final report

Project code: B.CCH.1025

Prepared by: Rosalind Gilbert, Diane Ouwerkerk and Athol Klieve
The University of Queensland and Department of Employment, Economic Development and Innovation

Date published: April 2012

PUBLISHED BY
Meat & Livestock Australia Limited
Locked Bag 991
NORTH SYDNEY NSW 2059

A genomic strategy to identify archaeal viruses in the rumen

Meat & Livestock Australia acknowledges the matching funds provided by the Australian Government to support the research and development detailed in this publication.

Abstract

Phage therapy is becoming increasingly important as a means of eradicating or controlling microbial populations and has been raised as a potential strategy to reduce methane emissions from ruminants. In this one-year project, the viral metagenome (virome) of the rumen was investigated. Virus particles were successfully purified and concentrated from relatively small volumes of bovine rumen fluid and sufficient quantities of viral DNA obtained to facilitate high-throughput sequencing, without the inclusion of additional amplification steps. Results of sequence analysis indicated that the rumen contains a highly diverse virus population, predominated by double-stranded DNA viruses (tailed bacteriophages of the order Caudovirales), with evidence of some changes occurring in the viral population in relation to the time of feeding. Viral marker genes (terminase large subunit) associated with previously sequenced archaeaphage (including the Siphoviruses, *Methanobacterium* phage ψM2 and *Methanothermobacter* phage ψM100) were detected within the virome dataset. Characterisation of the corresponding rumen microbial populations using a 16S rRNA gene approach also indicated the types of methanogens which may be infected or associated with these archaeaphage. Utilising a metagenomic approach has provided an ideal platform for identifying archaeaphage present within the rumen virome and has provided an insight into the advantages of adopting this approach for future studies examining the phage and viral populations of the rumen.

Executive Summary

Phage therapy strategies to control or eradicate microbial populations usually employ naturally occurring viruses, termed lytic phage, which specifically infect and eliminate the target (host) micro-organism. To date, very little is known about phages which may infect the methane-producing archaeal strains (methanogens) dominant within the rumen of Australian cattle and very few lytic phages able to infect methanogens (archaeaphage) have been isolated. Traditional culture-based attempts to isolate archaeaphage for rumen methanogens have proven to be time-consuming and largely unsuccessful. A culture-independent, metagenomic approach based on high through-put sequencing of all the viral genetic material (virome) present within the rumen would greatly increase our understanding and facilitate the detection of phages related to known, previously sequenced archaeaphages.

This report describes the work undertaken in a preliminary, one-year project to develop methodology and determine the structural and functional characteristics of the rumen virome, using a metagenomic approach. Techniques were developed for concentrating and purifying virus particles from rumen fluid samples and for extracting viral nucleic acids of sufficient quality and quantity to enable high through-put sequencing. Following the establishment of these techniques within our laboratory, an experiment was undertaken to characterise the virome and associated microbial populations of the rumen in a single steer (Brahman cross) and determine any changes which may occur in relation to the time after feeding, with rumen fluid samples obtained at feeding then at 4, 8 and 12 hours after feeding. Despite delays in obtaining sequencing data from the external sequencing provider, high through-put sequencing data was successfully obtained for two rumen virome samples (encompassing rumen DNA virus populations), four rumen bacterial population samples (based on bacterial 16S rRNA gene amplicon sequencing) and four rumen archaeal population samples (based on archaeal 16S rRNA gene amplicon sequencing).

Analysis of this sequence data was undertaken and the characteristics of the two rumen viromes and the directly associated microbial populations (bacterial and archaeal) determined. Both rumen viromes contained large proportions of sequences that could not be classified using existing sequence databases however the number of sequences obtained was sufficient to enable virus population identity to be predicted and individual sequences related (e-value threshold of 10^{-5}) to eight known archaeaphage to be identified within the larger virome sequence dataset (including sequences related to *Methanothermobacter* phage ψM100, *Natrialba* phage φCh1, *Halorubrum* phage HF2 and *Methanobacterium* phage ψM2).

In agreement with the only other published rumen virome study (Berg Miller *et al*. 2011), the virus populations were dominated by diverse populations of double-stranded DNA, tailed bacteriophages belonging to the virus order, Caudovirales. Both rumen viromes contained similar proportions of related phage families, however there were distinct variations in the proportion of individual phage populations occurring in response to time after feeding. Predominant phage types represented within the sequence data also included phages with host species identified as being highly dominant within the respective rumen bacterial populations. As a consequence of the Lucerne and Rhodes Grass diet being fed to the steer employed in the study, rumen bacterial populations were dominated by Firmicutes belonging to the order Clostridiales. In the 12 h virome dataset, for

example, viral sequences corresponding to 10 different *Clostridium* phages were detected, six of which had an abundance of > 1 % of the total virome population.

Taxonomic analysis of the two rumen virome sequence datasets confirmed the very high abundance of lysogenic phages, with a very large proportion of sequences obtained from the virus preparations (which were essentially free of microbial DNA contamination), directly attributed to genes also found within microbial genomes included in microbial sequence databases. For example, in the 4 h sample time point virome, which contained a total of 69,133 sequences, 25.25 % of sequences were related to bacterial genes and 0.15 % of sequences were related to archaeal genes. This confirms the findings of the only other rumen virome study (Berg Miller *et al.* 2011) and supports the previous culture-based studies which first suggested that the activity of lysogenic phages may play a major role in sustaining rumen phage populations and may actively contribute to horizontal gene transfer within the rumen microbial ecosystem (Klieve *et al*. 1989).

