthoMCL and were analysed using R statistical software (Section 2.23).

As a result, 1,824 orthologous clusters were revealed to be present in all animal genomes of *E. faecium* and were assigned as core genome (Figure 4.5), with 6,686 orthologous clusters assigned as the pan-genomes of animal *E. faecium* isolates. The pan-genome of animal *E. faecium* is open, since the number of orthologous clusters in the pan-genomes increased with each additional animal genome (Figure 4.6).

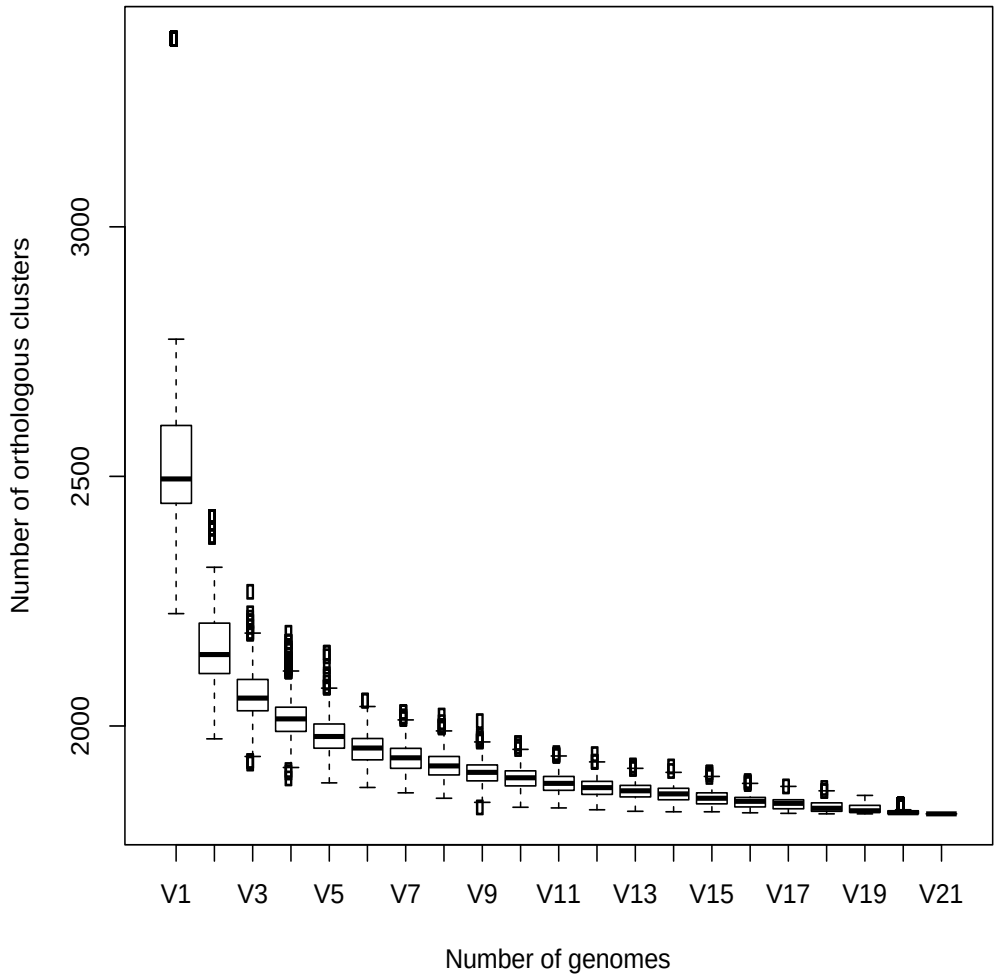


Figure 4.5: Core genome structure of animal *E. faecium.* The core genome is indicated for increasing numbers of sequenced animal *E. faecium*

genomes. Circles represent the number of core genes that exist when a particular genome is added. Black bars indicate median values.

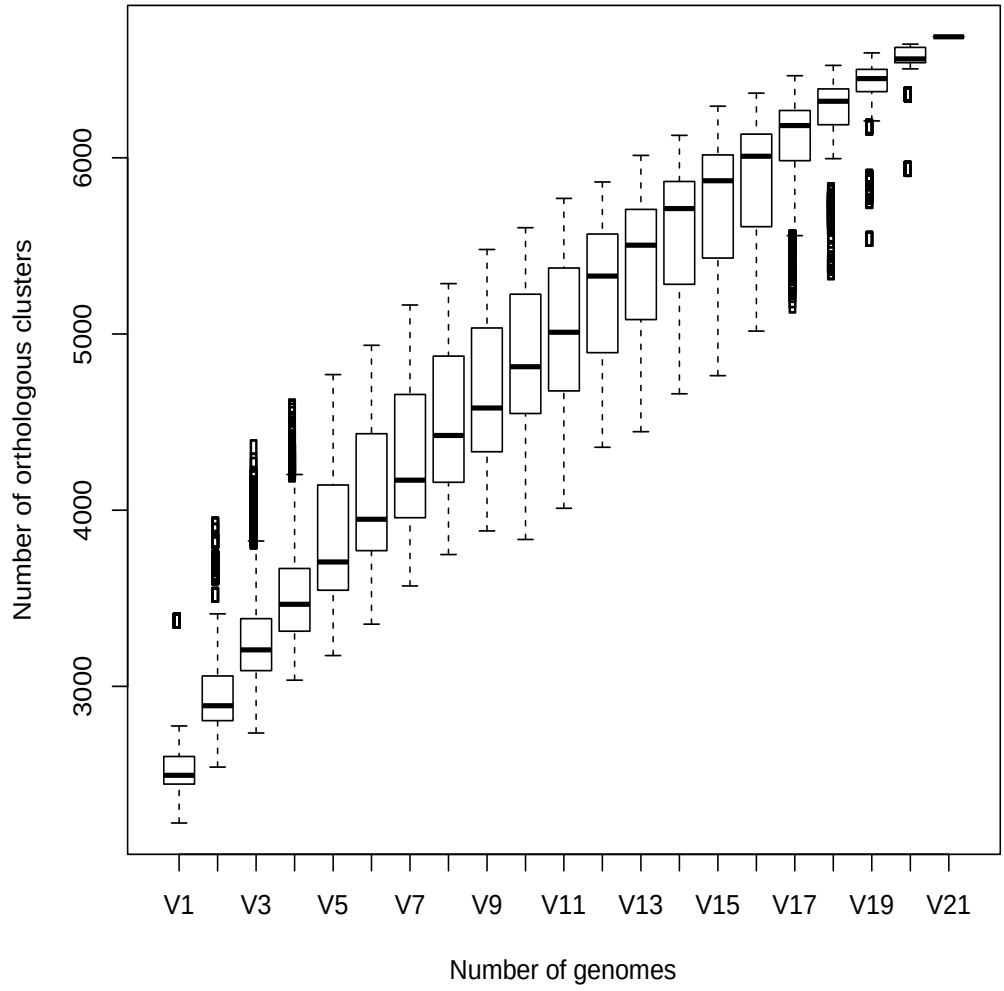


Figure 4.6: Pan-genome structure of animal *E. faecium.* The pan-genome is indicated for increasing numbers of sequenced animal *E. faecium* genomes. Circles represent the number of new genes that exist when a particular genome is added. Black bars indicate median values.

Table 4.3: Clusters of Orthologous Groups (COGs) of animal *E. faecium*. The table shows the categories numbers of COGs in the core genome of animal *E. faecium* and the percentage of each functional category compared with total COGs in the core genome. (-) indicates the absence of a category.

COG	COG Definition	Total	%
	Information storage and processing		
(J)	Translation	135	8.60
(K)	Transcription	129	8.22
(KL)	Atp-Dependent DNA helicase	1	0.06
(KT)	hydrolase_RelA	1	0.06
(L)	Replication,_recombination,_repair	97	6.18
(LU)	Protein involved in DNA mediated transformation	1	0.06
	Cellular processes		
(D)	Cell_cycle_control,_mitosis,_meiosis	19	1.21
(M)	Cell_wall/membrane_biogenesis	57	3.63
(MNOU)	Flagellum-specific muramidase which hydrolyzes the peptidoglycan layer to assemble the rod structure in the periplasmic space	1	0.06
(NOT)	daptor protein; enables recognition and targeting of proteins for proteolysis, involved in negative regulation of competence	1	0.06
(NOU)	Cleaves type-4 fimbrial leader sequence and methylates the N-terminal (generally Phe) residue protein	1	0.06
(O)	Posttranslational_modification,_protein_turnover,_chaperones	53	3.37
(T)	Signal_transduction_mechanisms	49	3.12
(U)	Intracellular_trafficking,_secretion	14	0.89
(V)	Defence _mechanisms	38	2.42
	Metabolism		
(C)	Energy_production,_conversion	63	4.01
(CP)	ABC transporter (permease)	1	0.06
(E)	Amino_acid_transport,_metabolism	122	7.77
(EGP)	Major facilitator superfamily protein	1	0.06
(EH)	D-Isomer specific 2-hydroxyacid dehydrogenase	1	0.06
(EQ)	Hydantoinase/Oxoprolinase	1	0.06
(ET)	-	1	0.06
(F)	Nucleotide_transport,_metabolism	66	4.20
(FG)	Histidine triad (HIT) protein	1	0.06
(FJ)	Deaminase	1	0.06
(G)	Carbohydrate_transport,_metabolism	161	10.26
(GK)	ROK family protein	1	0.06
(GKT)	Sugar:Hydrogen symporter protein	1	0.06
(GM)	NAD-dependent epimerase/dehydratase	3	0.19
(H)	Coenzyme_transport,_metabolism	39	2.48
(I)	Lipid_transport,_metabolism	40	2.54
(IQ)	Short-chain dehydrogenase/reductase	3	0.19
(P)	Inorganic_ion_transport,_metabolism	81	5.16
(Q)	Secondary_metabolites	13	0.82
	Poorly characterised		
(R)	General_function_prediction	147	9.36
(S)	Function_unknown	225	14.34
Grand Total		1569	86.63
Not in eggNOG db		216	11.92

Many of the functional categories of the animal core genome are associated with fundamental housekeeping functions. Approximately 14% of the core

orthologues have no known function. The functional categories of the animal core genome include carbohydrate transport and metabolism (10.26%), translation (8.60%), transcription (8.22%), replication, recombination and repair (6.18%) and amino acid transport and metabolism (7.77%) (Table 4.3).

Potentially, the presence of ill-defined orthologues in the sequenced animal *E. faecium* genomes are a marker of those genes acquired after the radiation of the genus and might be a strong indicator that these genes were not laterally-acquired. Animal *E. faecium* variably encode the pathways to utilise particular sugars, such as aldose, mannose mannitol, xylose, lactose, maltose and glucitol. In contrast, the uptake systems for fructose, galactitol, glucose, mannitol and galactose were found as core carbohydrate utilisation genes in animal *E. faecium.* The ability to generate energy from this range of sugars could be a requirement for successful colonisation of an animal host. Mobile genetic elements such as phage, plasmid and IS elements are present at low frequency in the core genome of animal *E. faecium* compared with the pan-genome of animal *E. faecium.* Variation was observed with the size of the core genome of the bird, pig and dog *E. faecium* sub-populations, (1897, 1990 and 2165) but might reflect the small numbers of strains for each host. This core genome size is larger than that of all *E. faecium* and approximately, 4%, 9% and 18.5% of the core genome of the bird, pig and dog is unique to these animal hosts (Supplemental File, S2).

Analysis of the pan-genome revealed that animal *E. faecium* contain 22.11% of genes with no known functional category. However, the two categories,

carbohydrate transport and metabolism, replication, recombination and repair comprise 14% and 15.35% of the pan-genome of the animal *E. faecium*, respectively (Table 4.4).

Table 4.4: Clusters of Orthologous Groups (COGs) of animal *E. faecium*. Table indicates the numbers of COGs in the pan-genome of animal *E. faecium* and the percentage of each functional category compared with total COGs in the pan-genome. (-) indicates the absence of a category.

COG	COG Definition	Total	%
	Information storage and processing		
(J)	Translation	19	0.82
(K)	Transcription	203	8.85
(KL)	Atp-Dependent DNA helicase	2	0.08
(KT)	hydrolase_RelA	1	0.04
(KOT)	Accessory gene regulator protein	1	0.04
(L)	Replication,_recombination,_repair	352	15.35
	Cellular processes		
(D)	Cell_cycle_control,_mitosis,_meiosis	26	1.13
(DJ)	Plasmid stabilisation system protein	1	0.04
(M)	Cell_wall/membrane_biogenesis	193	8.41
(N)	Cell_motility	3	0.13
(NU)	Mannosyl-Glycoprotein endo-beta-N	3	0.13
(O)	Posttranslational_modification,_protein_turnover,_chaperones	41	1.78
(T)	Signal_transduction_mechanisms	54	2.35
(U)	Intracellular_trafficking,_secretion	16	0.69
(V)	Defence _mechanisms	102	4.44
	Metabolism		
(C)	Energy_production,_conversion	55	2.39
(CT)	Adenylate/Guanylate	1	0.04
(E)	Amino_acid_transport,_metabolism	75	3.2
(EH)	D-Isomer specific 2-hydroxyacid dehydrogenase	1	0.04
(EQ)	Hydantoinase/Oxoprolinase	3	0.13
(F)	Nucleotide_transport,_metabolism	15	0.65
(FG)	Histidine triad (HIT) protein	1	0.04
(G)	Carbohydrate_transport,_metabolism	328	14.30
(GK)	ROK family protein	3	0.13
(GKT)	Sugar:Hydrogen symporter protein	5	0.21
(GM)	Nad-Dependent epimerase/dehydratase	13	0.56
(H)	Coenzyme_transport,_metabolism	26	1.13
(HI)	Citrate lyase	1	0.04
(I)	Lipid_transport,_metabolism	14	0.61
(IQ)	Short-chain dehydrogenase/reductase	6	0.26
(P)	Inorganic_ion_transport,_metabolism	71	3.09
(Q)	Secondary_metabolites	10	0.43
	Poorly characterised		
(R)	General_function_prediction	143	6.23
(RM)	Phosphatase	1	0.04
(S)	Function_unknown	507	22.11
Grand Total		2293	54.14
Not in eggNOG db		1942	45.85

Ascorbate, galactitol, rhamnose, ribulose, sucrose, sorbose, tagatose, and xylose carbohydrate utilisation genes were found in the pan-genome of animal *E. faecium.* Ascorbate uptake systems (PTS) were found variably in most (125/129) *E. faecium.* Two novel ascorbate PTS systems (ORTHOMCL2756 and ORTHOMCL329) were found in chicken and turkey E. *faecium* strains only (E0164 isolated from turkey, E1575 and E2134 isolated from chicken). The galactitol uptake systems (PTS) identified ORTHOMCL2056 and ORTHOMCL2309 are absent from bird strains, except for ostrich strain (E1576) and turkey strain (E0269) (Supplemental File, S2). Sorbose uptake systems (PTS) (ORTHOMCL4685) were absent in most (17/21) of animal *E faecium.*

In the pan-genome of animal *E. faecium* the frequency of replication, recombination and repair function genes is twice that of the same function in the core genome (15.35%). This is likely to be accounted for by the high number of mobile genetic element sequences in the pan-genome of animal strains of *E. faecium* and highlights extensive horizontal gene transfer in *E. faecium* (Table 4.4).

4.2.2.2 Relationships within animal *E. faecium*

Genomic comparisons were performed to investigate the relationship between the various animal *E. faecium* genomes. A presence / absence tree was produced by comparing the orthologous groups based on 6,686 accessory genes of animal *E. faecium* (Figure 4.7). The tree grouped most of

the animal strains from the same origin together in one clade, forming A, B, C and D.

Dog strains were grouped together forming clade A, however strain E1574 is very distinct from other dog strains and very similar to the poultry strain. Pig strains were grouped together forming clade C, but one strain is very different from the rest (E1578) and very similar to the bison strain (Figure 4.7). Most of the bird strains were grouped together in one branch forming clade D, however, a chicken strain (E429) and the ostrich (E1576) strain are very different from other bird strains using this methodology. Turkey strains (E0269 and E0164) are very similar to chicken strains (E0045 and E1575). The tree confirms species diversity between different animal hosts, whereby isolates have a set of genes that correlate with colonisation of their particular host.

A second analysis was performed by generating a neighbour-joining tree based on the distinction of 1,824 shared single copy orthologous groups. The aim is to model the evolutionary descent of the 21 animal sequenced *E. faecium* (Figure 4.8). The phylogenetic tree of animal *E. faecium* confirmed the outcomes of the overall gene content tree, which indicated species diversity associated with different animal hosts, whereby each strain appeared to have core genes that correlated with colonisation of their host. The bird (D), pig (C) and dog isolates have a core genome that appears specific to these hosts (Figure 4.7 and Figure 4.8).

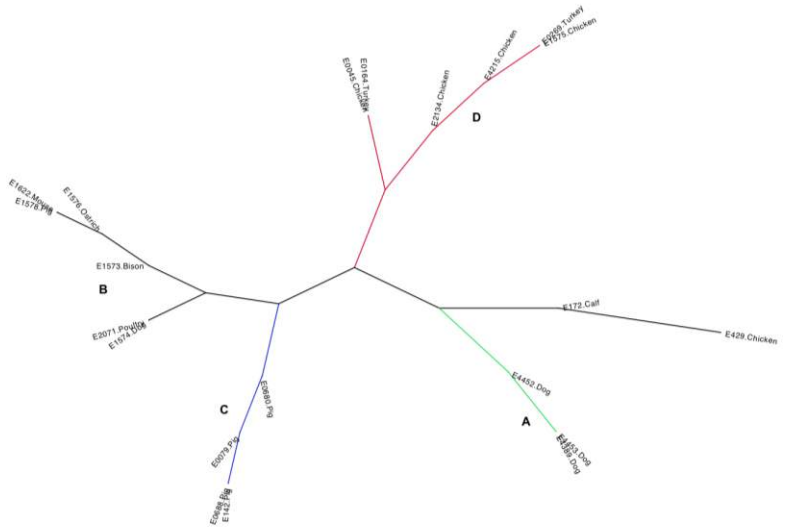


Figure 4.7: Overall gene content tree for animal *E. faecium.* The tree was generated from a comparison of the orthologous groups of publicly available animal *E. faecium* strains based on the overall gene content (presence/absence tree). Bird strains are highlighted in red, dog strains in green and pig strains in blue.

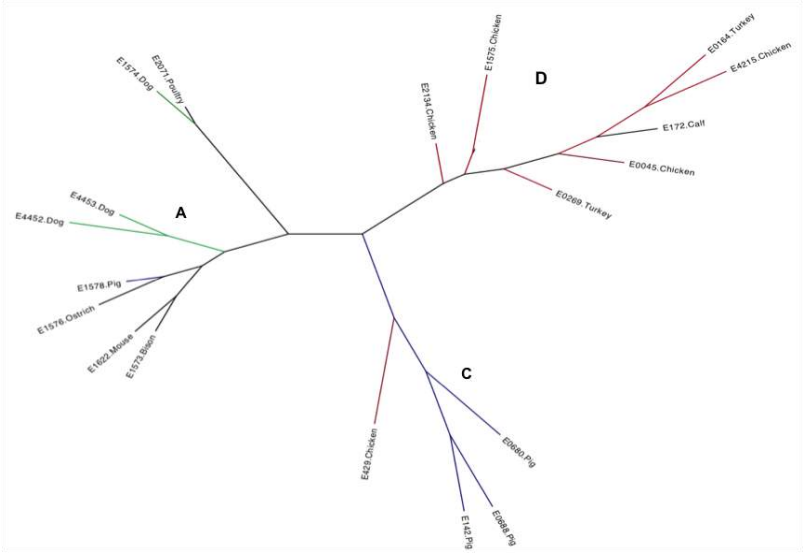


Figure 4.8: Neighbour-joining tree of *E. faecium.* The tree is based on the concatenated alignments of 1,824 shared single copy orthologous groups in 20 animal *E. faecium* genomes. Bootstrapping was performed with 1,000 replicates. The origins of the strains are indicated. Green indicates dog

origin, blue is pig origin and red is bird origin.

4.2.2.3 PhenoLink analyses of animal *E. faecium*

PhenoLink is a web-tool used to identify genetic links with phenotypes (section 2.18.1). PhenoLink analyses were performed using the *E. faecium* genomes to identify genes responsible for the clusters of different animal groups (chicken, pig and dog) and the CC17 group (Supplemental File, S3). The PhenoLink analyses were applied only to the 77 strains that were associated with source details in the NCBI database. Approximately, 117, 145 and 90 gene clades were identified as being responsible for the clade of chicken, pig, and dog strains of *E. faecium*, respectively. Separately, around 450 gene clusters were found to define the CC17 genotype clade. PhenoLink analyses of chicken, pig and dog strains of *E. faecium* identified that approximately 32, 40 and 30 % of the gene clusters responsible for the clades were hypothetical proteins, respectively.

The absence and presence of carbohydrate utilisation genes was associated with the generation of the animal group of *E. faecium.* Galactitol, mannose, L-rhamnose, lactose, galactose, xylulose, ascorbate, and fructose utilisation genes contributed to the clade of different animal group, bird, pig and dog.
In addition, mobile genetic elements such as phage, plasmid and IS elements were also associated with animal clades. For example putative phage encoded protein, plasmid recombination enzyme, putative transposon Tn*552* andtransposase IS30 family also linked to different animal phenotypes (Supplemental File, S3). See also discussion of prophage clusters in chapter

6. Proteinaceous toxins including bacteriocin piscicolin-126-precursor, bacteriocin class II with double glycine leader peptide, enterocin, lactococcin G processing protien and lactococcin A secretion protein LcnD were also involved in the clade of different animal *E. faecium.*

In addition, The presence and absence of several of hypothetical portions also contribute to the formation of the animal group of the animal *E. faecium* 39, 62 and 84 orthologues were found to be unique in bird, pig and dog, respectively. Approximately 49 % of the CC17 phenoLink genes are hypothetical proteins, with 12 % associated with mobile genetic elements. The distinction between CC17 and the other clinical strains of *E. faecium* was not clear suggesting that ac alternative relational tool might be required to dissect the precise drivers of clustering.

4.2.2.4 The novelty of animal *E. faecium* genomes used in this study

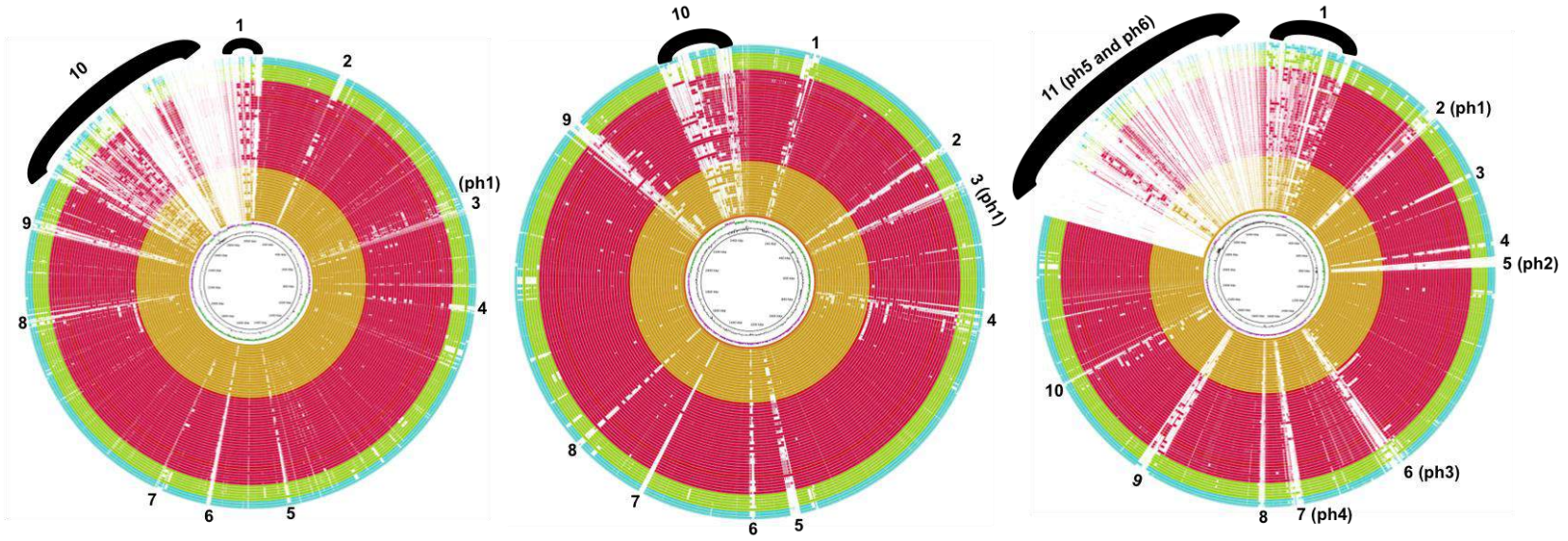
The genome assemblies from the animal *E. faecium* strains from calf, pig and chicken strain (E172, E142 and E429) were compared with those from 61 *E. faecium* genomes that are publicly available with full information. Genome maps reveal that the backbone gene content of *E. faecium* has synteny, which is highlighted by the near continuously coloured region that spans most of the chromosome. The three animal strains of *E. faecium* were used separately as references in the genome map to identify animal-specific regions (Table 4.5. Appendix).

When the E172 (calf) strain resulting from PacBio sequencing was used as the reference genome several novel regions were identified (Figure 4.9.A). Part of region A1 (from 0 to 27 kb) was found in 7 clinical strains including the Texas isolates (Figure 4.9.A) (Table 4.5. Appendix).

Region A2 (from 194 to 212 kb) appear to be a clinical-specific sequence because of the absence of these genes in the commensal isolates while present in most of the clinical strains plus two dog strains (E4453 and E4389) (Figure 4.9.A). In addition, a lactose utilisation operon was found in most of clinical and animal *E. faecium* but was absent from commensal strains excluding strain E1050 (region A4).

Region A6 and A7 seem to be an animal specific region by virtue of its absent from other *E. faecium* strains. A prophage is present in region A9 which shares similarity with a prophage found in bird strains (Figure 4.9.A).

Region A10 (from 2500 to 3000 kb) is likely to be not assembled so location of genes is unclear. The region is composed of plasmid and several heavy metal resistance genes (Figure 4.9.A) (Table 4.5. Appendix).



A. B. C.

Figure 4.9: Animal *E. faecium* genome maps. A. Circular map of predicted genome sequence from the comparator genome E172 (calf), B. Circular map of predicted genome sequence from strain E142 (pig). C. Circular map of predicted genome sequence from E429 (chicken). Genome comparisons are presented the predicted genome sequence from 61 human clinical strains, commensal and animal strains of *E. faecium*.

Key for the circular identifiers, moving from the centre circle outwards: CG content, GC-skew (G-C/G+C), reference genome, animal strains indicated in gold; chicken 2 (E1575), chicken 3 (E0045), chicken 4 (E2134), chicken 5 (E4215), Turkey 1 (E0269), Turkey 2 (E0164), Ostrich (E1576), Poultry (E2071), Pig 2 (E1578), Pig 3 (E0688), Pig 4 (E0679), Pig 5 (E0679), Dog 1 (E4452), Dog 2 (E4453), Dog 3 (E1574), Dog 4 (E4389), Bison (E1573), clinical isolates in red; DO, D344SRF, E1636, E1679, E1071, E1162, E1258, E1185, E1392, E1552, E0120, E1904 ,E1626, E2883, E2297, E1627, E1731, E1634, E6045, E1644, E6012, U0317, TX0133A, TX0133a01, TX0133a04, TX0133B, TX0133C, TX0082, 1231502, 1232408, 1230933, 1231410 and 1231501. Commensal strains (green); E1039, Com12, Com15, TX0130, E980 and E1050. Food and river water (blue); Cheese (E1604), Fish burger (E1613) and River water (E1630). The dark colour inside rings indicates 100% identity, light colour inside rings indicated 70% identity and grey colour inside rings indicates 50% identity. The clear parts on the rings indicate unique regions in the reference.

The genome of E142 (pig) is the smallest genome among the animal *E. faecium* (~2.5Mb) and the mega plasmid that is found in the chicken and calf genome is absent from this genome. These factors led to less interference in formation of the genome map (Figure 4.9. B) (Table 4.5. Appendix).

Region B2 (from 385 to 393 kb) exist in one pig strain E1578 and only strain 1, 231,410 of the clinical isolate. The prophage that is located in region B3 (from 452 to 468 kb), has homology with two pig strains (E0688 and E0679) and a clinical isolate E1552 (Figure 4.9. B).

A small plasmid is located in region B5 (from 1175 to 1192 kb) and parts of this plasmid have high similarity with sequences present in many clinical and animal isolates of *E. faecium*, but this plasmid is absent from commensal isolates (Table 4.5. Appendix). Region B6 from 1254 to 1262 kb is encoded in multiple animal and clinical strains of *E. faecium* including dog (E4453), bison, turkey (E0164), poultry, pig (E1578), E1071, E1552, E1627, E1634 and one commensal strain E1050 and notably nearly all of these strains were isolated from faeces and from the same geographic region (The Netherlands) (Figure 4.9. B).

Region B7 (from 1452 to 1461 kb) is present in all pig strains plus a clinical strain E1552. Region B8 (from 1586 to 1596 kb) presents in eight animal strains, nine clinical strains and the commensal strain E1050. Region C9 (from 2175 to 2205 kb) presents in one pig strain (E0688) and the clinical strain (E1552) (Figure 4.9. B). Region B10 contains the mega plasmid that share similarity with plasmids found in calf strain (E172) region A10 (Figure 4.9. A) (Table 4.5. Appendix).

By switching to strain E429 (chicken) genome as the reference several novel regions and low identity regions with other strains were clearly identified. Within these regions parts of the sequence were shared with several *E. faecium* genomes (Figure 4.9.C). The mosaic structure in region C1, that is

located from 0 to 200 kb, due to the absence of many parts of this region in most other *E. faecium* strains. The encoded functions strongly suggest that a mega plasmid is located in this region in the chicken strain (Figure 4.9.A); see also chapter 5, section 5.2.1.2 and Figure 5.3. The chicken strain assembly suggesting that this plasmid is integrated to the chromosome due to its located in the backbone.

A prophage in region C2 (from 415 to 455 kb) has similarity with a prophage found in another chicken strain (E1575), two dog strains (E4452, E4453), the bison and a poultry strain. Hypothetical proteins associated with this prophage have identity to proteins found in clinical isolates from bloodstream infection (TX0133A, TX0133a01, TX0133a04, TX0133B and TX0133C, 1232408, 1231410 and 1231501). Three commensal strains (E980, E1050 and E1039) also contain a genomic region with similarity to the phage in region C2 (Figure 4.9.A).

Region C3 (603 to 606 kb) absents from most *E. faecium* strains, being only in E1050 (commensal strain), plus the river water and 15 of the animal and clinical strains. Region C4 (from 769 to 777 kb) presents in other animal *E. faecium* strains (ostrich, dog (E1574), plus clinical (E1636, E1679, E0120, E1904), commensal (E1039, Com15) and food isolates (cheese, fish burger).

A prophage in region C6 (from 1340 to 1385 kb) is not present in most other *E. faecium*. However, it shares homology with only one animal-

associated strain (bison), commensal strains E1039, Com15 and E1050 and the clinical isolates (E1679 and E1904). In addition, the prophage in region C7 (from 1612 to 1619 kb) has homology with prophage that are found in the commensal strain E980 and the clinical strains (E1904 and E1731) only.

C11 contains numbers of poorly assembled region that contains unscaffolded contigs, so the location of genes in this region is unclear (See chapter 3). The region is about 670 kb in size and started from 2680 to 3350 kb. This region shares similarity with regions A10 and B10 in both calf and pig strains of *E. faecium* (Figure 4.9.A and B) (Table 4.5. Appendix).

4.3 Discussion

4.3.1 Core and pan-genome of *E. faecium*

The core genome size of *E. faecium* was estimated by van Schaik *et al* (2010) using seven strains to be 2172 (+/-) 20 genes which is much higher than the size of the core genome in this study (1,467 genes). The difference between the two estimates is a result of the number of strains that used being seven *E. faecium* genomes against 129 genomes in this study (Figure 4.1). As expected, the number of shared genes was reduced with addition of each new sequence (Tettelin, Masignani *et al*. 2005).

Lebreton *et al* (2013) stated that there is a slight difference in the core genome size of human infection isolates, including the clonal complex CC17 strains, which have larger core genomes (1,945 genes) than strains of

non-hospitalized humans strains (1,805 genes) or strains of a mixed group of animal and sporadic human infection (1,724 genes), which appears stable despite the claim that this is a very recent emergence of this CC17 group.

Comparing the proportion of this functional category in the core genome of *E. faecium,* animal isolates possess more carbohydrate transport and metabolism functions (10.26%) than other strain groups (Table 4.3). Carbohydrate transport and metabolism functions in *E. faecium* is very high when compared with other Gram-positive bacteria such as *Bacillus cereus* and *Bacillus subtilis* (0.07%) and reflects the capacity of *E. faecium* to utilise an array of carbon sources from plant origin (Alcaraz, Moreno-Hagelsieb *et al*. 2010). This finding reflects a specialisation in the metabolism of carbohydrates in animal *E. faecium* when compared to human isolates. It is well documented that animal isolates have considerably more genes for the degradation of carbohydrates. Fructose, mannitol, galactose and glucose uptake systems genes are found in the core genome of animal *E. faecium* plus they have the metabolic potential for the uptake and assimilation of plant-derived carbohydrates that exists in foodstuffs of their host.

Van Schaik *et al* (2010) estimated that almost 30% of the *E. faecium* genome seems to be accessory compared with an estimate here of 89%. It was confirmed that the *E. faecium* pan-genome is estimated to be broadly unlimited in size. van Schaik *et al* (2010) and Qin *et al* (2012) suggested that the open pan genome of *E. faecium* could be described by its capability

to assimilate foreign DNA into the gene pool. Since, *E. faecium* has a wide variability of ecological niches that it colonises and survives, and this life cycle might require a high degree of phenotypic adaptability. The wide variety of ecological niches has resulted in there being interaction of *E. faecium* with many non-pathogenic and pathogenic bacteria for example Bacilli, staphylococci, and streptococci, and extensive horizontal gene transfer between *E. faecium* and these bacteria has been documented. de Been *et al* (2013) also suggested that the *E. faecium* genome is highly plastic and limited barriers occur for the acquisition of foreign genetic elements, confirming high levels of recombination in *E. faecium*, which distinguished the existence of hybrid *E. faecium* strains. The significance of the open pan-genome is that the species has a high diversity of genes that could raise the fitness of the species in different environmental conditions. The increase of antibiotic resistance genes documented in clinical isolates and the *esp* gene that is located on a genomics island are well-described examples of the *E. faecium* gene pool that has been positively selected in the farm and clinical environments (van Schaik, Top *et al*. 2010, Qin, Galloway-Pena *et al*. 2012, Lebreton, van Schaik *et al*. 2013).

An open pan-genome was also shown in *Streptococcus agalactiae*, which is projected to contribute new genes when each new sequenced strain is added to the pool. In different species such as *Bacillus anthracis* the dynamics are distinctive and no predicted new genes were gained after when new sequenced strains added and its pan-genome can be fully described by four

genomes only, this is called a closed pan-genome (Tettelin, Masignani *et al*. 2005, Alcaraz, Moreno-Hagelsieb *et al*. 2010).

The presence of carbohydrate utilisation genes in the pan-genome of both *E. faecium* overall and in particular animal *E. faecium* is high. Some of these carbohydrate uptake system pathways appear to be novel for specific animal hosts. One possible scenario is that these carbohydrates could be present within feedstuffs thereby providing a direct selection for enteric bacteria that possess genes for their uptake and metabolism. The pan-genome of the animal *E. faecium* sub-populations (dog, pig and bird) have acquired genes that appear specific to each sub-population, including carbohydrate utilisation genes. These genes might be acquired to their genome from the food chain within a food promoter or from plant material.

The existence of certain carbohydrate uptake systems, such as glucose, might be associated with virulence of *E. faecium*, since glucose utilisation genes were found only in the clinical strains of *E. faecium* except one for animal strain isolated from dog. New habitat adaptation and the occurrence of new lineages relate directly to the gain and loss of genes. The occurrence of lateral gene transfer alters completed ancestral genome size (Dagan and Martin 2007) .

4.3.2 Phylogenetic and diversity of *E. faecium* genome

Core genome phylogenomics was achieved by comparing all the shared (orthologous) genes amongst all *E. feacium* isolates plus animal *E. faecium* isolates. The in-depth study of the core genome might answer relevant

evolutionary questions, for example what are the conserved genes within a different *E. faecium* sub-population range?

Phylogenetic analysis in this study, based upon core genes (Figure 4.3), gene content difference analysis (Figure 4.4) together with recent sequence studies of 16S rRNA and SNPs, indicates a clear and pronounced separation among community-associated, hospital-associated and animal-associated clades (Galloway-Pena, Roh *et al*. 2012, Qin, Galloway-Pena *et al*. 2012, Lebreton, van Schaik *et al*. 2013). In addition, there is a clear separation within *E. faecium* from different animal hosts (dog, chicken and pig) (Figure 4.7 and Figure 4.8).

