American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

http://asrjetsjournal.org/

Comparative Genomics of Enterococcus *Faecium* Bacteriophages

Ashwag Shami[a]*, Malcolm Horsburgh [b]

[a]*Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia*
[b]*The university of Liverpool, Liverpool, UK*
[a]*Email: AYshami@pnu.edu.sa*
[b]*Email: M.J.Horsburgh@liverpool.ac.uk*

Abstract

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 by the end of 2014. Comparative analysis of the seprophages 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, terminaselarge 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.

Keywords: *E. faecium*; Prophages; *Enterococcus*; Comparative Genomics.

1. Introduction

Bacteriophages that infected *Enterococcus* species were first identified around 70 years ago [1,2]. 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. Recently, phages that infect and lysogenise *E. faecalis* and *E. faecium* have been more extensively characterised [3]. 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*.

---

* Corresponding author.

These phages have an isometric head about 40 nm in size and a long non-contractile tail, ranging from 70 nm to 220 nm [4]. 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 [5]. The first non-tailed enterococcal phages were isolated by Mazaheri Nezhad Fard and his colleagues in 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 [6,7].

Phages play an important role in adding to the genome plasticity of *E. faecium* species . 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* [6]. 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* [8]. 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. 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.

2. Methodology

2.1 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 [9]. To predict phage-related genes in each genome, Artemis and BLAST were used to compare genes against the PHAST database.

2.2 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.

3. Results and Dissection

3.1 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 1). 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.

3.2 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 1 and Figure 1).

A cladogram tree (Figure 1) 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 1). 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 1).

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 1).

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* [7]. 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.