Interestingly, sequence data obtained during the course of this study for the bacterial and archaeal populations of the rumen contributed new information regarding the changes in microbial populations occurring in response to the time after feeding. These time-dependant changes in rumen microbial populations are most likely to be in direct response to the changing availability of substrates for methanogen growth. This was particularly evident for the archaeal community, given the lesser abundance and complexity of this microbial domain within the rumen, compared to the bacterial population. Despite populations of *Methanobrevibacter* being dominant at all sample time-points, the methanogen population increased in diversity over time, with increasing abundance of populations such as *Methanomicrobium*, *Thermogymnomonas* and *Methanimicrococcus*.

Using a metagenomic analysis approach has greatly increased our understanding of the biology of methanogens and their respective phages. It is anticipated that the methodology developed during the course of this project may be utilised for the analysis of rumen fluid samples collected from larger scale animal trials. It may also be employed for the metagenomic analysis of environmental samples suspected of containing relatively high concentrations of archaeaphage. This would facilitate the identification of samples which may be more conducive to the isolation of archaeaphage and archaeaphage-encoded gene products for future phage therapy-based methane reduction strategies.

Table of Contents

1.0 Background

Viruses, most of which infect micro-organisms, are the most abundant biological entities on earth (Edwards and Rohwer, 2005).The study of the viruses present in the rumen is complicated due to the fact that a single phylogenetic marker (like the 16S rRNA gene found in all bacteria and archaea) does not exist. This means the diversity and evolution of the viruses (including phages) present in the rumen can not be examined as readily as the microbial community.

Early studies of the rumen viral community employed electron microscopy to survey the types of viruses present on the basis of virus particle morphology and established that the rumen contained a dense, highly diverse population of viruses, predominated by tailed bacteriophages (Adams et al. 1966; Hoogenraad *et al*. 1967; Klieve and Bauchop, 1988). A later study by Klieve and Swain (1993) developed a Pulsed Field Gel Electrophoresis (PFGE) method to estimate the number of phage present in the rumen of sheep and established that the rumen contains a dynamic phage population that is maintained at high numbers by a significant, continual lysis of ruminal micro-organisms. Further investigations using PFGE methodology (Swain *et al*. 1996) were undertaken to determine the impact of nutritional and environmental factors on phage activity in the rumen of sheep, finding that phage populations followed a diurnal pattern in animals fed once daily. PFGE analysis however, did not allow individual phage or virus genes to be identified.

Metagenomic analysis of uncultured viral communities overcomes the limitations of PFGE, providing information and insights into the composition and structure of the viral communities present in the rumen. In addition, the rapid development of high through-put, next generation sequencing technologies has resulted in a marked decrease in the cost of sequencing, which enables large scale sequencing strategies to be undertaken.

Recent metagenomic studies reported in the literature, including the human gut microbiome and cow rumen microbiome (Qin *et al*. 2010; Hess *et al*. 2011), have greatly facilitated the study of phylogenetic composition and functional potential of complex microbial communities. To date however, there is only one published report describing a metagenomic analysis of the rumen virome (Berg Miller, 2011). In this study, the rumen virome was examined using samples obtained from 3 dairy (Holstein) cows, including one lactating cow. The methodology used to isolate viruses from the rumen fluid was based on the methods of Klieve and Swain (1993). This methodology was suited to the analysis of phage DNA by PFGE and enabled the extraction of viral DNA of sufficient purity for virome analysis, however the viral DNA was found to be contaminated with microbial DNA, as verified by 16S rRNA gene PCR. The current investigation has therefore addressed this methodology, introducing additional steps into the protocol used for the preparation of rumen virus DNA, to increase the yield of virus particles from rumen fluid and eliminate and check for contaminating microbial DNA.

In contrast to the previously published study (Berg Miller, 2011), the reported study examined the virome of multiple rumen fluid samples collected from a single steer over the course of a 12 h period. The study also used 16S rRNA gene sequence-based methods to identify the bacterial and archaeal (methanogen) communities directly associated with the rumen virus populations, enabling diurnal changes in the virome and associated microbial communities to be examined, an approach not previously reported in the literature.

This report encompasses results from three aspects of the one-year project:

(1) Development of methods to isolate and concentrate viral particles from rumen fluid and extract viral DNA. Method development also involved determining the volume of rumen fluid required in order to provide sufficient quantities of viral DNA to enable virus community analysis using a 454 pyrosequencing platform.

(2) An *in vivo* time-course experiment was undertaken to determine how viruses within the rumen influence the bacterial and archaeal (methanogen) populations. Results presented include an initial determination of the structural and functional attributes of the rumen virome, specifically the DNA virus populations corresponding to two rumen fluid collection time-points. Changes in the corresponding rumen bacterial and archaeal communities were also determined following analysis of 16S rRNA gene datasets, with distinct diurnal fluctuations observed for several microbial populations, including methanogenic archaea, over a 12 h time period after feeding.