The genomic data, in this study supports the recent studies that suggest the CC17 (clade A1) and commensal (clade C) appear to have formed a sub-population within the *E. faecium* species (Figure 4.3). It is also clear that these infectious isolates are not clonally associated with each other and have spread noticeably. In addition, analyses in this study confirmed that CC17 genotype cluster closely together and further away from the commensal isolates than the other infectious isolates, supporting the hypothesis that the CC17 genotype might represent a recently evolved genotype (van Schaik, Top *et al*. 2010, Lam, Seemann *et al*. 2012, Palmer, Godfrey *et al*. 2012, Qin, Galloway-Pena *et al*. 2012, de Been, van Schaik *et al*. 2013, Lebreton, van Schaik *et al*. 2013). In addition to the human strain evolution, animal isolates (clade B) also seem to have formed a sub-population (Lebreton, van Schaik *et al*. 2013). The timescales leading to this divergence is not clear.

Gene content analysis showed that strains designated as animal, clinical and commensal are different at the level of their genetic repertoire (Figure 4.4), however, each sub-population appears relatively closely related. Different sub-populations of animal *E. faecium*, including bird, pig and dog differ. For example, in terms of both their core genes and their overall gene content, *E. faecium* isolated from birds are grouped together in one clade (Figure 4.7 and Figure 4.8). Being grouped together as a specific sub-population might indicate that this clade of strains contain genes for colonising their bird host or some other aspect of their lifecycle (Figure 4.8). Large differences in gene content within the sub-populations of *E. faecium* species detected here indicate that, at the level of their core genome even in relatively closely-related isolates, the gain and/ or loss of mobile genetic elements is a major influence in shaping strain-specific properties (van Schaik, Top *et al*. 2010).

Two commensal strains E1050 and E1039 seem to represent hybrid genomes with clinical clade. Moreover, clinical strain 1,231,408 appears to be hybrid with the genome of commensal (Lebreton, van Schaik *et al*. 2013). One of the animal isolates from a pet dog (E4453) was associated with clade A1, which contains CC17 strains, which identifies potential links between hospital strains and household pets (de Regt, van Schaik *et al*. 2012). This link might occur by this strain having been transmitted to the dog from a human and being a transient coloniser or it could be a genuine resident of the dog.

Xylose utilisation genes together with fructose, galactitol, glucose, mannitol and galactose utilisation genes were found in the core genome of every animal *E. faecium.* These genes therefore represent a suite of animal *E. faecium* sugar utilisation mechanisms that may be required to colonise their animal host. The presence of these genes is likely to be a contributing factor that explains the separation of animal and human clades of *E. faecium.*

Two strains from this study isolated from calf (E172) and chicken (E429) appear to possess hybrid genomes. The calf strain (E172) contains a backbone genome must closely matching the bird clade. The chicken strain (E429) shows a very different backbone genome than other chicken strains in the bird clade, and a very similar backbone to the hybrid genome of the commensal strain E1093. Differences in the presence and absence of mobile genetic elements particularly phages and hypothetical proteins appear to explain the grouping of the chicken strain (E429) with E1039. This finding supports the hypothesis of van Schaik *et al* (2010) that even between relatively-closely related strains the repertoire of mobile genetic elements is a major influence in shaping strain-specific properties.

The geographic and the infection origin of *E. faecium* strains appear to play an important role in determining the separation of clades A, B and C (Figure 4.2 and Figure 4.4). As an example, strains isolated from Texas were grouped together in subgroup A2. With most of the strains in group A being isolated from blood and from USA. Group B contains animal, clinical, CC17 and river water isolates, mostly from Netherlands.

4.3.3 *E. faecium* sub-populations

Core and pan-genome analysis of *E. faecium* indicated that there are three main sub-populations of *E. faecium* species, including hospital-associated (clade A), animal-associated (clade B) and community-associated strains (clade C) and there are specific genes for these clades suggesting potentially unique gastrointestinal tract niches. In *E. faecalis* phylogenetic multiple analysis clades are not observed (Palmer, Godfrey *et al*. 2012, Kim and Marco 2014). Contrasting markedly with *E. faecium* where within sub-populations clear different sub-groups are present and these are associated with different animal hosts (bird, dog and pig) and the CC17 genotype (Figure 4.3 and Figure 4.8). Differences between individual orthologous clusters were compared to obtain genes that contributed to the genetic separation between the *E. faecium* clinical, commensal and animal isolates.

Several different functional categories were represented among the clinical isolates of *E. faecium.* Cell wall components that were found to be absent in the genome of the clinical *E. faecium* strains, such as capsular polysaccharide biosynthesis proteins, were positively correlated with the clinical group and these genes may play a role of increasing survival from innate defences such as opsonophagocytosis in the host thereby contributing to infection. The presence of particular lipoprotein may play a role in *E. faecium* virulence procedures. Study of lipoproteins in *E. faecalis* indicated that about 25% of the surface-associated proteins are lipoprotein with a potential involvement in *E. faecalis* virulence and producing candidates for vaccine production (Reffuveille, Leneveu *et al*. 2011).

The commensal group was found to have few mobile genetic elements and antibiotic resistance genes while enriched with genes encoding hypothetical (membrane) proteins and for capsule and vitamin biosynthesis, and sugar metabolism (Kim and Marco 2014). The absence of certain genes such as autolysin, which is a cell wall degrading protease that has the ability to alter host cell peptidoglycan plus a recombination protein, which might play a role in the acquisition of the antibiotic resistance (Qin, Singh *et al*. 1998, Boumghar-Bourtchai, Dhalluin *et al*. 2009) were found to drive the formation of the commensal group. However, how these genes might act to define strain group is unclear.

4.3.4 The novelty of animal *E. faecium* genomes used in this study

Comparison of the animal *E. faecium* with the other 58 *E. faecium* genomes with known source data revealed a mosaic-like structure, as previously described (Sillanpaa, Prakash *et al*. 2009, Qin, Galloway-Pena *et al*. 2012), revealing several highly variable regions. Some of these variable *E. faecium* regions are animal and clinical clade-specific (Figure 4.9). Notably, several regions on animal *E. faecium* genomes are absent or have low sequence identity in the commensal strains. Largely, mobile genetics element such as mega plasmid, phages and IS element are can fined to the variable regions of animal strains. Chapter 5 and 6 will examine these elements in details to characterise more fully these elements in animal and human E. *faecium* isolates.

The mosaic structure at the end of the three animal genomes was identified as a mega plasmid that encodes heavy metal resistance, antibiotic resistance and multible carbohydrate utilisation genes for mannose, trehalose, ribose, galactitol, mannose, L-rhamnose, lactose, galactose, xylulose, ascorbate, and sucrose. These sugar uptake genes were proposed previously as a potential reason for the separation of animal and clinical sub-groups of *E. faecium.* By analysis of the presence/absence of this novel region in the three animal *E. faecium* strains sequenced in this study it is clear that most of these carbohydrate utilisation genes were acquired via this mobile genetic elements. This confirms that horizontal gene transfer events have contributed significantly to the diversity of the *E. faecium* species, but in this case was not phylogenetic driver that distinguished clades.

Unique capsular polysaccharide synthesis proteins and other surface-acting proteins, such as sortase A and an LPXTG motif proteins were found in the 3 animal strains, which might have significant roles in virulence, such as adhesion immune defence and might be required to colonise their specific host (Qin, Galloway-Pena *et al*. 2012). Siezen *et al* (2006) suggested that novel genes encoding cell-surface proteins in Gram-positive bacteria signifying a niche-specific distribution (Siezen, Boekhorst *et al*. 2006).

Region C1 in the chicken strain (E429) (0 to 200 kb) which encodes the mega plasmid seems to be integrated into the chicken chromosome. This hypothesis will be tested in chapter 5 to identify integration of this plasmid. In addition, other mobile genetic elements present in the animal strains

together with antibiotic resistance will be compared with human *E. faecium* in chapter 5 to identify if these genes are similar or distinct to those carried by human isolates of *E. faecium.* Several novel regions in the three animal *E. faecium* are prophage and several could be animal specific. Comparative analysis of *E. faecium* prophages is explored in details in chapter 6.

Chapter Five: Mobile genetic elements in the genomes of *E. faecium* isolated from animals.

5.1 Introduction

"Horizontal genomics" is a new area of prokaryotic biology that investigates DNA sequences present in the chromosome that appear to have originated from other prokaryotes or eukaryotes. Plasmids, bacteriophages and transposons encode the capability to mobilise from one host to another (Frost, Leplae *et al.* 2005).

Galloway-Pena *et al* (2012) stated that the gain of mobile genetic elements carrying antibiotic resistance, virulence and/or fitness factors are the driving force behind the recent success of *E. faecium* as an opportunistic pathogen in hospitals. Investigations of gene clusters that are associated with vancomycin resistance and Tn*1546* in *E. faecium*, reported that horizontal gene transfer occurs between human and animal *E. faecium* isolates (Stobberingh, van den Bogaard *et al.* 1999, van den Bogaard, Willems *et al.* 2002).

In addition, the *esp* virulence gene is located on a large pathogenicity-associated island in *E. faecium* and this *esp* PAI can be transferred horizontally and inserts in a site-specific manner (Leavis, Top *et al.* 2004, van Schaik, Top *et al.* 2010). MGEs are transferred to human isolates and thereby add to the burden of the disease caused by *E. faecium*, for example, by transferring vancomycin resistance between bacteria. This capability is important to consider, since these genes were shown to be transferred to human isolates and to more virulent organisms such as *Staphylococcus aureus* (Qin, Galloway-Pena *et al.* 2012).

Specific aim

In this chapter comparative analysis of mobile genetic elements among *faecium* isolates will be determined to identify if those carried by animal isolates of *E. faecium* are similar to, or distinct from human isolates.

5.2 Results

5.2.1 Mobile genetics elements

5.2.1.1 Insertion sequence elements (IS)

The accessory genome of *E. faecium* has an extensive suite of transposable elements. The presence and the absence of insertion sequence elements and transposase orthologues in the pan-genome of *E. faecium* showed hierarchical clustering using a Pearson correlation algorithm (MeV-Section 2.18.1). Comparative analysis of IS elements in the pan-genome of *E. faecium*, including animal and human isolates, shows differences in the presence of these elements between the *E. faecium* sub-populations representing commensal, clinical and animal isolates (Figure 5.1).

IS elements were located in all *E. faecium* genomes including commensal isolates. However, there is a higher frequency of IS elements in the genomes of clinical isolates are absent in commensal isolates. Different sub-populations of *E. faecium* have particular IS elements the combination of which are unique to each, including the CC17 genotype and animal isolates (Figure 5.1).

From the comparative analysis of IS elements most of the CC17 isolates were grouped together (clade A1). The clinical blood strains isolated from Texas appear to share a unique set of IS elements (clade A2). Animal isolates of *E. faecium* were grouped into two distinct sets forming B1 and B2 and each group has a unique complement of IS elements (Figure 5.1).

IS30 was present in most *E. faecium* strains, including clinical, commensal and animal isolates. IS66 and IS605 were also commonly found only in clinical and animal *E. faecium* strains, however, IS66 was found in 19 isolates that mostly belong to the CC17 genotype, including two dog isolates, which suggests that IS66 could be a marker for this group genotype. IS605 is common to clinical and animal *E. faecium* (85 strains). IS2 was found only in four isolates of *E. faecium*; chicken (E429), calf (E172) pig (E0680) and a clinical strain (E1679) (Supplemental File, S1). A presence / absence tree of transposase orthologues in the pan genomes of *E. faecium* was generated and this groups together the chicken (E429) and calf (E172) strain, suggesting that they share a repertoire of IS that distinguishes these strains, which indicated that these sequences could either be novel to these strains or have been horizontally acquired (Figure 5.1). Generally, animal strains shared specific IS elements, for example, the IS elements present in most turkey, dog and chicken strains were grouped in a clade specific for these hosts.

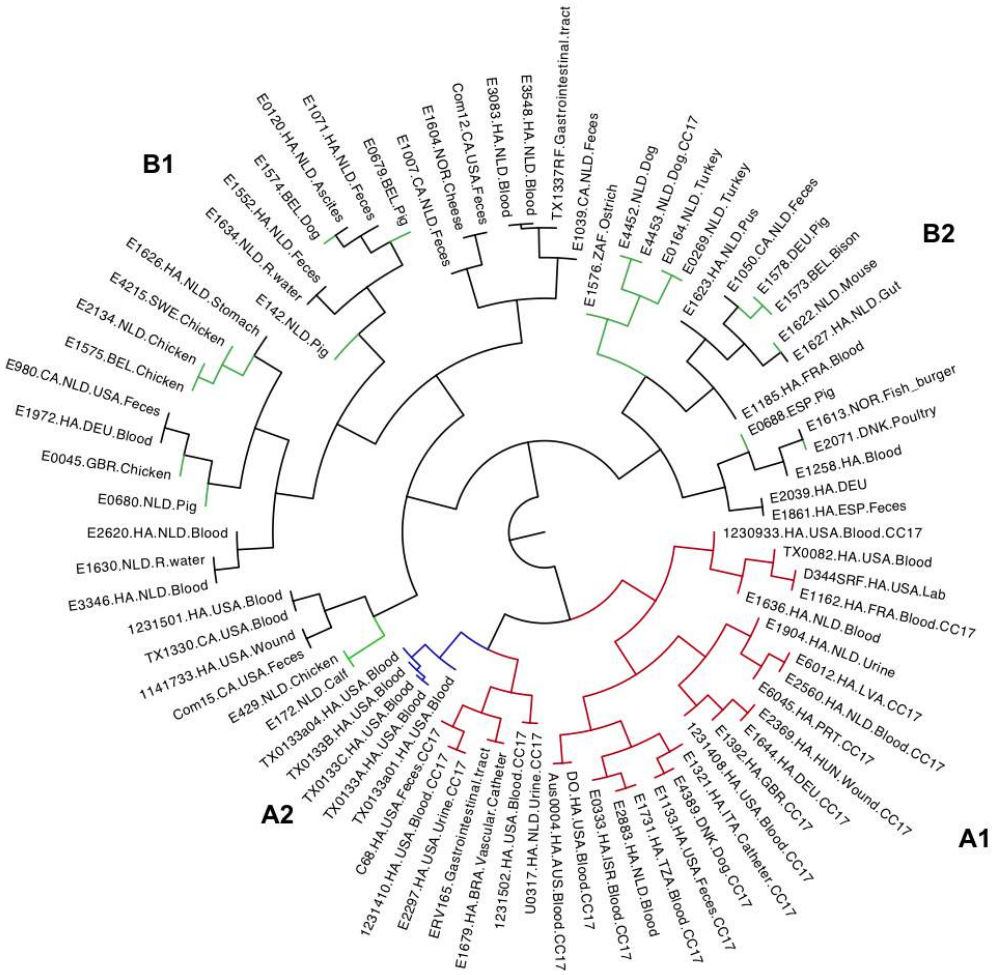


Figure 5.1: A presence and absence tree of transposase orthologues in *E. faecium.* The red clade indicates CC17 genotype isolates, blue indicates Texas strains, and green indicates animal isolates.

Further analysis was performed to investigate the unique IS elements in animal *E. faecium*. Using IS finder (Section 2.18.1) the number of IS elements in each genome was estimated at 180 (chicken, E429), 129 (calf, E172) and 45 (pig, E142) (Table 5.1). These IS elements revealed substantial homology with Gram-negative species including *Escherichia, Burkholderia*, *Pseudomonas* and *Xanthomonas* species, and Gram-positive

species including *Staphylococcus*, *Streptococcus*, *Bacillus* and *Lactobacillus* species.

Table 5.1: Insertion sequence elements in animal *E. faecium.* IS families in the three animal strains E429 (chicken), E172 (calf) and E142 (pig) according to the IS Finder database.

IS Family	IS group	Chicken (E429)	Calf (E172)	Pig (E142)
IS1	-	13	13	0
IS110	-	1	1	0
IS1182	-	7	4	1
IS1380	-	5	5	2
IS1595	ISPna2	1	1	1
IS1634	-	1	0	0
IS200/IS605	-	2	1	1
IS256	-	12	10	6
IS3	IS2/ IS3/ IS150	32	20	10
IS30	-	5	4	3
IS4	1S10/ IS231	8	6	0
IS5	IS5	47	26	0
IS6	-	28	27	14
IS607	-	1	1	1
IS982	-	2	2	2
ISAs1	-	2	2	0
ISL3	-	11	6	3
ISLre2	-	0	0	1
Grand Total		180	129	45

In particular, the IS1 and IS5 families share homology with elements of *E. coli* and *Pseudomonas aeruginosa,* respectively, and the IS6 family has homology with *Lactococcus lactis* and *Staphylococcus aureus.* A very similar number of IS elements in chicken, calf and pig isolates, were found to have homology with *E. faecium* and with other *Enterococcus* species, including *E. faecalis*, *E. hirae* and *E. casseliflavus*.

5.2.1.2 Plasmids

Many plasmids have been described in *Enterococcus* species that confer resistance to antimicrobials and heavy metals. To first investigate the extra-chromosomal plasmid content of the three animal strains of *E. faecium*, plasmid DNA was purified and visualised by gel electrophoresis. Three similarly sized plasmids were observed in the three animal strains, estimated at ~ 4.7 kb in size (Figure 5.2). The calf strain (E172) potentially contained at least one more plasmid of smaller size (~ 1.5 kb)

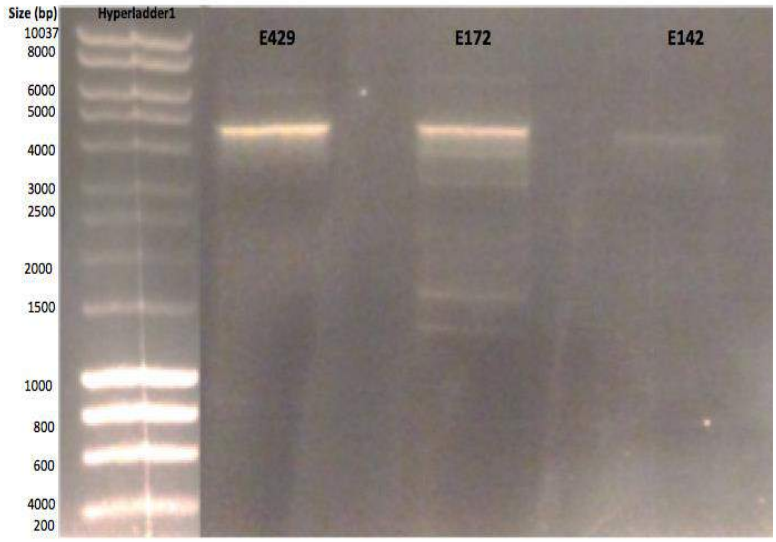


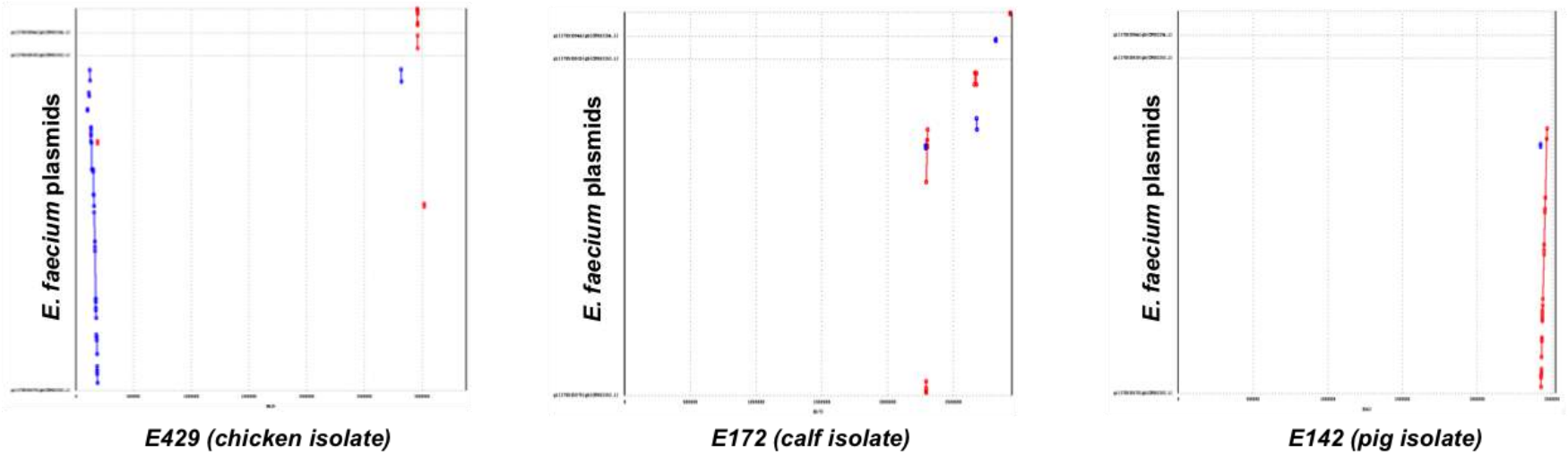
Figure 5.2: Gel-electrophoresis of plasmid DNA. Lanes from left to right: Hyperladder1; E429 (chicken strain); E172 (calf strain); E142 (pig strain).

To characterise the plasmid complement of the three animal strains *in silico* a comparative analysis was made with the 34 *E. faecium* plasmid sequences that were publicly available (Figure 5.3). This analysis indicated that the animal strains of *E. faecium* isolated from chicken and pig, each contain

DNA corresponding to mega-plasmids present in the closed genomes of *E. faecium* Aus0004, DO and strain Aus00085 (Figure 5.3). Strains E429 (chicken) and E142 (pig) appear to have the same mega plasmid, but located with a different synteny (scaffold 1 and 2, respectively). Strain E172 (calf) only possesses segments of this mega plasmid.

The plasmid sequence identified in animal isolates were found to have homology with strain DO plasmids (DO1 (CP003584.1, 36.26 Kb), DO2 (CP003585.1, 66.25 Kb), DO3 (CP003586.1, 251.93 Kb), strain Aus0004 plasmid Aus0004_p1 (CP003352.1, 56.52Kb) and strain Aus0085 plasmids P1 (CP006621.1, 130.72 Kb), P2 (CP006622.1, 67.31 Kb) and P3 (CP006623.1, 31 Kb).

The annotation of the DO, Aus0004 and Aus0085 identified plasmids that found in the complete genomes of *E. faecium*,,which have homology with animal isolates plasmid, reveal a variety of encoded functions, including toxin–antitoxin, sortase A and an *LPXTG* cell wall anchor protein. In addition, the plasmids contain genes encoding tetracycline resistance and multiple bacteriocin genes. Some of these genes may be found on plasmids but they are not necessarily plasmid genes.



A. **B.** **C.**

Figure 5.3: Comparative analysis of *E. faecium* plasmid sequences. Mummerplot analysis reveals homology between animal strain genomes (E429, E172 and E142) and 34 complete plasmid sequences retrieved from the NCBI database. (A) Plot identifies a mega plasmid within the assembled chicken genome (E429. (B) Plot revealing sequences homologous with plasmid in the calf strain (E172) and (C) the pig strain (E142), which appears to also have a mega plasmid.

Several of the novel animal genes (22 genes encoding hypothetical proteins) were located on a plasmid. Carbohydrate utilisation operons were identified in chapter 4 as being located on plasmids and these operons were identified with specificities for citrate, and ascorbate, resistance to heavy metal including lead, cadmium, zinc and mercury. These genes form the novel region C1 in the chicken genome map (Chapter 4 _Figure 4.10.C).

Analysis of plasmid genome content across all of the *E. faecium* genomes revealed relationships based on shared DNA sequences (Figure 5. 4). Genes carried by plasmids in animal *E. faecium* were found to be common across *E. faecium* strains, including the commensal isolates. The co-occurrence of the plasmid with animal and CC17 strains show strong association since most of the animal strains were located in a clade different from the CC17 strains, which suggested that animal strains contains plasmid genes specific for animal host.

Some of the plasmid genes, for example helix-destabilizing protein, helix-turn-helix domain protein and a sortase (surface protein transpeptidase) were found as core genes in *E. faecium* isolates (Figure 5. 4).

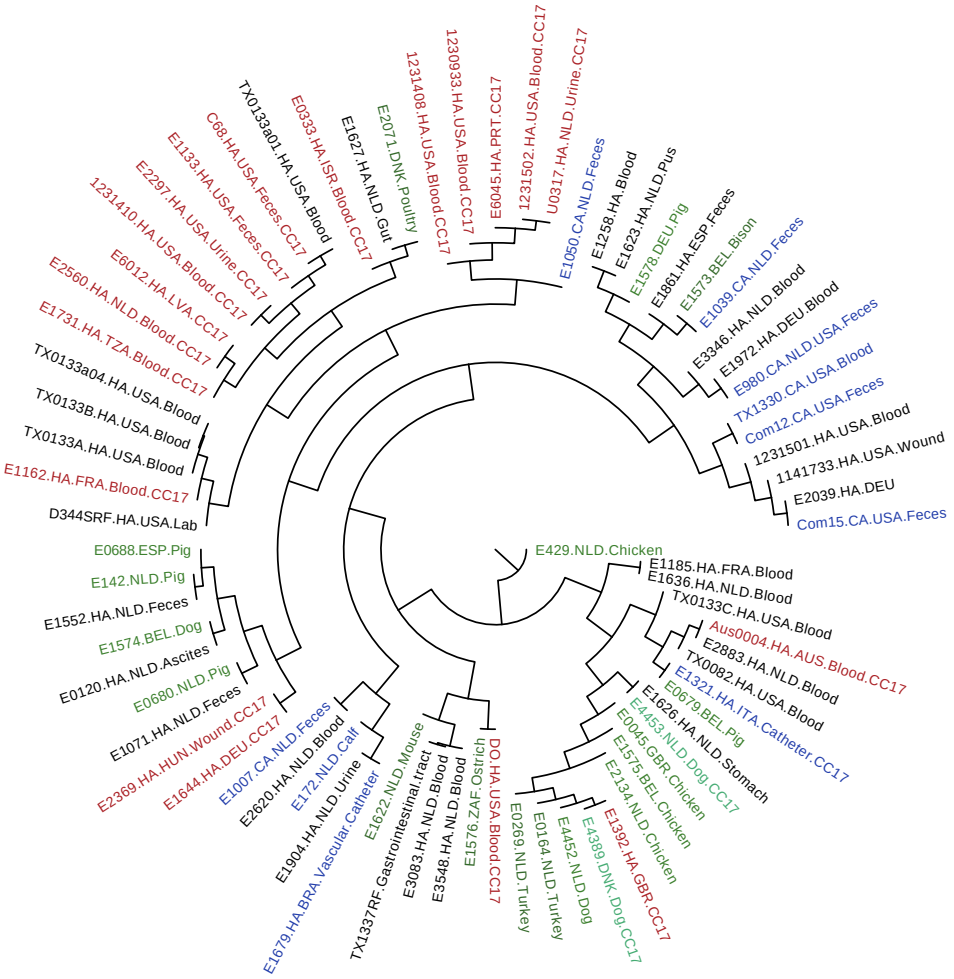


Figure 5.4: A presence and absence tree of plasmid orthologues in *E. faecium.* The red clade indicates CC17 genotype isolates, blue indicates commensal strains, green indicates animal isolates and black indicates other clinical isolates.

5.2.1.3 Bacteriophage

Phages have been described that were resident in *E. faecium* strains or that were shown to infect the species (Mazaheri Nezhad Fard, Barton *et al*. 2010, van Schaik, Top *et al*. 2010, Yasmin, Kenny *et al*. 2010, Galloway-Pena, Roh *et al*. 2012). Van Schaik *et al* (2010) indicated that the prophages that

have been induced from *E. faecium* are *Siphoviridae* and morphologically identical to prophages induced from *E. faecalis.*

The genome sequences of the *E. faecium* strains isolated from chicken (E429), calf (E172) and pig (E142) contain prophages. The genome size differences between the chicken strain and other two animal strains are mostly due to the acquisition of horizontally transferred of genetic material, and a major part of this derives from temperate bacteriophage. Six phage regions were found in chicken strain E429 compared with only one in calf and pig strains. *E. faecium* prophage are discussed in detail in chapter 6.

5.2.2 Investigating animal *E. faecium* genomes with regards to virulence, resistance and survival.

5.2.2.1 Virulence factors

BLAST analysis of candidate virulence factor genes present in human strains of *E. faecium* confirmed the presence of multiple virulence genes. The enterococcal surface protein (encoded by *esp)*, collagen adhesin precursor (encoded by *acm)*, secreted antigen SagA, pilus (encoded by *pilA* and *pilB*) and hemolysin (Table 5.2) are variably present among the three sequenced strains revealing that known virulence determinants reside in their genomes.

Table 5.2: Virulence factors in animal *E. faecium.*

Virulence gene product	E429 (chicken)	E172 (calf)	E142 (pig)
LPXTG surface protein	10	5	10
Collagen adhesin precursor	1	1	4
Esp	1	1	1
EspA	7	0	0
PilA	2	1	1
PilB	0	0	1
SagA	1	1	1
Hemolysin	3	3	3

The virulence proteins in animal *E. faecium* have 93 to 100% similarity with virulence genes in *E. faecium* as a whole, namely collagen adhesin precursor (AAN12397), PilB (ACI49665), PilA (ACI49671) and SagA (AF242196_3).

Collagen adhesin precursor gene was found in the chicken strain (position 520769-522466), calf strain (1474869-1477595) and with four copies (1504841-1506127, 1506662-1507090, 1507087-1507290 and 1507290-1507571) in the pig strain. The PilA gene is located in positions 178302-178679 and 180785-182434 in the chicken strain and 2429443-2431419 in the pig strain. Collagen adhesin precursor gene was found in most *E. faecium* isolates including clinical, commensal and animal. However, a novel collagen adhesin precursor homolog was found only in bird isolates and the calf strain (E172). PilB was found in the pig strain only (119904-121781). The SagA gene is located at positions 2654224-2655801, 1761135-1762700 and 1798805-1800376 in chicken, calf and pig, respectively. Hemolysin genes are located at 80020-801583, 971478-972131, and 2349596-2350903 in the chicken strain, 1051269-1051922,

1201273-1202649 and 2072066-2073367 in the calf strain and 1070554-1071207, 1229858-1231234 and 2099474-2100775 in the pig strain.

LPXTG family cell-wall anchored proteins were found in the three animal *E. faecium* genomes as multiple copies. At least 5 of these genes are novel since no significant similarity was found in the NCBI database which includes those from Gram-positive species including *Staphylococcus*, and *Lactobacillus* species. LPXTG in positions 3145123-3145713 and 3169914-3170672 in chicken strain share high level of similarity (89%) with LPTXG in *Lactobacillus brevis* and (98%) to Cna protein B-type domain protein in *Staphylococcus aureus*, respectively.

The gene encoding hyaluronidase was absent from the three animal *E. faecium* isolates, in contrast to its presence in all CC17 genotype isolates, confirming it represents a signature of this CC17 genotype. The gene encoding the enterococcal surface proteins Esp and EspA share low level of similarity (23 to 36%) with Esp (ZP_06678454) and cell wall surface anchor family protein EsbA (ZP_06702708) in *E. faecium* strains E1162 and U0317, respectively. The *Esp* gene is located at positions 1961868-1965038, 104424-107594 and 133123-135006 in the genomes of chicken, calf and pig, respectively.

5.2.2.2 Antibiotic resistance

Comparative analyses of antibiotic resistance genes among *E. faecium* isolates were previously reported by Qin *et al* (2012) and Lebreton *et al*

(2013) and revealed the widespread occurrence of antibiotic genes in *E. faecium* species (Table 5.3). A comparative analysis of antibiotic resistance genes in the three sequenced animal *E. faecium* isolates in this study was done by performing BLAST searches against antibiotic resistance sequence databases. Multiple antibiotic resistance genes were identified in the chicken (E429), calf (E172) and pig (E142) strain genomes (Table 5.3).

Table 5.3: Occurrence of antibiotic resistance genes in *E. faecium* isolates. Indicated genes encode resistance to antibiotics as follows: *ermA* and *ermB* (erythromycin), *lunB* (lincomycin), *aacA-aphD* (gentamycin), *aad6* (spectinomycin) and *aadE* (streptomycin); *cat* (chloramphenicol), *tetM* and *tetL*(tetracycline), *van A* (vancomycin type A), *van B* (vancomycin type B), *fos* (fosfomycin), *parC* and *g1rA* (fluoroquinolone and ciprofloxacin), *Pbp5-R* (ampicillin), *st* (streptothricin); *azlC* (azaleucine) ,*ble* (bleomycin), *fmtC* (oxacillin) and *vgb* (streptogramin). Red strains indicate clinical isolates, green indicates animal isolates and orange indicates commensal isolates. Unknown indicates information is not presented in the two analysis previously reported by Qin *et al* (2012) and Lebreton *et al* (2013).

Strain	*ermA*	*ermB*	lnuB	aac(6')-aph(2")	*aad6*	*aadE*	*cat*	*tetL*	*tetM*	vanA operon	*fos*
1_230_933	0	1	0	1	0	0	0	0	2	1	0
1_231_408	0	1	0	1	0	0	0	0	0	0	0
1_231_410	0	1	0	0	0	0	0	0	0	1	0
1_231_501	0	0	0	0	0	0	0	0	0	0	0
1_231_502	0	1	0	2	0	0	0	0	0	1	0
AUS0004	0	0	0	0	0	0	0	0	3	0	0
C68	0	1	0	1	0	0	0	0	0	0	0
D344SRF	0	1	0	0	0	0	0	1	2	0	0
E0120	0	2	2	0	0	0	0	1	0	1	0
E0333	0	1	0	0	0	0	0	0	0	2	0
E1039	0	0	0	0	0	0	0	0	0	0	0
E1071	0	1	2	0	0	1	0	1	0	1	0
E1133	0	1	0	0	0	0	0	1	0	0	0
E1162	0	0	0	0	0	0	0	1	0	0	0
E1185	0	1	0	0	0	0	1	0	2	0	0
E1258	0	0	0	0	0	0	0	0	0	0	0
E1321	0	1	0	0	0	0	0	1	0	0	0
E1392	0	1	0	1	0	0	0	0	2	0	0
E1552	0	1	0	0	0	0	0	0	2	1	0
E1623	0	0	0	0	0	0	0	0	0	0	0
E1626	0	1	0	0	0	0	0	0	2	0	0
E1627	0	1	0	0	0	0	0	1	0	0	0
E1634	0	0	0	0	0	0	0	0	0	0	0
E1636	0	0	0	0	0	0	0	0	2	0	0
E1644	0	1	0	1	0	0	0	0	0	1	0
E1679	1	1	0	1	1	0	0	0	0	1	1
E1731	0	1	0	1	0	0	0	1	0	0	0
E1861	0	0	0	0	0	0	0	0	0	0	0
E1904	1	2	0	1	1	0	0	0	0	0	0
E2039	0	0	0	0	0	0	0	0	0	0	0
E2297	0	1	0	1	0	0	0	0	0	1	0
E2369	0	1	0	1	0	0	0	0	0	1	0
E2620	0	0	0	0	0	0	0	0	0	0	0
E2883	0	0	0	0	0	0	0	1	3	0	0
E3083	0	0	0	0	0	0	0	0	0	0	0
E3346	0	0	0	0	0	0	0	0	0	0	0
E3548	0	0	0	0	0	0	0	0	0	0	0
E3548	0	2	2	3	0	0	0	0	0	0	0
E6012	0	1	2	1	0	0	0	1	0	2	0
E6045	0	1	2	1	0	0	0	1	0	2	0
LCTEF90	0	0	0	0	0	0	0	0	0	0	0
TC6	0	2	0	0	1	0	0	1	2	0	0
U0317	0	1	2	3	0	0	0	0	0	0	0
E0164	0	0	0	0	0	1	0	0	4	1	0
E4215	0	0	0	0	0	1	0	0	0	2	0
E0045	0	0	0	0	0	0	0	0	1	0	0
E0269	0	1	0	0	0	0	0	0	2	1	0
E2134	0	1	0	0	0	0	0	1	1	0	0
E1575	0	1	0	0	0	0	0	0	1	2	0
E0680	0	1	0	0	0	1	0	1	0	1	0
E0679	0	1	2	0	0	0	0	1	0	0	0
E2071	0	0	0	0	0	0	0	0	0	0	0
E0688	0	1	2	0	0	0	0	1	4	0	0
E1574	0	1	2	0	0	0	0	1	0	1	0
E1622	0	0	0	0	0	0	0	0	0	0	0
E1573	0	0	0	0	0	0	0	0	0	0	0
E1578	0	0	0	0	0	0	0	0	0	0	0
E4389	0	0	0	0	0	0	0	0	2	0	0
E1576	0	0	0	0	0	0	0	0	0	0	0
E4453	0	1	0	0	0	0	0	0	2	0	0
E4452	0	1	0	0	0	0	0	1	3	0	0
E429	1	3	0	1	8	0	0	1	2	1	1
E172	1	3	2	1	12	0	1	0	1	1	0
E142	1	2	0	1	21	2	0	0	2	1	0
E1050	0	0	0	0	0	0	0	0	0	0	0
E1972	0	0	0	0	0	0	0	0	0	0	0

E1007	0	0	0	0	0	0	0	0	0	0	0
Com15	0	0	0	0	0	0	0	0	0	0	0
E980	0	0	0	0	0	0	0	0	0	0	0
Com12	0	0	0	0	0	0	0	0	0	0	0
1_141_733	0	0	0	0	0	0	0	0	0	0	0

Strain	*parC*	*g1rA*	*Pbp5-R*	*st*	*azlC*	*ble*	*fmtC*	*vgb*
1_230_933	1	1	1	0	0	1	0	0
1_231_408	1	1	1	0	0	1	0	0
1_231_410	1	0	1	0	0	1	0	0
1_231_501	0	0	0	0	0	1	0	0
1_231_502	1	1	1	0	0	1	0	0
C68	1	1	1	0	0	1	0	0
D344SRF	0	0	0	0	0	1	0	0
E1039	0	0	1	0	0	1	0	0
E1071	0	0	1	0	0	0	0	0
E1162	0	0	1	0	0	0	0	0
E1636	0	0	1	0	0	0	0	0
E1679	0	0	1	1	0	0	0	0
U0317	1	1	1	0	0	0	0	0
E4453	2	6	3	0	2	unknown	unknown	unknown
E4452	3	6	2	1	1	unknown	unknown	unknown
E429	4	4	5	0	2	2	1	2
E172	3	2	2	0	2	2	1	2
E142	2	2	2	0	1	2	1	2
Com15	0	0	0	0	0	1	0	0
E980	0	0	0	0	0	1	0	0
Com12	0	0	0	0	0	1	0	0

Each of the sequenced animal *E. faecium* strains in this study is vancomycin resistant. To explore the nature of this resistance the *van* operons were identified by homology. In strain E429 (chicken) the *van* operon is about 7.6 kb in size (2874238bp - 2881898pb), with *van*Z located 398 kb distant to the operon (Figure 5.5). The operon is surrounded by mobile elements including transpoase TnA (Tn*1546*), a transcriptional regulator, Tn*916*, DNA topoisomerase and a tetracycline resistance gene is located 1.5kb upstream. Unexpectedly, a second copy of *van*R, *van*S and *van*Y are clustered together in an operon, 2.5 kb in size located about 2 Mb distant (647926- 650500).