Table 1: 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

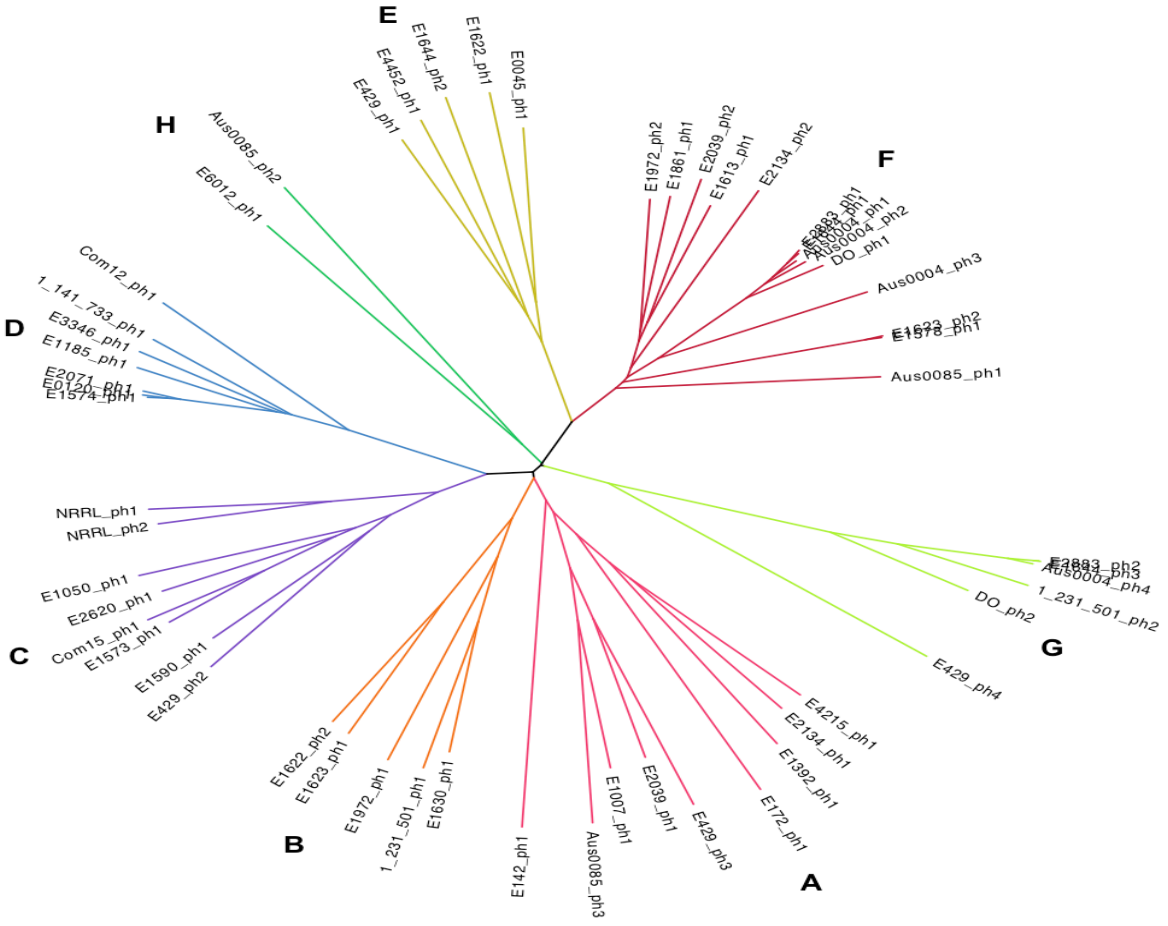

Figure 1: Cladogram tree of *E. faecium* prophages. The tree represents the cluster relationships for 56 *E. faecium* prophages present in the genomes of clinical, commensal, animal and food isolates.

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 [10]. 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 [11,12].

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 7). However, the protein sequences within clusters are highly conserved (Figure 2). Comparative analysis of 8 induced *E. faecalis* temperate phage identified by Yasmin and his colleagues (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 and his colleagues 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 [13]. Remarkably, *E. faecalis* polylysogeny has been described in a collection of clinical isolates, which carried up to 5 different inducible phages [7].

Protein similarities between the temperate *E. faecium* prophages suggested a low degree of similarity between the genomes at the nucleotide level (Figure 7). 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 [4].

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 2). 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 and his colleagues (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 [14]. 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, S1). Analyses performed by Villion and his colleagues (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 [15]. This observation was supported by Yasmin and his colleagues (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 [7].

3.3 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 2). 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 2).

Figure 2: 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.

3.4 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 [7,16]. 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 [17]. 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 [16,17].

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 [18].

A cladogram tree analysis generated using amino acid sequences of the integrases of the *E. faecium* prophage clusters (Figure 1 and Figure 2) identified multiple clades (Figure 3). 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. 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 1. For example cluster ORTHOMCL4499 and ORTHOMCL4377 were present in Integ2 and only ORTHOMCL2597 in Integ4.

The prophages represented by clusters (Figure 1) 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. faecium* hosts. The remaining prophages in clusters F have similar integrase sequence types (XerC family) (Integ1, 2) . Cluster C and D all have the same integrases sequence type (Integ2).