(3) A selection of seven individual phages from the Rumen Ecology Group phage culture collection, able to establish specific, lytic infection of the rumen bacterial strains, *Streptococcus bovis* 2B, *Prevotella ruminicola* ss *brevis* AR20 or *Ruminicoccus albus* AR67, were propagated with their respective host bacteria and examined by electron microscopy to confirm virus morphology and classification. Phage genomic DNA was then prepared in sufficient quantity for whole genome sequencing using a 454-pyrosequencing platform. This sequence data, still to be obtained from the external sequencing provider, will contribute to our understanding of the genetics of rumen phage, providing genome sequence data specific to phages from the rumen environment, currently lacking within the virus sequence reference databases. In addition, an initial sequence annotation of an inducible, lysogenic phage of the rumen bacteria *Prevotella ruminicola* ss *brevis* AR29 (φAR29) has been undertaken, to establish the bioinformatics capacity for phage annotation within our research group.

The laboratory-based and bioinformatics methodology developed during the course of this project may be applied to the examination of virus populations in rumen fluid samples obtained from larger scale experiments and has greatly advanced the current understanding of the bovine rumen virome and associated microbial populations. It has successfully utilised metagenomic analysis tools to establish the presence of viruses related to known archaeaphages within bovine rumen fluid samples, through the identification of key marker genes. Marker genes examined include those involved in virus particle processing and packaging which are highly conserved between viral groupings for example, the *Caudovirales*-specific terminase large subunit (TerL) gene (Sullivan *et al.* 2009). The Eukaryote, archaeal and large DNA virus-specific, DNA polymerase family B (PolB) gene (Monnier *et al*. 2008) was also utilised as an effective marker gene for the identification of archaeaphage-related sequences contained within the larger virome dataset.

Utilising a metagenomic, whole-virome approach, followed by sequence-specific analysis incorporating the use of conserved phage marker genes, has provided an ideal platform for identifying viruses related to known archaeaphage in rumen fluid samples. It has also provided an insight into the technical requirements of adopting this approach for future studies examining the phage and viral populations of the rumen.

2.0 Project Objective

To gain a better understanding of ruminal archaeaphage to enhance the ability to recognise and isolate these viruses for use in an archaeaphage therapy strategy.

3.0 Methods

3.1 *Method Development: Preparation of viral nucleic acid from rumen fluid*

An initial experiment was undertaken to optimise viral nucleic acid extraction methods and determine the volume of rumen fluid needed to provide sufficient viral particles for DNA and RNA extractions and subsequent 454 pyrosequencing. Rumen fluid was obtained from a single steer fitted with a rumen fistula, maintained on a diet consisting of 4 kg Lucerne Hay (85-90 % dry matter[DM]) and 4 kg Rhodes Grass Hay (90 % DM). DNA from virus particles was extracted using the methods outlined by Klieve and Gilbert (2005).

Additional steps were also introduced in an attempt to maximise viral DNA yield and purity as outlined in Table 1. Each of the eight different experimental conditions was tested with a 5 ml volume of rumen fluid. One of the main experimental conditions examined was the pre-incubation of rumen fluid at 70 ˚C for 20 min. A step originally included in rumen phage preparations in order to eliminate nuclease activity (Klieve & Swain, 1993). The bovine rumen fluid was then clarified by centrifugation to remove plant debris and microbial cells removed by successive filtrations through 0.45 µm and 0.22 µm low protein binding filter units (Millipore). The absence of microbial cells was confirmed by microscopy and the absence of microbial growth following inoculation of the 0.22 µm filtered bovine rumen fluid into anaerobic and aerobic Rumen Fluid (RF) media (Klieve *et al.* 1989), incubated at 37 ˚C for 7 days.

Other additional steps examined included the treatment of filtered rumen fluid with RNase-free DNase (Qiagen) to degrade contaminating microbial DNA and the addition of PEG 8000 and NaCl to final concentrations of 10 % (w/v) and 0.5 M respectively, to enhance virus particle precipitation. Virus particles were then concentrated by ultracentrifugation at 23 600 rpm at 4 ˚C for 2 h (Optima L-100 XP Ultracentrifuge, Beckman Coulter), and the virus pellet resuspended in TE buffer (pH 7.6). The suspension was treated with RNase before the virus particle protein coat was removed by the addition of proteinase K (Klieve and Gilbert, 2005). The DNA was then extracted by repeated phenol: chloroform: isoamylalcohol (25:24:1) extractions and precipitated using 100 % ethanol and sodium acetate. Viral DNA concentrations were quantified by agarose gel electrophoresis and spectroscopy (Nanodrop 8000, Thermo Scientific).

Table 1. Method steps tested to maximise viral DNA yield and purity (treatment employed [+])

Test sample	*70°C pre-incubation of rumen fluid*	*DNase treatment*	*PEG 8000/NaCl precipitation*
1	+	+	+
2	+	+	
3	+		+
4	+		
5		+	+
6		+	
7			+
8			

3.2 *Time course experiment: Diurnal fluctuations in the virome and corresponding microbial populations*

Rumen fluid was collected from a single steer (commercial Brahman cross breed, approximately 75 % *Bos indicus*) fitted with a rumen fistula, housed individually within a feedlot pen and maintained on a Lucerne and Rhodes grass hay diet (detailed above) fed once daily. Rumen fluid was collected via aspiration from the rumen fistula prior to feeding (t= 0 h) then at 4, 8 and 12 h after feeding. Immediately after collection, the rumen fluid was placed on ice and for the purposes of virome analysis, stored at -20 °C. At each collection time-point, rumen fluid obtained for the purposes of microbial community 16S rRNA gene analysis was sub-sampled into two replicate 1 ml volumes, centrifuged at 11 000 x *g* for 10 min, the supernatant discarded and the cell pellet stored at -20°C.