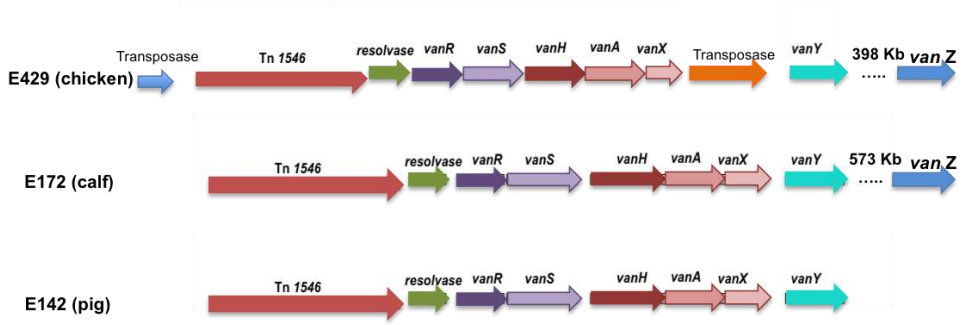


Figure 5.5: Vancomycin resistance genes in animal *E. faecium*. The arrows show a similar Tn*1546* linked operon that is composed of 6 *van* genes (*van*R, S, H, A, X, and Y).

In the calf strain (E172), the Van operon is smaller 5.987 kb (2514921bp - 2520908bp), with *vanZ* is located 573 kb distant. The operon in the pig strain is a similar size as the calf strain operon (located 24141042bp-2408056bp), however, *vanZ* is absent.

Mutations in the *gyrA* or *parC* subunit genes that are responsible for fluoroquinolone and ciprifloxacin resistance were found in the three animal *E. faecium* strains. The described amino acid change E to K occurs in codon 88 of the *gyrA* gene and amino acid change E to K in occurs in codon 86 of *parC*. Fluoroquinolone, streptothricin and azaleucine resistance were found only in the animal strains, which might reflect that these antibiotics are used in animal husbandry. Gentamicin resistance was also found in the three animal isolates (Table 5.3)

5.2.2.3 Genomics Island

The IslandViewer server (Section 2.18.1) was used to identify Genomic Islands (GI) (Section 1.12.4) in the chromosome sequence of calf strain E172 derived from PacBio sequencing. The E172 genome harbours multiple genomic islands with 21 regions totalling 369 kb (Table 5.4). The GIs in the calf strain bunched at the end of the genome. This differs markedly from the GIs that are found in the clinical *E. faecium* strains DO and Aus0004 which are spread across their genome. Within the E172 genome the GI region corresponded with novel mega plasmid that also was also found in the chicken strain (E429) in this study (Region C11- Figure 4.7). Several of the GI regions in the calf strain are unique (Figure 4.7 A).

The pathogenicity island carrying the *esp* gene is absent from the DO genome and from all the three animal *E. faecium* (Qin, Galloway-Pena *et al.* 2012). There are 13 other possible genomic island regions totalling 107 kb present in the (DO strain) genome. Which mainly encode carbohydrate utilisation genes and IS elements. For example, operons for the utilisation of mannose/fructose/sorbose, glucose and fructose/mannitol were found in five of the identified GI regions.

Table 5.4: GI associated with animal *E. faecium* isolated from calf using PacBio sequencing platform. GI regions, position, size of GI and the key genes presented in each region.

GI	Position	Size	Key genes in GI region
1	7647-23306	15659	Mobile element protein, sugar utilisation operon (beta-1,3-glucosyltransferase, UDP-glucose dehydrogenase, dTDP-glucose 4,6-dehydratase, glycosyltransferase) capsular polysaccharide biosynthesis proteins and Lipid carrier
2	369360-375322	5962	Cell wall surface anchor family proteins, Cell wall surface anchor family proteins, LPXTG motif and mobile element protein
3	768523-787902	19379	Ribosomal proteins and ribose 5-phosphate isomerase A
4	2492183-2505846	13663	Transposase, IS204/IS1001/IS1096/IS1165, sugar utilisation operon (UDP-glucose dehydrogenase, PTS system, galactose-inducible IIB component / PTS system, galactose-inducible IIC component, PTS system IIA component, dTDP-glucose 4,6-dehydratase) and mobile element proteins
5	2516838-2533509	16671	Integrases, mobile element proteins, PTS system, cellobiose-specific IIC component, Putative hydrolase, haloacid dehalogenase family and Alcohol dehydrogenase
6	2552058-2561564	9506	Mobile element proteins and hypothetical proteins, endonuclease and type I restriction-modification system, restriction subunit R
7	2572986-2583610	10624	Heavy metals operon (Lead, cadmium, zinc and mercury transporting ATPase, Copper-translocating P-type ATPase, Copper chaperone) glucose 1-dehydrogenase, replication-associated protein RepB, mobile element proteins and hypothetical proteins
8	2583977-2591668	7691	Multicopper oxidase, Lead, cadmium, zinc and mercury transporting ATPase, Copper-translocating P-type ATPase, Transposase, IS4 and Phosphate regulon transcriptional regulatory protein PhoB (SphR)
9	2612029-2625907	13878	Sucrose operon (Sucrose permease, major facilitator superfamily, Sucrose-6-phosphate hydrolase, Fructokinase, Sucrose operon repressor ScrR, LacI family) phage-related proteins, ntegrase/recombinase core domain family and choloylglycine hydrolase.
10	2641977-2650278	8301	Cellobiose operon (PTS system, cellobiose-pecific IIA, IIB, IIC components) Beta-glucosidase, 6-phospho-beta-glucosidase, sugar kinase
11	2660624-2669004	8380	Transposase, IS204/IS1001/IS1096/IS1165, histidinol-phosphatase, Two-component response regulator VncR, Ferric iron ABC transporter, iron-binding protein, Methionine ABC transporter ATP-binding protein
12	2683060-2769945	86885	Phage integrase, integrase, integrase/recombinase, core domain family, mobile element proteins, chromosome (plasmid) partitioning protein ParA / Sporulation initiation inhibitor protein Soj, Putative peptidoglycan bound protein (LPXTG motif), Antiadhesin Pls, binding to squamous nasal epithelial cells, Clumping factor ClfB, fibrinogen binding protein, resolvase, Beta-lactamase repressor BlaI, Beta-lactamase regulatory sensor-transducer BlaR1, Phage protein, Cytosolic protein containing multiple CBS domains, replication initiator protein A
13	2781409-2797952	16543	Zeta toxin genes, chloramphenicol acetyltransferase, glutamate synthase (NADPH), Type II restriction enzyme, methylase subunit YeeA.
14	2805558-2825811	20253	Chromosome partitioning ATPase, Site-specific recombinase, DNA invertase Pin related protein, trehalose operon transcriptional repressor, trehalose-6-phosphate hydrolase, transporter.
15	2856053-2866089	10036	Hypothetical proteins, Methionine synthase II (cobalamin-independent), Transcriptional regulator PadR family

16	2874325-2886841	12516	Fumarate reductase flavoprotein subunit, mannose-6-phosphate isomerase, catalase, gamma-aminobutyrate:alpha-ketoglutarate aminotransferase
17	2910173-2926108	15935	Activator of the mannose operon (transcriptional antiterminator), BglG family, D-alanyl-D-alanine carboxypeptidase, alcohol dehydrogenase, phage infection proteins, N-acetylmuramoyl-L-alanine amidase, family 4
18	2926129-2934524	8395	Glyoxylate reductase, L-lysine permease, Cobalt-zinc-cadmium resistance protein, two-component sensor histidine kinase, regulation of D-alanyl-lipoteichoic acid biosynthesis, DltR, Glycine betaine ABC transport system, ATP-binding protein OpuAA
19	2934660-2987933	53273	Vancomycin A resistance operon(TnpA transposase, resolvase, Vancomycin response regulator VanR, Sensor histidine kinase VanS, D-lactate dehydrogenase VanH , D-alanine--D-lactate ligase VanA), Glycine betaine ABC transport system, ATP-binding protein OpuAA, OpuAB, OpuAC, Chromosomal replication initiator protein DnaA, Two-component response regulator SA14-24
20	2989525-3005194	15669	Vancomycin B resistance operon, chromosome partitioning protein ParA, similar to plasmid replication protein, replication control protein PrgN

In the genome of Aus0004, 15 genomic island regions totalling about 262 kb were found. One genomic island of 60 kb was uniquely present in the Aus0004 strain when compared with 22 other strains by Lam *et al.* (2012). In this study the GIs in Aus0004 strain were identified to mainly encode tetracycline and *van* type B resistance, the *esp* gene plus mannose and sucrose utilisation operons.

5.3 Discussion

5.3.1 Insertion sequence elements

Insertion sequence elements (IS) and transposases are the foremost mobile genetic elements in *E. faecium* (Qin, Galloway-Pena *et al*. 2012).

Comparative genomics of entire *E. faecium* chromosomes has provided insights into the mobile genetic elements that are present in the *E. faecium* DNA pool. The IS correspond to the major MGEs in clinical enterococcal isolates, and commonly discovered IS-families represent IS3, IS6, IS30,

IS256, ISL3, IS4, IS66, IS110, IS200/IS605, IS982, IS1182 and IS1380. IS16 is widespread in clinical *E. faecium*, and has also been identified in clinical *E. faecalis* strains and as a fragment of pRUM-like plasmids (Hegstad, Mikalsen *et al.* 2010). IS elements are also the most noticeable group of genes enriched in all CC17 strains and the majority of hospital-associated strains (van Schaik, Top *et al.* 2010, Qin, Galloway-Pena *et al.* 2012).

Transposable elements may provide genome plasticity by facilitating recombination between homologous transposable elements generating rearrangements in chromosomal and plasmid DNA (Heaton, Discotto *et al.* 1996). Frost *et al.* (2005) suggested that chromosomal deletions and rearrangements can also result from activates co-localised with mobile genetics elements, such as transposases and site-specific recombinases as well as homologous recombination systems of the host. IS elements, for example, ISEfm1, IS1251, IS66, and ISEfa10 were proposed as a reason for the inversion in the genome of the complete genome of *E. faecium* (Aus0004) and also with three animal isolates described in chapter 3 (Figure 3.4).

CC17 genotype isolates appear to have unique transposase-related genes, which might contribute to the virulence of these strains (Figure 5.1). Leavis, *et al.* (2007) described that IS are proposed to contribute to the success of this genetic sub-population in its competition with other enterococci in hospital settings, generating a novel globally spread nosocomial subspecies.

Future work could explore whether IS element contribute animal host colonisation.

The ISL3 and ISEf1 families were the most commonly observed IS types in the human strain (Aus0004), and ISEf1 is also common in *E. faecalis* (Lam, Seemann *et al.* 2012). ISEf1 is absent in the three animal *E. faecium* genomes isolated from chicken (E429), calf (E172) and pig (E142). Contrastingly, IS3, IS6 and IS256 families were the most common of these elements observed in the animal strains, although the IS3 and IS256 elements were also prominent among the hospital clade (Leavis, Willems *et al.* 2007). Enrichment of specific IS elements in the genome of bacterial sub-species has been recognised previously. Yao *et al.* 2005 demonstrated that in the clinical strains of *S. epidermidis*, IS256 is existent in multiple copies, where it might improve genome flexibility of biofilm-forming and multiresistant strains.

The variable presence of insertion elements distinguishes hospital-associated from human commensal and animal *E. faecium* strains and could be used diagnostically. IS16 was exclusively spread only among the clonal complex of hospital-associated CC17 strains (Werner, Fleige *et al.* 2011). In addition, IS66 is mostly found in CC17 genotype strains from human and animal sources. Qin *et al* (2012) suggested that IS elements IS16, ISEnfa3, IS3, IS911, IS116/IS110/IS902 and IS66 have the potential to be used as a molecular screen to identify clinical *E. faecium*, although IS3 and IS110 were found in the animal strains of *E. faecium* isolated from chicken (E429),

calf (E172) and pig (E142) (Table 5.1). The presence of these IS elements in the animal strains might reflect an association of the animal *E. faecium* with the clinical isolates.

Most of the IS elements and transposons present in the three animal *E. faecium* are co-located with genomic islands (Tables 5.4). In addition, most of the IS elements in the animal strain are unique and found in novel regions (Figure 4.7). The association of these elements with GI and novel region in the genome map might reflect horizontal transfer of these genes from different species, since several of these IS elements revealed substantial homology with both Gram-negative genera, including *Escherichia, Burkholderia, Pseudomonas* and *Xanthomonas* species, and Gram-positive genera, including *Staphylococcus, Streptococcus, Bacillus* and *Lactobacillus* species.

5.3.1.2 Plasmid

Enterococcus species harbour plasmids which often mediate resistance to antimicrobials and heavy metals, provide enhance virulence and/or encode DNA repair mechanisms (Arias, Panesso *et al.* 2009, Garcia-Migura, Hasman *et al.* 2009). The mega-plasmids identified in chicken (E429) and pig (E142) harbour genes encode potential adhesi with the presence of sortase A and an *LPXTG* cell wall anchor protein. It is known that *LPXTG* surface proteins may play a significant role in the pathogenesis of *E. faecium* in hospital-related infections (Hendrickx, van Wamel *et al.* 2007, Lam, Seemann *et al.* 2012).

The mega-plasmid was found in the genome of chicken, calf and pig, which is unique to these strains (Figure 4.7), and it encodes heavy metal resistance genes for resistance to lead, cadmium zinc and mercury.

The mega-plasmid (56kb) is apparently integrated into the chromosome of the chicken *E. faecium* strain (E429). Due to the homology between plasmids and the genome an occurrence of a single homologous recombination event can integrate a complete plasmid into the chromosome (Heap, Ehsaan *et al.* 2012). Homologous recombination following transformation will potentially occur if plasmids are incapable of replication in a specific host. These insertion incidents have been widely detected in *E. faecalis*, *E. coli*, *B. subtilis*, *S. pneumoniae*, *L. plantarum* and *L. lactis subsp. lactis* (Casey, Daly *et al.* 1991).

5.3.2 Distribution of genes encoding MSCRAMM-like proteins, putative virulence genes and antibiotic resistance determinants

A previous study by Qin *et al* (2012) reported that 15 genes encoding LPXTG family cell wall-anchored proteins with MSCRAMM-like features were present in the complete genome of E. *faecium* (TX16). The LPXTG family cell wall-anchored proteins present in the three animal strains are novel or share homology with other Gram-positive species such as *Staphylococcus* and *Lactobacillus* species.

Qin *et al* (2012) identified that in 21 *E. faecium* draft genomes, all of the MSCRAMM-encoding genes were broadly dispersed, excluding (*esbA*), which was only present in HA-clade isolates. Multiple copies of *esbA*-like

genes were also found with low sequence identity (25-37%) in the three animal *E. faecium* genomes in this study, possibly indicating they are novel MSCRAMMs. Enterococcal surface protein (Esp) and collagen-binding adhesin (Acm) contribute to colonisation and infection, however recent studies have determined that Esp is not fundamental for infection in murine infection models (Heikens, Leendertse *et al.* 2009). An *esp*-like gene was found in the three animal *E. faecium* genomes but the low percentage identity (24%), possibly indicating it is distinct. Collagen adhesin genes with percentage identity ranging from 61% to 100% were found in the three animal strains. This gene is present as a pseudogene in all of the *E. faecium* commensal isolates except 1,141,733 in Qin *et al* (2012) study and *acm* pseudogenes were also found in clinical *E. faecium* that do not belong to CC17 genotype.

The presence and absence of 19 antibiotic resistance genes across 72 *E. faecium* isolates including clinical, animal and commensal was also searched. These data correspond to previously published frequency data for a smaller set of isolates (Qin, Galloway-Pena *et al.* 2012, Lebreton, van Schaik *et al.* 2013). Comparative analysis of antibiotic resistance revealed that commensal, animal and clinical isolates have clear differences in terms of their resistance profile. All of the clinical and animal isolates have multiple resistance determinants, excluding strains 1,231,501 and E1039. The clinical strain (1,231,501) lacks all antibiotic resistances including pbp5-R, may have lost genes through recombination and acquired pbp5-S. Certainly, 1,231,501 was shown to be a hybrid of clinical and commensal

genomes by Palmer, *et al* (2012) and the (hybrid) region including pbp5-S, which could clarify the origin of pbp5-S in this strain.

In contrast, Qin *et al* (2012) stated that all of the commensal-associated isolates (1,141,733, Com12, Com15, E980 and TX1330) lacked genes for antibiotic resistance to chloramphenicol, erythromycin, streptomycin, spectinomycin, gentamycin, vancomycin, ciprofloxacin and ampicillin. Strain E1039, which is a commensal isolate, but genetically closer to the clinical strains, has an ampicillin resistance gene. In 2013, same analysis applied by Lebreton *et al* to two other commensal isolates (E1050 and E1007) showed their resistance to streptomycin and spectinomycin, while E1050 also encoded resistance to fosfomycin.

Disease treatment and growth promotion could explain the multiple antimicrobial resistance of most *E. faecium* isolates, including animals strains. The delivery of low levels of antimicrobials has apparently resulted in considerable colonisation of animals with antibiotic resistant bacteria, such as *E. coli* strains and acquisition of resistance in *E. coli* in the intestinal flora of the farmers has been described (Marshall and Levy 2011, Lebreton, van Schaik *et al.* 2013). Aarestrup (2000) reported that resistance to streptothricin antibiotics has been described in Gram-negative bacteria as a result of using nourseothricin as an antimicrobial feed promoter in industrial animal farms in Germany. In addition, resistance to streptogramins may be related to the use of virginamycin, as a feed promoter combined in agriculture for animal food production. Virginamycin use was prevented in

Denmark in 1998 followed by the rest of the EU in 1999. Virginamycin resistance was identified in this study in all three animal *E. faecium* and these strains were isolated from the same geographic region (The Netherlands) and resistance might also have arisen from the historic use of this antibiotic as a feed promoter in Dutch agriculture.

5.3.3 Genomic Islands

The GIs that were found in animal *E. faecium* confirmed the hypothesis of Juhas, *et al* (2009) that the GIs comprise a family of mobile elements including conjugative transposons and prophages. GIs with functions that improve the fitness of the bacteria such as carbohydrate utilisation genes were found in *E. faecium* and could have been directly or indirectly positively selected (Hacker and Carniel 2001).

Genomic island analysis by codon usage bias and composition variation showed that E172 has 21 GIs, although animal *E. faecium* also possesses a large number of mobile elements in the region of the GIs, suggesting that most of the genomic variable loci in the three animal *E. faecium* isolated from chicken, calf and pig were acquired via lateral gene transfer, possibly through mobile elements such as transposons (Table5.4). In addition, the presence mobile elements in the GIs also give clues as to how these segments entered. As previously stated, the Pathogenicity Island of *E. faecalis* is littered with sequences that are related to mobile genetic elements (McBride, Coburn *et al.* 2009).

Pyrosequencing constructed genome analyses of numbers of medical importance microorganisms have been shown presence of genomic islands for the improvement of pathogenicity and the diversity inside single bacterial species (van Schaik, Top *et al.* 2010). Numbers of virulence and antibiotic resistance genes were identify in the genomic islands of *E. faecium*. For example, *esp* gene, is carried on a large pathogenicity island between 13.8, 64, 68 and 104 kb in size. The pathogenicity island was found in four strains of *E. faecium* Aus0004, E1162, E1679 and U0317. However, this GI is absent in the three animal *E. faecium* isolated from chicken (E429), calf (E172) and pig (E142).

GIs that present in the three animal *E. faecium* the complete genome DO and Aus0004 are encode a complete pathway for several carbohydrate utilisation, for example, cellobiose, galactose, fructose, sorbose, sucrose suggested that animal *E. faecium* are capable of using these carbohydrates as a carbon source. Carbohydrate utilisation pathways in GIs are traits supposed to contribute to pathogenicity or altering *E. faecalis* relationship with the host (McBride, Coburn *et al.* 2009). The presence of these carbohydrate utilisation genes in the pan genome of *E. faecium* and in GIs suggested these genes are acquired through lateral gene transfer.

Two GIs were identified to encode the vancomycin resistance type A and B and several plasmid replication proteins in animal strains suggesting that these GIs are forming the mega plasmid in the animal *E. faecium.* In

addition, these explain the localisation of these GIs in the end of the animal genomes assembly. This study defined groups of genes that have been combined into the GIs in animal *E. faecium* and provides indication for the acquisition of these parts of the genome as mobile functional elements. However, this study did not investigate the transfer of the GI genes between *E. faecium* isolates, the description of these regions may allow more targeted analysis of transfer focusing on movement of specific regions of the GI. Investigation of these regions as functional part might offer clues to their influences to fitness.

Chapter Six: Comparative genomics of *E. faecium* bacteriophages.

6.1 Introduction

Bacteriophages that infected *Enterococcus* species were first identified around 70 years ago (Clark and Clark 1927, Evans 1934). Images of enterococcal phages were captured by Rogers and Sarles using electron microscopy and they stated that the enterococcal phages seemed to have icosahedral heads and long non-contractile tails (Rogers and Sarles 1963). Recently, phages that infect and lysogenise *E. faecalis* and *E. faecium* have been more extensively characterised (Duerkop, Palmer *et al*. 2014).

So far, the induced prophages of *Enterococcus* were all *Siphoviridae* and temperate phages isolated from *E. faecium* are morphologically identical to prophages from *E. faecalis*. These phages have an isometric head about 40 nm in size and a long non-contractile tail, ranging from 70 nm to 220 nm (van Schaik, Top *et al*. 2010). However, diverse phages are capable of infecting *Enterococcus* and comprise phages related to the *Siphoviridae* as well as non-tailed phages with icosahedral shaped capsids (Brede, Snipen *et al*. 2011). The first non-tailed enterococcal phages were isolated by Mazaheri Nezhad Fard *et al* (2010) and included polyhedral, filamentous, and pleomorphic (PFP) phages that are likely to be virulent (lytic).

Within the Firmicute phylum of Gram-positive bacteria, temperate phages are important vectors for the horizontal transfer of virulence genes (Yasmin, Kenny *et al*. 2010). Phages play an important role in adding to the genome plasticity of *E. faecium* species (Lam, Seemann *et al*. 2012). The ability of enterococcal phages to mediate transduction can transfer antibiotic genes

between different *Enterococcus* species, including *E. faecalis*, *E. faecium*, *E. gallinarum*, *E. hirae*, and *E. casseliflavus* (Fard, Barton *et al*. 2010).

The complete genomes of *E. faecium* TX16 (DO) and Aus0004 encoded two and three phage-like sequences, respectively. The phages found in DO strains have similarity with ORFs in hospital-associated strains but low similarity with ORFs of community-associated strains. The phages found in Aus0004 are present in all CC17 genotype genomes but they are variably present in other *E. faecium* isolates. These phages of DO and Aus0004 share high similarity with phage genes found in species of other genera, including *Clostridium*, *Listeria*, *Lactobacillus* and *Staphylococcus* (Lam, Seemann *et al*. 2012, Qin, Galloway-Pena *et al*. 2012).

The presence of *E. faecium* phages in most clinical isolates potentially indicates an association of the phages with either virulence or the transfer of antibiotic resistance. Multiple, sequenced *E. faecium* genomes are available in public databases, however a rigorous bioinformatic analysis of the many prophage sequences using the multitude of available genomes remains to be performed. Moreover, the presence/absence of prophages across different *E. faecium* genomes has not been determined.

Specific aims

In this chapter prophage-related sequences will first be identified in the genomes of animal *E. faecium* isolated from chicken, calf and pig and characterised. Comparative genomics of *E. faecium* prophage from the

publicly available genomes will be then performed to understand the relationships between different phages. In addition, the potential carriage of cargo genes that might be associated with virulence or fitness of this species will be determined.

6.2 Results

6.2.1 Bacteriophage induction and distribution

Plaque assays were performed to investigate the presence of inducible phages in the sequenced genomes of the three animal *E. faecium* isolates. All three strains are expected to be lysogens since the individual *E. faecium* genome contain three functional prophage genome sequences for E429 (chicken) and one for both E142 (calf) and E172 (pig). Induction of prophage into the lytic cycle in the three strains resulted in released phages for strains E429 (chicken) and E172 (calf) as determined by mitomycin C (4 μg ml^{-1}) treatment of the strains to produced cell lysates that were used to infect animal and human isolates as indicator strains (Table 6.1) in spot plaque tests. Lysis was confirmed as being phage-derived by plating for individual plaques on each indicator strain (data not shown). The absence of lysis with several indicator strains following infection with E429 and E142 lysate might result from the absence of a cognate receptor or homoimmunity.

Table 6.1: Phage lysis of *E. faecium* indicator strains. Phage lysis of a panel of isolates using filter-sterilised lysates produced after addition of mitomycin C to strains E429, E172 and E142. (-) indicates absence of plaques (+) indicates presence of plaques and not tested (X).

Indicator strains	Lysates strains			Indicator strain source
	E429	E172	E142	
E429 (LIV1072)	**X**	+	-	Chicken
E172 (LIV1071)	-	**X**	-	Calf
E142 (LIV1070)	+	+	**X**	Pig
LIV66	+	+	-	TX16, Endocarditis isolate
LIV153	-	+	-	VanA resistant strain
LIV294	+	+	-	Chicken faeces
LIV296	+	+	-	Jaguar faeces- Chester zoo
LIV297	+	+	-	Mouth swab
LIV298	+	+	-	Mouth swab
LIV299	+	+	-	Irish rodent faeces
LIV302	+	+	-	Dog faeces
LIV303	-	+	-	Mouth swab

6.2.2 Phage and bacteriocin differentiation

Both phage and bacteriocin production by the calf strain (E172) were evident after addition of mitomycin C (4 µg ml^{-1}). However, spot tests also showed clear zones when supernatant of E172 was tested prior to addition of mitomycin C. To investigate whether this cell lysis was free phage-derived or due to bacteriocin, supernatant samples were tested from across a growth curve. Phage was differentiated from bacteriocin using the following procedures: (i) spot test, (ii) size exclusion centrifugation and (iii) individual plaque assay. The spot test showed a clear zone of cell lysis using filter-sterilised supernatant of E172 strain. The clarity of the cell lysis was maximal after 4 hours, (Figure 6.1) equivalent to mid-exponential growth phase. Size exclusion filtration of pre-induction supernatant using a

Centricon plus-20 column followed by plaque assay of filtrate resulted in cell lysis. In contrast, plaque assay of filtrate showed no individual plaques, confirming bacteriocin production as the reason for cell lysis in the absence of mitomycin C.

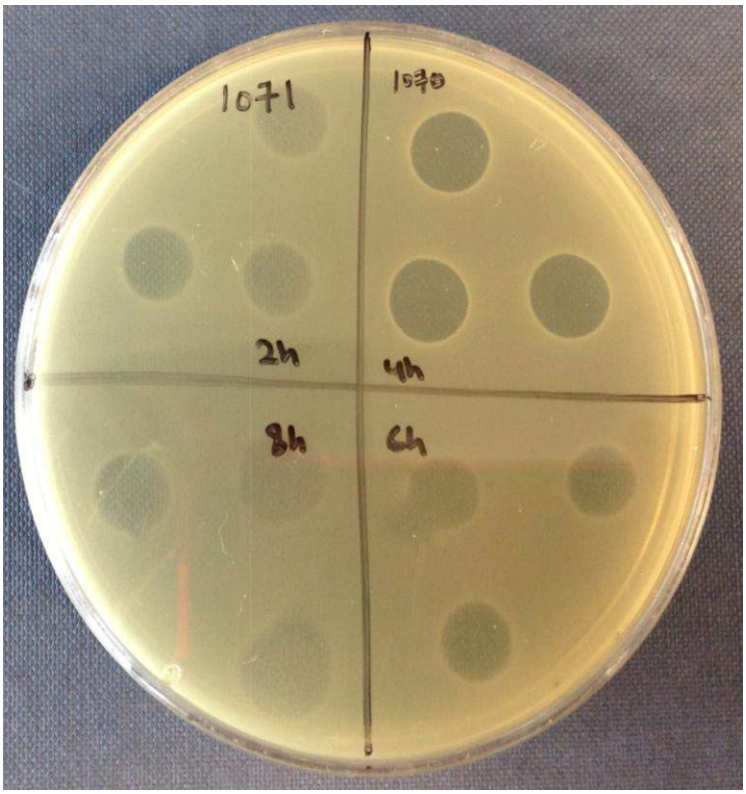


Figure 6.1: Production of bacteriocin by *E. faecium* E172 (calf). Supernatant from E172 (calf) was tested for lysis of the indicator strain E142 (pig). Bacteriocin production peaks after 4 hours growth at 37C.

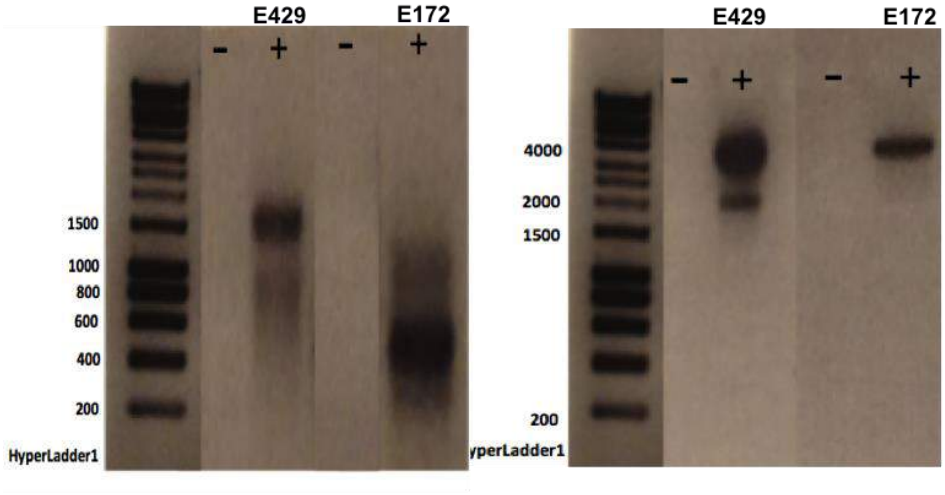
Genome analysis of animal *E. faecium* strains identified that the strain E429 (chicken) genome has four genes encoding enterocin A immunity (ORTHOMCL2886, ORTHOMCL1870, ORTHOMCL4691 and ORTHOMCL4113), two genes encoding lactococcin G processing and

transport ATP binding protein LagD (ORTHOMCL4805 and ORTHOMCL5192) and a class II sec-dependent bacteriocin gene (ORTHOMCL2657). The E142 (pig) genome encodes genes, encoding enterocin A immunity (ORTHOMCL2223), a lactococcin G processing and transport ATP binding protein LagD (ORTHOMCL2613) and a lactococcin A secretion protein LcnD (ORTHOMCL2614) and one gene encoding a Class II sec-dependent bacteriocin (ORTHOMCL2657), plus the bacteriocin piscicolin (ORTHOMCL2212). All these genes were absent from the calf genome, suggesting that a novel bacteriocin might be encoded by this strain given the demonstrated activity and the failure to identify bacteriocin homologues or that the matching region was not present in the sequence output.

6.2.3 Transduction using identified phages

Mitomycin C induction lysates produced from the chicken and calf strains were tested for their ability to package and transduce chromosomal and extra-chromosomal DNA. The *E. faecium* donor strain E142 (pig) contains both chromosome and plasmid-located antibiotic genes and it was infected with the cell-free phage lysates to screen for transductions. Two antibiotic resistance genes were tested for transduction into recipient indicator *E. faecium* cells of LIV299 and LIV303: tetracycline resistance encoded by pM7M2 (NC_016009); chromosomal ampicillin resistance gene E142-SEPT09050 located at position 877649:880039.

Transduction successfully produced antibiotic resistant colonies for the two markers attempted. To identify whether successful transduction of the identified antibiotic genes had occurred, PCR amplification of each corresponding resistance gene was performed (Figure 6.2). PCR amplification identified that the 4 kb tetracycline resistance amplicon was present in both donor and recipient. In contrast, the ampicillin resistance that was selected could only be confirmed to be due to the donor as the 1.5 kb locus following transduction using the strain E429 phage lysate.



A. B.

Figure 6.2: PCR amplification of antibiotic resistance genes after transduction using animal *E. faecium* phage. (A) Ampicillin (1.5 kb) resistance locus amplified from strain LIV299 transductants isolated from the chicken (E429) and calf (E172) strains. (B) Tetracycline (4kb) resistance locus amplified from strain LIV303 transductants isolated from the chicken (E429) and calf (E172) strain lysis of strain E142 bearing ampicillin and tetracycline resistance. (-) indicates strains prior to transduction with the

absence of antibiotic resistance, (+) indicates PCR amplicon after transduction of the antibiotic resistance.

6.2.4 Animal *E. faecium* bacteriophages

6.2.4.1 General genome features of animal *E. faecium* phages

The sequenced genomes of the three vancomycin-resistant *E. faecium* strains isolated from chicken (E429), calf (E172) and pig (E142) harbour multiple phage-related sequences. Six phage regions were found in the chicken strain (E429) and one each in calf (E172) and pig (E142) (Table 6.2).

Table 6.2: Phage-related sequences of sequenced animal *E. faecium.*

Strain	Phage position	Size (Kb)	No. Of ORFs	GC%
E429_ph1	412480-460595	48.1	70	36.7
E429_ph2	1347483-1395061	47.5	60	36.9
E429_ph3	1589043-1629766	40.7	55	37.6
E429_ph4	1992956-2009130	16.1	46	38
E429_cp1	3023847-3041052	17.2	21	44.7
E429_cp2	3009148-3080914	71.7	105	44
E172_ph1	486654-506555	19.9	27	37.5
E142_ph1	433557-468604	35	41	37.3

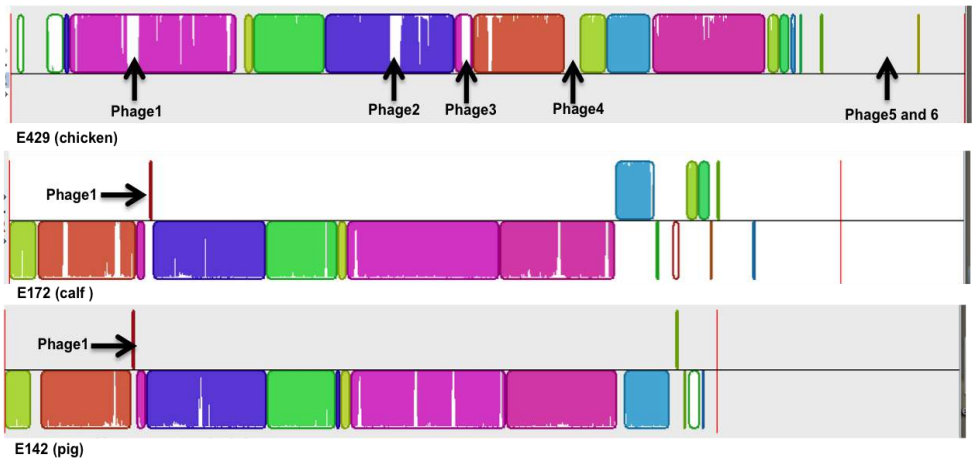


Figure 6.3: Genome alignment of animal *E. faecium*. The E429 (chicken) DNA sequence was used as a reference DNA sequence to which E172 (calf) and E142 (pig) were aligned and compared. White space within the locally collinear blocks in the chicken strain corresponds with phage regions and the coloured areas represent the similarity in the DNA sequences. Phage 1 in calf and pig share tail proteins with phage 3 in chicken genome.

The identified animal *E. faecium* phage sequences are very diverse and range in size from 17 to 48 kb double-stranded DNA (dsDNA), with an average GC% content of 36 - 44 mol% and between 21 and 105 coding sequences (Table 6.2).

6.2.4.2 Organisation of animal prophage genomes

The number of predicted ORFs identified per phage genome correlates with the phage size. Phages from the chicken *E. faecium* have the largest genome size and they encode 105, 70, 60, 55 and 46 putative genes, whereas 41 ORFs were predicted for the phage genomes of E142_ph1.

The chicken and calf *E. faecium* strains also have small regions of phage-related sequence of 17.2 and 19.9 kb, which might represent cryptic phage. Protein sequences deduced from putative ORFs were screened for homology with proteins from sequence databases using BLASTP in the PHAge **S**earch **T**ool (PHAST) (Section 2.18.1). Significant database matches and preliminary functional assignments are listed in supplemental file, S4.