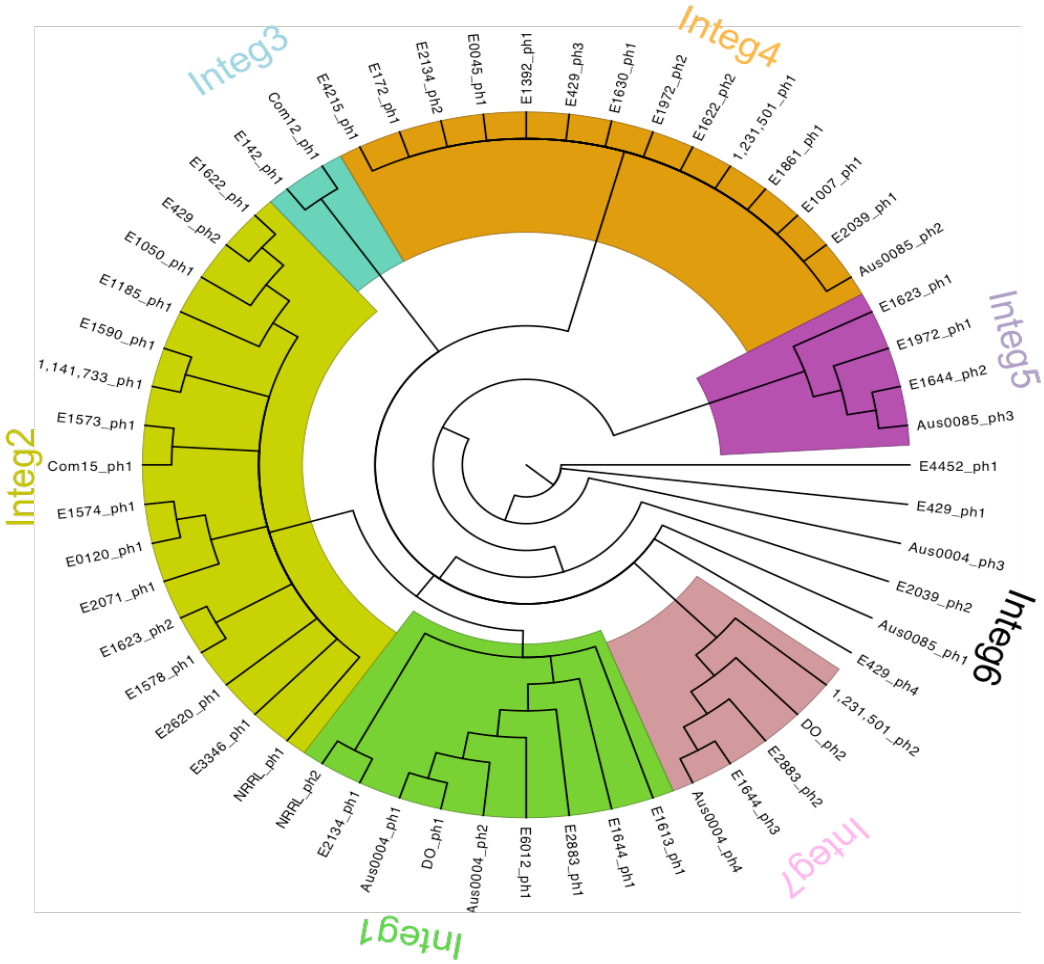

Figure 3: Cladogram tree of *E. faecium* prophage integrases. The cladogram is based on the alignment of integrases amino acid sequences and represents the relationship between *E. faecium* prophage integrases.

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 [18]. 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 [18,19].

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 3). The tyrosine recombinase family is common in *Streptococcus suis* prophages [16], Mycobacteriophage [20], *Listeria* prophages [18] and *Staphylococcus aureus* [21]. However, the integrases of *E. faecalis* were reported to be serine recombinase family members [7]. 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 [3] (Figure 3).

3.5 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, S1). 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 [16]. 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.

3.6 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 4). 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 3), prophages belonging to cluster C and D are highly similar but their terminases show marked variation (Figure 4). The portal protein gene was identified in 37 phage genomes but was not evident in nineteen.

Terminase is an enzyme necessary for the packaging of dsDNA into the progeny phages [22]. 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 [23]. 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 4). 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 [22] and a similar packaging mechanism can be inferred for most of the phage sequence types A-F, H.

3.7 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 5). These tail proteins were grouped into 7 different, that matched the clusters determined by supported the comparative genome analysis (Table 6.3 and Figure 1). 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.

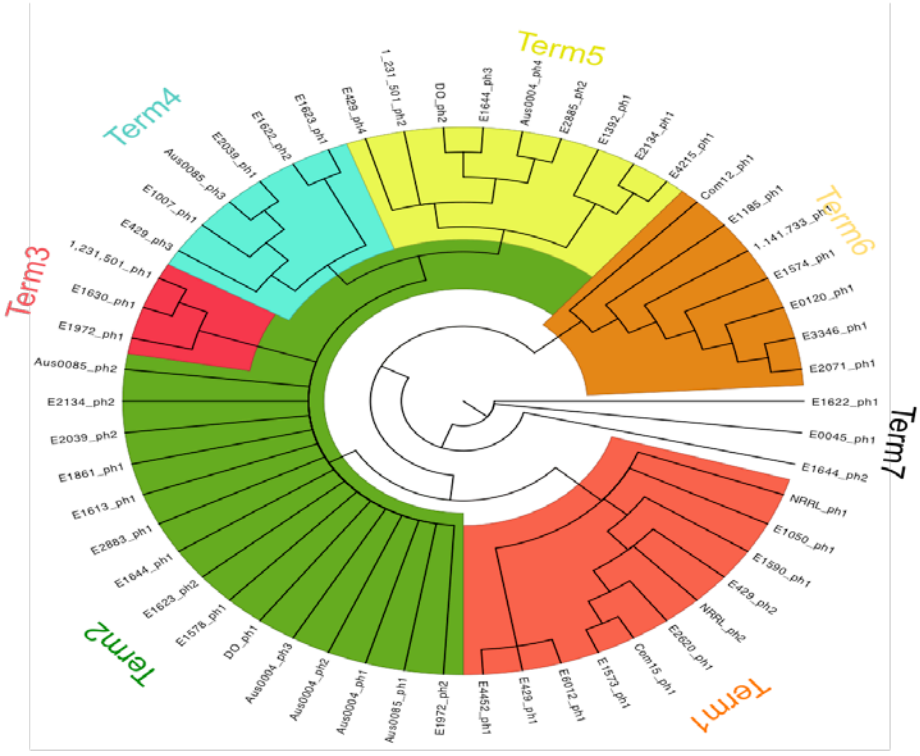

Figure 4: Cladogram tree of the large terminase subunits of *E. faecium* prophages. The tree is based on an alignment of the amino acid sequence of 54 terminases.