Preparation of viral community DNA

For each rumen fluid collection time-point, virome DNA was prepared from a 30 ml volume of rumen fluid using the methods described above, excluding the pre-incubation at 70 °C. DNA for virome analysis was quantified by gel electrophoresis and spectroscopy using a Nanodrop 8000 and a Cubit quantification system (Invitrogen). The absence of contaminating microbial DNA was confirmed by PCR testing using universal bacterial 16S rRNA gene primers (Lane, 1991). Viral DNA was sequenced directly by an external sequencing provider (Australian Genome Research Facility Ltd) using the Genome Sequencer GS FLX Titanium platform (Roche).

Preparation of viral community RNA

Viral particles to be used for RNA extractions from each rumen fluid collection time-point were separated from a 30 ml volume of rumen fluid, concurrently with those used for virome DNA preparation. This involved clarification and filtration of the rumen fluid, followed by concentration of viral particles by ultracentrifugation and treatment with RNase and DNase as outlined above. Viral RNA was extracted using the QIAamp Viral RNA Mini Kit (Qiagen) using the protocols outlined by the manufacturer. The resulting products were treated with a mixture of DNases

including Turbo DNase (Applied Biosciences) and Baseline Zero (Epicentre), then purified by phenol: chloroform: isoamylalcohol extraction. Viral RNA concentrations were quantified by agarose gel electrophoresis and spectroscopy (Nanodrop 8000). The ThermoScript RT-PCR system (Invitrogen) was then used to construct an initial cDNA transcript from the RNA. Transcript cDNA was then amplified using random hexamers, dNTPs and Phi29 DNA polymerase (GenomiPhi kit, GE). Resultant DNA was then purified using a DNA quick-spin column (Qiagen) and final DNA concentrations determined by agarose gel electrophoresis and spectroscopy (Nanodrop 8000). The absence of contaminating microbial DNA was confirmed by PCR testing using universal bacterial 16S rRNA gene primers (Lane, 1991). Final DNA preparations were sequenced directly on a quarter section of a 454-pyrosequencing plate by an external sequencing provider (Australian Genome Research Facility Ltd) using the Genome Sequencer GS FLX Titanium platform (Roche).

Bacterial and Archaeal community analysis

DNA for bacterial and archaeal community analyses at each time point (t= 0, 4, 8, 12 h) were prepared using a modification of the method of Yu & Forster (2005). Briefly, this involved cell lysis by repeated bead beating in TE buffer with SDS and NaCl at final concentrations of 2 % (w/v) and 250 mM respectively. This was followed by ammonium acetate treatment to remove unwanted proteins and polysaccharides, RNase treatment and purification using the QIAamp DNA kit (Qiagen) using the protocol outlined by the manufacturer.

PCR was used to amplify the respective 16S rRNA genes of the bacterial and archaeal communities for amplicon library construction and subsequent pyrosequencing. For each rumen fluid sample, four sets of primers with specific identifying barcodes (Appendix 2, Table A2.1.) were used for the amplification of partial bacterial 16S rRNA genes from nt position: 341 – 787 (DasSarma and Fleischmann, 1995; Watanabe *et al*. 2001). In addition, for each rumen fluid sample, one set of barcoded primers (Table A2.1) was used to amplify partial archaeal 16S rRNA genes from nt position: 340 – 1000 (Gantner *et al.* 2010). Gel extraction purification of PCR products was performed using the QIAquick Gel Extraction kit using the protocol outlined by the manufacturer. DNA concentrations of purified PCR products were quantified by agarose gel electrophoresis and spectroscopy using the Qubit quantification system (Invitrogen).

Final PCR product preparations were sequenced directly on a quarter section of a 454-pyrosequencing plate by an external sequencing provider (Australian Genome Research Facility Ltd) using the Genome Sequencer GS FLX Titanium platform (Roche).

Analysis of virome data, DNA viruses

Virus DNA preparations from two rumen fluid collection time-points were sent for 454-pyrosequencing and each respective dataset analysed using the same programs and corresponding parameters. Sequences were trimmed and quality checked to remove sequence artifacts (quality score: quality window average 25, quality window size 50 bp, minimum length 100bp, maximum ambiguity 0, maximum homop 10) using mothur (Schloss *et al*. 2009).

Functional attributes of the rumen virome were determined using the metagenomic analysis tools employed by MG-RAST (http://metagenomics.anl.gov/) and

WebCarma 1.0 (Gerlach and Stoye, 2011). The MG-RAST data analysis used predominantly BLAST-based methods for taxonomic classification against seed subsystems. Most taxonomic and functional gene-based comparisons were undertaken within the WebCARMA 1.0 software pipeline, utilising both BLAST and Hidden markov model-based, HMMER3 outputs to increase gene classification accuracy. Internal default parameters were used and environmental gene tags (EGTs) utilised for most graphical presentations. These EGTs represented translated reads with matches to PFAM protein families. Further analyses of individual genes within the datasets were undertaken using the PFAM and Gene Ontology (GO) identification numbers.