In general, the majority of prophage genes encode proteins that have homology with phage proteins from the sequence databases. A summary of the most significant protein functions identified in the genomes are outlined below. The genomic structure of animal *E. faecium* phages is displayed in Figure 6.4 as direct output from the PHAST server.

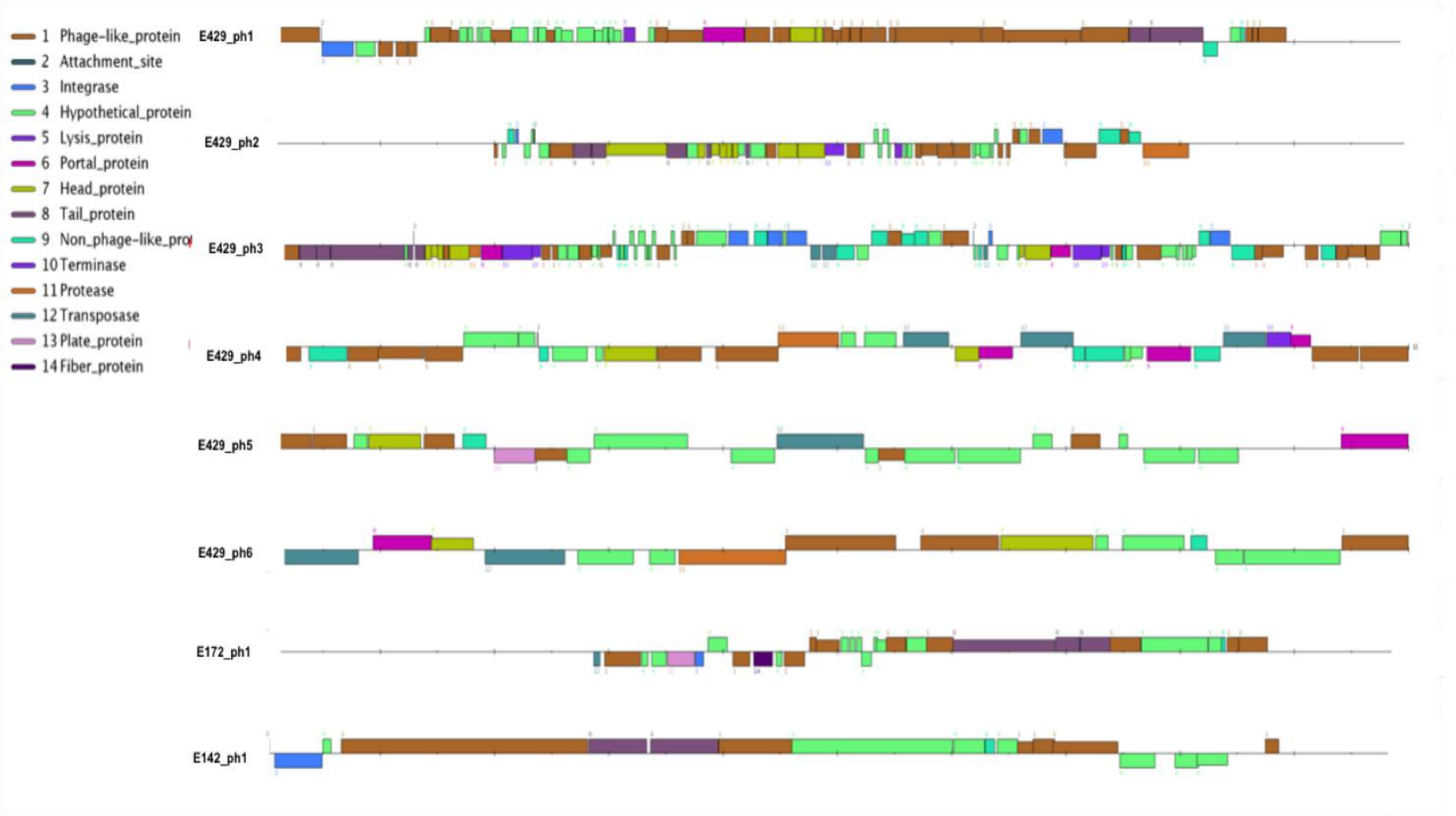

1 Phage-like_protein
2 Attachment_site
3 Integrase
4 Hypothetical_protein
5 Lysis_protein
6 Portal_protein
7 Head_protein
8 Tail_protein
9 Non_phage-like_pro
10 Terminase
11 Protease
12 Transposase
13 Plate_protein
14 Fiber_protein
E429_ph1
E429_ph2
E429_ph3
E429_ph4
E429_ph5
E429_ph6
E172_ph1
E142_ph1

Figure 6.4: Functional annotation comparison of *E. faecium* phage elements from the three animal strains according to PHAST database. Phages E429_ ph1, ph2, ph3, ph4, ph5 and ph6 are present in strain E429 (chicken); phage E172_ ph1 is present in strain E172 (calf) and phage and E142_ ph1 is present in strain E142 (pig). Modular organisation is highlighted with different colours and numbers to reveal grouped functions associated with the phage lifecycle, Brown (1) for phage-like protein; dark green (2) for attachment site; sky blue (3) for integrase; light green (4) for hypothetical protein; purple (5) for lysis proteins; magenta (6) for portal protein; mustard (7) for head proteins; medium purple (8) for tail proteins; turquoise (9) for non-phage-like proteins; deep violet (10) for terminase; orange (11) for protease ; marine blue (12) for transposase; and light pink (13) for plate proteins.

In *silico* analysis of potential functionality indicates that chicken phages, have sufficient composition for integration/excision, DNA replication and capsid/tail morphogenesis to generate functional virions, either alone or synergistically with other phage. However, E142_ph1 was found in pig genome, does not have sufficient composition for integration/excision (Figure 6.4 and Supplemental File, S4.). Animal *E. faecium* phage-related genes have high similarity to phage found in species of other genera not only *Enterococcus* phage for example *Lactococcus, Lactobacillus, Streptococcus, Listeria* and *Bacillus* phage. The hypothetical proteins in the phage regions have high similarity with *E. faecium* strains Aus0004 and NRRL.

6.2.5 Comparative genomic analysis *E. faecium* bacteriophage

6.2.5.1 General features of *E. faecium* phage genome

Thirty-nine strains of *E. faecium* out of 139 available from the NCBI genome database revealed the presence of 56 prophage-like elements. These identified putative prophages were functionally investigated using *in silico* analyses. The phage genomes dataset comprises prophage-like elements from 12 animal strains, 15 clinical strains including two strains from the CC17 genotype, 4 commensal strains, 2 food strains, a strain isolated from river water and 3 strains of unknown source.

The prophage genomes range in size from 13.9 to 55.1 kb, with an average G + C content of 35% to 37.9% and show considerable variation encoding between 17 to 72 ORFs (Table 6.3). These ORFs revealed substantial sequence similarity with sequences in the PHAST databases. The majority of the ORFs carried by the *E. faecium* prophages are organised to be transcribed in one direction, whereas the lysogeny module was typically transcribed in the opposite direction.

6.2.5.2 Genome clustering: gene content analysis

Based on gene content of whole-genome alignments, the 56 prophage sequences were classified into 8 different clusters. The main purpose of clustering the *E. faecium* phage genomes was to determine relationships among genes and modules that might have been exchanged between phage genomes by lateral gene transfer and which is likely to produce their mosaic architecture.

The phage cluster identifiers are presented in Table 6.3. Cluster A contains Aus0085_ph3, E1007-ph1, E1392-ph1, E2039_ph1, E2134_ph1, E4215_ph1, E142_ph1, E172_ph1 and E429_ph3. Cluster B contains 1,231,501_ph1, E1622_ph2, E1623_ph1, E1630_ph1, and E1972_ph1. Cluster C contains Com15_ph1, E1050-ph1, E1573_ph1, E1590_ph1, E2620_ph1, E429_ph2, NRRL_ph1 and NRRL_ph2. Cluster D contains E1185-ph1, E0120_ph1, Com12-ph1, E2071_ph1, E1574_ph1, 1,141,733_ph1 and E3346_ph1. Cluster E contains E1644_ph2, E4452_ph1, E429_ph1, E0045_ph1 and E1622_ph1. Most of the cluster A, B, D and E prophages are present in animal *E. faecium* isolated from chickens (E429 and E0045), dog (E4452) and mouse (E1622) plus one clinical strain belonging to CC17 (E1644). Cluster F contains Aus0004_ph1, Aus0004_ph2, Aus0004_ph3, Aus0085_ph1, DO_ph1, E1578_ph1, E1613_ph1, E1623_ph2, E1644_ph1, E1861_ph1, E1972_ph2, E2039_ph2 and E2883_ph1. Most of the cluster F prophages are present in clinical isolates including one strain belong to CC17 (E1644_ph1), Cluster G contains E429_ph4, DO_ph2, 1,231,501_ph2, and Aus0004_ph4, E1644_ph3 and E2883_ph2 and cluster H contains Aus0085_ph2 and E6012_ph1. Most of the prophages in clusters A and C are from commensal and animal isolates. Cluster B and D are mixed clusters that contain prophages isolated from clinical, commensal, animal and river water (Table 6.3 and Figure 6.5).

Table 6.3: Genometrics of prophage-related sequences of *E. faecium.* The 56 phage genomes were retrieved from 39 isolates of *E. faecium*.

Prophage	Phage location	Size (Kb)	No. of ORFs	GC%	Group	Source
Aus0085_ph3	2455417:2491948	36.5	54	37.9	A	Unknown
E1007-ph1	1299495:1344452	44.9	68	37.4	A	Commensal
E1392-ph1	694822:740020	45.1	70	37.1	A	Unknown
E2039_ph1	91409:136931	45.5	70	36.7	A	Clinical
E2134_ph1	425367:466596	41.2	65	37.5	A	Chicken
E4215_ph1	184650:226771	42.1	59	37.7	A	Chicken
E142_ph1	433557:468604	35	41	37.3	A	Pig
E172_ph1	486654:506555	19.9	27	37.5	A	Calf
E429_ph3	1589043:1629766	40.7	55	37.6	A	Chicken
1,231,501_ph1	536501:583886	47.3	71	36.3	B	Clinical
E1622_ph2	792009:835344	43.3	62	35.9	B	Mouse
E1623_ph1	337845:381585	43.7	61	36.2	B	Clinical
E1630_ph1	220718:265025	44.3	72	36.5	B	River water
E1972_ph1	460219:503311	43	69	36.7	B	Clinical
Com15_ph1	738612:773660	34.3	48	36	C	Commensal
E1050-ph1	1147537:1184635	37.1	51	36	C	Commensal
E1573_ph1	138216:175262	37	54	36.2	C	Bison
E1590_ph1	182184:225277	42.9	61	36.2	C	Unknown
E2620_ph1	1053933:1092651	38.7	53	35.8	C	Clinical
NRRL_ph1	1164025:1207440	43.4	61	35.9	C	Food
NRRL_ph2	1889100:1925416	36.3	54	36	C	Food
E429_ph2	1347483:1395061	47.5	60	36.9	C	Chicken
E1185-ph1	831195:867404	36	55	36.7	D	Clinical
E0120_ph1	573663:610140	36.4	54	36.4	D	Clinical
Com12-ph1	516386:553835	35.9	47	35.1	D	Commensal
E2071_ph1	715129:755872	40.7	57	36.3	D	Poultry
E1574_ph1	526208:565655	39.4	56	36.4	D	Dog
1,141,733_ph1	832928:871079	36.9	53	35.9	D	Clinical
E3346_ph1	469734:510315	40.4	57	36.9	D	Clinical
E1644_ph2	2184725:2220527	35.8	58	37.4	E	Clinical CC17
E4452_ph1	2586336:2630564	44.2	66	36.8	E	Dog
E429_ph1	412480:460595	48.1	70	36.7	E	Chicken
E0045_ph1	522912:567869	44.9	63	36.4	E	Chicken
E1622_ph1	549160:590470	41.3	52	36.2	E	Mouse
Aus0004_ph1	824093:864998	40.9	67	35.4	F	Clinical
Aus0004_ph2	1456511:1496444	39.9	65	35.6	F	Clinical
Aus0004_ph3	2397865:2437393	395	64	36.1	F	Clinical
Aus0085_ph1	785758:840919	55.1	85	36.2	F	Clinical
DO_ph1	821000:858000	37	59	35.9	F	Clinical
E1578_ph1	1158179:1199732	41.5	63	35.4	F	Pig
E1613_ph1	301205:339194	37.9	60	35.4	F	Food
E1623_ph2	621815:661019	39.2	59	35.4	F	Clinical
E1644_ph1	774244:815311	41	67	35.4	F	Clinical CC17
E1861_ph1	756909:796923	40	64	35	F	Clinical
E1972_ph2	524415:562485	38	55	35.1	F	Clinical
E2039_ph2	164944:203986	37.7	55	35.8	F	Clinical
E2883_ph1	524837:567202	42.3	66	35.5	F	Clinical
E2134_ph2	1274188:1322221	48	52	35.3	F	Chicken
Do_ph2	2072323-2089135	16.8	25	36.7	G	Clinical
E429_ph4	1992956:2009130	16.1	46	38	G	Chicken
Aus0004_ph4	2159576-2174179	14.6	19	36.5	G	Clinical
1,231,501_ph2	241734-255551	13.8	17	36.3	G	Clinical
E2883_ph2	1735348-1750156	14.8	19	36.4	G	Clinical
E1644_ph3	1961837-1976645	14.8	19	36.4	G	Clinical CC17
Aus0085_ph2	2215833:2252096	36.2	58	35.2	H	Unknown
E6012_ph1	357820:399130	41.3	68	35.5	H	Clinical CC17

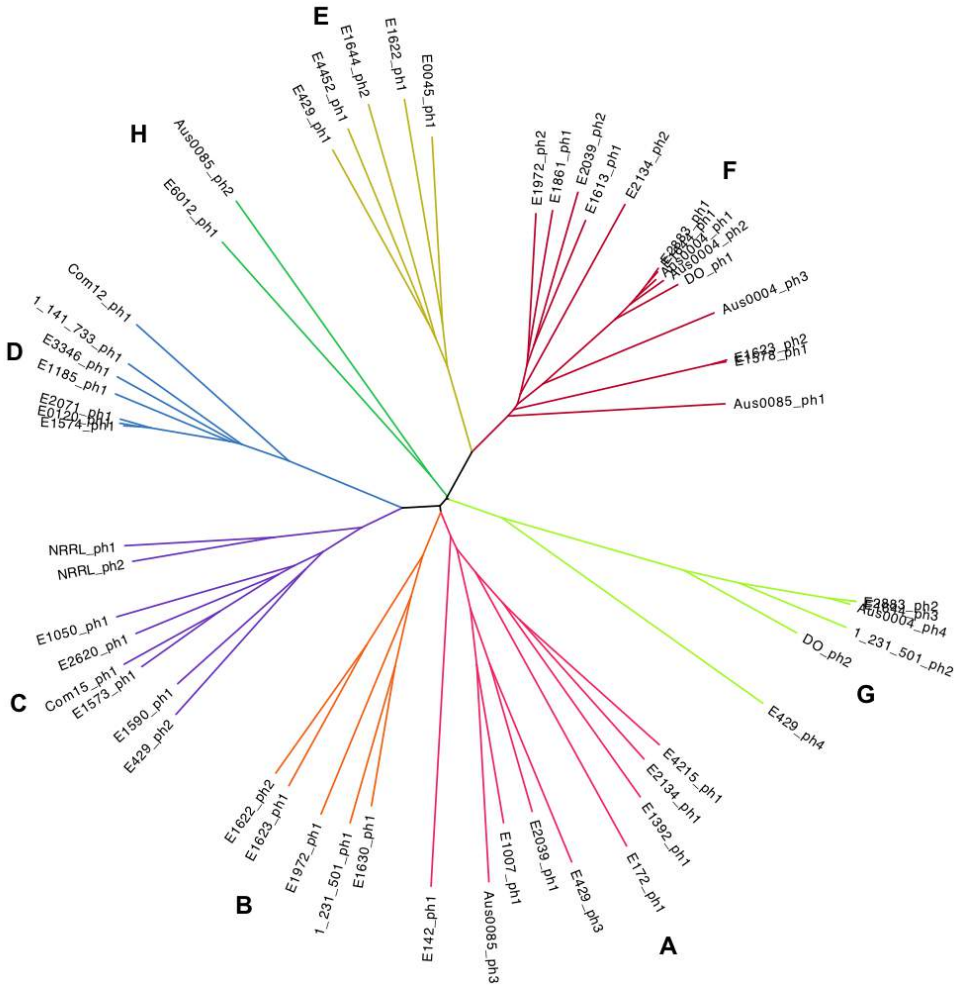


Figure 6.5: Cladogram tree of *E. faecium* prophages. The tree represents the cluster relationships for 56 *E. faecium* prophages present in the genomes of clinical, commensal, animal and food isolates.

A cladogram tree (Figure 6.5) reveals there are clear relationships between the identified prophage genome clusters. Several pairs of clusters are observed to be derived from the same ancestor, for example, clusters A and B, C and D, plus E and F are sister clades. Clusters G, H include prophage genomes from different ancestors. While distantly related, most of the phage

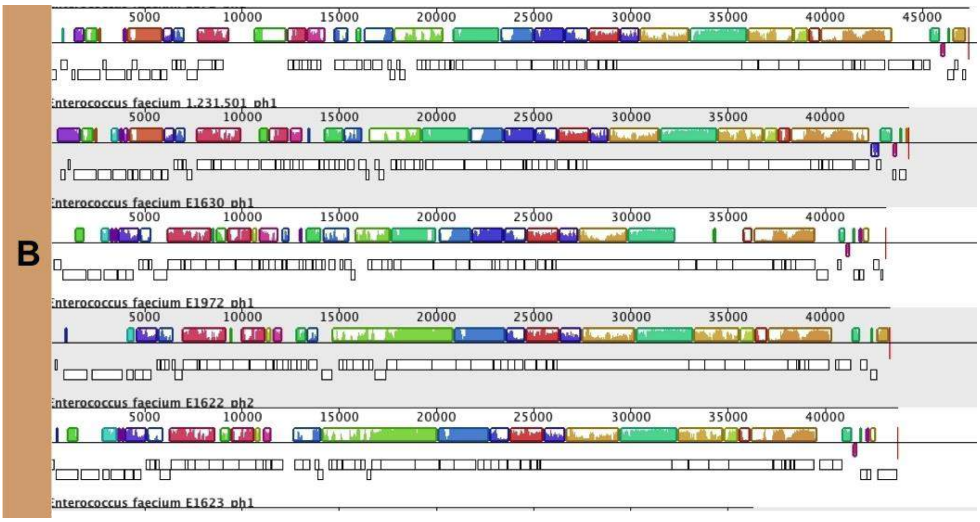
genomes in clusters A and E are prophages present in animal *E. faecium* isolates. Cluster F mainly contains prophages present in clinical *E. faecium* isolates, however, two strains isolated from a pig (E1578_ph1) and from a food (E1613_ph1) were also grouped in this cluster (Figure 6.5).

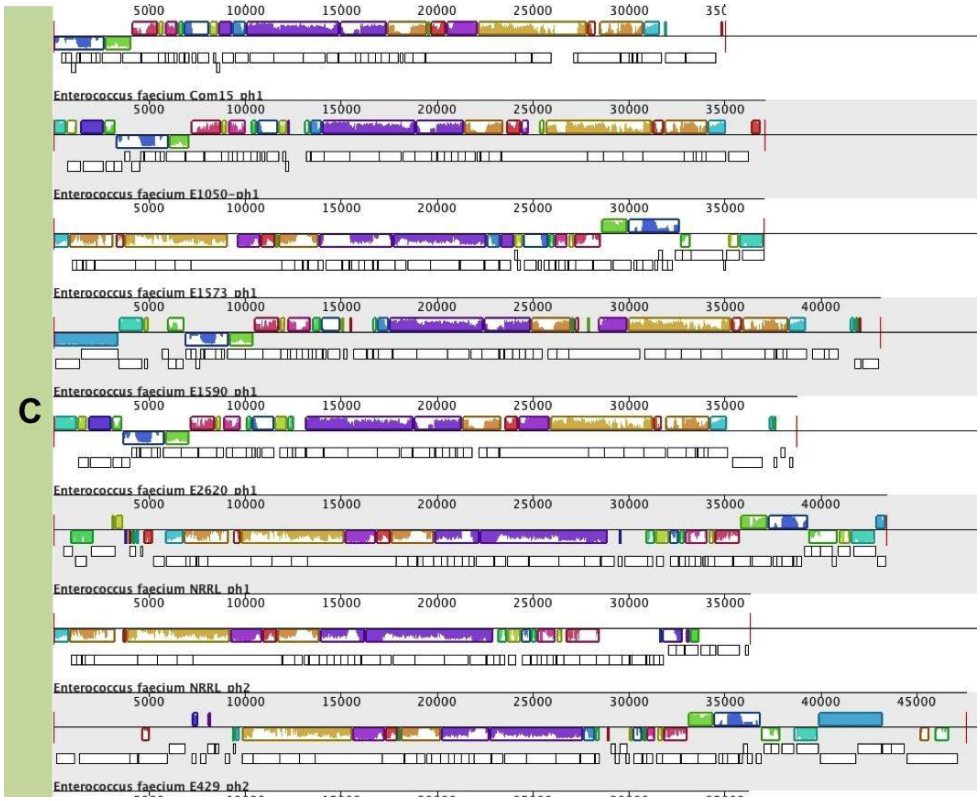
Several examples of phage genomes that were resident in the same host were also found to be grouped together and to share high similarity with each other. For example, Aus0004_ph1, Aus0004_ph2, and Aus0004_ph3 are clustered together in group F and NRRL_ph1 and NRRL_ph2 are clustered together in group C. In contrast, high similarity in prophage genomes was not evident between prophages found in the chicken strain (E429), which contains six prophage sequences and they were each located in separate clusters formed from different ancestors. Prophages found in clinical strains that belong to the CC17 genotype were grouped into four different clusters, E, F, G and H that are formed from the same ancestor (Figure 6.5).

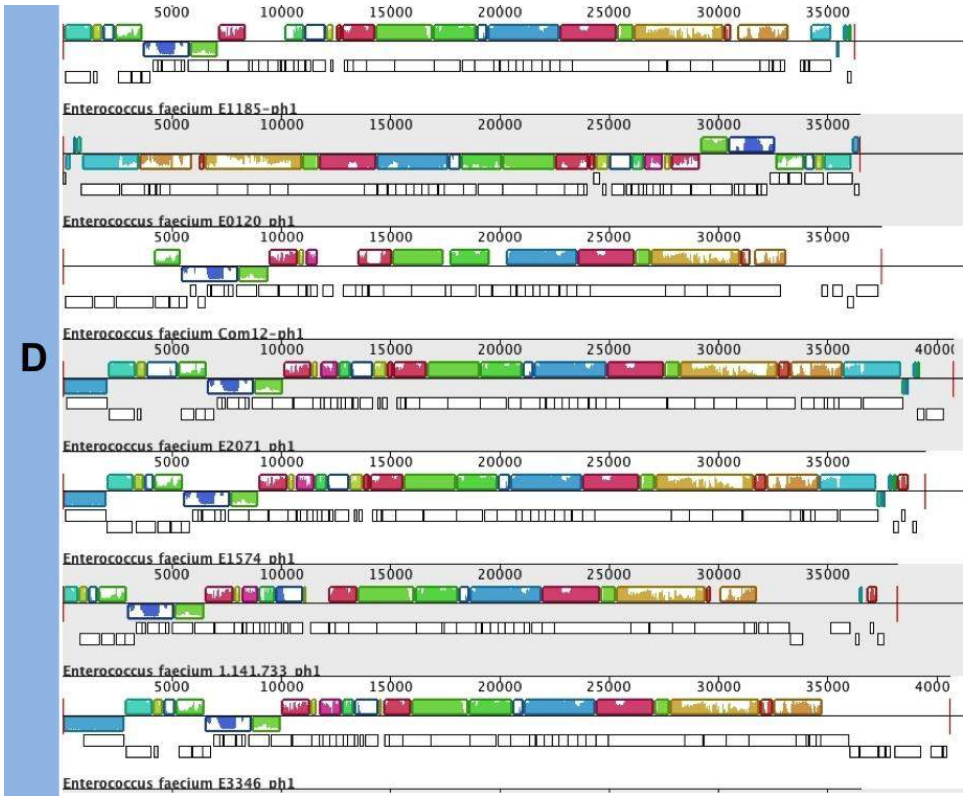
6.2.5.3 Genome clustering: pairwise prophage genome analyses

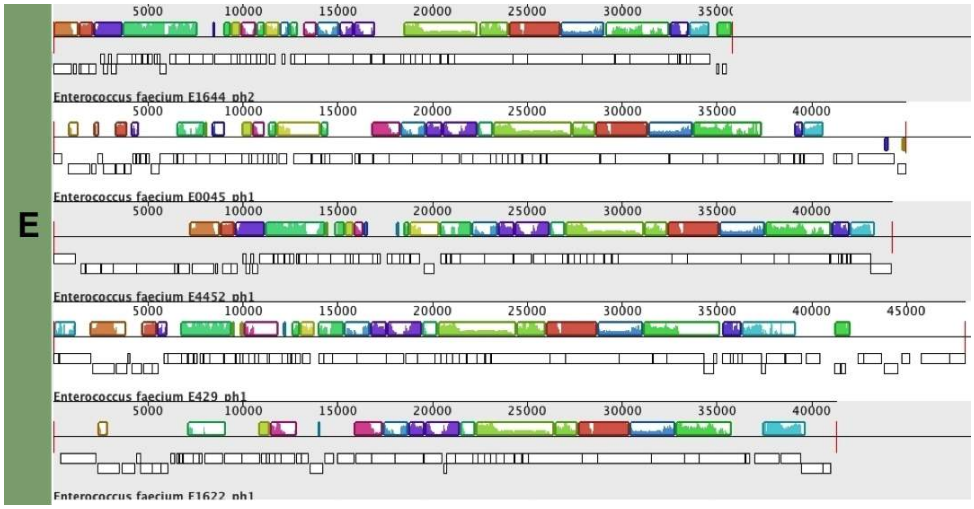
A progressive Mauve multiple alignment was used to identify locally collinear blocks (LCBs) (Section 2.18.1) of conserved sequence segments. Among the *E. faecium* prophage genomes, those in cluster C and D share a considerable number of LCBs (Figure 6.6). While the other prophages in clades A, B, E, F, G and H share fewer related blocks of sequences, they also differ in their overall sequence from each other. All prophages revealed a highly mosaic-like structure and the Mauve analysis proved useful for

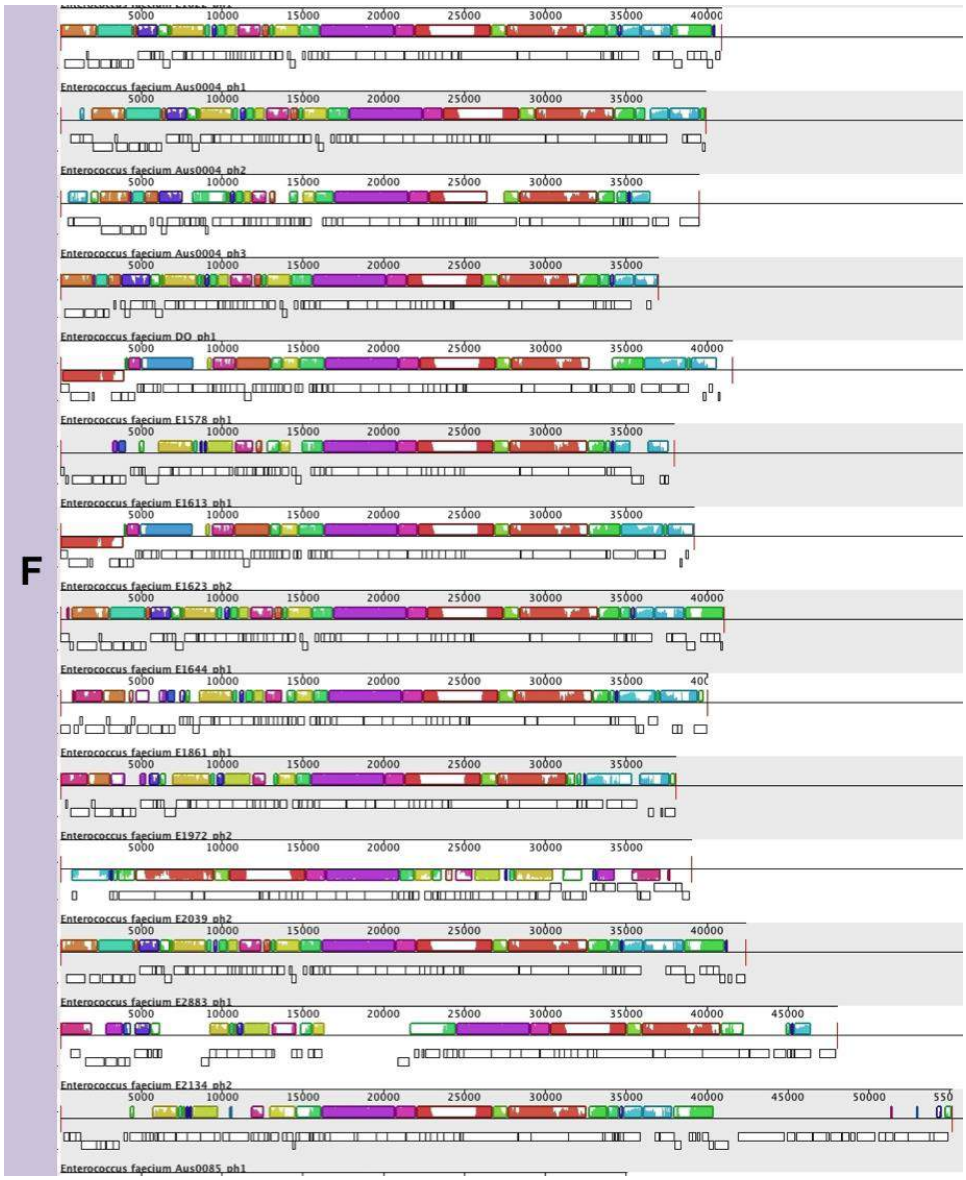
displaying segments of similarity between more distantly related genomes, as well as revealing potentially newly-acquired genes among more closely-related genomes. For example, the phage genomes in cluster F clearly illustrate high identity with each other and the locations of the LCBs are well-conserved. Potential newly acquired genes were identified as mobile elements portions and hypothetical proteins (Figure 6.6).


B
5000
10000
15000
20000
25000
30000
35000
40000
45000
Enterococcus faecium 1,231,501_ph1
Enterococcus faecium E1630_ph1
Enterococcus faecium E1972_ph1
Enterococcus faecium E1622_ph2
Enterococcus faecium E1623_ph1


C
5000
10000
15000
20000
25000
30000
35000
40000
45000
Enterococcus faecium Com15_ph1
Enterococcus faecium E1050-ph1
Enterococcus faecium E1573_ph1
Enterococcus faecium E1590_ph1
Enterococcus faecium E2620_ph1
Enterococcus faecium NRRL_ph1
Enterococcus faecium NRRL_ph2
Enterococcus faecium E429_ph2


D
5000
10000
15000
20000
25000
30000
35000
40000
Enterococcus faecium E1185-ph1
Enterococcus faecium E0120_ph1
Enterococcus faecium Com12-ph1
Enterococcus faecium E2071_ph1
Enterococcus faecium E1574_ph1
Enterococcus faecium 1.141.733_ph1
Enterococcus faecium E3346_ph1


E
5000
10000
15000
20000
25000
30000
35000
40000
45000
Enterococcus faecium E1644_ph2
Enterococcus faecium E0045_ph1
Enterococcus faecium E4452_ph1
Enterococcus faecium E429_ph1
Enterococcus faecium E1622_ph1


F
Enterococcus faecium Aus0004_ph1
Enterococcus faecium Aus0004_ph2
Enterococcus faecium Aus0004_ph3
Enterococcus faecium DO_ph1
Enterococcus faecium E1578_ph1
Enterococcus faecium E1613_ph1
Enterococcus faecium E1623_ph2
Enterococcus faecium E1644_ph1
Enterococcus faecium E1861_ph1
Enterococcus faecium E1972_ph2
Enterococcus faecium E2039_ph2
Enterococcus faecium E2883_ph1
Enterococcus faecium E2134_ph2
Enterococcus faecium Aus0085_ph1

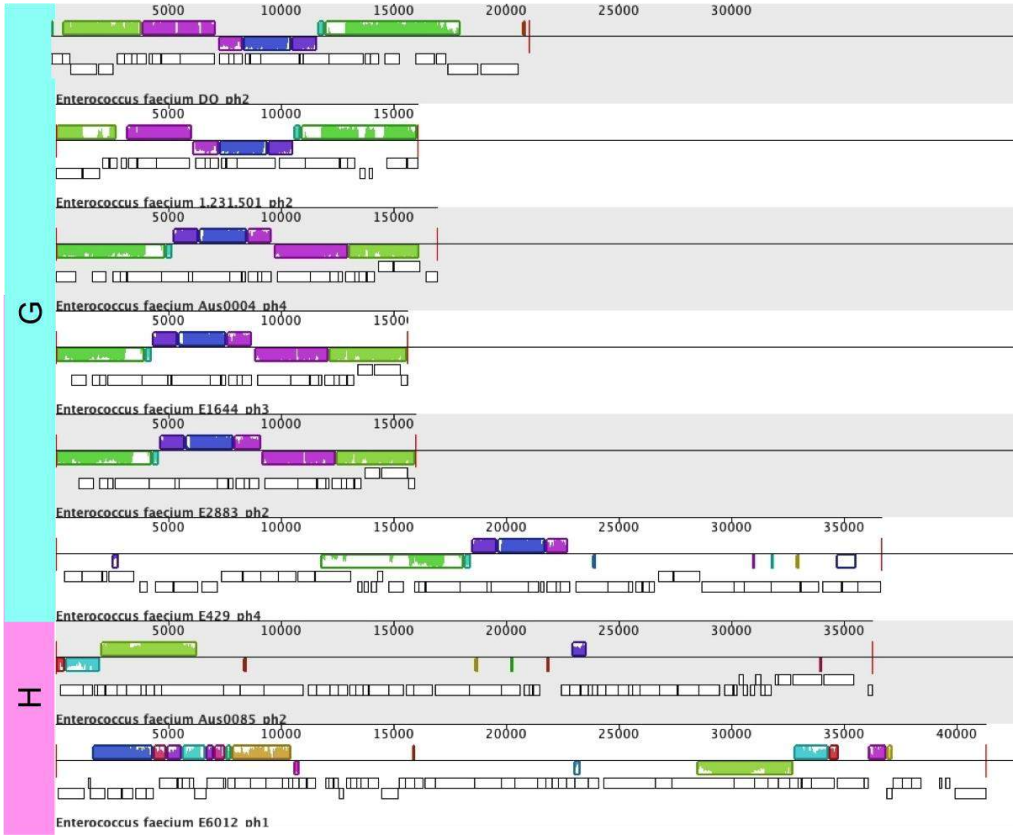


Figure 6.6: Mauve alignment of *E. faecium* phage genomes. Protein alignments of each of 56 *E. faecium* phage genome clusters displayed as segments of similarity between genomes. The strength of the relationship is represented by colour blocks.

6.2.5.4 Lysogeny module of *E. faecium* prophages

The overall organisation of the prophage lysogeny modules across the *E. faecium* phages is similar to temperate phages found in *E. faecalis* and other low G + C Gram-positive bacteria (Yasmin, Kenny *et al*. 2010, Tang, Bossers *et al*. 2013). The first transcriptional unit of the phage (i.e. as it appears on the host chromosome) is typically the integrase region. Genes encoding integrases, transcriptional regulators belonging to the Cro/cI and

SinR repressor family, were identified in the analysed lysogeny clusters. Phages have two repressor proteins. One is essential for maintenance of lysogenic and the other for the control of the lytic cycle of growth. The first repressor called cI silences transcription of the other phage genes and maintains lysogen immunity to superinfection by other phages. Cro is the second repressor and it functions midway in the lytic cycle to turn down expression of the early genes encoding Cro itself and the cI repressor gene (Johnson, Meyer *et al*. 1978). The SinR repressor belongs to the group of Sin (sporulation inhibition) proteins of *Bacillus subtilis*. The SinR protein structure contains two domains: a dimerisation domain stabilised by a hydrophobic core and a DNA-binding domain similar to domains of the bacteriophage 434 cI and Cro proteins that control prophage induction.

Transcriptional regulators belonging to the SinR family are encoded in all the prophage genomes of cluster C and D. Transcriptional regulators belonging to the Cro/cI family of repressor are present in several phage genomes in clusters F, (3/14; E1613_p1, E1861_ph1 and E2039_ph2). Most of the prophages in cluster E (4/5) have transcriptional regulators belonging to the repressor (Cro/cI), while E1622_ph1 has a SinR-like transcriptional regulator. All the prophage in cluster G have a distinct repressor from a different family that shares very high similarity (E-value 1.00E-11) with the cI-like repressor present in *Lactococcus* phage bIL311.

Antirepressors are small proteins which provide an alternative induction strategy for prophages by binding to lysogen maintenance repressors and they were identified in twenty-one *E. faecium* prophage genomes.

Antirepressors-like proteins were identified in cluster F (9/14), 5/9 in cluster A, 2/5 of prophages belonging to cluster E (E1644_ph2 and E4452_ph1) and 4/6 prophages belonging to cluster G. Antirepressors were absent from clusters C and D prophage genomes.