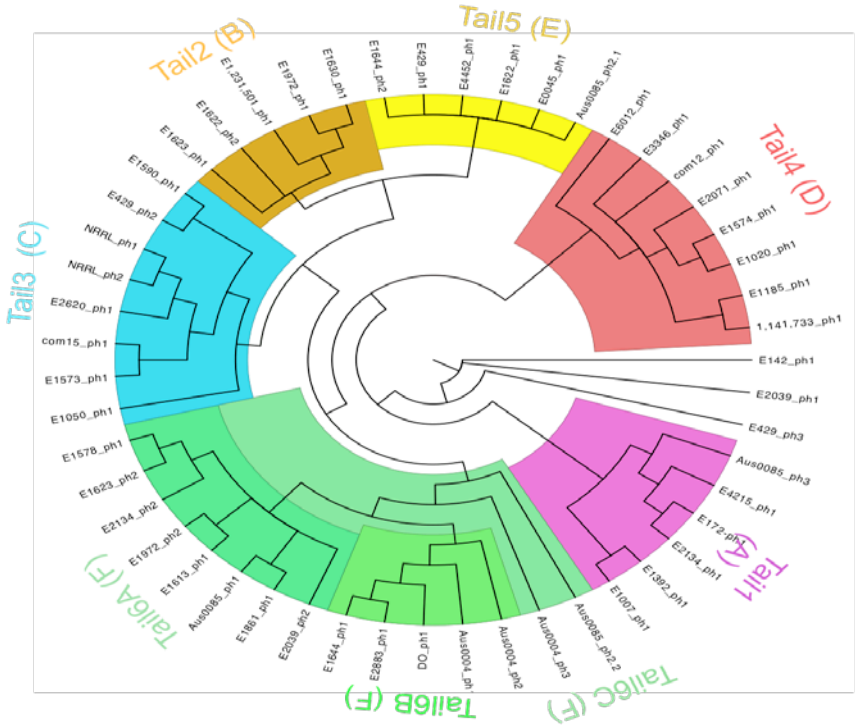

Figure 5: Cladogram tree of the tail protein of *E. faecium* prophages. The alignment of the amino acid sequence of 51 tail proteins reveals differences between *E. faecium* prophages producing distinct groupings.

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 [22]. 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, S1).

E. faecium prophage tail proteins indicate clear differences between the prophage clusters (Figure 5) 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) [24,25]. Genome sequences are not sufficient to definitively classify *E. faecium* prophage as *Siphoviridae* using electron microscopy will be required for confirmation.

3.8 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) [6].

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 . Three clades of *E. faecium* prophages seem to have a different sequence type of holin (Holin 1, 2 and 4) (Figure 6). 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).

The activity of endolysin and holin are significant factors for progeny phages to disrupt the host cell at the end of the lytic cycle [26]. 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 . 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. Holin-endolysin system are

typically used by bacteriophages with large genomes, while a single lysis protein is commonly used by bacteriophages with small genomes [26,27].