Virus-specific analysis of the trimmed, filtered virome datasets was undertaken using the metagenomic processing pipeline, Metavir (Roux *et al.* 2011). Programs utilised included Genome Relative Abundance and Average Size (GAAS, Angly *et al*. 2009) using Blastx searches against the NCBI reference sequence complete viral genome proteins (release 2012-3-07). Phylogenetic comparisons based on several virus-specific marker genes (Table 2.) were undertaken and cluster analysis performed using the complete T4 and T12 virome datasets and 98% similarity on 35 base-pairs. Rarefaction curves (based on 98 % threshold) were also calculated for comparison of the rumen virome datasets and virome data from other microbial ecosystems (Table 3.). Only viromes that contain more than 50,000 sequences with a sequence length of over 100 bp on average were available for virome comparison by BLAST within Metavir. The three publically available datasets (Berg Miller *et al*. 2010) corresponding to the rumen virome of three Holstein dairy cows (MG-RAST I.D. numbers 4450679.3, 4450680.3 and 4450678.3) were obtained from MG-RAST, submitted to Metavir and utilised for virome comparison.

Table 2. Description of viral marker genes used for phylogenetic comparisons within the Metavir toolbox including the gene name, the viral group identified with the specific marker and viral protein function.

Gene	Viral group	Protein function	Reference
Rep	Circo-Nano-Geminiviridae	Replication	Rosario *et al*. 2009
VP1	Microviridae	Major capsid	Desnues *et al*. 2008
G20	T4-like Bacteriophages	Capsid assembly	Dorigo *et al*. 2004
GP23	T4-like Bacteriophages	Major capsid	Comeau *et al*. 2008
T7gp17	Podoviridae	Phage tail fiber	unpublished
TerL	Caudovirales	Terminase Large subunit	Sullivan *et al.* 2009
PolB	dsDNA viruses	DNA Polymerase family B	Monnier *et al*. 2008
MCP	Large eukaryotic dsDNA viruses	Major capsid	Rowe *et al*. 2011
RdRP	RNA viruses	RNA-dependant RNA polymerase	Culley *et al*. 2003

Table 3. Description of virome datasets employed for virome diversity comparisons undertaken within Metvir (Roux *et al*. 2011).

Virome dataset name	Ecosystem	Reference
Human Faeces A	Eukaryote	Kim *et al*. 2011
Human Gut L3d2	Eukaryote	Kim *et al.* 2011
Seawater, Gulf of Mexico	Marine	Angly *et al*. 2005
Seawater, Sargasso sea	Marine	Angly *et al.* 2005
Freshwater Lake Bourget	Freshwater	Roux *et al.* 2012
Freshwater Lake Pavin	Freshwater	Roux *et al*. 2012
Hypersaline saltern high 12075	Hypersaline	Dinsdale *et al*. 2008
Hypersaline saltern high 111605	Hypersaline	Dinsdale *et al.* 2008
Hot Spring Octopus	Hyperthermophile	Shoenfeld *et al*. 2008
Rumen viromes Lact-6993, Dry-7887 and Cull-7664	Eukaryote (bovine rumen)	Berg Miller *et al*. 2011

Analysis of virus community diversity (alpha and beta diversity)

Initial alpha diversity and taxon abundance was determined using the Genome relative Abundance and Average Size program (GAAS, Angly *et al*. 2009) incorporated within the Metavir toolbox.

Prior to further estimations of virome relative abundance, the datasets corresponding to the two rumen fluid collection time-points, T4 and T12, were filtered to remove replicate sequences and cluster consensus files created using cdhit_454 (Huang *et al*. 2010). Cross-contig spectra were produced with the Circonspect program (Angly *et al*. 2006) including self versus self as controls. The community structure and alpha diversity, including measures of richness, evenness and the Shannon-Wiener index, were modelled using PHACCS (Angly *et al*. 2005), based on the Circonspect contig spectra and GAAS average genome lengths for each virome dataset estimated using the lognormal rank-abundance model.

A Monte-Carlo based comparison of viral community diversity (beta-diversity) between the two virome datasets was undertaken using Maxiphi (Angly *et al*. 2006). Control cross-contig spectra (self versus self) were used to optimise the genome size by testing values within 50 % of the predicted input average genome length (60 kb), with the optimal average genome length estimated to be 74 kb and the relative quasi-likelihoods of cross-contig spectrums and best percentages (shared and permuted) calculated.

Analysis of bacterial and archaeal 16S rRNA gene raw sequence data

The raw sequence data was received from AGRF in the form of two files HI2F9CS03.sff and HIF6V2303.sff. The data contained in these files was extracted into .fasta and .qual files for analysis using the bioinformatics pipeline mothur (Schloss *et al.* 2009). Further analysis was conducted using the Quantitative Insights into Microbial Ecology (QIIME) open source software package (Caporaso *et al*. 2010a). Map files were established for the bacterial sequences and archaeal sequences (Appendix 2, Table A2.2). These map files were used to split the raw sequences into bacterial and archaeal sequences. The raw sequences were screened to remove sequences that did not meet the parameters outlined in Table 4.

Table 4. QIIME parameters used to split the raw sequences into individual libraries.

Parameter	Bacterial sequences	Archaeal Sequences
Minimum sequence length	430	300
Maximum sequence length	470	500
Quality score	10	10
Window size	50	50
Maximum homopolymer run	6	6
Number mismatches in primer	0	0

Bacterial 16S rRNA gene analysis

Four sets of sequences were obtained for each time point (0, 4, 8 and 12 h) and from these Operational Taxonomic Units (OTUs) were picked using the QIIME software pipeline based on a sequence similarity within the reads of 97 %. The most abundant sequence was selected within each representative sequence and aligned using PyNAST (Caporaso *et al.* 2010b). The aligned OTUs were screened for chimeras (OTUs comprised of two or more phylogenetically distinct parent sequences) using ChimeraSlayer (Caporaso *et al.* 2010a). The OTUs were then filtered to remove the identified chimeric sequences and to remove all OTUs that did not have a representative present in all four sets of sequences. The four sets were then pooled into one set of sequences for each time point.