Integrases in the studied *E. faecium* phage genomes all belong to the site-specific tyrosine (XerC) family, which utilise a catalytic tyrosine to mediate strand cleavage (Groth and Calos 2004). A cladogram tree analysis generated using amino acid sequences of the integrases of the *E. faecium* prophage clusters (Figure 6.5 and Figure 6.6) identified multiple clades (Figure 6.7).

While the prophage integrases present as seven different clades (labelled Integ1 to Integ7) they all belong to the tyrosine XerC family. The differences between the clades represent minor (Supplemental File, S5). The pan-genome of *E. faecium* reveals 15 different sequence types of the tyrosine XerC family, however, only 7 are represented in the genomes of *E. faecium* used in the phage comparison (ORTHOMCL4499, ORTHOMCL2990, ORTHOMCL4377, ORTHOMCL2597, ORTHOMCL2459, ORTHOMCL2787 and ORTHOMCL2561, (Supplemental File, S1). The integrases clusters were spread non-uniformly between the 7 prophage clades shown in figure 6.5. For example cluster ORTHOMCL4499 and ORTHOMCL4377 were present in Integ2 and only ORTHOMCL2597 in Integ4.

The prophages represented by clusters (Figure 6.5) have a cluster specific integrase sequence types. In contrast to those from cluster F which comprises multiple integrase sequence type (Integ1, 2 and 6). These integrases of clade Integ6 differ from other *E. faecium* phage integrases and might represent recombinases enabling phage to infect widely across *E. faeciun* hosts. The remaining prophages in clusters F have similar integrase sequence types (XerC family) (Integ1, 2) (Supplemental File, S5). Cluster C and D all have the same integrases sequence type (Integ2).

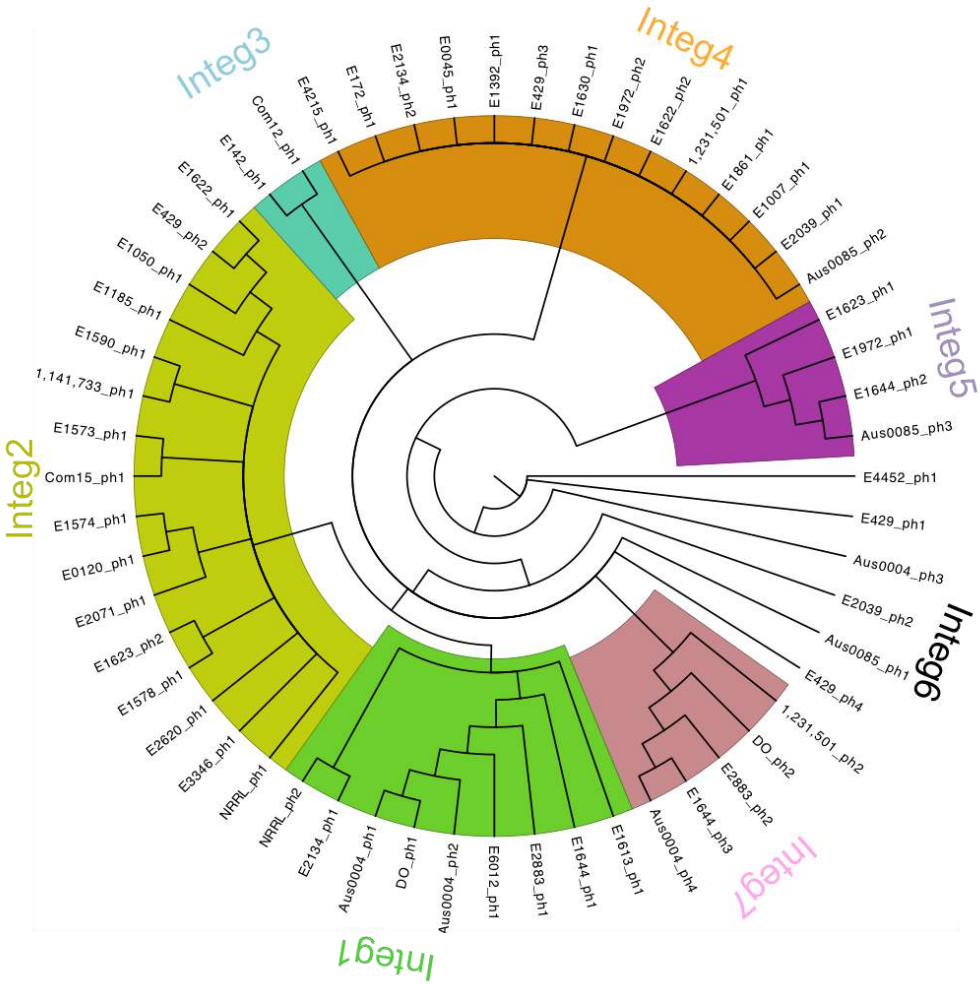


Figure 6.7: Cladogram tree of *E. faecium* prophage integrases. The cladogram is based on the alignment of integrases amino acid sequences and represents the relationship between *E. faecium* prophage integrases.

6.2.5.5 Replication module

The replication module of the identified prophages was typically bordered on one side by the lysogeny module and on the other side by the packaging module. ORFs with significant sequence similarity to proteins involved in DNA replication were identified in all 56 *E. faecium* prophage genomes (Supplemental File, S4).

The majority of the replication modules contain a gene encoding a putative single-strand DNA binding protein (SSB). No significant sequence similarity was shown between the SSB across the phage clusters A to H. SSB was encoded in four out of five prophage genomes in cluster B, 5/8 from cluster C and 2/7 from cluster D within cluster E SSB proteins shared high amino acid sequence similarity, excluding strain E0045. Most of the prophages in cluster F have a gene encoding an SSB excluding prophages E2314_ph2 and E1861_ph1, which both encode the same distinct SSB.

It has previously been described that many bacteriophages code for their own SSB meaning they do not rely on those encoded by their host (Tang, Bossers *et al*. 2013). Multiple examples were identified here of *E. faecium* prophages that lacked a gene encoding a DNA binding protein, suggesting that they depend on SSB encoded by their hosts. Phage replication initiation

and membrane attachment functions together with phage-associated recombinase proteins are encoded in most of *E. faecium* prophages in the replication module. The absent of some of these genes in several prophages reveals a requirement for DNA replication functions for their lifecycle.

6.2.5.6 Packaging module

Most of the packaging modules in the *E. faecium* phage genomes identified here are principally comprised of three genes encoding the small and large subunits of the terminase and the portal protein. In 23 of the prophages the terminase is encoded by a single gene while in 31 the terminase gene appeared as two ORFs (small and large subunits). No terminase gene was identified in two animal prophages E142_ph1, and E172_ph1. A cladogram tree based on the amino acid sequences of the terminases (large subunits) revealed that the integrases of the *E. faecium* prophage clusters are discriminated into seven different clades (Figure 6.8).

The terminase protein sequences of all prophages in clusters D and F share high similarity and were grouped in clades Term6 and Term2, respectively (Supplemental File, S6). All prophages in cluster C were grouped together as a clade sister group to three prophages present in chicken (E429), dog (E4452) and a clinical prophages CC17 genotype (E6012) (Term1) that contains two different clades derived from a common ancestor. The Term7 clade contains highly diverse protein sequences. Based on prophage intergrase sequence analyses (Figure 6.7), prophages belonging to cluster C and D are highly similar but their terminases show marked variation (Figure

6.8). The portal protein gene was identified in 37 phage genomes but was not evident in nineteen.

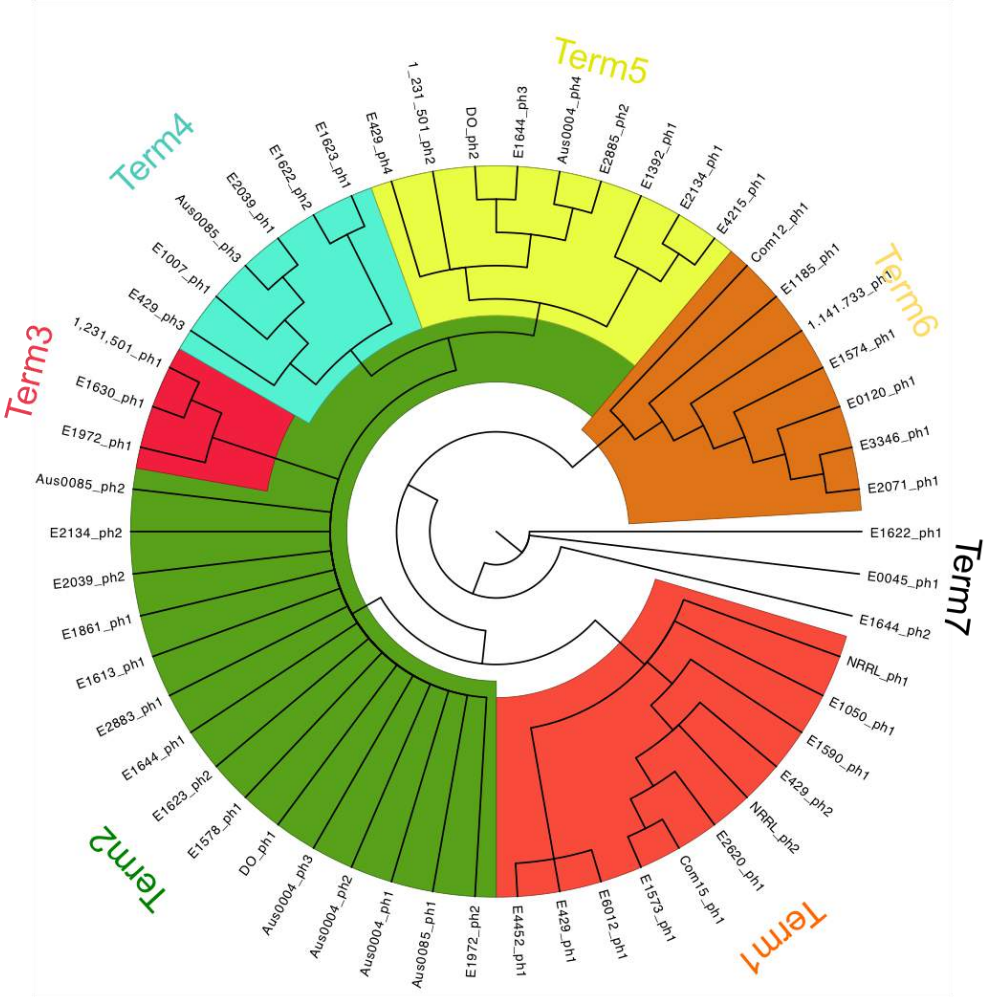


Figure 6.8: Cladogram tree of the large terminase subunits of *E. faecium* prophages. The tree is based on an alignment of the amino acid sequence of 54 terminases.

6.2.5.7 Morphology module

In all of the *E. faecium* prophages analysed, the head morphogenesis and the tail structural genes are the largest modules. These major capsid and tail portions show high similarity to proteins of the same annotated functions of *Listeria*, *Lactobacillus*, *Staphylococcus*, *Paenibacillus*, *Mycobacterium*, *Enterococcus* and *Lactococcus* bacteriophages (Supplemental File, S7).

The majority of the *E. faecium* prophages contained two or three putative tail proteins, including the major and the minor tail proteins. Tail proteins were not encoded in all of the prophage present in cluster G, however, head-tail joining proteins and head-tail adaptor proteins were present in this cluster which will serve as functional replacements. These proteins share very high similarity (E-value 1.00 E-08) with head-tail joining proteins found in *E. faecalis* prophage EFRM31 (NC_015270).

A cladogram tree based on the amino acid sequences of the phage tail length tape-measure protein, which is encoded by the largest ORF of this module, indicated that *E. faecium* prophages comprise different major tail proteins (Figure 6.9). These tail proteins were grouped into 7 different, that matched the clusters determined by supported the comparative genome analysis (Table 6.3 and Figure 6.5).

Cluster B prophages encoded the longest phage tail tape-measure gene (6.44 kb) while cluster A prophages possessed tail genes ranging from 2.50 to 3.39 kb. The tail tape-measure gene in cluster E is ~ 3.11 kb, cluster F is ~3.43 kb cluster C is ~ 4.71 kb and cluster D is ~ 3.48 kb in size, further

highlighting the heterogeneity of this major structural component of the virion.

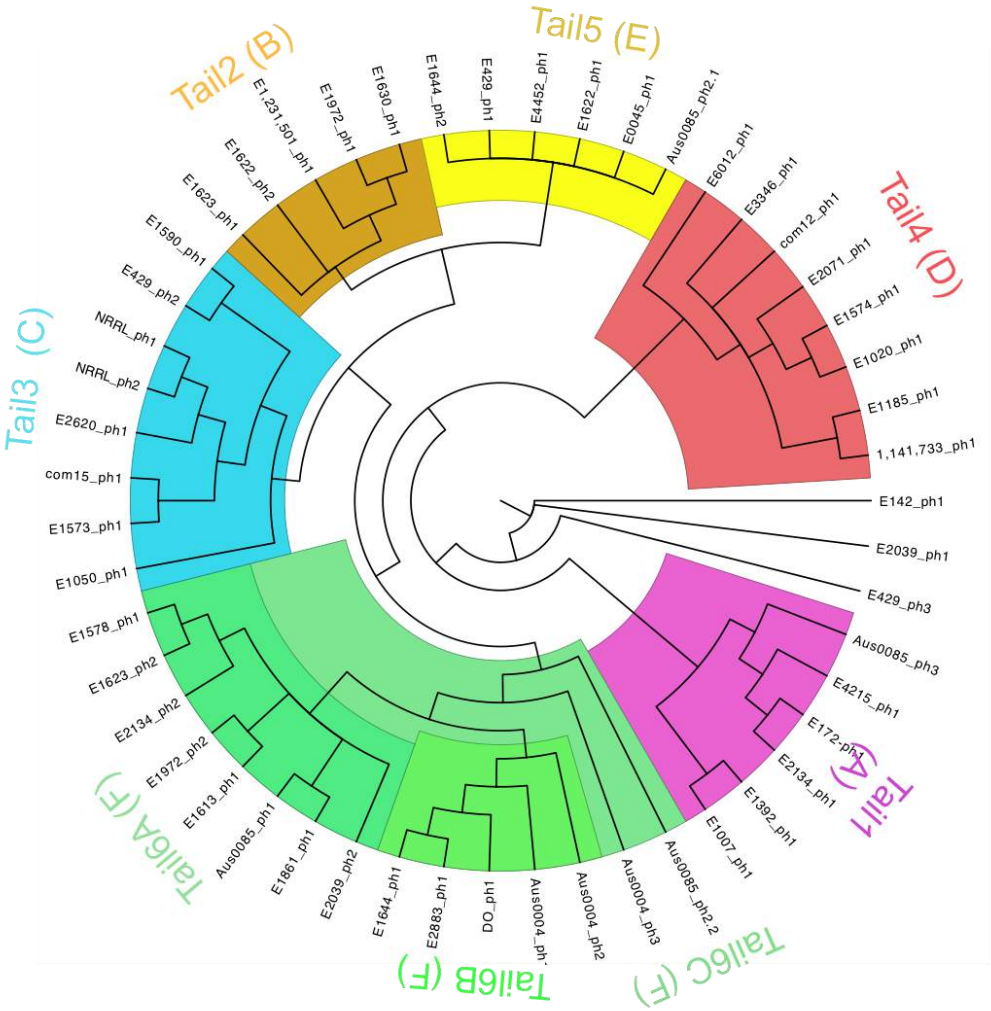


Figure 6.9: Cladogram tree of the tail protein of *E. faecium* prophages. The alignment of the amino acid sequence of 51 tail proteins reveals differences between *E. faecium* prophages producing distinct groupings.

6.2.5.8 Lysis module

The lysis modules of the *E. faecium* prophages mainly consist of a holin. Four prophages: 1,141,733_ph1, E4452_ph1, E1578_ph1 and E172_ph1 contain endolysin genes. Prophage E429_ph3 and Com12-ph1 contain Hydrolase genes.

All prophages of cluster G do not encode lysis module genes, which suggest they encode different unidentified lysis systems or they are reliant on that produced by other resident phage or phage-like elements to complete their lytic cycle (Fard, Barton *et al*. 2010) .

A cladogram was produced using the phage holin amino acid sequences which revealed that 27 prophages have the same holin (Holin3) and these genes have very high similarity with a holin described in *E. faecalis* temperate bacteriophages (Figure 6.10). Three clades of *E. faecium* prophages seem to have a different sequence type of holin (Holin 1, 2 and 4) (Figure 6.11). Seven prophages possess two genes encoding holins with both genes adjacent to each other. According to the PHAST database the Holin1 clade have very high sequence identity (E-value range from 2.00E-26 to 8.00E-26) to *E. faecalis* phiFL4A and phiEf11 holins. Most of the phage holins that form clade Holin2 have homology (E-value 6.00E-12) with the *Lactococcus* phage ul36 holin. The Holin4 sequences have high similarity with a holin found in *E. faecalis* EF62phi (E-value 8E-48).

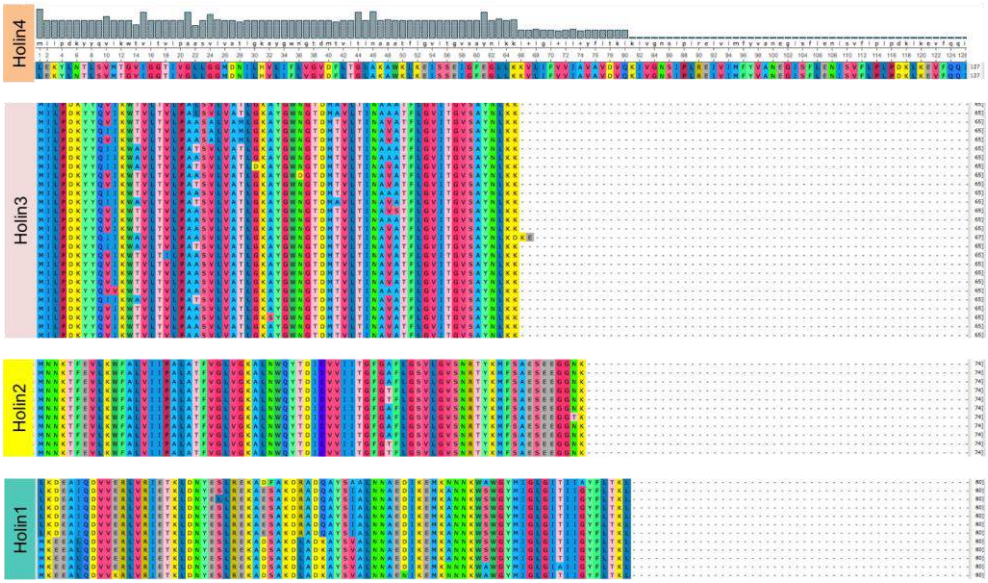


Figure 6.10: Multiple alignments of *E. faecium* prophage holins. The protein alignment indicates high sequence conservation within 4 main holin clusters.

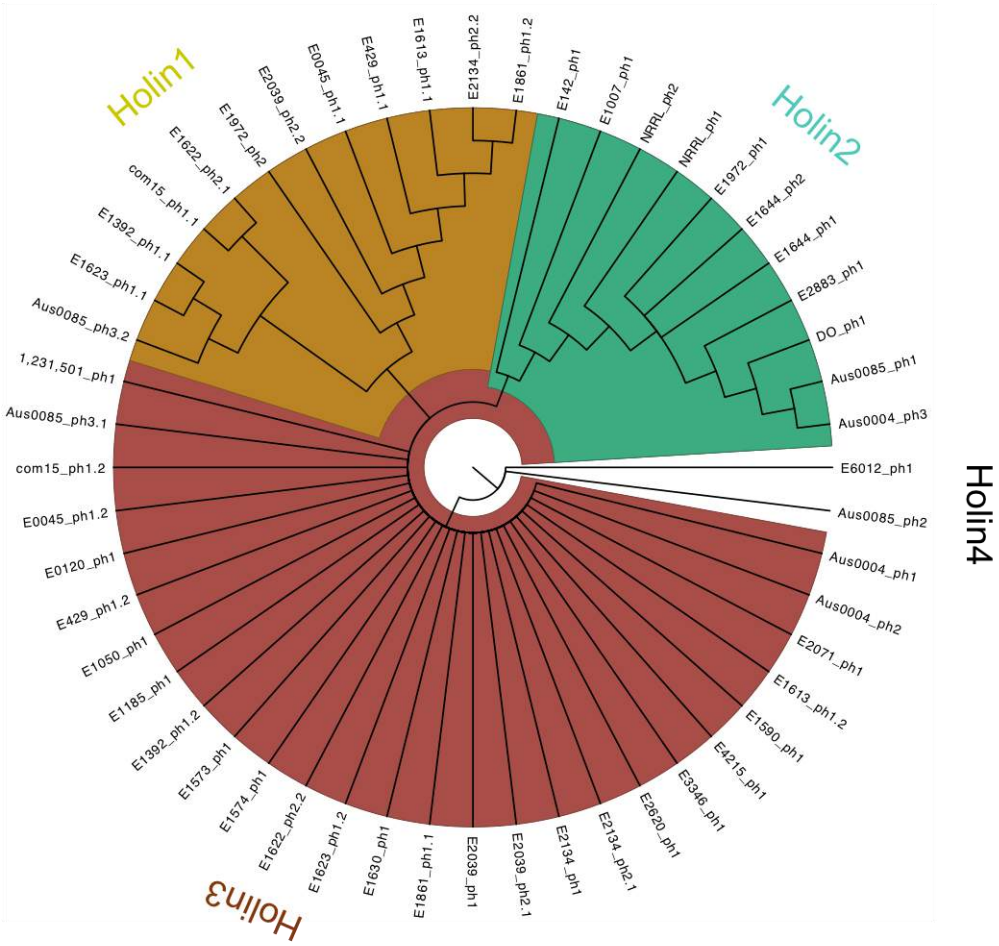


Figure 6.11: Cladogram tree of *E. faecium* prophage holins. Based on the alignment of 52 amino acid sequence of the holin protein, *E. faecium* prophages have 4 different families of holin. The Holin 4 protein sequences are nearly identical.

6.2.6 Cluster diversity and newly-acquired genes

An alternative perspective on the cluster relationships was sought by investigating the conserved sequences in *E. faecium* prophages. The presence locally collinear blocks of sequence was identified using Mauve alignment of representative prophages of each sequence type. The genome

alignments identified common regions (blocks) across multiple phage sequence types and these regions show diversity. However, there are many regions that are specific to each cluster (Figure 6.12).

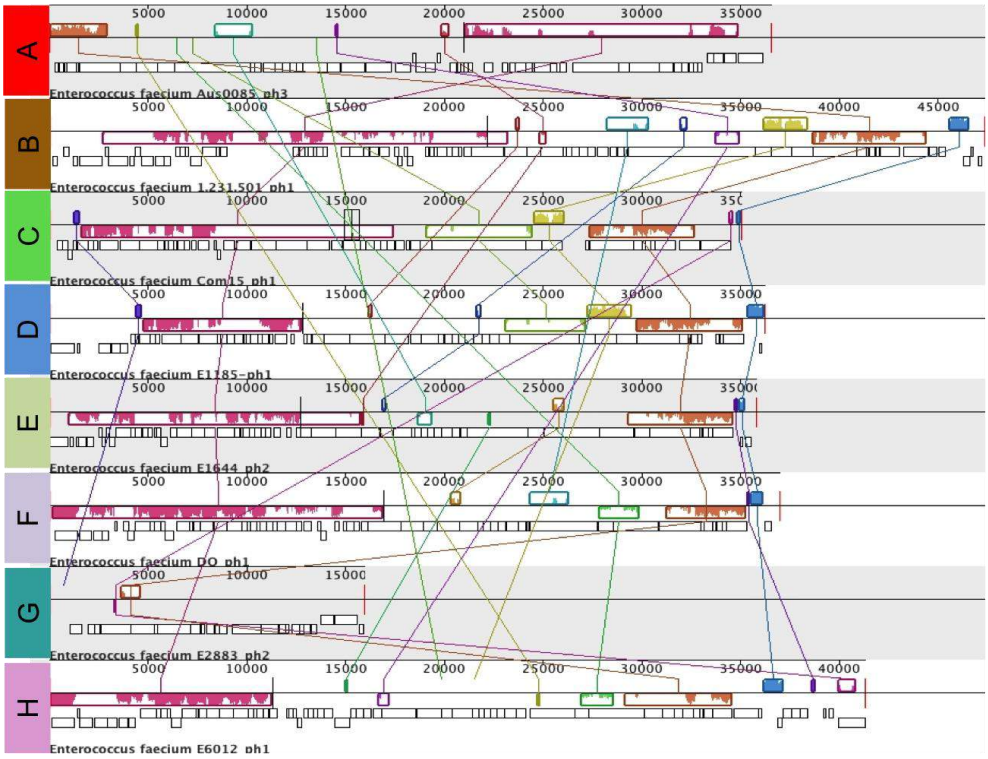


Figure 6.12: Mauve alignment of 9 *E. faecium* prophage type genomes. Pairwise alignment of one prophage genome of each of *E. faecium* pophage clusters A, B, C, D, E, F, G and H displays a low degree of similarity between the prophage genomes and highlighted diversity. The strength of the relationship is represented by coloured region.

A benefit of the genome clustering described above is that it potentially enables the identification and classification of those genes that are most expected to have been exchanged horizontally. Since each cluster contains common genes that exist in all cluster members, genes are revealed that are

present in only a subset of the genomes. The lack of conservation could be a consequence of gene loss from genomes or the recent acquisition by horizontal genetic exchange. Although both possibilities could account for genes that exist in only one genome, these genes are more likely to be recently acquired. When genes exist in a single prophage genome of one cluster and are presented in one or more prophage belonging to other clusters these genes need to be studied carefully to explore the origins of these genes.

Looking at shared genes between prophage types also identifies colocalised genes that are equally present, which might further supports horizontal transfer between phage types. For example, a hypothetical protein that is located in rightmost genomic segments was found in all prophage clusters excluding cluster G. Another hypothetical protein from the leftmost genome region is associated with cluster E and F only. Most of the unique genes of individual prophage genomes within the clusters represent small hypothetical proteins, which might be host specific or arise from geographical or environmental influences.

6.2.7 Identification of putative phage attachment sites

To identify both the phage and the bacterial attachment sites, genomic sequences flanking the putative prophages were analysed for the existence of short directly repeated sequences using Unipro UGENE (Section 2.18.1). Phage attachment sites (attP) are commonly located in the proximity of the phage integrase gene (5′ end) and of the lysin (3′ end). The corresponding

genomic regions of the *E. faecium* prophage were searched and putative phage attachment sites for the majority of the prophages were identified (Table 6.4).

Twenty-eight of the 56 prophage genome insertion sites revealed defined genome ends ranging from 15 to 93 bp in size. The attachment sites for the remaining phage could not be determined using this approach. The attP sites of E1007_ph1, E1630_ph1 and E1622_ph2 share an identical 17 bp although there are potentially an extra two base pairs in the E1622_ph2 site. The phage integrase proteins for both of these prophages share homology and are grouped together (Integ3 clade) in the integrase tree (Figure 6.7). In addition, three prophages belonging to G sequence type share an identical 27 bp attP site and share the same integrase sequence group (Iinteg7) (Table 6.4). This identifies that as is expected specific integrases use equivalent core attachment sites for their insertion. *In silico* analysis of the cluster G prophages revealed that they are inserted at the same relative location in the genome of the host *E. faecium* with each being bordered by genes encoding a sulfite exporter TauE/SafE, a putative acytyltransferase, a citrate transporter (*citS)* adjacent to the 3′ lysis module and a holliday junction resolvase adjacent to the 5′ integrase.

Table 6.4: Putative attachment sites attP of *E. faecium* prophages.

Prophage	Group	Ends	Putative phage attachment site (5′–3′)	Attachment site position
E1007-ph1	A	17-base 3′	**ACTCCCGCCGTCTCCAT**	(1301608..1301624,1343909..1343925)
Aus0085_ph3	A	61-base 3′	**ACTCTTAATCAGCGGGTCGCGGG TTCGAGCCCCTCACGGCCCATTG GGTGCCAAACCCACG**	(2447984..2448044,2491648..2491708)
E2039_ph1	A	22-base 3′	**TTGAATGCCATTTTGAATGCCA**	(92227..92248,135581..135602)
E1630_ph1	B	15-base 3′	**ACTCCCGCCGTCTCCAT**	(220657..220673,263743..263759)
E1622_ph2	B	19-base 3′	**CGACTCCCGCCGTCTCCAT**	(792457..792475,834546..834564)
E1972_ph1	B	81-base 3′	**GTTTTTAACAAAAAA**	(457389..457403,500357..500371)
E1623_ph1	B	36-base 3′	**TTTTTTTGTTATCTGTTTTTTTA TATTAACGATTTC**	(338198..338233,380462..380497)
E1573_ph1	C	15-base 3′	**CCTTGCTACTTCTTACTTTCTTC**	(134740..134762,175329..175351)
Com15_ph1	C	21-base 3′	**GAAGAAGAAAGTAAGAAGTAG**	(734115..734135,774551..774571)
E1050-ph1	C	15-base 3′	**TGGCTCTTTTTTTAT**	(1139566..1139580,1181591..1181605)
E1185-ph1	D	17-base 3′	**AAGAAGTAGCAAGGTTT**	(831187..831203,874769..874785)
Com12-ph1	D	15-base 3′	**GATGAACTTCCTTTA**	(512345..512359,553082..553096)
E1574_ph1	D	15-base 3′	**TACTAATACTTCTAC**	(516241..516255,560853..560867)
E0045_ph1	E	22-base 3′	**AAATCCTGTACCTTCCTTATAT**	(523576..523597,563714..563735)
E1622_ph1	E	15-base 3′	**GATATCATGGAGAAT**	(546639..546653,588124..588138)
E1644_ph2	E	19-base 3′	**TACATCATACCGCCCATCA**	(2184823..2184841,2220814..2220832)
E429_ph1	E	15-base 3′	**TTTTTGAAAAAAATA**	(411434..411448,452588..452602)
E2883_ph1	E	32-base 3′	**AAATAACCCCTGTATCCTTTGCG GTACAGGGG**	(525152..525183,566117..566148)
E1623_ph2	F	16-base 3′	**AAGAAGCCTTCATGGC**	(622172..622187,660867..660882)
E1644_ph1	F	93-base 3′	**ATAAGTAGACATTGTAGTTTCTA AACTGCTATGTCCTAAACGTTTC GATACGCTAAGTATATTTACTCC TTGATAAAGTAAAATAGATGCATG**	(775389..775481,815621..815713)
E1578_ph1	F	15-base 3′	**ATTCTCCATGATATC**	(1158831..1158845,1199033..1199047)
Aus0004_ph2	F	93-base 3′	**CATGCATCTATTTTACTTTATCAAG GAGTAAATATACTTAGCGTATCGAA ACGTTTAGGACATAGCAGTTTAGAA ACTACAATGTCTACTTAT**	(1454034..1454126,1494228..1494320)
DO_ph2	G	27-base 3′	**TCTATTCTTCTTCTTCCGCCATGAAT G**	(2072323..2072349,2087130..2087156)
Aus0004_ph4	G	22-base 3′	**ATGGCATACAATATGGCATACA**	(2159576..2159597,2174179..2174200)
1231501_ph2	G	22-base 3′	**ATTGTATGCCAT**	(241734..241745, 255551..255562)
E2883_ph2	G	27-base 3′	**TCTATTCTTCTTCTTCCGCCATGAAT G**	(1735348..1735374,1750156..1750182)
E1644_ph3	G	27-base 3′	**TCTATTCTTCTTCTTCCGCCATGAAT G**	(1961837..1961863,1976645..1976671)

6.2.8 Identification of *E. faecium* phage cargo genes

Previous reports of *Siphoviridae* from low G+C Gram-positive bacteria have revealed that there is frequent carriage of cargo genes in converting phages. These genes are commonly located at the distal 3′-end of the phage genomes of staphylococci, lactococci and listeria (Canchaya, Proux et al. 2003, Brussow, Canchaya et al. 2004). To identify potential cargo genes of *E. faecium* phages, BLASTP was used to identify genes located distal to the holing gene and prior to the 3′ attachment site sequence repeat (attR).

With seven prophages there were identifiable no genes located beyond the holin gene to influence host fitness or virulence (Table 6.5). Phage cargo genes were identified in the remaining 27 *E. faecium* phages encoded multiple distinct hypothetical proteins, tRNA, transposase, cold shock protein (CspC) and an integrase core domain protein (Figure 6.13). Cluster G prophages encode a VirE (virulence-associated protein E) domain protein that is also encoded in *E. faecalis* temperate phages. This gene is located in centrally in these prophage genomes indicate a role in replication and might have no association with *E. faecium* virulence.

Table 6.5: Cargo genes in converting prophages of *E. faecium*.

Prophage	Cargo genes encode
E0045_ph1 **E1050-ph1** **Com12-ph1** **DO_ph2** **Aus0004_ph4** **1,231,501** **E2883_ph2** **E1644_ph3**	No lysogenic conversion
E1007-ph1 **E1622_ph1** **E1185-ph1** **E1630_ph1** **E2039_ph1**	Hypothetical proteins and cold shock protein (*cspC*)
E1644_ph2 **E1972_ph1** **E1623_ph1** **E1623_ph2** **E1578_ph1** **Com15_ph1** **E1573_ph1**	tRNA _met and hypothetical protein
E429_ph1 **E1622_ph2**	Hypothetical proteins
E2883_ph1 **E1644_ph1** **Aus0004_ph2**	tRNA _met, transposase, integrase core domain protein and hypothetical protein
E1574_ph1	Hypothetical proteins, tRNA _met, transposase, cold shock protein *(cspC)*, transcriptional regulator *ygaV*, molecular chaperone Hsp31 and glyoxalase 3, NAD dependent epimerase/dehydratase family protein 3-demethyl ubiquinone-9-3 methyltransferase and TraX protein.
Aus0085_ph3	N-acetylmuramoly_L_alanine amidase , tRNA _met, transposase, and hypothetical proteins

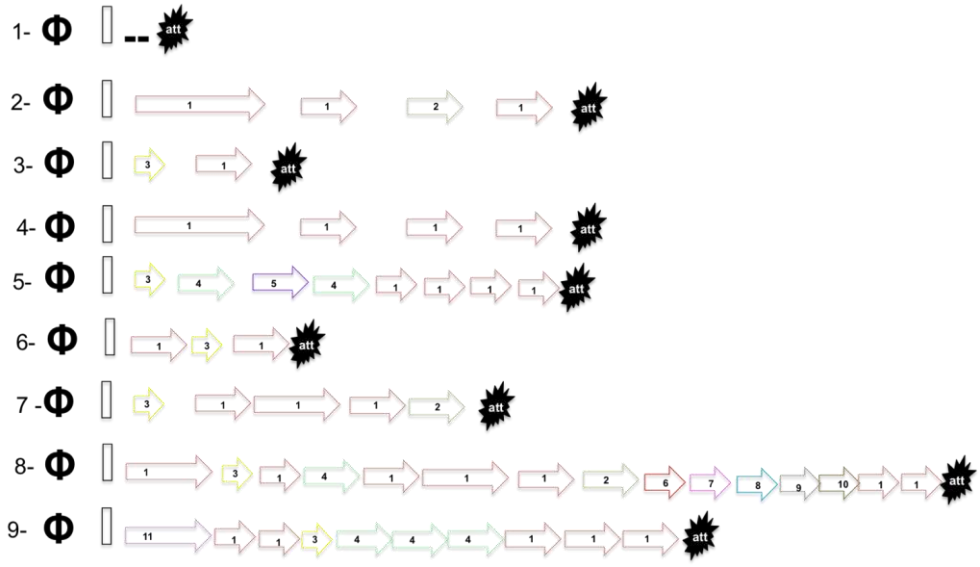


Figure 6.13: Cargo genes in converting prophages of *E. faecium.* Model 1 indicates no lysogenic conversion. The arrow numbers indicate (1) hypothetical protein; (2) cold shocked protein *cspc* (3) tRNA-met; (4) transposase; (5) integrase core domain protein; (6) transcriptional regulator *ygaV;* (7) molecular chaperone Hsp31 and glyoxalase 3; (8) NAD dependent epimerase/dehydratase family protein; (9) 3-demethyl ubiquinone-9-3 methyltransferase; (10) TraX protein; (11) N-acetylmuramolyL-alanine amidase.

Hsp31 encoded in E1574_ph1 by *hchA* is known as a heat-inducible molecular chaperone in *E. coli* (Subedi, Choi *et al*. 2011). Sastry *et al* (2002) indicated that Hsp31 is a homodimeric molecular chaperone that is conserved in pathogenic eubacteria and fungi. The translin-associated factor X (TRAX) protein plays roles in key cellular processes, such as DNA recombination and spatial and temporal expression of mRNA, and in siRNA processing (Gupta and Kumar 2012). Five *E. faecium* prophages contained

insertion sequence elements such as ISEfa8, ISEnfa3 and IS30 family plus a transposases, which might influence host fitness or virulence by altering gene expression or directing recombination or contribute to mobilisation.

6.2.9 *E. faecium* cryptic phage

Eleven of the 39 *E. faecium* host genomes that contain prophage used in the comparison are polylysogens, which harbour multiple prophages and phage-like elements (cryptic phage). For example the described chicken *E. faecium* genome harbours six prophage regions including three intact prophages and 2 cryptic phages (Figure 5.14). Five of the eleven phage are small in comparison with the other *E. faecium* phages (17.2 kb to 33.3 kb) with an average G + C content of 34.21% to 43.07%. The genomes of all the cryptic phages encoded a total of 12 to 105 ORFs (Table 6.6). These cryptic phages have significant sequence similarity to *E. faecalis*, *Lactobacillus*, *Lactococcus* and *Listeria* phages.