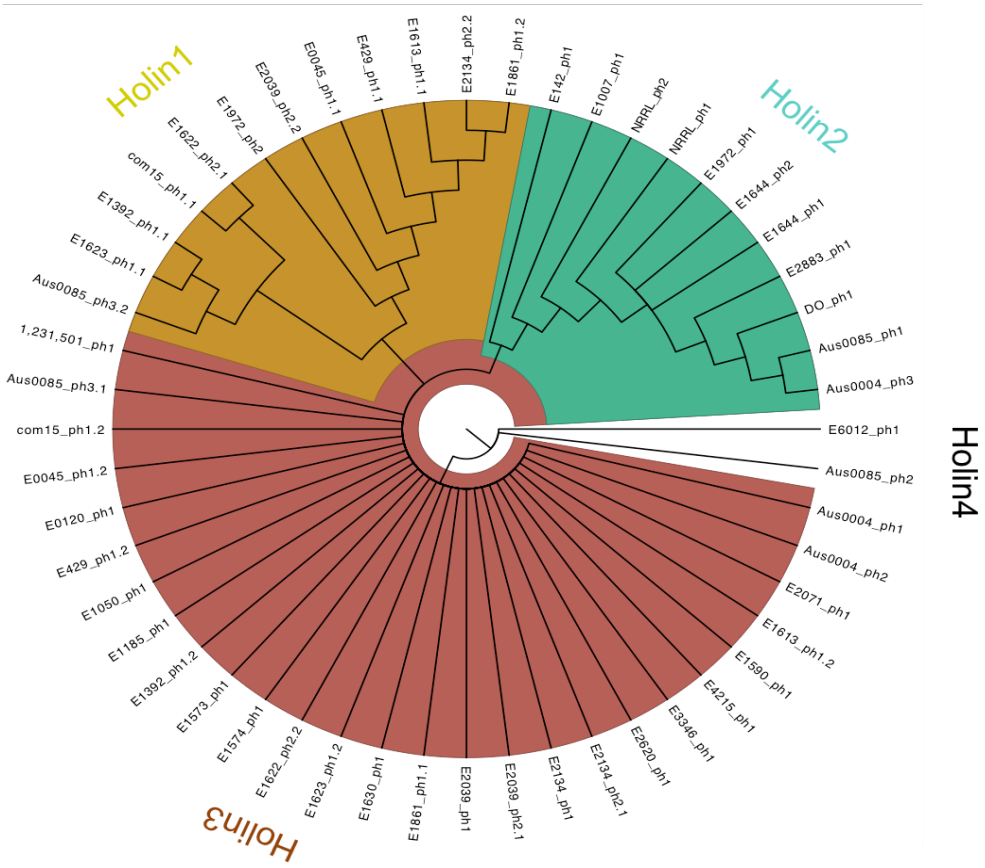

Figure 6: Cladogram tree of *E. faecium* prophage holins. Based on the alignment of 52 amino acid sequence of the holin protein, *E. faecium* prophages have 4 different families of holin. The Holin 4 protein sequences are nearly identical.

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, S1). Phage holins that form clade Holin1 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 and his colleagues (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 [25].

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 [28] 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 and his colleagues (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 [29].