An OTU heatmap was generated to display both the OTU counts per sample as well as a colour-based representation of the OTUs contribution, with dark blue representing a low contribution through to red representing high contribution. An OTU network was created and visualised using the Cytoscape program (Smoot *et. al*, 2011) where large white diamonds represent individual rumen fluid samples, small white diamonds represent OTUs and different coloured lines represent OTUs present in different samples. The OTUs were also summarised by taxa as stacked bar graphs using the taxonomic information assigned to each representative sequence by RDP classifier (Cole *et al*. 2009).

To further assess any time-related differences occurring within the bacterial community structure, all bacterial communities detected corresponding to each of the four sampling time points, were analysed for beta diversity (QIIME) using

phylogenetically aware measures of diversity (both weighted and unweighted unifrac). A jack-knifed replicate (bootstrapped), unweighted pair group method with arithmetric mean (UPGMA) phylogenetic tree of the total bacterial OTU sequences was then built using the QIIME default parameters. Additional measures of bacterial community alpha diversity (evenness and the Shannon Wiener Index) were also calculated (QIIME) for the sequences corresponding to each sample time point. For determination of the observed species (OTUs) as a measure of species richness, the sequence data was normalised between samples with the minimum number of bacterial sequences for all samples estimated (from rarefaction analysis) to be 8060.

Archaeal 16S rRNA gene sequence analysis

A single set of archaeal sequences was obtained for each time point (0, 4, 8 and 12 h) and quality filtered (denoised) based on a sequence similarity within the reads of 97 %, prior to OTU determination using the QIIME software pipeline (Caporaso *et al.* 2010a). The most abundant sequence was selected within each representative sequence and aligned using PyNAST (Caporaso *et al.* 2010b). A total of 69 aligned OTUs were obtained and screened for chimeras (OTUs comprised of two or more phylogenetically distinct parent sequences), which were then removed using ChimeraSlayer (Caporaso *et al*. 2010a). An OTU heatmap was generated to displays both the OTU counts per sample as well as a colour-based representation of the OTUs contribution, with dark blue representing low contribution through to red representing high contribution. An OTU network was created and visualised using the Cytoscape program (Smoot *et. al*, 2011) where large white diamonds represented individual rumen fluid samples, small white diamonds represent OTUs and different coloured lines represent OTUs present in different samples. The OTUs were also summarised by taxa as stacked bar graphs using the taxonomic information assigned to each representative sequence by RDP classifier (Cole *et al*. 2009)

To further assess any time-related differences occurring within the archaeal community structure, all archaeal communities detected corresponding to each of the four sampling time points, were analysed for beta diversity using phylogenetically aware measures of diversity (both weighted and unweighted unifrac). A jack-knifed replicate (bootstrapped), unweighted pair group method with arithmetric mean (UPGMA) phylogenetic tree of the total bacterial OTU sequences was then built using the QIIME default parameters. Additional measures of archaeal community alpha diversity (evenness and the Shannon Wiener Index) were also calculated (QIIME) for the sequences corresponding to each sample time point. For determination of the observed species (OTUs) as a measure of species richness, the sequence data was normalised between samples with the minimum number of archaeal sequences for all samples estimated (from rarefaction analysis) to be 899.

Transmission electron microscopy

Subsamples (10 µl volume) of concentrated phage stocks were deposited on copper square mesh grids (200 mesh, Proscitech) coated with nitrocellulose and incubated at room temperature for 5 min before removing excess fluid. Grids were then negatively stained with 1 % ammonium molybdate, pH 6.8 (Sigma-Aldrich) for 5 min. Images were obtained using a JEOL 1400 transmission electron microscope. Micrograph images were obtained with a digital camera (Gatan).

3.3 *Preparation of individual phage DNA for sequence analysis*

The complete draft sequence of a lysogenic phage present in our culture collection, φAR29 (rumen bacterial host *Prevotella ruminicola* AR29) was obtained from Dr Shawn Seet (PhD thesis, 2005).

Lytic phages chosen for sequence analysis (Table 5.) were propagated from frozen stocks by infection with their respective bacterial host in liquid lysis culture (Klieve, 2005). Phage particles were purified and concentrated by ultracentrifugation as described above and DNA extracted following enzymatic digestion with Proteinase K to remove the phage protein coat (Klieve and Gilbert, 2005).

For phage φRa02 only, phage DNA was sheared by sonication to 2 – 5 kb length fragments, end-repaired, and ligated into *Sma*1 cut *puc*19 vectors (Fermentas). Fragments were then transformed into DH5α competent cells (α-Select Gold Efficiency; Bioline). Colonies containing phage DNA fragments were selected and subcultured using MacConkey Agar$^{+\ AMP}$ to make a clone library. DNA from each clone library was purified using the alkaline lysis mini-prep kit (Qiagen) and cut with the restriction enzyme *Bam* HI (Roche) and fragment sizes determined by agarose gel electrophoresis. Phage clone libraries were submitted to Macrogen Inc. for sequencing.