The cryptic phages encode between 2 to 5 functional phage proteins. The presence of lysogeny, packaging, morphology and lysis modules vary considerably. All cryptic prophages lack replication genes (Table 6.7). For example, head and tail morphogenesis modules essential for capsid formation as well as genes involved in packaging and lysis exist in the genome of E1574_cp1, for example. However, it lacks an integrase which suggests that this might represent a remnant phage. In addition, as further examples only head morphogenesis and portal genes are present in E429_cp1 and E429_cp2, while head and tail morphogenesis modules are

present in E0120_cp1. Phage E0120_cp2 and E1573_cp1 encode integrase, Cro repressor, head and lysis proteins. Functional incomplete life cycle gene sets suggest that these phage are either defective or belong to phage-related chromosomal islands (PRCIs) predicted previously in Gram-positive bacteria and recently reported in *E. faecalis* by Matos *et al* (2013).

Phage-like element associated genes (found within cryptic phage regions) could play a role for improve the fitness or the virulence of the bacteria. N-acetylmuramoyl-L-alanine amidase, which is an enzyme from the family of cell wall hydrolases was encoded by E0120_cp1 phage and E1972_cp1; a choloylglycine hydrolase family gene was present in E1972_cp1; an ATP-binding cassette transporter was encoded by E1573_cp1 and E1972_cp1; a transposase and cold shock protein were encoded by E0120_cp2; envelope glycoprotein, copper chaperone, serine protease, IS5 transposase were found in chicken cryptic phages E429_cp1 and E429_cp2 and CRISPR-associated protein Csn1 family gene was present in E429_cp2 (Supplemental File, S8).

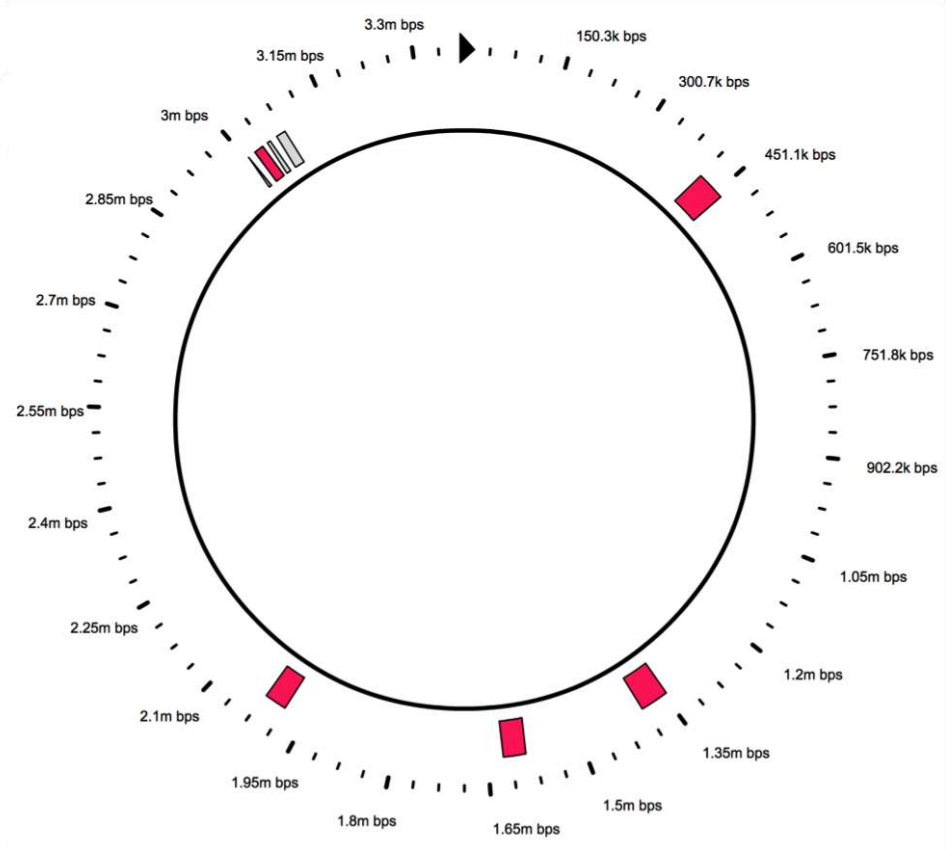


Figure 6.14: Genome of *E. faecium* isolated from chicken (E429). The presence of prophage and cryptic phages are indicated in the genome with red blocks indicating the genome of prophages and grey indicating the genomes of cryptic phage.

Table 6.6: Genometrics of cryptic phage related sequences of *E. faecium.* Seven cryptic phage genomes were identified in 5 strains of *E. faecium*.

Cryptic phage	Size (kb)	CDS	Region position	GC%
E1020_cp1	26	30	217194-243199	43.07
E1020_cp2	18.5	15	846146-864738	34.21
E1573_cp1	30.6	19	345381-376039	37.64
E1574_cp1	18.7	12	801318-820051	35.41
E1972_cp1	33.3	41	602973-636351	38.13
E429_cp1	17.2	21	3023847-3041052	44.7
E429_cp2	71.7	105	3009148-3080914	44

Table 6.7: Predicted phage life-cycle functions present in *E. faecium* cryptic phages.

Cryptic phage	Repressor	Intergrase	Terminases	Portal	Head	Tail	lysis
E1020_cp1	-	-	+	-	+	+	-
E1020_cp2	+	+	-	-	-	+	+
E1573_cp1	+	+	-	-	-	+	+
E1574_cp1	-	-	+	+	+	+	+
E1972_cp1	-	+	-	-	-	+	+
E429_cp1	-	-	-	+	+	-	-
E429_cp2	-	-	-	+	+	-	-

6.3 Discussion

6.3.1 Bacteriophage of animal *E. faecium* strains

Bacteriophages are capable of transferring virulence and antimicrobial resistance genes to new hosts through generalised transduction (Yasmin, Kenny *et al*. 2010, Mazaheri Nezhad Fard, Barton *et al*. 2011). Temperate phages that infect *E. faecium* and *E. faecium* prophages have been described in several of studies but there has been no intensive characterisation using comparative genomics approaches (Mazaheri Nezhad Fard, Barton *et al*. 2010, van Schaik, Top *et al*. 2010, Galloway-Pena, Roh *et al*. 2012).

The presence of inducible temperate bacteriophages in the genomes of three animal strains of *E. faecium* isolates (chicken (E429), calf (E172) and pig (E142)) were first examined by subjecting them to physical and chemical inducing agents. Plaque assays were performed using the induced culture supernatants and determined that the chicken (E429) and calf (E172) isolates could be induced into lytic cycle. The failure to identify released bacteriophage from strain E142 (pig) could be due to the indicator strains

being lysogens and thus immune or they might not have the receptors for prophage binding. *In silico* analysis of the E142 prophage genome suggested that the failure to identify released bacteriophage from strain E142 was more likely due to the absence of a packaging module.

Genomic diversity in *E. faecium* was reported to be related to the phage and phage-like sequences present in the accessory genome of *E. faecium* strains (van Schaik, Top *et al*. 2010). This previous study reported the CDS of expected phage origins contribute between 2.3% (E1071) and 5.2% (E980) of the total number of CDS of a genome. In the study of the three animal isolates presented here CDS of predicted phage origins makeup between 2.5% (E429), 0.7% (E172) and 0.68% (E142) of the total number of CDS of the genome of the three animal strains, which is lower for strain E172 and E142 then the previous report. In addition to the three potential prophage regions in the chicken strain it was found that > 88.9 kb of phage-related genes were also identified in the chromosome, mostly at the end of the genome, assembly indicating that the genome may not be ordered correctly (Figure 6.14).

Two and three phages have been described in the closed genomes of *E. faecium* TX16 (DO) and Aus0004, respectively. Both phages resident in the DO strain have similarity with likely prophage regions within most clinical strains as well as some commensal E. *faecium* strains. The prophages present in strain Aus0004 share high similarity with phages presented not only in many other enterococci, but also other low G+C Firmicutes such as

Clostridium, *Listeria*, *Lactobacillus* and *Staphylococcus* (Lam, Seemann *et al*. 2012, Qin, Galloway-Pena *et al*. 2012). Similarly phage and phage-related regions of the three sequenced animal strains share high similarity with prophage of *Enterococcus*, *Lactococcus, Staphylococcus* and *Listeria* species.

The prophages isolated from the sequenced chicken and calf strains were tested for their ability to package and transduce chromosomal and extra-chromosomal DNA. Transduction was effectively achieved for a plasmid encoded tetracycline gene with several phages. Yasmin *et al* (2010) demonstrated that *E. faecalis* bacteriophage could mobilise plasmid and chromosomal antibiotic resistance genes. The study here supports that this role could also apply to *E. faecium* phages and further studies will be needed to determine the extent that transduction contributes to lateral gene transfer relative to well-described roles of plasmid-encoded conjugation mechanisms.

6.3.2 Comparative genomic analysis *E. faecium* prophage

6.3.2.1 General features of *E. faecium* prophage genomes

In silico analysis was applied to identify 56 *E. faecium* prophage from 39 strains on the basis that their sequences contained both integrase and lysin genes. These *E. faecium* prophage genomes comprised between 17 to 72 ORFs and their size ranged from 13.9 to 55.1 kb with 35% to 37.9% average G + C content (Table 6.3).

The organisation of *E. faecium* prophage is very comparable and the protein coding sequences form equivalent functional clusters similar to temperate bacteriophages of *E. faecalis* (Yasmin, Kenny *et al*. 2010). The majority of ORFs presented in the *E. faecium* prophage genomes were transcribed in one direction, whereas the lysogeny module was generally transcribed in the opposite direction.

Phage classification is more complicated since there is no single gene that exists in all phages upon which a general scheme could be based. As a result, several research groups have suggested different classification schemes for the taxonomy of these viruses (Adriaenssens, Edwards *et al*. 2014). One approach established by Rohwer and Edwards (2002) using a grouping of completely sequenced phages is to draw a phage proteomic tree based on protein distances. Another approach is produced by the documentation of mechanisms leading to the connection between groups of phages. This scheme was used for classification based on shared genes in which each phage is characterised by its membership to a set of clusters (Lima-Mendez, Van Helden *et al*. 2008).

Using protein sequence of the overall gene content of *E. faecium* prophage genomes and comparative genomics to identify clusters, the prophage genome were assigned to 8 different clusters which share a very low degree of DNA identity (Figure 6.12). However, the protein sequences within clusters are highly conserved (Figure 6.6). Comparative analysis of 8 induced *E. faecalis* temperate phage identified by Yasmin *et al* (2010)

revealed four different phage groups (ΦFL1, ΦFL2, ΦFL3, and ΦFL4) and more than 97% sequence identity within three phage groups (ΦFL1A to C, ΦFL2A and B, and ΦFL3A and B). Two groups, ΦFL1 and ΦFL2 share a high degree of DNA identity (87 to 88%), which is spread throughout their genome. The major difference between these groups exists in in the region between the transcribed clusters of genes with putative functions in DNA replication and packaging. This region contains different genes encoding proteins with high levels of sequence identity to those encoded by the EF_1417-EF1489 (phage03) and EF_2084-EF_2145 (phage05) regions of the *E. faecalis* V583 genome sequence (Lepage, Brinster *et al*. 2006). The V583 phage03 and phage05 regions seem to be complete prophages, suggesting that hybrid phage genomes in *E. faecalis* were generated by recombination. The chromosome of V583 has seven prophage-like elements (V583-pp1 to V583-pp7). In addition, one prophage (pp2) is found as a part of the core genome of *E. faecalis* isolates (Matos, Lapaque *et al*. 2013). Remarkably, *E. faecalis* polylysogeny has been described in a collection of clinical isolates, which carried up to 5 different inducible phages (Yasmin, Kenny *et al*. 2010).

Protein similarities between the temperate *E. faecium* prophages suggested a low degree of similarity between the genomes at the nucleotide level (Figure 6.12). The results of pairwise DNA alignments revealed only very small regions of nucleotide identity. This indicates that each *E. faecium* phage type represents possibly novel DNA, consequently lysogeny is driving the genomic diversity of their host strains (van Schaik, Top *et al*. 2010).

In contrast, within clusters that define the *E. faecium* prophage types there is very high similarity, and yet the H cluster prophage, is clearly a distant relative (Figure 6.6). A possible explanation is this cluster has recently acquired the ability to infect *E. faecium*. It remains to be seen if other prophage genomes that are distinct from the *E. faecium* prophage types revealed here are isolated in the future, which will allow grater analysis of phage diversity and evolution.

The major similarity between the 8 prophage clusters is within hypothetical phage proteins that are located in the rightmost (3') region of the genomes. Juhala *et al* (2000) indicated that Siphoviridae show strong conservation of the order of virion structure and assembly genes and highlighted a lack of horizontal exchange between the groups of structural genes. Comparative genome analysis of the *E. faecium* prophages using the PHAST database identified that the *E. faecium* prophages share high similarity with segments of *Listeria, Lactobacillus, Enterococcus* and *Lactococcus* prophages. This sequence identity is confined mostly to the morphogenesis and lysis modules (Supplemental File, S4). Analyses performed by Villion *et al* (2009) revealed that the virulent lactococcal phage encodes a morphogenesis module that is similar to the *E. faecalis* V583 prophage and considered that recombination could happen between phages infecting these low G+C bacteria. This observation was supported by Yasmin *et al* (2010) when they reported identities between prophages of lactococci and *E. faecalis.* The comparative analysis of *E. faecium* prophage lends further support to this hypothesis of intergeneric exchange and shows that this has

occurred between multiple different phage types and bacterial species plus there is likely to be a flux of genes also between enterococcal species.

6.3.2.2 Functional module of *E. faecium* prophages

The identified *E. faecium* prophages show genetic functionality necessary for integration/excision, DNA replication and capsid/tail morphogenesis to produce functional virions. The first unit of the phage (i.e. as it appears on the host chromosome) is the integrase region, which is typically leftward transcribed and it is necessary for phage genome integration and excision from the bacterial chromosome during its temperate life cycle. Site-specific recombination between DNA sequences corresponding to the phage attachment site (attP) and the bacterial attachment site (attB) are mediated by phage integrase enzymes (Groth and Calos 2004). Enterococcal bacteriophage integrase was previously indicated to present a site-specific recombination amongst a phage attachment site (*attP*) and a host attachment site (*attB*) in its host, following two new hybrid sites, *attL* and *attR*. The *att* sites typically contain a core sequence, which is short between 2 bp to >10 bp and it is same between all the *att* sites in the identical phage system. The core sequence identifies and bind regions that integrases or accessory factors (Groth and Calos 2004, Park, Lim *et al*. 2007).

The putative integrases of the 56 prophages within the 8 phage types belong to the tyrosine integrase recombinase family and possess near identical amino acid sequences (Figure 6.7). The tyrosine recombinase family is common in *Streptococcus suis* prophages (Tang, Bossers *et al*. 2013),

Mycobacteriophage (Hatfull, Jacobs-Sera *et al*. 2010) *Listeria* prophages (Groth and Calos 2004) and *Staphylococcus aureus* (Goerke, Pantucek *et al*. 2009). However, the integrases of *E. faecalis* were reported to be serine recombinase family members (Yasmin, Kenny *et al*. 2010). Hirano *et al* (2011) indicated that integrases could use other accessory proteins such as recombination directionality factors and mediate prophage integration and excision. Based upon a cladogram tree of *E. faecium* prophage integrases, the clusters corresponding to phage types A-J clearly have distinct integrases sequences (Figure 6.7).

Terminase is an enzyme necessary for the packaging of dsDNA into the progeny phages (Kutter and Sulakvelidze 2005). The packaging modules identified in most of the *E. faecium* phage genomes here are principally comprised of three genes, encoding the small and large subunits of terminase and the portal protein. Terminases are responsible for the identification of their phage DNAs, ATP-dependent cleavage of the DNA concatemer and packaging of the DNA molecules into the blank capsid shells over the portal protein (Fujisawa and Minagawa 1986). Amino acid sequences alignments of the terminases large subunit, showed that the terminases of most of *E. faecium* prophages appeared to be highly conserved across prophage types clusters. The large terminase subunits of animal *E. faecium* phage including chicken E429 and E0045, a dog E4452 and mouse genome E1622 are share similarity with each other (Figure 6.8). Most of the animal *E. faecium* prophages appear to possess unique lysogeny and packaging modules, suggesting that their lifecycle in their animal host

strain needs a specific phage functional module. The portal gene was absent in nineteen *E. faecium* prophages and the reason for this is unclear. If these phages are capable of entering the lytic lifecycle they would need functional complementation by another portal protein. The eight temperate phages identified in *E. faecalis* as being inducible into the lytic lifecycle each contain putative terminase and portal protein functions, consistent with capsid packaging of DNA being achieved using the head-full mechanism (Yasmin, Kenny *et al*. 2010) and a similar packaging mechanism can be inferred for most of the phage sequence types A-F, H.

Major and minor head proteins and the scaffold protein are significant structural factors absolutely required for morphogenesis of the icosahedral capsid. Base plate and tail fibers are variable components of the tail tip that facilitate adhesion to the bacterial host surface and enzymatic degradation of the peptidoglycan (Kutter and Sulakvelidze 2005). In all *E. faecium* prophages identified here the head morphogenesis and tail structure proteins were identified and the tail represents the largest module. The major capsid and tail proteins of the *E. faecium* prophage shared high level sequence identity with proteins of *Listeria*, *Lactobacillus*, *Staphylococcus*, *Paenibacillus*, *Mycobacterium*, *Enterococcus* and *Lactococcus* bacteriophages (Supplemental File, S4).

E. faecium prophage tail proteins indicate clear differences between the prophage clusters (Figure 6.9) and the tail gene size in ranges from 2.5 kb to 6.4 kb. The bacteriophage tail is used to identify a suitable host and ensure

effective genome delivery to the cell cytoplasm. Tail morphology has been used previously as the basis for the classification of Caudovirales phages. Three different families of Caudovirales were identified according to their tail morphology, *Myoviridae* have a complex contractile tail (e.g., T4 and Mu); the *Podoviridae* have a short noncontractile tail (e.g., P22 and T7); and the *Siphoviridae*, characterized by their long noncontractile tail (e.g., lactococcal phages) (Veesler and Cambillau 2011, Fokine and Rossmann 2014). Genome sequences are not sufficient to definitively classify *E. faecium* prophage as *Siphoviridae* using electron microscopy will be required for confirmation.

The activity of endolysin and holin are significant factors for progeny phages to disrupt the host cell at the end of the lytic cycle (Bernhardt, Wang *et al*. 2002). The products of the holin and endolysin genes typically perform the fundamental functions of the lysis module of temperate bacteriophages. The small holins accumulate in the membrane and at the end of the lytic cycle from pores that permeabilise the membrane, while the endolysin molecules accumulate at the cytosol until the pores are produced to reach the cell wall, where they hydrolyse peptidoglycan (Wang, Smith *et al*. 2000). Three classes of holin can be defined according to their number of potential transmembrane domains. Class I, II and III members can form three, two and one transmembrane domains, respectively (Wang, Smith *et al*. 2000). Holin-endolysin system are typically used by bacteriophages with large genomes, while a single lysis protein is commonly used by bacteriophages with small genomes (Bernhardt, Wang *et al*. 2002).

The majority of the lysis modules in the identified *E. faecium* prophages comprise one holin. However, prophages 1,141,733_ph1, E4425_ph1 and E172_ph1 also contain endolysin genes and lysis gene is absent in the prophages that forming cluster G. Most of these holins have homology with holin found in *E. faecalis* temperate bacteriophages (Supplemental File, S4). Phage holins that form clade Holin1 (Figure 6.10) have homology with holins of *Lactococcus* phage ul36 and *E. faecalis* phiFL4A and phiEf11. The high level of conservation indicates recombination might occur between *E. faecium* prophage and these species or they share a common ancestor. The location of the holin gene is within a region that is known to be influenced extensively by horizontal gene transfer. Fokine *et al* (2014) stated that the mosaic boundaries of prophage that are seen in pairwise comparisons of genomes are taken to be the locations of illegitimate (non-homologous) recombination in their ancestry.

The Cladogram trees of the functional module of *E. faecium* prophages has great genome rearrangement. Prophage form cluster G share similarity in most of the structural genes with *Enterococcus faecalis* phage (EFRM31). Aus0004_ph2 share similarity in DNA packaging/ head and tail morphogenesis module with *Listeria* phage 2389. While Aus0004_ph3 share similarity in DNA packaging/ head and tail morphogenesis module with *Listeria* phage 2389 and the lysis with *E. faecalis* (EF62phi). This suggested that prophage genomics analysis might present recombinant phages combining structural genes from different phage families as seen.

Recombination in phage genomes is not rare; it was also presented in Gram-negative bacteria *Salmonella*, *Shigella Flexneri* and *Pseudomonas aeruginosa* phage and plant pathogen *Xylella fastidiosa* phage that used in Canchaya *et al* (2003) study. *Pseudomonas aeruginosa* phage contains of a P2-like tail gene of *Myoviridae* cluster separated by a lysis cassette from a lambda-like tail gene cassette. However, Shinomiya (1984) stated that superinfecting *Pseudomonas* phage PS17 presented phenotypic mixing with pyocin R2, consequentially stretched the host range for PS17, but genetic recombination was not detected and this might be due to natural or engineered phage resistance mechanisms. In addition, Durmaz et al (2000) identified that several lactococcal phages can be escaped from regulate by swapping part of their genome with DNA from prophages or prophage remnants, which they encountered in the infected cell. These explanations obviously establish that prophage DNA is the raw material for both phage and bacterial evolution.

6.3.2.3 *E. faecium* prophage genome diversity

E. faecium prophage genomics supported the hypotheses of the modular theory of phage evolution. According to Botstein (1980) phage genomes are groups of functionally related genes (mosaics of modules) that are able to recombine in genetic exchanges among distinct phages infecting the same cell. Juhala *et al* (2000) declared that recombination basically happens everywhere and the evident modular structure is instead the result of selection eliminating all genetic recombinations that do not lead to viable phage arrangements. Selection would also limit all recombinations that are

less competitive than the present phage types.

In silico analysis of the *E. faecium* prophage genomes suggested many of the prophages could be defective and apparently in a dynamic process of gradual decay. Genetic recombination between *E. faecium* phages can lead to new chimeric phage types. The leftmost regions that contain the structure and assembly genes show grater conservation than the rightmost genomic segments in *E. faecium* prophages (Figure 6.6). It is important to notice that the degree of *E. faecium* prophages type diversity does not only reflect the number of genomes present. Based on the protein alignment analysis of the main structural genes in the prophage genomes (integrases, terminase large subunit, tail protein and holin), high diversity in the protein sequence of these structure genes was found among the *E. faecium* prophages.

Multiple unique genes were also found in *E. faecium* prophages. Unique genes in each cluster, including genes that belong to phage structure, were identified when one prophage of each cluster was aligned (Figure 6.12). Each of the clusters comprises a minimum of 20% of cluster-specific genes; prophage genome-specific genes cluster H shows no obvious relationship with any of the other clusters.

6.3.4 *E. faecium* prophage cargo

Many temperate phages integrated into the genome of bacterial pathogens encode genes associated with virulence phenotypes such as intracellular survival, invasion and toxin production, which are not essential for phage

viability (Perkins, Kingsley *et al*. 2009). Cargo regions in low G+C Firmcutes phages are characteristically located at the end of the phage opposite from integrase (Bobay, Touchon *et al*. 2013). Investigation of *E. faecium* prophage cargo regions indicate that 19 0f the 26 prophages regions contain potential lysogenic conversion genes. However, the analysis of *E. faecium* phage cargo was based on draft sequence assembly, which may or may not be correct as missasemblies could cause cargo genes to associated with the wrong phage genes.

Notably, cold shock protein (CspC), tRNA, transposase and integrase core domain. These genes might influence host fitness or virulence, or contribute in the mobilisation of converting activities found in this terminal phage genome region (Yasmin, Kenny *et al*. 2010). Cold shock protein genes were also described as being encoded on prophages of *E. faecalis*. Their maintenance in several phage elements in both *E. faecium* and *E. faecalis* might indicated there is selection for their function in the life-cycle of their hosts and/or there is frequent recombination between phages of both species.

The role of IS elements and transposase in *E. faecium* virulence were described in several studies which suggested that these phage encoded elements could influence their host. Temperate phages can modulate bacterial fitness or virulence in at least three ways: introduction of fitness factors, gene disruption, and lysis-mediated competitiveness. The import of fitness factors (lysogenic conversion) presents new traits to the host by

offering genes that are not essential for the phage life cycle (Brussow, Canchaya *et al*. 2004).

Bailly-Bechet *et al* (2007) indicated that the main difference among phages with tRNAs and those without any tRNAs is the genome length: phages holding tRNAs are considerably longer (average lengths ~70 and ~30 kb, respectively). The first report that phages carry tRNA genes was made over 40 years ago in T4 phage (Weiss, Hsu *et al*. 1968). Extensive study of their role by Wilson (1973) identified that the deletion of these genes caused lower burst sizes and reduced protein synthesis. tRNAs also afford integration points for phages, plasmids, and pathogenicity islands. It was proposed that phage-encoded tRNAs could also be important for understanding the role of phages in bacterial evolution given that large eukaryotic viruses comprise other elements of the translation machinery, such as tRNA synthetases (Raoult, Audic *et al*. 2004).

Examination of the *E. faecium* phage genomes reveals a potential virulence gene present in a prophage from cluster G. Virulence-associated protein E (vapE) contributes to the type IV secretion pathway (Zhao, Sagulenko *et al*. 2001). VapE was first recognized in *Dichelobacter nodosus* and part of this protein was reported to be associated with virulence in *D. nodosus* (Bloomfield, Whittle *et al*. 1997). Recently, the mechanism by which VapE affects virulence has not yet been determined (Ma, Geng *et al*. 2013). The presence of an integrase gene (XerC) closely upstream of *vapE*, might link bacteriophages in the evolution and transfer of these bacterial virulence

elements in swine streptococcosis. Moreover, a *vapE*-like gene has also been identified in a pathogenicity island of *Staphylococcus aureus*, and in phages of *Vibrio parahaemolyticus* and *Streptococcus pneumoniae* and *Enterococcus faecalis* (Romero, Croucher *et al*. 2009, Yasmin, Kenny *et al*. 2010). The contribution of the *vapE* gene to the virulence of *Enterococcus* remains to be clarified.

The study of phage-encoded virulence factors among *E. faecium* strains is more limited compared with their description in several other low-GC Gram-positive pathogenic bacteria e.g. staphylococci. Nevertheless, the transducing abilities of the animal *E. faecium* prophages in chicken E429 and calf E172 genome together with the shared sequence homology with those infecting low-GC Gram-positive bacteria, hints at a potential role in the transfer of genetic information between different genera. This study also demonstrated that animal *E. faecium* prophages can transfer antibiotic resistance genes in enterococci such as tetracycline (*tetM*). Given that the *E. faecium* isolates used in this study were resistant to many antibiotics (Table 4.8) in observance with earlier reports (e.g. Klare, Konstabel *et al*. 2003), a large number of antibiotic resistance genes could potentially be mobilised by transduction.

6.3.5 Cryptic phage

Genome analysis of the *E. faecium* isolates identifies polylysogenic hosts. The phage-like elements are not likely to all be functional for the production of progeny without the existence of helper elements. Nevertheless, they do

contain multiple functional genes. Polylysogeny frequently leads to phenomena whereby prophage impact bacterial host behaviour (Wang, Kim *et al*. 2010, Matos, Lapaque *et al*. 2013). For example, Phage Related Chromosomal Islands (PRCIs) of several Gram-positive bacteria are mobile genetic elements, primarily defined as *S. aureus* pathogenicity islands (SaPIs) (Matos, Lapaque *et al*. 2013). Infection by a helper phage or by induction of an endogenous prophage drives excision of SaPIs from the bacterial chromosome (Ubeda, Maiques *et al*. 2008).

The cryptic phages in the genomes of the animal *E. faecium* strains might also function as helper phage and thereby contribute to fitness or pathogenic traits. For example, genes located on cryptic phage (E429_cp2) encode function such as hydrolase, transposase, IS5 and copper chaperone. Interestingly, genes that are known as an immune mechanism against phage (CRISPR-associated protein Csn1 family) are also encoded by this cryptic phage for example.

Complex interactions between V583 *E. faecalis* phages were described by Matos *et al* (2013). Three levels of phage interactions were identified: phage-related chromosomal island can hijacks other phage capsids and interferes with infectivity; phages can utilise a temperature-dependent inhibition of other phage excisions; finally, phage can block excision of others phages. Further studies will be needed to determine the extent of interactions between *E. faecium* prophages and cryptic phages.

Chapter Seven: Conclusions and Future Work.

7.1 Conclusions

This study aimed to generate, collate and interpret information from the genome sequencing of *E. faecium* to answer several key questions. Firstly, are strains from animals very different from human isolates and have they acquired genes specific for colonising an animal host? Secondly, which mobile genetic determinants are present in animal strains of *E. faecium* and are these common to or distinct from those in human isolates?

The data presented here from phylogenomics analyses reveals discrimination of isolates into clades, which broadly grouped strains of animal and human origin. Identification of genes specific for host colonisation remains unresolved although genes pertaining to particular clades were identified and these could be further characterised to examine their role in colonisation.

This study has described sequencing, assembly, annotation and homology of three animal strains of *E. faecium* isolated from chicken, calf and pig. Two types of sequencing methods were used to complete the genomes of the animal isolates; 454 sequencing platform with PCR amplification attempt gap closure in the genome of chicken *E. faecium*; and PacBio sequencing which generated a near complete genome of *E. faecium* isolated from calf.

Comparative analysis of animal and human isolates of *E. faecium* demonstrated that *E. faecium* species share the same core genome. However, in strains that are relatively closely related the presence and

absence of mobile genetic elements is the major influence in shaping strain-specific properties. Relationship analysis using all the publicly available genomes indicates a pronounced separation of isolates into community, hospital and animal-associated clades, supporting previous studies. In addition, it was evident that strains of *E. faecium* isolated from different sub-populations including the Clonal Complex 17, clinical, commensal and animals, including bird, pig and dog sub-groups were related to each other and mostly grouped in same clade in the phylogenetic tree, but with some exceptions. Notably, most *E. faecium* strains isolated from the same geographic region or infection source were grouped together.

Plasmids, IS, transposons and prophages are abundant in most *E. faecium* isolates. IS elements are the most noticeable group of genes enriched in all CC17 strains and the majority of hospital-associated strains. Animal and clinical *E. faecium* isolates share multiple IS elements, for example the IS3 and IS256 families were most frequent in animal strains, although these elements were also present within the hospital clade. In this study, a mega plasmid was identified in the genomes of the sequenced chicken, calf and pig *E. faecium* isolates, which is specific to these strains. A second mega plasmid identified in the sequenced chicken and pig genomes was also present in the humans isolate genomes. Comparative genomic analyses were applied to 56 prophage identified from 39 *E. faecium* strains retrieved on the basis that their sequences contained both integrase and lysin genes. The prophages were discriminated into eight different sequence types A to H. The majority of the prophages in clusters A and C are from commensal and

animal isolates. Cluster B and D sequences are mixed clusters that contain prophages isolated from clinical, commensal, animal and river water sources while most of those from cluster F are present in clinical isolates including.

The association of IS and prophages with genomic islands (GIs) and novel regions in the genome maps likely reflects horizontal transfer of these genes between different species, since these elements had considerable homology with both Gram-negative genera, including *Escherichia, Burkholderia, Pseudomonas* and *Xanthomonas* species, and Gram-positive genera, including *Staphylococcus, Streptococcus, Bacillus, Listeria, Lactococcus, Lactobacillus* and *Paenibacillus* species. Several of these mobile elements were unique to the animal strains sequenced as the main body of this study.

7.2 Future work

Short sequence reads and the assembly of complex genomes such as those of *E. faecium* remains a challenge. Most commonly, the high frequency of repeat sequences add additional complexity and as observed with strains studied here they confounded assembly. Repeated sequences of DNA bring difficulties when attempting to infer relative locations in the genome corresponding to reads, and it is suggested they happen far more often in real genomes than they would in a sequence of independently randomly produced bases (Henson, Tischler *et al*. 2012). These well-described problems are additional to correcting read errors and considering heterozygosity, while staying within the limits of practical computability,

thereby making assembly more difficult and complex (Henson, Tischler *et al*. 2012). Genomic rearrangements due to repeat sequences increase the complexity of the *E. faecium* genome (Ferrarini, Moretto *et al*. 2013). Accordingly, the 454 sequencing platform combined with *de novo* assembly approaches fail to completely resolve assembly of the animal *E. faecium* genomes.

Network assembly processes will be required for future study of animal *E. faecium* and the species more broadly. Mismatches between the *in vitro* and *in silico* analysis of mobile genetic elements and the integration of the mega plasmid into the precise assembly of the multiple bacteriophages present in the chicken *E. feacium* genome require further study to be fully explained. The basis for the mega plasmid integration into the chicken *E. feacium* chromosome needs to be explored to rule out potential errors in genome assembly.

Many assembly issues would be resolved with further use of the Pacific Biosciences RS (PacBio) platform, which was successfully applied here to sequence the *E. faecium* calf strain E172. The PacBio long-read sequencing platform provides advantages for assembly of this species, due to increased read length and equitable genome coverage making it possible to assemble genome sequence data with few or no gaps by generating longer contigs (Ferrarini, Moretto *et al*. 2013).

Several reported phylogenomic studies using limited number of *E. faecium* genomes, supported an initial report of a primary phylogenetic split in the *E. faecium* population, which separates human commensal isolates as a clade distinct from animal and human clinical isolates in a separate clade (Galloway-Pena, Roh *et al*. 2012, Palmer, Godfrey *et al*. 2012). In the study presented here, it was found that nosocomial *E. faecium* strains are clustered into two subgroups instead of one. Animal *E. faecium* isolates were discriminated into one subgroup that contain a small number of nosocomial *E. faecium* strains, suggesting different evolutionary traits for emergent clinical and animal isolates, and these findings support those reported by Willems (2012) and Lebreton *et al* (2013).

The study of MGE is challenging since there are many complications with annotating MGE sequences and therefore as a whole they are poorly annotated, particularly as part of bacterial-genome sequencing projects. For example, few phages have previously been well characterised in *E. faecium* and only recently one complete phage genome (IME-EFm1) was reported (Wang, Wang *et al*. 2014). The narrow sequence homology among functionally equivalent phage-encoded proteins complicates the study of their function (Pedulla, Ford *et al*. 2003). There is a requirement for developments in bioinformatics of MGEs to identify their unique features.

Pathway analysis to generate effective metabolism reconstructions remains incomplete due to a lack of knowledge. These gaps include carbohydrate utilisation. A genome scale construction of animal, clinical and commensal

E. faecium metabolism would allow examination of several challenging research questions about niche specialisation. Moreover, properties such as pathway redundancy and growth burden of pathways contributing to colonisation and virulence. Study of carbohydrate utilisation in animal and human *E. faecium* will help to determine the carbohydrates required for host colonisation, their relative utilisation and contribution to host adaptation.

References

Acinas, S. G., L. A. Marcelino, V. Klepac-Ceraj and M. F. Polz (2004). "Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons." J Bacteriol **186**(9): 2629-2635.

Adriaenssens, E. M., R. Edwards, J. H. E. Nash, P. Mahadevan, D. Seto, H.-W. Ackermann, R. Lavigne and A. M. Kropinski (2014). "Integration of genomic and proteomic analyses in the classification of the Siphoviridae family." Virology.

al Jeshi, A. (1999). "Moclobemide response in obsessive-compulsive disorder." Can J Psychiatry **44**(3): 285.

Al-Faleh, F. Z., M. Al-Jeffri, S. Ramia, R. Al-Rashed, M. Arif, M. Rezeig, I. Al-Toraif, M. Bakhsh, A. Mishkkhas, O. Makki, H. Al-Freihi, S. Mirdad, A. AlJuma, T. Yasin, A. Al-Swailem and A. Ayoola (1999). "Seroepidemiology of hepatitis B virus infection in Saudi children 8 years after a mass hepatitis B vaccination programme." J Infect **38**(3): 167-170.

Alcaraz, L. D., G. Moreno-Hagelsieb, L. E. Eguiarte, V. Souza, L. Herrera-Estrella and G. Olmedo (2010). "Understanding the evolutionary relationships and major traits of Bacillus through comparative genomics." BMC Genomics **11**: 332.

Allen, T. E., N. D. Price, A. R. Joyce and B. O. Palsson (2006). "Long-range periodic patterns in microbial genomes indicate significant multi-scale chromosomal organization." PLoS Comput Biol **2**(1): e2.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Alvarez-Elcoro, S. and M. J. Enzler (1999). "The macrolides: erythromycin, clarithromycin, and azithromycin." Mayo Clin Proc **74**(6): 613-634.

Archimbaud, C., N. Shankar, C. Forestier, A. Baghdayan, M. S. Gilmore, F. Charbonne and B. Joly (2002). "In vitro adhesive properties and virulence factors of Enterococcusfaecalis strains." Res Microbiol **153**(2): 75-80.

Arias, C. A., D. Panesso, K. V. Singh, L. B. Rice and B. E. Murray (2009). "Cotransfer of antibiotic resistance genes and a hylEfm-containing virulence plasmid in Enterococcus faecium." Antimicrob Agents Chemother **53**(10): 4240-4246.

Arthur, M., P. Reynolds and P. Courvalin (1996). "Glycopeptide resistance in enterococci." Trends Microbiol **4**(10): 401-407.