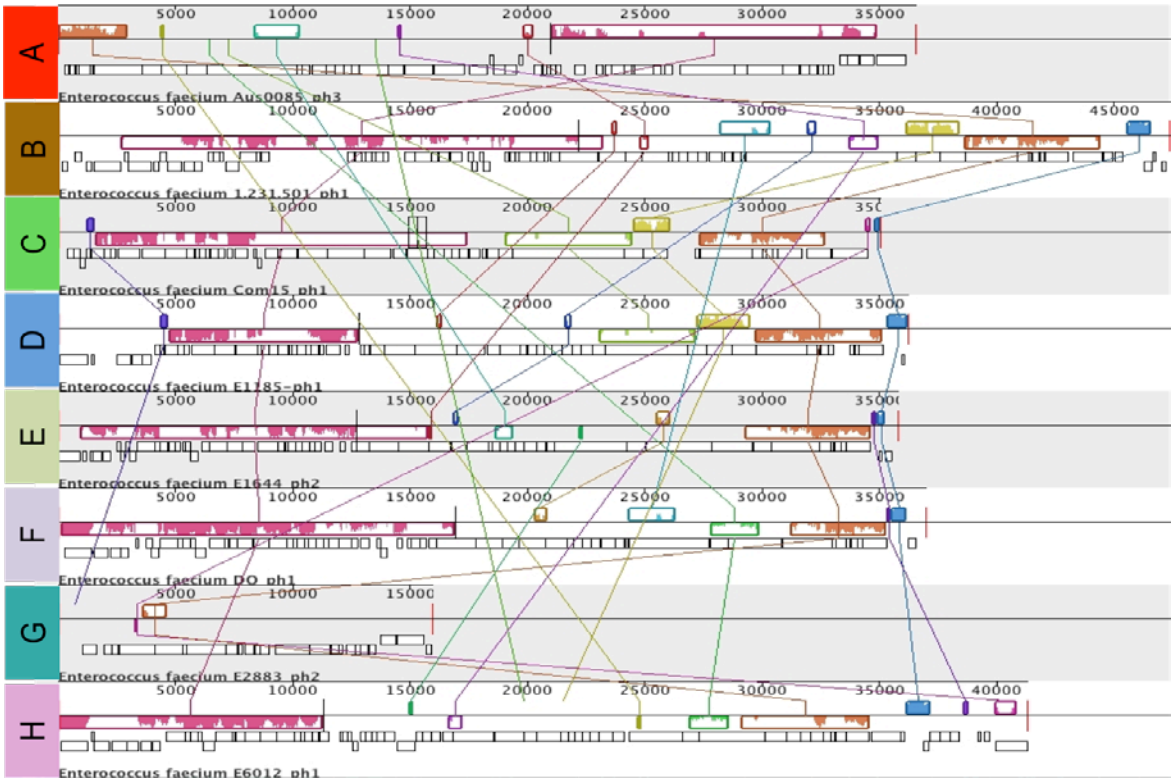

Figure 7: Mauve alignment of 9 *E. faecium* prophage type genomes. Pairwise alignment of one prophage genome of each of *E. faecium* pophage clusters A, B, C, D, E, F, G and H displays a low degree of similarity between the prophage genomes and highlighted diversity. The strength of the relationship is represented by coloured region.

3.9 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 7).

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.

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 and his colleagues (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 [14,30].

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 2). 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 7). 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.

3.10 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 2). 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 3). 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 and his colleagues (2013) [13]. 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, S2).

Table 2: 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 3: 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	-	-	-	+	+	-	-

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 [13, 31]. For example, Phage Related Chromosomal Islands (PRCIs) of several Gram-positive bacteria are mobile genetic elements, primarily defined as *S. aureus* pathogenicity islands (SaPIs). Infection by a helper phage or by induction of an endogenous prophage drives excision of SaPIs from the bacterial chromosome [13, 32].

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 and his colleagues (2013) [13]. 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.

4. Conclusion

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 study of mobile genomics elements (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 [9]. The narrow sequence homology among functionally equivalent phage-encoded proteins complicates the study of their function [33]. There is a requirement for developments in bioinformatics of MGEs to identify their unique features.

References

[1] Clark, P. and Clark, A. A Bacteriophage Active Against a Virulent Hemolytic Streptococcus. Experimental Biology and Medicine, 24(7), pp.635-639(1927).

[2] Evans, A. The Prevalence Of Streptococcus Bacteriophage. Science, 80(2063), pp.40-41 (1934).

[3] Duerkop BA, Palmer KL, Horsburgh MJ. Enterococcal Bacteriophages and Genome Defense. Feb 11. In: Gilmore MS, Clewell DB, Ike Y, et al., editors. Enterococci: From Commensals to Leading Causes of Drug Resistant Infection [Internet]. Boston: Massachusetts Eye and Ear Infirmary; 2014-. Available from: http://www.ncbi.nlm.nih.gov/books/NBK190419/ (2014)

[4] 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. "Pyrosequencing-based comparative genome analysis of the nosocomial pathogen Enterococcus faecium and identification of a large transferable pathogenicity island." BMC Genomics **11**: 239 (2010).

[5] Brede, D., Snipen, L., Ussery, D., Nederbragt, A. and Nes, I. Complete Genome Sequence of the Commensal Enterococcus faecalis 62, Isolated from a Healthy Norwegian Infant. Journal of Bacteriology, 193(9), pp.2377-2378 (2011).

[6] Mazaheri Nezhad Fard, R., M. D. Barton and M. W. Heuzenroeder. "Novel Bacteriophages in Enterococcus spp." Curr Microbiol **60**(6): 400-406 (2010).

[7] Yasmin, A., J. G. Kenny, J. Shankar, A. C. Darby, N. Hall, C. Edwards and M. J. Horsburgh. "Comparative genomics and transduction potential of Enterococcus faecalis temperate bacteriophages." J Bacteriol **192**(4): 1122-1130 (2010).

[8] Galloway-Pena, J., J. H. Roh, M. Latorre, X. Qin and B. E. Murray. "Genomic and SNP analyses demonstrate a distant separation of the hospital and community-associated clades of Enterococcus faecium." PLoS One **7**(1): e30187 (2012).

[9] 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. "Characterization and complete genome sequence analysis of novel bacteriophage IME-EFm1 infecting Enterococcus faecium." J Gen Virol **95**(Pt 11): 2565-2575 (2014).

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

[11] Rohwer, F. and Edwards, R.. The Phage Proteomic Tree: a Genome-Based Taxonomy for Phage. Journal of Bacteriology, 184(16), pp.4529-4535 (2002).