All other phages were propagated from frozen stocks by infection with their respective bacterial host in liquid lysis culture (Klieve, 2005). Phage particles were then purified, concentrated and DNA prepared as described above. Purified phage DNA concentration was then determined by spectrophotometry and phage DNA quality checked for bacterial contamination confirmed by PCR testing using universal bacterial 16S rRNA gene primers (Lane, 1991). Phage identity was also checked prior to submission for sequencing, by determining the phage DNA restriction enzyme fingerprint pattern and comparing this pattern with any published or previously reported patterns. Where phage genomic DNA was known or suspected of being ssDNA, purified DNA was amplified using random hexamers, dNTPs and phi29 DNA polymerase (Genomiphi kit, GE).

Phage DNA concentrations were quantified by agarose gel electrophoresis and spectroscopy (Nanodrop 8000, Thermo Scientific). Prepared phage genomic DNA was then submitted for shotgun cloning and sequencing by an external sequencing provider (Australian Genome Research Facility Ltd) on a quarter 454-pyrosequencing plate using the Genome Sequencer GS FLX Titanium platform (Roche).

Table 5. Description of phages chosen for genome sequence analysis.

Phage name	Bacterial host	Reference
φAR29	*Prevotella ruminicola* AR29	Dr Shawn Seet, PhD thesis (2005); Klieve *et al*. (1989)
φRa01	*Ruminicoccus albus* AR67	Klieve *et al*. (2004)
φRa02	*Ruminicoccus albus* AR67	Klieve *et al*. (2004)
Φ149-112-5	*Ruminicoccus albus* AR67	B.CCH.1007 Final report
φBrb01	*Bacteroides ruminicola* AR20	Klieve *et al*. (1991)
φBrb02	*Bacteroides ruminicola* AR20	Klieve *et al*. (1991)
2BV	*Streptococcus bovis* 2B	Iverson and Millis (1967)
φSb01	*Streptococcus bovis* 2B	Klieve and Bauchop, (1991)
φSb02	*Streptococcus bovis* 2B	Klieve *et al*. (1999)

Analysis of individual phage sequence data

Annotated phage sequence data for φAR29 was initially analysed to identify genes of interest using the Rapid Annotation using Subsystem Technology (RAST http://rast.nmpdr.org/), the Integrated Services for Genomic Analysis pipeline (Hemmerich *et al*. 2010 http://isga.cgb.indiana.edu/Home) and the tRNAscan-SE Search Server (Schattner *et al*. 2005 http://lowelab.ucsc.edu/tRNAscan-SE/). Further analysis involved translation of the nucleotide sequence and identification of gene function using the Basic Local Alignment Search Tool (http://blast.ncbi.nlm.nih.gov/) to compare sequences with those currently available in publically available genomic databases including the virus-specific ncbi dsDNA virus proteins database (NCBI VOG) and the phage Seed subsystems containing reference sequences of 1910 viruses (http://www.phantome.org/PhageSeed/seedviewer.cgi).

4.0 Results

4.1 Method development: Preparation of viral nucleic acid from rumen fluid

Methods were adapted from established methodology (Klieve and Gilbert 2005) to isolate and concentrate viral particles from rumen fluid and extract viral DNA, free of contaminating plant and/or microbial DNA. Method development also involved determining the minimum volume of rumen fluid required in order to provide sufficient quantities of viral DNA to enable virus community analysis using a 454-pyrosequencing Genome Sequencer GS FLX Titanium platform (Roche).

An initial experiment to test several methodologies used rumen fluid collected from a single steer. This experiment verified that relatively small quantities of rumen fluid (5 ml volume), particularly when treated with PEG 8000/NaCl to facilitate particle precipitation, could generate viral DNA concentrations of sufficient quantity (Table 6.) and quality for submission to external sequencing providers.

Table 6. Methodology development: viral DNA concentrations obtained from bovine rumen fluid (treatment employed [+]).

Test sample	70°C pre-incubation of rumen fluid	DNase treatment	PEG 8000/NaCl precipitation	Viral DNA concentration (ng/µl)
1	+	+	+	37.1
2	+	+		88.2
3	+		+	152.1
4	+			51.8
5		+	+	149.6
6		+		83.8
7			+	121.7
8				99.3

4.2 Time course experiment: Diurnal fluctuations in the virome and corresponding microbial populations

Following initial method development, an *in vivo* time-course experiment was undertaken to determine how phage influence bovine rumen microbial populations, including the bacterial and archaeal (methanogen) populations. Rumen fluid samples obtained were prepared for virome sequence analysis, with quality controls including DNA concentration and PCR testing for bacterial DNA contamination. This PCR test was based on detection of the bacterial universal 16S rRNA gene (Lane, 1991), with the negative result obtained indicating that the DNA sample submitted for sequencing did not contain significant quantities of bacterial DNA.

Sequencing took longer than anticipated due to delays incurred by the external sequencing provider (letter to verify this delay included in Appendix 3) however high through-put sequencing data was successfully obtained for further analysis for two rumen virome samples (encompassing rumen DNA virus populations), four rumen bacterial population samples (based on bacterial 16S rRNA gene amplicon sequencing) and four rumen archaeal population samples (based on archaeal 16S

rRNA gene amplicon sequencing). Two additional virome samples (encompassing rumen RNA virus populations) have been submitted for sequencing (2nd March, 2012) with results still pending.

4.2.1 Attributes of the rumen virome and diurnal fluctuations in virus populations

Virome Sequence data

Virome sequence data was obtained for rumen fluid collected from a single steer at four hours (T4) and 12 hours (T12) after feeding, with a combined total of 161 085 raw sequences obtained. This data set was trimmed and filtered to remove possible sequencing artefacts, and subdivided according to sample time point with the resultant data sets containing 69 133 and 63 046 sequences >100bp in length for each of the T4 and T12 sample time-points, respectively (Table 7.).