Bager, F., F. M. Aarestrup, M. Madsen and H. C. Wegener (1999). "Glycopeptide resistance in Enterococcus faecium from broilers and pigs following discontinued use of avoparcin." Microb Drug Resist **5**(1): 53-56.

Bailly-Bechet, M., M. Vergassola and E. Rocha (2007). "Causes for the intriguing presence of tRNAs in phages." Genome Res **17**(10): 1486-1495.

Barrangou, R., C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero and P. Horvath (2007). "CRISPR provides acquired resistance against viruses in prokaryotes." Science **315**(5819): 1709-1712.

Barrett, S. J. and P. H. Sneath (1994). "A numerical phenotypic taxonomic study of the genus Neisseria." Microbiology **140 (Pt 10)**: 2867-2891.

Bayjanov, J. R., D. Molenaar, V. Tzeneva, R. J. Siezen and S. A. van Hijum (2012). "PhenoLink--a web-tool for linking phenotype to ~omics data for bacteria: application to gene-trait matching for Lactobacillus plantarum strains." BMC Genomics **13**: 170.

Bentley, S. D. and J. Parkhill (2004). "Comparative genomic structure of prokaryotes." Annu Rev Genet **38**: 771-792.

Bentorcha, F., G. De Cespedes and T. Horaud (1991). "Tetracycline resistance heterogeneity in Enterococcus faecium." Antimicrob Agents Chemother **35**(5): 808-812.

Bernhardt, T. G., I. N. Wang, D. K. Struck and R. Young (2002). "Breaking free: "protein antibiotics" and phage lysis." Res Microbiol **153**(8): 493-501.

Besemer, J. and M. Borodovsky (2005). "GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses." Nucleic Acids Res **33**(Web Server issue): W451-454.

Bessen, D. E. and A. Kalia (2002). "Genomic localization of a T serotype locus to a recombinatorial zone encoding extracellular matrix-binding proteins in Streptococcus pyogenes." Infect Immun **70**(3): 1159-1167.

Billstrom, H., B. Lund, A. Sullivan and C. E. Nord (2008). "Virulence and antimicrobial resistance in clinical Enterococcus faecium." Int J Antimicrob Agents **32**(5): 374-377.

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

Bloomfield, G. A., G. Whittle, M. B. McDonagh, M. E. Katz and B. F. Cheetham (1997). "Analysis of sequences flanking the vap regions of Dichelobacter nodosus: evidence for multiple integration events, a killer system, and a new genetic element." Microbiology **143 (Pt 2)**: 553-562.

Blot, M., J. Meyer and W. Arber (1991). "Bleomycin-resistance gene derived from the transposon Tn5 confers selective advantage to Escherichia coli K-12." Proc Natl Acad Sci U S A **88**(20): 9112-9116.

Bonten, M. J., R. Willems and R. A. Weinstein (2001). "Vancomycin-resistant enterococci: why are they here, and where do they come from?" Lancet Infect Dis **1**(5): 314-325.

Borgmann, S., D. M. Niklas, I. Klare, L. T. Zabel, P. Buchenau, I. B. Autenrieth and P. Heeg (2004). "Two episodes of vancomycin-resistant Enterococcus faecium outbreaks caused by two genetically different clones in a newborn intensive care unit." Int J Hyg Environ Health **207**(4): 386-389.

Boto, L. (2010). "Horizontal gene transfer in evolution: facts and challenges." Proc Biol Sci **277**(1683): 819-827.

Botstein, D. (1980). "A theory of modular evolution for bacteriophages." Ann N Y Acad Sci **354**: 484-490.

Boumghar-Bourtchai, L., A. Dhalluin, B. Malbruny, S. Galopin and R. Leclercq (2009). "Influence of recombination on development of mutational resistance to linezolid in Enterococcus faecalis JH2-2." Antimicrob Agents Chemother **53**(9): 4007-4009.

Bourgogne, A., D. A. Garsin, X. Qin, K. V. Singh, J. Sillanpaa, S. Yerrapragada, Y. Ding, S. Dugan-Rocha, C. Buhay, H. Shen, G. Chen, G. Williams, D. Muzny, A. Maadani, K. A. Fox, J. Gioia, L. Chen, Y. Shang, C. A. Arias, S. R. Nallapareddy, M. Zhao, V. P. Prakash, S. Chowdhury, H. Jiang, R. A. Gibbs, B. E. Murray, S. K. Highlander and G. M. Weinstock

(2008). "Large scale variation in Enterococcus faecalis illustrated by the genome analysis of strain OG1RF." Genome Biol **9**(7): R110.

Bozdogan, B., R. Leclercq, A. Lozniewski and M. Weber (1999). "Plasmid-mediated coresistance to streptogramins and vancomycin in Enterococcus faecium HM1032." Antimicrob Agents Chemother **43**(8): 2097-2098.

Breuner, A., L. Brondsted and K. Hammer (1999). "Novel organization of genes involved in prophage excision identified in the temperate lactococcal bacteriophage TP901-1." J Bacteriol **181**(23): 7291-7297.

Brussow, H., C. Canchaya and W. D. Hardt (2004). "Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion." Microbiol Mol Biol Rev **68**(3): 560-602, table of contents.

Budzik, J. M. and O. Schneewind (2006). "Pili prove pertinent to enterococcal endocarditis." J Clin Invest **116**(10): 2582-2584.

Burrus, V. and M. K. Waldor (2004). "Shaping bacterial genomes with integrative and conjugative elements." Res Microbiol **155**(5): 376-386.

Camargo, I. L., M. S. Gilmore and A. L. Darini (2006). "Multilocus sequence typing and analysis of putative virulence factors in vancomycin-resistant and vancomycin-sensitive Enterococcus faecium isolates from Brazil." Clin Microbiol Infect **12**(11): 1123-1130.

Canny, G. O. and B. A. McCormick (2008). "Bacteria in the intestine, helpful residents or enemies from within?" Infect Immun **76**(8): 3360-3373.

Carvalho Mda, G., A. G. Steigerwalt, R. E. Morey, P. L. Shewmaker, L. M. Teixeira and R. R. Facklam (2004). "Characterization of three new enterococcal species, Enterococcus sp. nov. CDC PNS-E1, Enterococcus sp.

nov. CDC PNS-E2, and Enterococcus sp. nov. CDC PNS-E3, isolated from human clinical specimens." J Clin Microbiol **42**(3): 1192-1198.

Casey, J., C. Daly and G. F. Fitzgerald (1991). "Chromosomal integration of plasmid DNA by homologous recombination in Enterococcus faecalis and Lactococcus lactis subsp. lactis hosts harboring Tn919." Appl Environ Microbiol **57**(9): 2677-2682.

Centers for Disease, C. and Prevention (1993). "Nosocomial enterococci resistant to vancomycin--United States, 1989-1993." MMWR Morb Mortal Wkly Rep **42**(30): 597-599.

Cheng, S., F. K. McCleskey, M. J. Gress, J. M. Petroziello, R. Liu, H. Namdari, K. Beninga, A. Salmen and V. G. DelVecchio (1997). "A PCR assay for identification of Enterococcus faecium." J Clin Microbiol **35**(5): 1248-1250.

Chibani-Chennoufi, S., M. L. Dillmann, L. Marvin-Guy, S. Rami-Shojaei and H. Brussow (2004). "Lactobacillus plantarum bacteriophage LP65: a new member of the SPO1-like genus of the family Myoviridae." J Bacteriol **186**(21): 7069-7083.

Chlebicki, M. P. and A. Kurup (2008). "Vancomycin-resistant enterococcus: a review from a Singapore perspective." Ann Acad Med Singapore **37**(10): 861-869.

Chopin, M. C., A. Chopin and E. Bidnenko (2005). "Phage abortive infection in lactococci: variations on a theme." Curr Opin Microbiol **8**(4): 473-479.

Chow, J. W. (2000). "Aminoglycoside resistance in enterococci." Clin Infect Dis **31**(2): 586-589.

Chu, V. T., R. Gottardo, A. E. Raftery, R. E. Bumgarner and K. Y. Yeung (2008). "MeV+R: using MeV as a graphical user interface for Bioconductor applications in microarray analysis." Genome Biol **9**(7): R118.

Clark, A. J., W. Inwood, T. Cloutier and T. S. Dhillon (2001). "Nucleotide sequence of coliphage HK620 and the evolution of lambdoid phages." J Mol Biol **311**(4): 657-679.

Clewell, D. B. (1990). "Movable genetic elements and antibiotic resistance in enterococci." Eur J Clin Microbiol Infect Dis **9**(2): 90-102.

Clewell, D. B., S. E. Flannagan and D. D. Jaworski (1995). "Unconstrained bacterial promiscuity: the Tn916-Tn1545 family of conjugative transposons." Trends Microbiol **3**(6): 229-236.

Coburn, P. S., C. M. Pillar, B. D. Jett, W. Haas and M. S. Gilmore (2004). "Enterococcus faecalis senses target cells and in response expresses cytolysin." Science **306**(5705): 2270-2272.

Cohan, F. M. (2001). "Bacterial species and speciation." Syst Biol **50**(4): 513-524.

Connor, N., J. Sikorski, A. P. Rooney, S. Kopac, A. F. Koeppel, A. Burger, S. G. Cole, E. B. Perry, D. Krizanc, N. C. Field, M. Slaton and F. M. Cohan (2010). "Ecology of speciation in the genus Bacillus." Appl Environ Microbiol **76**(5): 1349-1358.

Coque, T. M., R. J. Willems, J. Fortun, J. Top, S. Diz, E. Loza, R. Canton and F. Baquero (2005). "Population structure of Enterococcus faecium causing bacteremia in a Spanish university hospital: setting the scene for a future increase in vancomycin resistance?" Antimicrob Agents Chemother **49**(7): 2693-2700.

Dagan, T. and W. Martin (2007). "Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution." Proc Natl Acad Sci U S A **104**(3): 870-875.

Daubin, V., E. Lerat and G. Perriere (2003). "The source of laterally transferred genes in bacterial genomes." Genome Biol **4**(9): R57.

de Been, M., W. van Schaik, L. Cheng, J. Corander and R. J. Willems (2013). "Recent recombination events in the core genome are associated with adaptive evolution in Enterococcus faecium." Genome Biol Evol **5**(8): 1524-1535.

de Regt, M. J., W. van Schaik, M. van Luit-Asbroek, H. A. Dekker, E. van Duijkeren, C. J. Koning, M. J. Bonten and R. J. Willems (2012). "Hospital and community ampicillin-resistant Enterococcus faecium are evolutionarily closely linked but have diversified through niche adaptation." PLoS One **7**(2): e30319.

de Vaux, A., G. Laguerre, C. Divies and H. Prevost (1998). "Enterococcus asini sp. nov. isolated from the caecum of donkeys (Equus asinus)." Int J Syst Bacteriol **48 Pt 2**: 383-387.

Del Papa, M. F. and M. Perego (2008). "Ethanolamine activates a sensor histidine kinase regulating its utilization in Enterococcus faecalis." J Bacteriol **190**(21): 7147-7156.

Delihas, N. (2011). "Impact of small repeat sequences on bacterial genome evolution." Genome Biol Evol **3**: 959-973.

DeLisle, S. and T. M. Perl (2003). "Vancomycin-resistant enterococci: a road map on how to prevent the emergence and transmission of antimicrobial resistance." Chest **123**(5 Suppl): 504S-518S.

Depardieu, F., P. E. Reynolds and P. Courvalin (2003). "VanD-type vancomycin-resistant Enterococcus faecium 10/96A." Antimicrob Agents Chemother **47**(1): 7-18.

Donkor, E. S. (2013). "Sequencing of bacterial genomes: principles and insights into pathogenesis and development of antibiotics." Genes (Basel) **4**(4): 556-572.

Doolittle, W. F. (1999). "Lateral genomics." Trends Cell Biol **9**(12): M5-8.

Durmaz, E. and T. R. Klaenhammer (2000). "Genetic analysis of chromosomal regions of Lactococcus lactis acquired by recombinant lytic phages." Appl Environ Microbiol **66**(3): 895-903.

Eaton, T. J. and M. J. Gasson (2002). "A variant enterococcal surface protein Esp(fm) in Enterococcus faecium; distribution among food, commensal, medical, and environmental isolates." FEMS Microbiol Lett **216**(2): 269-275.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-1797.

Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach and

S. Turner (2009). "Real-time DNA sequencing from single polymerase molecules." Science **323**(5910): 133-138.

el Amin, N. A., S. Jalal and B. Wretlind (1999). "Alterations in GyrA and ParC associated with fluoroquinolone resistance in Enterococcus faecium." Antimicrob Agents Chemother **43**(4): 947-949.

Ellegaard, K. M., L. Klasson, K. Naslund, K. Bourtzis and S. G. Andersson (2013). "Comparative genomics of Wolbachia and the bacterial species concept." PLoS Genet **9**(4): e1003381.

Elsner, H. A., I. Sobottka, D. Mack, M. Claussen, R. Laufs and R. Wirth (2000). "Virulence factors of Enterococcus faecalis and Enterococcus faecium blood culture isolates." Eur J Clin Microbiol Infect Dis **19**(1): 39-42.

English, A. C., S. Richards, Y. Han, M. Wang, V. Vee, J. Qu, X. Qin, D. M. Muzny, J. G. Reid, K. C. Worley and R. A. Gibbs (2012). "Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology." PLoS One **7**(11): e47768.

Fabretti, F., C. Theilacker, L. Baldassarri, Z. Kaczynski, A. Kropec, O. Holst and J. Huebner (2006). "Alanine esters of enterococcal lipoteichoic acid play a role in biofilm formation and resistance to antimicrobial peptides." Infect Immun **74**(7): 4164-4171.

Farrell, D. J., R. E. Mendes, J. E. Ross, H. S. Sader and R. N. Jones (2011). "LEADER Program results for 2009: an activity and spectrum analysis of linezolid using 6,414 clinical isolates from 56 medical centers in the United States." Antimicrob Agents Chemother **55**(8): 3684-3690.

Farrelly, V., F. A. Rainey and E. Stackebrandt (1995). "Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species." Appl Environ Microbiol **61**(7): 2798-2801.

Ferrarini, M., M. Moretto, J. A. Ward, N. Surbanovski, V. Stevanovic, L. Giongo, R. Viola, D. Cavalieri, R. Velasco, A. Cestaro and D. J. Sargent (2013). "An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome." BMC Genomics **14**: 670.

Fisher, K. and C. Phillips (2009). "The ecology, epidemiology and virulence of Enterococcus." Microbiology **155**(Pt 6): 1749-1757.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick and et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." Science **269**(5223): 496-512.

Fokine, A. and M. G. Rossmann (2014). "Molecular architecture of tailed double-stranded DNA phages." Bacteriophage **4**(1): e28281.

Fontana, R., M. Ligozzi, F. Pittaluga and G. Satta (1996). "Intrinsic penicillin resistance in enterococci." Microb Drug Resist **2**(2): 209-213.

Fraser, C. M., J. A. Eisen, K. E. Nelson, I. T. Paulsen and S. L. Salzberg (2002). "The value of complete microbial genome sequencing (you get what you pay for)." J Bacteriol **184**(23): 6403-6405; discusion 6405.

Fraser, C. M. and R. D. Fleischmann (1997). "Strategies for whole microbial genome sequencing and analysis." Electrophoresis **18**(8): 1207-1216.

Frost, L. S., R. Leplae, A. O. Summers and A. Toussaint (2005). "Mobile genetic elements: the agents of open source evolution." Nat Rev Microbiol **3**(9): 722-732.

Fujisawa, H. and T. Minagawa (1986). "[DNA packaging by double stranded DNA phages]." Uirusu **36**(2): 185-193.

Galloway-Pena, J., J. H. Roh, M. Latorre, X. Qin and B. E. Murray (2012). "Genomic and SNP analyses demonstrate a distant separation of the hospital and community-associated clades of Enterococcus faecium." PLoS One **7**(1): e30187.

Galloway-Pena, J. R., S. R. Nallapareddy, C. A. Arias, G. M. Eliopoulos and B. E. Murray (2009). "Analysis of clonality and antibiotic resistance among early clinical isolates of Enterococcus faecium in the United States." J Infect Dis **200**(10): 1566-1573.

Garcia-Migura, L., H. Hasman and L. B. Jensen (2009). "Presence of pRI1: a small cryptic mobilizable plasmid isolated from Enterococcus faecium of human and animal origin." Curr Microbiol **58**(2): 95-100.

Garcia-Vallve, S., A. Romeu and J. Palau (2000). "Horizontal gene transfer in bacterial and archaeal complete genomes." Genome Res **10**(11): 1719-1725.

Gholizadeh, Y. and P. Courvalin (2000). "Acquired and intrinsic glycopeptide resistance in enterococci." Int J Antimicrob Agents **16 Suppl 1**: S11-17.

Goerke, C., R. Pantucek, S. Holtfreter, B. Schulte, M. Zink, D. Grumann, B. M. Broker, J. Doskar and C. Wolz (2009). "Diversity of prophages in

dominant Staphylococcus aureus clonal lineages." J Bacteriol **191**(11): 3462-3468.

Goldberg, S. M., J. Johnson, D. Busam, T. Feldblyum, S. Ferriera, R. Friedman, A. Halpern, H. Khouri, S. A. Kravitz, F. M. Lauro, K. Li, Y. H. Rogers, R. Strausberg, G. Sutton, L. Tallon, T. Thomas, E. Venter, M. Frazier and J. C. Venter (2006). "A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes." Proc Natl Acad Sci U S A **103**(30): 11240-11245.

Grissa, I., G. Vergnaud and C. Pourcel (2007). "CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats." Nucleic Acids Res **35**(Web Server issue): W52-57.

Groth, A. C. and M. P. Calos (2004). "Phage integrases: biology and applications." J Mol Biol **335**(3): 667-678.

Gupta, G. D. and V. Kumar (2012). "Identification of nucleic acid binding sites on translin-associated factor X (TRAX) protein." PLoS One **7**(3): e33035.

Haaber, J., G. M. Rousseau, K. Hammer and S. Moineau (2009). "Identification and characterization of the phage gene sav, involved in sensitivity to the lactococcal abortive infection mechanism AbiV." Appl Environ Microbiol **75**(8): 2484-2494.

Hacker, J. and E. Carniel (2001). "Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes." EMBO Rep **2**(5): 376-381.

Hacker, J. and J. B. Kaper (2000). "Pathogenicity islands and the evolution of microbes." Annu Rev Microbiol **54**: 641-679.

Hammerum, A. M., S. E. Flannagan, D. B. Clewell and L. B. Jensen (2001). "Indication of transposition of a mobile DNA element containing the vat(D) and erm(B) genes in Enterococcus faecium." Antimicrob Agents Chemother **45**(11): 3223-3225.

Hammerum, A. M., V. Fussing, F. M. Aarestrup and H. C. Wegener (2000). "Characterization of vancomycin-resistant and vancomycin-susceptible Enterococcus faecium isolates from humans, chickens and pigs by RiboPrinting and pulsed-field gel electrophoresis." J Antimicrob Chemother **45**(5): 677-680.

Hatfull, G. F., D. Jacobs-Sera, J. G. Lawrence, W. H. Pope, D. A. Russell, C. C. Ko, R. J. Weber, M. C. Patel, K. L. Germane, R. H. Edgar, N. N. Hoyte, C. A. Bowman, A. T. Tantoco, E. C. Paladin, M. S. Myers, A. L. Smith, M. S. Grace, T. T. Pham, M. B. O'Brien, A. M. Vogelsberger, A. J. Hryckowian, J. L. Wynalek, H. Donis-Keller, M. W. Bogel, C. L. Peebles, S. G. Cresawn and R. W. Hendrix (2010). "Comparative genomic analysis of 60 Mycobacteriophage genomes: genome clustering, gene acquisition, and gene size." J Mol Biol **397**(1): 119-143.

Heap, J. T., M. Ehsaan, C. M. Cooksley, Y. K. Ng, S. T. Cartman, K. Winzer and N. P. Minton (2012). "Integration of DNA into bacterial chromosomes from plasmids without a counter-selection marker." Nucleic Acids Res **40**(8): e59.

Heaton, M. P., L. F. Discotto, M. J. Pucci and S. Handwerger (1996). "Mobilization of vancomycin resistance by transposon-mediated fusion of a VanA plasmid with an Enterococcus faecium sex pheromone-response plasmid." Gene **171**(1): 9-17.

Hegstad, K., T. Mikalsen, T. M. Coque, G. Werner and A. Sundsfjord (2010). "Mobile genetic elements and their contribution to the emergence of antimicrobial resistant Enterococcus faecalis and Enterococcus faecium." Clin Microbiol Infect **16**(6): 541-554.

Heikens, E., M. Leendertse, L. M. Wijnands, M. van Luit-Asbroek, M. J. Bonten, T. van der Poll and R. J. Willems (2009). "Enterococcal surface protein Esp is not essential for cell adhesion and intestinal colonization of Enterococcus faecium in mice." BMC Microbiol **9**: 19.

Heikens, E., K. V. Singh, K. D. Jacques-Palaz, M. van Luit-Asbroek, E. A. Oostdijk, M. J. Bonten, B. E. Murray and R. J. Willems (2011). "Contribution of the enterococcal surface protein Esp to pathogenesis of Enterococcus faecium endocarditis." Microbes Infect **13**(14-15): 1185-1190.

Hendrickx, A. P., M. J. Bonten, M. van Luit-Asbroek, C. M. Schapendonk, A. H. Kragten and R. J. Willems (2008). "Expression of two distinct types of pili by a hospital-acquired Enterococcus faecium isolate." Microbiology **154**(Pt 10): 3212-3223.

Hendrickx, A. P., M. van Luit-Asbroek, C. M. Schapendonk, W. J. van Wamel, J. C. Braat, L. M. Wijnands, M. J. Bonten and R. J. Willems (2009). "SgrA, a nidogen-binding LPXTG surface adhesin implicated in biofilm formation, and EcbA, a collagen binding MSCRAMM, are two novel adhesins of hospital-acquired Enterococcus faecium." Infect Immun **77**(11): 5097-5106.

Hendrickx, A. P., W. J. van Wamel, G. Posthuma, M. J. Bonten and R. J. Willems (2007). "Five genes encoding surface-exposed LPXTG proteins are

enriched in hospital-adapted Enterococcus faecium clonal complex 17 isolates." J Bacteriol **189**(22): 8321-8332.

Hendrix, R. W., G. F. Hatfull and M. C. M. Smith (2003). "Bacteriophages with tails: chasing their origins and evolution." Research in Microbiology **154**(4): 253-257.

Henson, J., G. Tischler and Z. Ning (2012). "Next-generation sequencing and large genome assemblies." Pharmacogenomics **13**(8): 901-915.

Holden, M. T., H. Hauser, M. Sanders, T. H. Ngo, I. Cherevach, A. Cronin, I. Goodhead, K. Mungall, M. A. Quail, C. Price, E. Rabbinowitsch, S. Sharp, N. J. Croucher, T. B. Chieu, N. T. Mai, T. S. Diep, N. T. Chinh, M. Kehoe, J. A. Leigh, P. N. Ward, C. G. Dowson, A. M. Whatmore, N. Chanter, P. Iversen, M. Gottschalk, J. D. Slater, H. E. Smith, B. G. Spratt, J. Xu, C. Ye, S. Bentley, B. G. Barrell, C. Schultsz, D. J. Maskell and J. Parkhill (2009). "Rapid evolution of virulence and drug resistance in the emerging zoonotic pathogen Streptococcus suis." PLoS One **4**(7): e6072.

Hollenbeck, B. L. and L. B. Rice (2012). "Intrinsic and acquired resistance mechanisms in enterococcus." Virulence **3**(5): 421-433.

Huycke, M. M., D. F. Sahm and M. S. Gilmore (1998). "Multiple-drug resistant enterococci: the nature of the problem and an agenda for the future." Emerg Infect Dis **4**(2): 239-249.

Hyman, P. and S. T. Abedon (2010). "Bacteriophage host range and bacterial resistance." Adv Appl Microbiol **70**: 217-248.

Jacob, F. and J. Monod (1961). "Genetic regulatory mechanisms in the synthesis of proteins." J Mol Biol **3**: 318-356.

Jaurin, B. and S. Normark (1983). "Insertion of IS2 creates a novel ampC promoter in Escherichia coli." Cell **32**(3): 809-816.

Jensen, L. B., N. Frimodt-Moller and F. M. Aarestrup (1999). "Presence of erm gene classes in gram-positive bacteria of animal and human origin in Denmark." FEMS Microbiol Lett **170**(1): 151-158.

Jett, B. D., M. M. Huycke and M. S. Gilmore (1994). "Virulence of enterococci." Clin Microbiol Rev **7**(4): 462-478.

Johnson, A., B. J. Meyer and M. Ptashne (1978). "Mechanism of action of the cro protein of bacteriophage lambda." Proc Natl Acad Sci U S A **75**(4): 1783-1787.

Johnson, A. P. (1994). "The pathogenicity of enterococci." J Antimicrob Chemother **33**(6): 1083-1089.

Johnston, N. J., T. A. Mukhtar and G. D. Wright (2002). "Streptogramin antibiotics: mode of action and resistance." Curr Drug Targets **3**(4): 335-344.

Juhala, R. J., M. E. Ford, R. L. Duda, A. Youlton, G. F. Hatfull and R. W. Hendrix (2000). "Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages." J Mol Biol **299**(1): 27-51.

Juhas, M., J. R. van der Meer, M. Gaillard, R. M. Harding, D. W. Hood and D. W. Crook (2009). "Genomic islands: tools of bacterial horizontal gene transfer and evolution." FEMS Microbiol Rev **33**(2): 376-393.

Kanoh, S. and B. K. Rubin (2010). "Mechanisms of action and clinical application of macrolides as immunomodulatory medications." Clin Microbiol Rev **23**(3): 590-615.

Kayaoglu, G. and D. Orstavik (2004). "Virulence factors of Enterococcus faecalis: relationship to endodontic disease." Crit Rev Oral Biol Med **15**(5): 308-320.

Kazmierczak, M. J., M. Wiedmann and K. J. Boor (2005). "Alternative sigma factors and their roles in bacterial virulence." Microbiol Mol Biol Rev **69**(4): 527-543.

Keeling, P. J. and J. D. Palmer (2008). "Horizontal gene transfer in eukaryotic evolution." Nat Rev Genet **9**(8): 605-618.

Kehoe, L. E., J. Snidwongse, P. Courvalin, J. B. Rafferty and I. A. Murray (2003). "Structural basis of Synercid (quinupristin-dalfopristin) resistance in Gram-positive bacterial pathogens." J Biol Chem **278**(32): 29963-29970.

Khalil, M., Y. Al-Mazrou, M. Al-Howasi and M. Al-Jeffri (1999). "Immunogenicity of FDA DTP versus WHO DTP." Ann Saudi Med **19**(5): 417-419.

Klare, I., D. Badstubner, C. Konstabel and W. Witte (1999). "Identification of enterococci and determination of their glycopeptide resistance in German and Austrian clinical microbiology laboratories." Clin Microbiol Infect **5**(9): 535-539.

Klare, I., C. Konstabel, D. Badstubner, G. Werner and W. Witte (2003). "Occurrence and spread of antibiotic resistances in Enterococcus faecium." Int J Food Microbiol **88**(2-3): 269-290.

Klare, I., C. Konstabel, S. Mueller-Bertling, G. Werner, B. Strommenger, C. Kettlitz, S. Borgmann, B. Schulte, D. Jonas, A. Serr, A. M. Fahr, U. Eigner and W. Witte (2005). "Spread of ampicillin/vancomycin-resistant Enterococcus faecium of the epidemic-virulent clonal complex-17 carrying

the genes esp and hyl in German hospitals." Eur J Clin Microbiol Infect Dis **24**(12): 815-825.

Klare, I., G. Werner and W. Witte (2001). "Enterococci. Habitats, infections, virulence factors, resistances to antibiotics, transfer of resistance determinants." Contrib Microbiol **8**: 108-122.

Koch, S., M. Hufnagel, C. Theilacker and J. Huebner (2004). "Enterococcal infections: host response, therapeutic, and prophylactic possibilities." Vaccine **22**(7): 822-830.

Koeppel, A. F., J. O. Wertheim, L. Barone, N. Gentile, D. Krizanc and F. M. Cohan (2013). "Speedy speciation in a bacterial microcosm: new species can arise as frequently as adaptations within a species." ISME J **7**(6): 1080-1091.

Kotra, L. P., J. Haddad and S. Mobashery (2000). "Aminoglycosides: perspectives on mechanisms of action and resistance and strategies to counter resistance." Antimicrob Agents Chemother **44**(12): 3249-3256.

Kuhn, I., A. Iversen, L. G. Burman, B. Olsson-Liljequist, A. Franklin, M. Finn, F. Aarestrup, A. M. Seyfarth, A. R. Blanch, X. Vilanova, H. Taylor, J. Caplin, M. A. Moreno, L. Dominguez, I. A. Herrero and R. Mollby (2003). "Comparison of enterococcal populations in animals, humans, and the environment--a European study." Int J Food Microbiol **88**(2-3): 133-145.

Kuhn, I., A. Iversen, M. Finn, C. Greko, L. G. Burman, A. R. Blanch, X. Vilanova, A. Manero, H. Taylor, J. Caplin, L. Dominguez, I. A. Herrero, M. A. Moreno and R. Mollby (2005). "Occurrence and relatedness of vancomycin-resistant enterococci in animals, humans, and the environment in different European regions." Appl Environ Microbiol **71**(9): 5383-5390.

Kunisawa, T. (1992). "Synonymous codon preferences in bacteriophage T4: a distinctive use of transfer RNAs from T4 and from its host Escherichia coli." J Theor Biol **159**(3): 287-298.

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

Kusumoto, M., T. Ooka, Y. Nishiya, Y. Ogura, T. Saito, Y. Sekine, T. Iwata, M. Akiba and T. Hayashi (2011). "Insertion sequence-excision enhancer removes transposable elements from bacterial genomes and induces various genomic deletions." Nat Commun **2**: 152.

Kutter, E. and A. Sulakvelidze (2005). Bacteriophages: Biology and Applications: Molecular Biology and Applications. Boca Raton, CRC Press.

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

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson and D. G. Higgins (2007). "Clustal W and Clustal X version 2.0." Bioinformatics **23**(21): 2947-2948.

Lawrence, J. G. and H. Hendrickson (2003). "Lateral gene transfer: when will adolescence end?" Mol Microbiol **50**(3): 739-749.

Lebreton, F., W. van Schaik, A. M. McGuire, P. Godfrey, A. Griggs, V. Mazumdar, J. Corander, L. Cheng, S. Saif, S. Young, Q. Zeng, J. Wortman, B. Birren, R. J. Willems, A. M. Earl and M. S. Gilmore (2013). "Emergence

of epidemic multidrug-resistant Enterococcus faecium from animal and commensal strains." MBio **4**(4).

Leclerc, H., L. A. Devriese and D. A. Mossel (1996). "Taxonomical changes in intestinal (faecal) enterococci and streptococci: consequences on their use as indicators of faecal contamination in drinking water." J Appl Bacteriol **81**(5): 459-466.

Leclercq, R. (1997). "Enterococci acquire new kinds of resistance." Clin Infect Dis **24 Suppl 1**: S80-84.

Leclercq, R., S. Dutka-Malen, A. Brisson-Noel, C. Molinas, E. Derlot, M. Arthur, J. Duval and P. Courvalin (1992). "Resistance of enterococci to aminoglycosides and glycopeptides." Clin Infect Dis **15**(3): 495-501.

Lepage, E., S. Brinster, C. Caron, C. Ducroix-Crepy, L. Rigottier-Gois, G. Dunny, C. Hennequet-Antier and P. Serror (2006). "Comparative genomic hybridization analysis of Enterococcus faecalis: identification of genes absent from food strains." J Bacteriol **188**(19): 6858-6868.

Lester, C. H., D. Sandvang, S. S. Olsen, H. C. Schonheyder, J. O. Jarlov, J. Bangsborg, D. S. Hansen, T. G. Jensen, N. Frimodt-Moller, A. M. Hammerum and D. S. Group (2008). "Emergence of ampicillin-resistant Enterococcus faecium in Danish hospitals." J Antimicrob Chemother **62**(6): 1203-1206.

Levesque, C., L. Piche, C. Larose and P. H. Roy (1995). "PCR mapping of integrons reveals several novel combinations of resistance genes." Antimicrobial Agents and Chemotherapy **39**(1): 185-191.

Li, L., C. J. Stoeckert, Jr. and D. S. Roos (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome Res **13**(9): 2178-2189.

Li, S., H. Fan, X. An, H. Fan, H. Jiang, Y. Chen and Y. Tong (2014). "Scrutinizing virus genome termini by high-throughput sequencing." PLoS One **9**(1): e85806.

Lima-Mendez, G., J. Van Helden, A. Toussaint and R. Leplae (2008). "Reticulate representation of evolutionary and functional relationships between phage genomes." Mol Biol Evol **25**(4): 762-777.

Llacer, J. L., L. M. Polo, S. Tavarez, B. Alarcon, R. Hilario and V. Rubio (2007). "The gene cluster for agmatine catabolism of Enterococcus faecalis: study of recombinant putrescine transcarbamylase and agmatine deiminase and a snapshot of agmatine deiminase catalyzing its reaction." J Bacteriol **189**(4): 1254-1265.

Lowe, B. A., J. D. Miller and M. N. Neely (2007). "Analysis of the polysaccharide capsule of the systemic pathogen Streptococcus iniae and its implications in virulence." Infect Immun **75**(3): 1255-1264.

Ludwig, W., E. Seewaldt, R. Kilpper-Balz, K. H. Schleifer, L. Magrum, C. R. Woese, G. E. Fox and E. Stackebrandt (1985). "The phylogenetic position of Streptococcus and Enterococcus." J Gen Microbiol **131**(3): 543-551.

Lund, B., I. Adamsson and C. Edlund (2002). "Gastrointestinal transit survival of an Enterococcus faecium probiotic strain administered with or without vancomycin." Int J Food Microbiol **77**(1-2): 109-115.

Luo, C., S. T. Walk, D. M. Gordon, M. Feldgarden, J. M. Tiedje and K. T. Konstantinidis (2011). "Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species." Proc Natl Acad Sci U S A **108**(17): 7200-7205.

Magi, G., R. Capretti, C. Paoletti, M. Pietrella, L. Ferrante, F. Biavasco, P. E. Varaldo and B. Facinelli (2003). "Presence of a vanA-carrying pheromone response plasmid (pBRG1) in a clinical isolate of Enterococcus faecium." Antimicrob Agents Chemother **47**(5): 1571-1576.

Makinen, P. L., D. B. Clewell, F. An and K. K. Makinen (1989). "Purification and substrate specificity of a strongly hydrophobic extracellular metalloendopeptidase ("gelatinase") from Streptococcus faecalis (strain 0G1-10)." J Biol Chem **264**(6): 3325-3334.

Malachowa, N. and F. R. DeLeo (2010). "Mobile genetic elements of Staphylococcus aureus." Cell Mol Life Sci **67**(18): 3057-3071.

Mannu, L., A. Paba, E. Daga, R. Comunian, S. Zanetti, I. Dupre and L. A. Sechi (2003). "Comparison of the incidence of virulence determinants and antibiotic resistance between Enterococcus faecium strains of dairy, animal and clinical origin." Int J Food Microbiol **88**(2-3): 291-304.

Manson, J. M., L. E. Hancock and M. S. Gilmore (2010). "Mechanism of chromosomal transfer of Enterococcus faecalis pathogenicity island, capsule, antimicrobial resistance, and other traits." Proc Natl Acad Sci U S A **107**(27): 12269-12274.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H.

Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley and J. M. Rothberg (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.

Marraffini, L. A. and E. J. Sontheimer (2008). "CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA." Science **322**(5909): 1843-1845.

Marshall, B. M. and S. B. Levy (2011). "Food animals and antimicrobials: impacts on human health." Clin Microbiol Rev **24**(4): 718-733.

Matos, R. C., N. Lapaque, L. Rigottier-Gois, L. Debarbieux, T. Meylheuc, B. Gonzalez-Zorn, F. Repoila, F. Lopes Mde and P. Serror (2013). "Enterococcus faecalis prophage dynamics and contributions to pathogenic traits." PLoS Genet **9**(6): e1003539.

Mazaheri Nezhad Fard, R., M. D. Barton and M. W. Heuzenroeder (2010). "Novel Bacteriophages in Enterococcus spp." Curr Microbiol **60**(6): 400-406.

Mazaheri Nezhad Fard, R., M. D. Barton and M. W. Heuzenroeder (2011). "Bacteriophage-mediated transduction of antibiotic resistance in enterococci." Lett Appl Microbiol **52**(6): 559-564.

McArthur, A. G., N. Waglechner, F. Nizam, A. Yan, M. A. Azad, A. J. Baylay, K. Bhullar, M. J. Canova, G. De Pascale, L. Ejim, L. Kalan, A. M. King, K. Koteva, M. Morar, M. R. Mulvey, J. S. O'Brien, A. C. Pawlowski, L. J. Piddock, P. Spanogiannopoulos, A. D. Sutherland, I. Tang, P. L. Taylor, M. Thaker, W. Wang, M. Yan, T. Yu and G. D. Wright (2013). "The comprehensive antibiotic resistance database." Antimicrob Agents Chemother **57**(7): 3348-3357.

McBride, S. M., P. S. Coburn, A. S. Baghdayan, R. J. Willems, M. J. Grande, N. Shankar and M. S. Gilmore (2009). "Genetic variation and evolution of the pathogenicity island of Enterococcus faecalis." J Bacteriol **191**(10): 3392-3402.