[12] Lima-Mendez, G., J. Van Helden, A. Toussaint and R. Leplae. "Reticulate representation of evolutionary and functional relationships between phage genomes." Mol Biol Evol **25**(4): 762-777

(2008).

[13] Matos, R. C., N. Lapaque, L. Rigottier-Gois, L. Debarbieux, T. Meylheuc, B. Gonzalez-Zorn, F. Repoila, F. Lopes Mde and P. Serror. "Enterococcus faecalis prophage dynamics and contributions to pathogenic traits." PLoS Genet **9**(6): e1003539 (2013).

[14] Juhala, R. J., M. E. Ford, R. L. Duda, A. Youlton, G. F. Hatfull and R. W. Hendrix. "Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages." J Mol Biol **299**(1): 27-51 (2000).

[15] Villion, M. Bacteriophages of Lactobacillus. Frontiers in Bioscience, Volume(14), p.1661 (2009).

[16] Tang, F., A. Bossers, F. Harders, C. Lu and H. Smith. "Comparative genomic analysis of twelve Streptococcus suis (pro)phages." Genomics **101**(6): 336-344(2013).

[17] Johnson, A., B. J. Meyer and M. Ptashne. "Mechanism of action of the cro protein of bacteriophage lambda." Proc Natl Acad Sci U S A **75**(4): 1783-1787(1978).

[18] Groth, A. and Calos, M. Phage Integrases: Biology and Applications. Journal of Molecular Biology, 335(3), pp.667-678 (2004).

[19] Park, M. O., K. H. Lim, T. H. Kim and H. I. Chang. "Characterization of site-specific recombination by the integrase MJ1 from enterococcal bacteriophage phiFC1." J Microbiol Biotechnol **17**(2): 342-347 (2007).

[20] 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. "Comparative genomic analysis of 60 Mycobacteriophage genomes: genome clustering, gene acquisition, and gene size." J Mol Biol **397**(1): 119-143 (2010).

[21] Goerke, C., R. Pantucek, S. Holtfreter, B. Schulte, M. Zink, D. Grumann, B. M. Broker, J. Doskar and C. Wolz . "Diversity of prophages in dominant Staphylococcus aureus clonal lineages." J Bacteriol **191**(11): 3462-3468(2009).

[22] Kutter, E. and A. Sulakvelidze. Bacteriophages: Biology and Applications: Molecular Biology and Applications. Boca Raton, CRC Press (2005).

[23] Hamada, K., Fujisawa, H. and Minagawa, T. A defined in vitro system for packaging of bacteriophage T3 DNA. Virology, 151(1), pp.119-123 (1986).

[24] Veesler, D. and Cambillau, C. A Common Evolutionary Origin for Tailed-Bacteriophage Functional Modules and Bacterial Machineries. Microbiology and Molecular Biology Reviews, 75(3), pp.423-433 (2011).

[25] Fokine, A. and Rossmann, MMolecular architecture of tailed double-stranded DNA phages. Bacteriophage, 4(2), p.e28281 (2014).

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

[27] Wang, I. N., D. L. Smith and R. Young (2000). "Holins: the protein clocks of bacteriophage infections." Annu Rev Microbiol 54: 799-825 (2002).

[28] Canchaya, C., Proux, C., Fournous, G., Bruttin, A. and Brussow, H. Prophage Genomics. Microbiology

and Molecular Biology Reviews, 67(3), pp.473-473 2003).

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

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

[31] Wang, X., Y. Kim, Q. Ma, S. H. Hong, K. Pokusaeva, J. M. Sturino and T. K. Wood "Cryptic prophages help bacteria cope with adverse environments." Nat Commun **1**: 147(2010).

[32] Ubeda, C., E. Maiques, P. Barry, A. Matthews, M. A. Tormo, I. Lasa, R. P. Novick and J. R. Penades."SaPI mutations affecting replication and transfer and enabling autonomous replication in the absence of helper phage." Mol Microbiol **67**(3): 493-503(2008)..

[33] Pedulla, M., Ford, M., Karthikeyan, T., Houtz, J., Hendrix, R., Hatfull, G., Poteete, A., Gilcrease, E., Winn-Stapley, D. and Casjens, S. Corrected Sequence of the Bacteriophage P22 Genome. Journal of Bacteriology, 185(4), pp.1475-1477 (2003).