McCoy, R. C., R. W. Taylor, T. A. Blauwkamp, J. L. Kelley, M. Kertesz, D. Pushkarev, D. A. Petrov and A. S. Fiston-Lavier (2014). "Illumina TruSeq Synthetic Long-Reads Empower De Novo Assembly and Resolve Complex, Highly-Repetitive Transposable Elements." PLoS One **9**(9): e106689.

Mingeot-Leclercq, M. P., Y. Glupczynski and P. M. Tulkens (1999). "Aminoglycosides: activity and resistance." Antimicrob Agents Chemother **43**(4): 727-737.

Mohamed, J. A. and D. B. Huang (2007). "Biofilm formation by enterococci." J Med Microbiol **56**(Pt 12): 1581-1588.

Moritz, E. M. and P. J. Hergenrother (2007). "Toxin-antitoxin systems are ubiquitous and plasmid-encoded in vancomycin-resistant enterococci." Proc Natl Acad Sci U S A **104**(1): 311-316.

Mundt, J. O. (1961). "Occurrence of Enterococci: Bud, Blossom, and Soil Studies." Appl Microbiol **9**(6): 541-544.

Murray, B. E. (1992). "Beta-lactamase-producing enterococci." Antimicrob Agents Chemother **36**(11): 2355-2359.

Murray, B. E. and B. Mederski-Samaroj (1983). "Transferable beta-lactamase. A new mechanism for in vitro penicillin resistance in Streptococcus faecalis." J Clin Invest **72**(3): 1168-1171.

Nallapareddy, S. R., K. V. Singh and B. E. Murray (2008). "Contribution of the collagen adhesin Acm to pathogenesis of Enterococcus faecium in experimental endocarditis." Infect Immun **76**(9): 4120-4128.

Nallapareddy, S. R., K. V. Singh, J. Sillanpaa, D. A. Garsin, M. Hook, S. L. Erlandsen and B. E. Murray (2006). "Endocarditis and biofilm-associated pili of Enterococcus faecalis." J Clin Invest **116**(10): 2799-2807.

Nallapareddy, S. R., K. V. Singh, J. Sillanpaa, M. Zhao and B. E. Murray (2011). "Relative contributions of Ebp Pili and the collagen adhesin ace to host extracellular matrix protein adherence and experimental urinary tract infection by Enterococcus faecalis OG1RF." Infect Immun **79**(7): 2901-2910.

Naser, S., F. L. Thompson, B. Hoste, D. Gevers, K. Vandemeulebroecke, I. Cleenwerck, C. C. Thompson, M. Vancanneyt and J. Swings (2005). "Phylogeny and identification of Enterococci by atpA gene sequence analysis." J Clin Microbiol **43**(5): 2224-2230.

National Nosocomial Infections Surveillance, S. (2004). "National Nosocomial Infections Surveillance (NNIS) System Report, data summary from January 1992 through June 2004, issued October 2004." Am J Infect Control **32**(8): 470-485.

Nes, I. F., D. B. Diep and H. Holo (2007). "Bacteriocin diversity in Streptococcus and Enterococcus." J Bacteriol **189**(4): 1189-1198.

Neuhaus, F. C. and W. G. Struve (1965). "Enzymatic Synthesis of Analogs of the Cell-Wall Precursor. I. Kinetics and Specificity of Uridine Diphospho-N-Acetylmuramyl-L-Alanyl-D-Glutamyl-L-Lysine:D-Alanyl-D-Alanine Ligase (Adenosine Diphosphate) from Streptococcus Faecalis R." Biochemistry **4**: 120-131.

Nies, D. H. (2003). "Efflux-mediated heavy metal resistance in prokaryotes." FEMS Microbiol Rev **27**(2-3): 313-339.

Nowlan, S. S. and R. H. Deibel (1967). "Group Q streptococci. I. Ecology, serology, physiology, and relationship to established enterococci." J Bacteriol **94**(2): 291-296.

Okhravi, N., P. Adamson, M. M. Matheson, H. M. Towler and S. Lightman (2000). "PCR-RFLP-mediated detection and speciation of bacterial species causing endophthalmitis." Invest Ophthalmol Vis Sci **41**(6): 1438-1447.

Okonechnikov, K., O. Golosova and M. Fursov (2012). "Unipro UGENE: a unified bioinformatics toolkit." Bioinformatics **28**(8): 1166-1167.

Palmer, K. L. and M. S. Gilmore (2010). "Multidrug-resistant enterococci lack CRISPR-cas." MBio **1**(4).

Palmer, K. L., P. Godfrey, A. Griggs, V. N. Kos, J. Zucker, C. Desjardins, G. Cerqueira, D. Gevers, S. Walker, J. Wortman, M. Feldgarden, B. Haas, B. Birren and M. S. Gilmore (2012). "Comparative genomics of enterococci: variation in Enterococcus faecalis, clade structure in E. faecium, and defining characteristics of E. gallinarum and E. casseliflavus." MBio **3**(1): e00318-00311.

Park, M. O., K. H. Lim, T. H. Kim and H. I. Chang (2007). "Characterization of site-specific recombination by the integrase MJ1 from enterococcal bacteriophage phiFC1." J Microbiol Biotechnol **17**(2): 342-347.

Parkhill, J. (2000). "In defense of complete genomes." Nat Biotechnol **18**(5): 493-494.

Pedulla, M. L., M. E. Ford, J. M. Houtz, T. Karthikeyan, C. Wadsworth, J. A. Lewis, D. Jacobs-Sera, J. Falbo, J. Gross, N. R. Pannunzio, W. Brucker, V. Kumar, J. Kandasamy, L. Keenan, S. Bardarov, J. Kriakov, J. G. Lawrence, W. R. Jacobs, Jr., R. W. Hendrix and G. F. Hatfull (2003). "Origins of highly mosaic mycobacteriophage genomes." Cell **113**(2): 171-182.

Pennisi, E. (1998). "EVOLUTION: Genome Data Shake Tree of Life." Science **280**(5364): 672-674.

Pepper, K., C. Le Bouguenec, G. de Cespedes and T. Horaud (1986). "Dispersal of a plasmid-borne chloramphenicol resistance gene in streptococcal and enterococcal plasmids." Plasmid **16**(3): 195-203.

Perkins, T. T., R. A. Kingsley, M. C. Fookes, P. P. Gardner, K. D. James, L. Yu, S. A. Assefa, M. He, N. J. Croucher, D. J. Pickard, D. J. Maskell, J. Parkhill, J. Choudhary, N. R. Thomson and G. Dougan (2009). "A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi." PLoS Genet **5**(7): e1000569.

Piekarska, K., R. Gierczynski, M. Lawrynowicz-Paciorek, M. Kochman and M. Jagielski (2008). "Novel gyrase mutations and characterization of

ciprofloxacin-resistant clinical strains of Enterococcus faecalis isolated in Poland." Pol J Microbiol **57**(2): 121-124.

Prange, A., R. Chauvistre, H. Modrow, J. Hormes, H. G. Truper and C. Dahl (2002). "Quantitative speciation of sulfur in bacterial sulfur globules: X-ray absorption spectroscopy reveals at least three different species of sulfur." Microbiology **148**(Pt 1): 267-276.

Prestel, E., S. Salamitou and M. S. DuBow (2008). "An examination of the bacteriophages and bacteria of the Namib desert." J Microbiol **46**(4): 364-372.

Price, M. N., P. S. Dehal and A. P. Arkin (2010). "FastTree 2--approximately maximum-likelihood trees for large alignments." PLoS One **5**(3): e9490.

Purnell, S. E., J. E. Ebdon and H. D. Taylor (2011). "Bacteriophage lysis of Enterococcus host strains: a tool for microbial source tracking?" Environ Sci Technol **45**(24): 10699-10705.

Qin, X., J. R. Galloway-Pena, J. Sillanpaa, J. H. Roh, S. R. Nallapareddy, S. Chowdhury, A. Bourgogne, T. Choudhury, D. M. Muzny, C. J. Buhay, Y. Ding, S. Dugan-Rocha, W. Liu, C. Kovar, E. Sodergren, S. Highlander, J. F. Petrosino, K. C. Worley, R. A. Gibbs, G. M. Weinstock and B. E. Murray (2012). "Complete genome sequence of Enterococcus faecium strain TX16 and comparative genomic analysis of Enterococcus faecium genomes." BMC Microbiol **12**: 135.

Qin, X., K. V. Singh, Y. Xu, G. M. Weinstock and B. E. Murray (1998). "Effect of disruption of a gene encoding an autolysin of Enterococcus faecalis OG1RF." Antimicrob Agents Chemother **42**(11): 2883-2888.

Ramsey, M., A. Hartke and M. Huycke (2014). The Physiology and Metabolism of Enterococci. Enterococci: From Commensals to Leading Causes of Drug Resistant Infection. M. S. Gilmore, D. B. Clewell, Y. Ike and N. Shankar. Boston.

Raoult, D., S. Audic, C. Robert, C. Abergel, P. Renesto, H. Ogata, B. La Scola, M. Suzan and J. M. Claverie (2004). "The 1.2-megabase genome sequence of Mimivirus." Science **306**(5700): 1344-1350.

Reffuveille, F., C. Leneveu, S. Chevalier, Y. Auffray and A. Rince (2011). "Lipoproteins of Enterococcus faecalis: bioinformatic identification, expression analysis and relation to virulence." Microbiology **157**(Pt 11): 3001-3013.

Reinert, R. R., G. Conrads, J. J. Schlaeger, G. Werner, W. Witte, R. Lutticken and I. Klare (1999). "Survey of antibiotic resistance among enterococci in North Rhine-Westphalia, Germany." J Clin Microbiol **37**(5): 1638-1641.

Reynolds, P. E. (1989). "Structure, biochemistry and mechanism of action of glycopeptide antibiotics." Eur J Clin Microbiol Infect Dis **8**(11): 943-950.

Reynolds, P. E. and P. Courvalin (2005). "Vancomycin resistance in enterococci due to synthesis of precursors terminating in D-alanyl-D-serine." Antimicrob Agents Chemother **49**(1): 21-25.

Ribeiro, F. J., D. Przybylski, S. Yin, T. Sharpe, S. Gnerre, A. Abouelleil, A. M. Berlin, A. Montmayeur, T. P. Shea, B. J. Walker, S. K. Young, C. Russ, C. Nusbaum, I. MacCallum and D. B. Jaffe (2012). "Finished bacterial genomes from shotgun sequence data." Genome Res **22**(11): 2270-2277.

Rice, L. B. (1998). "Tn916 family conjugative transposons and dissemination of antimicrobial resistance determinants." Antimicrob Agents Chemother **42**(8): 1871-1877.

Rice, L. B. (2001). "Emergence of vancomycin-resistant enterococci." Emerg Infect Dis **7**(2): 183-187.

Roberts, A. P., M. Chandler, P. Courvalin, G. Guedon, P. Mullany, T. Pembroke, J. I. Rood, C. J. Smith, A. O. Summers, M. Tsuda and D. E. Berg (2008). "Revised nomenclature for transposable genetic elements." Plasmid **60**(3): 167-173.

Roberts, M. C. and S. L. Hillier (1990). "Genetic basis of tetracycline resistance in urogenital bacteria." Antimicrob Agents Chemother **34**(2): 261-264.

Rohwer, F. and R. Edwards (2002). "The Phage Proteomic Tree: a genome-based taxonomy for phage." J Bacteriol **184**(16): 4529-4535.

Romero, P., N. J. Croucher, N. L. Hiller, F. Z. Hu, G. D. Ehrlich, S. D. Bentley, E. Garcia and T. J. Mitchell (2009). "Comparative genomic analysis of ten Streptococcus pneumoniae temperate bacteriophages." J Bacteriol **191**(15): 4854-4862.

Rothberg, J. M., W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, J. H. Leamon, K. Johnson, M. J. Milgrew, M. Edwards, J. Hoon, J. F. Simons, D. Marran, J. W. Myers, J. F. Davidson, A. Branting, J. R. Nobile, B. P. Puc, D. Light, T. A. Clark, M. Huber, J. T. Branciforte, I. B. Stoner, S. E. Cawley, M. Lyons, Y. Fu, N. Homer, M. Sedova, X. Miao, B. Reed, J. Sabina, E. Feierstein, M. Schorn, M. Alanjary, E. Dimalanta, D. Dressman, R. Kasinskas, T. Sokolsky, J. A. Fidanza, E. Namsaraev, K. J. McKernan,

A. Williams, G. T. Roth and J. Bustillo (2011). "An integrated semiconductor device enabling non-optical genome sequencing." Nature **475**(7356): 348-352.

Rothberg, J. M. and J. H. Leamon (2008). "The development and impact of 454 sequencing." Nat Biotechnol **26**(10): 1117-1124.

Rouch, D. A., M. E. Byrne, Y. C. Kong and R. A. Skurray (1987). "The aacA-aphD gentamicin and kanamycin resistance determinant of Tn4001 from Staphylococcus aureus: expression and nucleotide sequence analysis." J Gen Microbiol **133**(11): 3039-3052.

Rudy, C., K. L. Taylor, D. Hinerfeld, J. R. Scott and G. Churchward (1997). "Excision of a conjugative transposon in vitro by the Int and Xis proteins of Tn916." Nucleic Acids Res **25**(20): 4061-4066.

Safford, C. E., J. M. Sherman and H. M. Hodge (1937). "Streptococcus salivarius." J Bacteriol **33**(3): 263-274.

Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe and M. Smith (1977). "Nucleotide sequence of bacteriophage phi X174 DNA." Nature **265**(5596): 687-695.

Sanger, F., S. Nicklen and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A **74**(12): 5463-5467.

Santagati, M., F. Campanile and S. Stefani (2012). "Genomic diversification of enterococci in hosts: the role of the mobilome." Front Microbiol **3**: 95.

Sastry, M. S., K. Korotkov, Y. Brodsky and F. Baneyx (2002). "Hsp31, the Escherichia coli yedU gene product, is a molecular chaperone whose activity is inhibited by ATP at high temperatures." J Biol Chem **277**(48): 46026-46034.

Schleifer, K. H., R. Kilpper-Balz, J. Kraus and F. Gehring (1984). "Relatedness and classification of Streptococcus mutans and "mutans-like" streptococci." J Dent Res **63**(8): 1047-1050.

Schneider, D. and R. E. Lenski (2004). "Dynamics of insertion sequence elements during experimental evolution of bacteria." Res Microbiol **155**(5): 319-327.

Scott, J. R. and G. G. Churchward (1995). "Conjugative transposition." Annu Rev Microbiol **49**: 367-397.

Serruto, D., L. Serino, V. Masignani and M. Pizza (2009). "Genome-based approaches to develop vaccines against bacterial pathogens." Vaccine **27**(25-26): 3245-3250.

Shankar, N., A. S. Baghdayan and M. S. Gilmore (2002). "Modulation of virulence within a pathogenicity island in vancomycin-resistant Enterococcus faecalis." Nature **417**(6890): 746-750.

Shankar, V., A. S. Baghdayan, M. M. Huycke, G. Lindahl and M. S. Gilmore (1999). "Infection-derived Enterococcus faecalis strains are enriched in esp, a gene encoding a novel surface protein." Infect Immun **67**(1): 193-200.

Shinomiya, T. (1984). "Phenotypic mixing of pyocin R2 and bacteriophage PS17 in Pseudomonas aeruginosa PAO." J Virol **49**(2): 310-314.

Siezen, R., J. Boekhorst, L. Muscariello, D. Molenaar, B. Renckens and M. Kleerebezem (2006). "Lactobacillus plantarum gene clusters encoding putative cell-surface protein complexes for carbohydrate utilization are conserved in specific gram-positive bacteria." BMC Genomics **7**: 126.

Sillanpaa, J., V. P. Prakash, S. R. Nallapareddy and B. E. Murray (2009). "Distribution of genes encoding MSCRAMMs and Pili in clinical and natural populations of Enterococcus faecium." J Clin Microbiol **47**(4): 896-901.

Silver, S. and L. T. Phung (1996). "Bacterial heavy metal resistance: new surprises." Annu Rev Microbiol **50**: 753-789.

Simjee, S., D. G. White, J. Meng, D. D. Wagner, S. Qaiyumi, S. Zhao, J. R. Hayes and P. F. McDermott (2002). "Prevalence of streptogramin resistance genes among Enterococcus isolates recovered from retail meats in the Greater Washington DC area." J Antimicrob Chemother **50**(6): 877-882.

Singh, K. V., S. R. Nallapareddy and B. E. Murray (2007). "Importance of the ebp (endocarditis- and biofilm-associated pilus) locus in the pathogenesis of Enterococcus faecalis ascending urinary tract infection." J Infect Dis **195**(11): 1671-1677.

Smith, F. R., C. F. Niven and J. M. Sherman (1938). "The Serological Identification of Streptococcus Zymogenes with the Lancefield Group D." J Bacteriol **35**(4): 425-428.

Son, J. S., S. Y. Jun, E. B. Kim, J. E. Park, H. R. Paik, S. J. Yoon, S. H. Kang and Y. J. Choi (2010). "Complete genome sequence of a newly isolated lytic bacteriophage, EFAP-1 of Enterococcus faecalis, and antibacterial activity of its endolysin EFAL-1." J Appl Microbiol **108**(5): 1769-1779.

Sorensen, T. L., M. Blom, D. L. Monnet, N. Frimodt-Moller, R. L. Poulsen and F. Espersen (2001). "Transient intestinal carriage after ingestion of

antibiotic-resistant Enterococcus faecium from chicken and pork." N Engl J Med **345**(16): 1161-1166.

Starikova, I., M. Al-Haroni, G. Werner, A. P. Roberts, V. Sorum, K. M. Nielsen and P. J. Johnsen (2013). "Fitness costs of various mobile genetic elements in Enterococcus faecium and Enterococcus faecalis." J Antimicrob Chemother **68**(12): 2755-2765.

Stewart, C. R., S. R. Casjens, S. G. Cresawn, J. M. Houtz, A. L. Smith, M. E. Ford, C. L. Peebles, G. F. Hatfull, R. W. Hendrix, W. M. Huang and M. L. Pedulla (2009). "The genome of Bacillus subtilis bacteriophage SPO1." J Mol Biol **388**(1): 48-70.

Stobberingh, E., A. van den Bogaard, N. London, C. Driessen, J. Top and R. Willems (1999). "Enterococci with glycopeptide resistance in turkeys, turkey farmers, turkey slaughterers, and (sub)urban residents in the south of The Netherlands: evidence for transmission of vancomycin resistance from animals to humans?" Antimicrob Agents Chemother **43**(9): 2215-2221.

Su, Z., B. Ning, H. Fang, H. Hong, R. Perkins, W. Tong and L. Shi (2011). "Next-generation sequencing and its applications in molecular diagnostics." Expert Rev Mol Diagn **11**(3): 333-343.

Subedi, A., C. Ubeda, R. P. Adhikari, J. R. Penades and R. P. Novick (2007). "Sequence analysis reveals genetic exchanges and intraspecific spread of SaPI2, a pathogenicity island involved in menstrual toxic shock." Microbiology **153**(Pt 10): 3235-3245.

Subedi, K. P., D. Choi, I. Kim, B. Min and C. Park (2011). "Hsp31 of Escherichia coli K-12 is glyoxalase III." Mol Microbiol **81**(4): 926-936.

Sueoka, N. (1962). "On the genetic basis of variation and heterogeneity of DNA base composition." Proc Natl Acad Sci U S A **48**: 582-592.

Sullivan, J. T., J. R. Trzebiatowski, R. W. Cruickshank, J. Gouzy, S. D. Brown, R. M. Elliot, D. J. Fleetwood, N. G. McCallum, U. Rossbach, G. S. Stuart, J. E. Weaver, R. J. Webby, F. J. De Bruijn and C. W. Ronson (2002). "Comparative sequence analysis of the symbiosis island of Mesorhizobium loti strain R7A." J Bacteriol **184**(11): 3086-3095.

Summers, A. O. (2006). "Genetic linkage and horizontal gene transfer, the roots of the antibiotic multi-resistance problem." Anim Biotechnol **17**(2): 125-135.

Tamames, J. (2001). "Evolution of gene order conservation in prokaryotes." Genome Biol **2**(6): RESEARCH0020.

Tang, C. and D. Holden (1999). "Pathogen virulence genes--implications for vaccines and drug therapy." Br Med Bull **55**(2): 387-400.

Tang, F., A. Bossers, F. Harders, C. Lu and H. Smith (2013). "Comparative genomic analysis of twelve Streptococcus suis (pro)phages." Genomics **101**(6): 336-344.

Teng, F., M. Kawalec, G. M. Weinstock, W. Hryniewicz and B. E. Murray (2003). "An Enterococcus faecium secreted antigen, SagA, exhibits broad-spectrum binding to extracellular matrix proteins and appears essential for E. faecium growth." Infect Immun **71**(9): 5033-5041.

Tenson, T., M. Lovmar and M. Ehrenberg (2003). "The mechanism of action of macrolides, lincosamides and streptogramin B reveals the nascent peptide exit path in the ribosome." J Mol Biol **330**(5): 1005-1014.

Tettelin, H. and T. Feldblyum (2009). Bacterial Genome Sequencing Molecular epidemiology of microorganisms. D. A. Caugant. New York, Humana Press. **551**.

Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. Deboy, T. M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli and C. M. Fraser (2005). "Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome"." Proc Natl Acad Sci U S A **102**(39): 13950-13955.

Tettelin, H., K. E. Nelson, I. T. Paulsen, J. A. Eisen, T. D. Read, S. Peterson, J. Heidelberg, R. T. DeBoy, D. H. Haft, R. J. Dodson, A. S. Durkin, M. Gwinn, J. F. Kolonay, W. C. Nelson, J. D. Peterson, L. A. Umayam, O. White, S. L. Salzberg, M. R. Lewis, D. Radune, E. Holtzapple, H. Khouri, A. M. Wolf, T. R. Utterback, C. L. Hansen, L. A. McDonald, T. V. Feldblyum, S. Angiuoli, T. Dickinson, E. K. Hickey, I. E. Holt, B. J. Loftus, F. Yang, H. O. Smith, J. C. Venter, B. A. Dougherty, D. A. Morrison, S. K. Hollingshead and C. M. Fraser (2001). "Complete genome sequence of a virulent isolate of Streptococcus pneumoniae." Science **293**(5529): 498-506.

Teuber, M., F. Schwarz and V. Perreten (2003). "Molecular structure and evolution of the conjugative multiresistance plasmid pRE25 of Enterococcus faecalis isolated from a raw-fermented sausage." Int J Food Microbiol **88**(2-3): 325-329.

Thomson, J. M. and R. A. Bonomo (2005). "The threat of antibiotic resistance in Gram-negative pathogenic bacteria: beta-lactams in peril!" Curr Opin Microbiol **8**(5): 518-524.

Top, J., R. Willems and M. Bonten (2008). "Emergence of CC17 Enterococcus faecium: from commensal to hospital-adapted pathogen." FEMS Immunol Med Microbiol **52**(3): 297-308.

Touchon, M., C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl, P. Bidet, E. Bingen, S. Bonacorsi, C. Bouchier, O. Bouvet, A. Calteau, H. Chiapello, O. Clermont, S. Cruveiller, A. Danchin, M. Diard, C. Dossat, M. E. Karoui, E. Frapy, L. Garry, J. M. Ghigo, A. M. Gilles, J. Johnson, C. Le Bouguenec, M. Lescat, S. Mangenot, V. Martinez-Jehanne, I. Matic, X. Nassif, S. Oztas, M. A. Petit, C. Pichon, Z. Rouy, C. S. Ruf, D. Schneider, J. Tourret, B. Vacherie, D. Vallenet, C. Medigue, E. P. Rocha and E. Denamur (2009). "Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths." PLoS Genet **5**(1): e1000344.

Touchon, M. and E. P. Rocha (2007). "Causes of insertion sequences abundance in prokaryotic genomes." Mol Biol Evol **24**(4): 969-981.

Tsiodras, S., H. S. Gold, E. P. Coakley, C. Wennersten, R. C. Moellering, Jr. and G. M. Eliopoulos (2000). "Diversity of domain V of 23S rRNA gene sequence in different Enterococcus species." J Clin Microbiol **38**(11): 3991-3993.

Ubeda, C., E. Maiques, P. Barry, A. Matthews, M. A. Tormo, I. Lasa, R. P. Novick and J. R. Penades (2008). "SaPI mutations affecting replication and transfer and enabling autonomous replication in the absence of helper phage." Mol Microbiol **67**(3): 493-503.

van Belkum, A., S. Scherer, L. van Alphen and H. Verbrugh (1998). "Short-sequence DNA repeats in prokaryotic genomes." Microbiol Mol Biol Rev **62**(2): 275-293.

van den Bogaard, A. E. and E. E. Stobberingh (1999). "Antibiotic usage in animals: impact on bacterial resistance and public health." Drugs **58**(4): 589-607.

van den Bogaard, A. E., R. Willems, N. London, J. Top and E. E. Stobberingh (2002). "Antibiotic resistance of faecal enterococci in poultry, poultry farmers and poultry slaughterers." J Antimicrob Chemother **49**(3): 497-505.

van Schaik, W., J. Top, D. R. Riley, J. Boekhorst, J. E. Vrijenhoek, C. M. Schapendonk, A. P. Hendrickx, I. J. Nijman, M. J. Bonten, H. Tettelin and R. J. Willems (2010). "Pyrosequencing-based comparative genome analysis of the nosocomial pathogen Enterococcus faecium and identification of a large transferable pathogenicity island." BMC Genomics **11**: 239.

van Sorge, N. M., J. N. Cole, K. Kuipers, A. Henningham, R. K. Aziz, A. Kasirer-Friede, L. Lin, E. T. Berends, M. R. Davies, G. Dougan, F. Zhang, S. Dahesh, L. Shaw, J. Gin, M. Cunningham, J. A. Merriman, J. Hutter, B. Lepenies, S. H. Rooijakkers, R. Malley, M. J. Walker, S. J. Shattil, P. M. Schlievert, B. Choudhury and V. Nizet (2014). "The classical lancefield

antigen of group a Streptococcus is a virulence determinant with implications for vaccine design." Cell Host Microbe **15**(6): 729-740.

Vankerckhoven, V., T. Van Autgaerden, C. Vael, C. Lammens, S. Chapelle, R. Rossi, D. Jabes and H. Goossens (2004). "Development of a multiplex PCR for the detection of asa1, gelE, cylA, esp, and hyl genes in enterococci and survey for virulence determinants among European hospital isolates of Enterococcus faecium." J Clin Microbiol **42**(10): 4473-4479.

Veesler, D. and C. Cambillau (2011). "A common evolutionary origin for tailed-bacteriophage functional modules and bacterial machineries." Microbiol Mol Biol Rev **75**(3): 423-433, first page of table of contents.

Vergis, E. N., N. Shankar, J. W. Chow, M. K. Hayden, D. R. Snydman, M. J. Zervos, P. K. Linden, M. M. Wagener and R. R. Muder (2002). "Association between the presence of enterococcal virulence factors gelatinase, hemolysin, and enterococcal surface protein and mortality among patients with bacteremia due to Enterococcus faecalis." Clin Infect Dis **35**(5): 570-575.

Walk, S. T., E. W. Alm, L. M. Calhoun, J. M. Mladonicky and T. S. Whittam (2007). "Genetic diversity and population structure of Escherichia coli isolated from freshwater beaches." Environ Microbiol **9**(9): 2274-2288.

Walsh, C. T., S. L. Fisher, I. S. Park, M. Prahalad and Z. Wu (1996). "Bacterial resistance to vancomycin: five genes and one missing hydrogen bond tell the story." Chem Biol **3**(1): 21-28.

Wang, I. N., D. L. Smith and R. Young (2000). "Holins: the protein clocks of bacteriophage infections." Annu Rev Microbiol **54**: 799-825.

Wang, X., Y. Kim, Q. Ma, S. H. Hong, K. Pokusaeva, J. M. Sturino and T. K. Wood (2010). "Cryptic prophages help bacteria cope with adverse environments." Nat Commun **1**: 147.

Wang, Y., W. Wang, Y. Lv, W. Zheng, Z. Mi, G. Pei, X. An, X. Xu, C. Han, J. Liu, C. Zhou and Y. Tong (2014). "Characterization and complete genome sequence analysis of novel bacteriophage IME-EFm1 infecting Enterococcus faecium." J Gen Virol **95**(Pt 11): 2565-2575.

Wegener, H. C., M. Madsen, N. Nielsen and F. M. Aarestrup (1997). "Isolation of vancomycin resistant Enterococcus faecium from food." Int J Food Microbiol **35**(1): 57-66.

Weiss, S. B., W. T. Hsu, J. W. Foft and N. H. Scherberg (1968). "Transfer RNA coded by the T4 bacteriophage genome." Proc Natl Acad Sci U S A **61**(1): 114-121.

Werner, G., C. Fleige, U. Geringer, W. van Schaik, I. Klare and W. Witte (2011). "IS element IS16 as a molecular screening tool to identify hospital-associated strains of Enterococcus faecium." BMC Infect Dis **11**: 80.

Werner, G., B. Hildebrandt and W. Witte (2001). "Aminoglycoside-streptothricin resistance gene cluster aadE-sat4-aphA-3 disseminated among multiresistant isolates of Enterococcus faecium." Antimicrob Agents Chemother **45**(11): 3267-3269.

Wicken, A. J., S. D. Elliott and J. Baddiley (1963). "The identity of streptococcal group D antigen with teichoic acid." J Gen Microbiol **31**: 231-239.

Willems, R. J., W. P. Hanage, D. E. Bessen and E. J. Feil (2011). "Population biology of Gram-positive pathogens: high-risk clones for

dissemination of antibiotic resistance." FEMS Microbiol Rev **35**(5): 872-900.

Willems, R. J., J. Top, N. van Den Braak, A. van Belkum, H. Endtz, D. Mevius, E. Stobberingh, A. van Den Bogaard and J. D. van Embden (2000). "Host specificity of vancomycin-resistant Enterococcus faecium." J Infect Dis **182**(3): 816-823.

Willems, R. J., J. Top, W. van Schaik, H. Leavis, M. Bonten, J. Siren, W. P. Hanage and J. Corander (2012). "Restricted gene flow among hospital subpopulations of Enterococcus faecium." MBio **3**(4): e00151-00112.

Willems, R. J. and W. van Schaik (2009). "Transition of Enterococcus faecium from commensal organism to nosocomial pathogen." Future Microbiol **4**(9): 1125-1135.

Willems, R. J. L., W. Homan, J. Top, M. van Santen-Verheuvel, D. Tribe, X. Manzioros, C. Gaillard, C. M. J. E. Vandenbroucke-Grauls, E. M. Mascini, E. van Kregten, J. D. A. van Embden and M. J. M. Bonten (2001). "Variant esp gene as a marker of a distinct genetic lineage of vancomycinresistant Enterococcus faecium spreading in hospitals." The Lancet **357**(9259): 853-855.

Wilson, J. H. (1973). "Function of the bacteriophage T4 transfer RNA's." J Mol Biol **74**(4): 753-757.

Wirth, R. (1994). "The sex pheromone system of Enterococcus faecalis. More than just a plasmid-collection mechanism?" Eur J Biochem **222**(2): 235-246.

Witte, W., R. Wirth and I. Klare (1999). "Enterococci." Chemotherapy **45**(2): 135-145.

Yao, Y., D. E. Sturdevant, A. Villaruz, L. Xu, Q. Gao and M. Otto (2005). "Factors characterizing Staphylococcus epidermidis invasiveness determined by comparative genomics." Infect Immun **73**(3): 1856-1860.

Yap, W. H., Z. Zhang and Y. Wang (1999). "Distinct types of rRNA operons exist in the genome of the actinomycete Thermomonospora chromogena and evidence for horizontal transfer of an entire rRNA operon." J Bacteriol **181**(17): 5201-5209.

Yasmin, A., J. G. Kenny, J. Shankar, A. C. Darby, N. Hall, C. Edwards and M. J. Horsburgh (2010). "Comparative genomics and transduction potential of Enterococcus faecalis temperate bacteriophages." J Bacteriol **192**(4): 1122-1130.

Zankari, E., H. Hasman, S. Cosentino, M. Vestergaard, S. Rasmussen, O. Lund, F. M. Aarestrup and M. V. Larsen (2012). "Identification of acquired antimicrobial resistance genes." J Antimicrob Chemother **67**(11): 2640-2644.

Zhang, X., F. L. Paganelli, D. Bierschenk, A. Kuipers, M. J. Bonten, R. J. Willems and W. van Schaik (2012). "Genome-wide identification of ampicillin resistance determinants in Enterococcus faecium." PLoS Genet **8**(6): e1002804.

Zhang, X., J. Top, M. de Been, D. Bierschenk, M. Rogers, M. Leendertse, M. J. Bonten, T. van der Poll, R. J. Willems and W. van Schaik (2013). "Identification of a genetic determinant in clinical Enterococcus faecium strains that contributes to intestinal colonization during antibiotic treatment." J Infect Dis **207**(11): 1780-1786.

Zhao, Z., E. Sagulenko, Z. Ding and P. J. Christie (2001). "Activities of virE1 and the VirE1 secretion chaperone in export of the multifunctional VirE2 effector via an Agrobacterium type IV secretion pathway." J Bacteriol **183**(13): 3855-3865.

Zhou, Y., Y. Liang, K. H. Lynch, J. J. Dennis and D. S. Wishart (2011). "PHAST: a fast phage search tool." Nucleic Acids Res **39**(Web Server issue): W347-352.

Appendix

Table 4.5: The novel regions in animal *E. faecium* genomes used in this study.

Region	Calf (E172)	Pig (E142)	Chicken (E429)
1	Tagatose and glucose utilisation operons and lipid carrier	Type I restriction-modification system restriction subunits R and M, site-specific recombination, two cell wall surface anchor family proteins and sortase A (LPXTG specific)	Copper uptake genes, heavy metal genes (lead, cadmium, zinc, and mercury transporting ATPase, IS elements (ISSdy1, Tn*916* and ISEc9), citrate fermentation, maltose utilisation operon, sucrose utilisation operon , several phage integrases, sortase A LPXTG specific, tetracycline resistance gene (*tetM*) and replication proteins (*repA*)
2	Mobile element proteins, site-specific recombinase (phage integrase family) and replication initiation factor	Polysaccharide biosynthesis proteins CpsF and CpsM and membrane protein involved in the export of O-antigen teichoic acid lipoteichoic acids	Prophage
3	Mobile element proteins, site-specific recombinase, integrase/recombunase core domain family, probable cadmium transporting ATPase (EC 3.6.3.3), transcriptional regulators (TetR and lclR family) and L-rhamonose utilisation operon	Prophage	Cluster for agmatine (decarboxylated arginine) catabolis
4	Lactose utilisation operon	Hypothetical membrane proteins, a sorbitol utilisation operon, hydrolase and a protease	Sugar transferase genes and genes encoding a membrane protein involved in the export of O-antigen teichoic acid, lipoteichoic acids and transposases IS204/IS1001/IS1096/IS1165
5	rRNA operon, cluster for agmatine (decarboxylated arginine) catabolism	Tetracycline resistance and Tn*916*	unique hypothetical proteins (11 genes) phage related integrases, ATP/GTP-binding proteins and DNA or RNA helicase of superfamily II
6	Membrane protein involved in the export of O-antigen teichoic acid lipoteichoic acids, capsular polysaccharide biosynthesis protein and beta lactamase	Membrane protein O-antigen, beta-lactamases and glycosyl transferase	Prophage
7	ascorbate utilisation operon and several transposases	Transcriptional regulators of the TetR and MerR family proteins, a putative hydrolase and six hypothetical proteins	Prophage
8	Several prophage genes, superinfection immunity protein, mobile element proteins and a several transposases	Protease IV (EC 3.4.21), a bacteriocin export accessory protein and an ABC transporter	A unique transposase, hydrolase, set-specific recombinases, integrase and membrane protein involved in the export of O-antigen teichoic acid, lipoteichoic acids

9	Prophage	Two sucrose utilisation operons	Tagatose and lactose utilisation operons, plasmid proteins, phage integrases and transposase IS204/IS1001/IS1096/IS1165
10	Cobalt-zinc-cadmium resistance proteins and lead, cadmium, zinc, mercury and copper translocating ATPase and multicopper oxidase. hypothetical proteins, plasmid genes (*repA, repB*) and carbohydrate (mannose, trehalose, ribose and sucrose) utilisation operon, IS elements, (LPXTG) cell wall surface anchor protein, sortase A, surface protein transpeptidase, extracellular proteins and antibiotic resistance genes such as vancomycin type A and B resistance operon, tetracycline resistance, beta lactamase and restriction-modification system	Cobalt-zinc- cadmium resistance proteins and lead, cadmium, zinc, mercury and copper translocating , plasmid genes (*repA, repB*), carbohydrate (mannose, trehalose, ribose and sucrose) utilisation operon, IS elements, (LPXTG) cell wall surface anchor protein, sortase A, surface protein transpeptidase, extracellular proteins and antibiotic resistance genes such as vancomycin type A and B resistance operon, tetracycline resistance, beta lactamase and restriction-modification system	Glutamate decarboxylase (EC4.1.1.15), glutamate/gamma- aminobutyrate antiporter
11	-	-	blue copper oxidase CueO precursor, iron-sulfur cluster assembly protein SufB, cadmium resistance proteins and lead, cadmium, zinc, mercury and copper translocating ATPase , hypothetical proteins, plasmid genes (*repA*), IS element, Tn*916*, cell wall surface anchor protein, sortase (surface protein transpeptidase), vancomycin type A and B resistance operon, tetracycline resistance genes, and extracellular proteins