Table 7. Summary of pyrosequencing data processing for the viral metagenomes corresponding to two rumen fluid collection time-points after feeding (T4 = 4 h; T12 = 12 h).

	Sample time-point dataset	
	T4	**T12**
Total number of quality filtered sequences >100bp	69 133	63 046
average sequence length (bp)	248.99	241.70
Sequence length standard deviation	89.58	87.27
Average GC content (%)	41.36	44.23
Average GC ratio	1.56	1.37

Structural and functional characteristics of the rumen virome and diurnal fluctuations

The dataset was first analysed using the virus-specific metagenomic processing pipeline, Metavir (Roux *et al.* 2011). Species richness was initially determined using rarefaction techniques. In ecology, rarefaction techniques allow the calculation of species richness from the results of sampling and the generation of rarefaction curves. In this case, if the curve generated for a viral community flattens as the number of sampled sequences increases, additional sampling of sequences is unlikely to yield additional virus family identification. Rarefaction curves for the two virome datasets T4 and T12 are presented in Figure A1.1 (Appendix 1.). These curves indicated that neither dataset encompassed all the virus diversity within the rumen virome community. A comparison of the T4 and T12 dataset rarefaction results with rarefaction results for a selection of published virome datasets was then undertaken using the same methodology. Rarefaction curves obtained indicated that even in virome datasets containing considerably higher numbers of sequences (up to 10 times more sequences) complete virus sequence coverage was not always achieved when sequences were examined based on rarefaction (Figure A1.2). These results indicate that for virome analysis, rarefaction curves may need to be treated with caution, with rarefaction being limited in its application to virome data

due to the considerable number of under-represented (rare) virus types encountered within most environmental samples.

Virus population diversity and taxonomic composition was determined using the Genome Relative Abundance and Average Size (GAAS) program (Angly *et al*. 2009) incorporated into Metavir. This tool was particularly useful for designating the rumen virus sequences obtained for each time-point to known virus and phage taxonomic groupings, utilising a virus-specific database (NCBI reference sequence, complete viral genome proteins, release 2012-3-07). Both virome datasets contained considerable numbers of unclassified virus sequences, with 8.84 % of the T4 virome dataset and 8.88 % of the T12 virome dataset with significant hits (GAAS analysis based on threshold e-value 10^{-5}) to viruses encompassed within the NCBI reference sequence complete viral genome proteins database (release 2012-3-07).

Each virome was dominated by sequences (94 % of identified virus sequences) associated with tailed, dsDNA phages, belonging to the Caudovirales. This large grouping includes phages which may also be morphologically-distinguished on the basis of tail structure, into the virus families, Siphoviridae, Myoviridae and Podoviridae. The rumen viromes were found to be predominated by members of the Siphoviridae (long, non-contractile tails), accounting for 63 % of the sequences designated as Caudovirales (example of Caudovirales taxonomic composition, Figures 1. and 2.). Further taxonomic examination enabled identification of the most abundant viral sequences for each virome (listed in Table 9.) as a percentage of the total number of viruses (6112 total viruses for T4; 5546 total viruses for T12), when using a threshold of 10^{-5} on e-value, within the GAAS program (Metavir).

Predominant phage types represented within the sequence data included phages for host species identified as being highly dominant within the respective rumen bacterial populations (Results section 4.2.2). For example, viral sequences corresponding to 10 different *Clostridium* phages were detected within the T12 virome, six of which had an abundance of > 1% of the total virome population (Table 9.). Interestingly for the T12 virome, viral sequences corresponding to 21 different *Lactoccocus* phages were detected, with only 2 having >1 % abundance, yet isolates of this bacterial genus were under-represented within the matching bacterial dataset.

Several rumen virome sequences were found to be closely related to archaeaphage present within the virus reference sequence database (Table 10.). These included archaeaphage with host species predominantly belonging to the phylum Euryarchaeota which includes many methanogens and halophiles which can grow in high salt conditions. The exceptions were sequences corresponding to viruses of the *Sulfolobus* and *Acidianus*, which belong to the phylum Crenarchaeota. Of the archaeaphage sequences detected, most corresponded to tailed archaeaphage belonging to the Myoviridae (for example, *Natrialba* phage PhiCh1 and *Halorubrum* phage HF2) and Siphoviridae (for example, *Methanobacterium* phage psiM1 and Archaeal BJ1 virus).

Table 8. Relative taxonomic classification to virus family level for each virome dataset obtained for the 4 h and 12 h sample time points (T4 and T12 respectively), as determined by similarity to known DNA and protein sequences (GAAS analysis based on threshold e-value 10^{-5}, Metavir).

Virus classification*	% of virus sequences	
	T4 virome	T12 virome
dsDNA virus, no RNA stage	**94.0**	**94.0**
Others**	**6.0**	**6.0**
Caudovirales(Siphoviridae)	63.0	63.0
Caudovirales(Myoviridae)	24.0	22.0
Caudovirales(Podoviridae)	5.0	4.0
Unclassified Caudovirales	3.0	5.0
Mimiviridae	0.7	0.6
Phycodonaviridae	0.9	1.0
Unclassified dsDNA phages	1.7	3.0
Other dsDNA viruses	1.7	1.4

* Homology to known sequences assigned to viruses (threshold e-value 10^{-5})

** Others includes sequences designated as belonging to ssDNA viruses, RNA viruses and other unclassified viruses

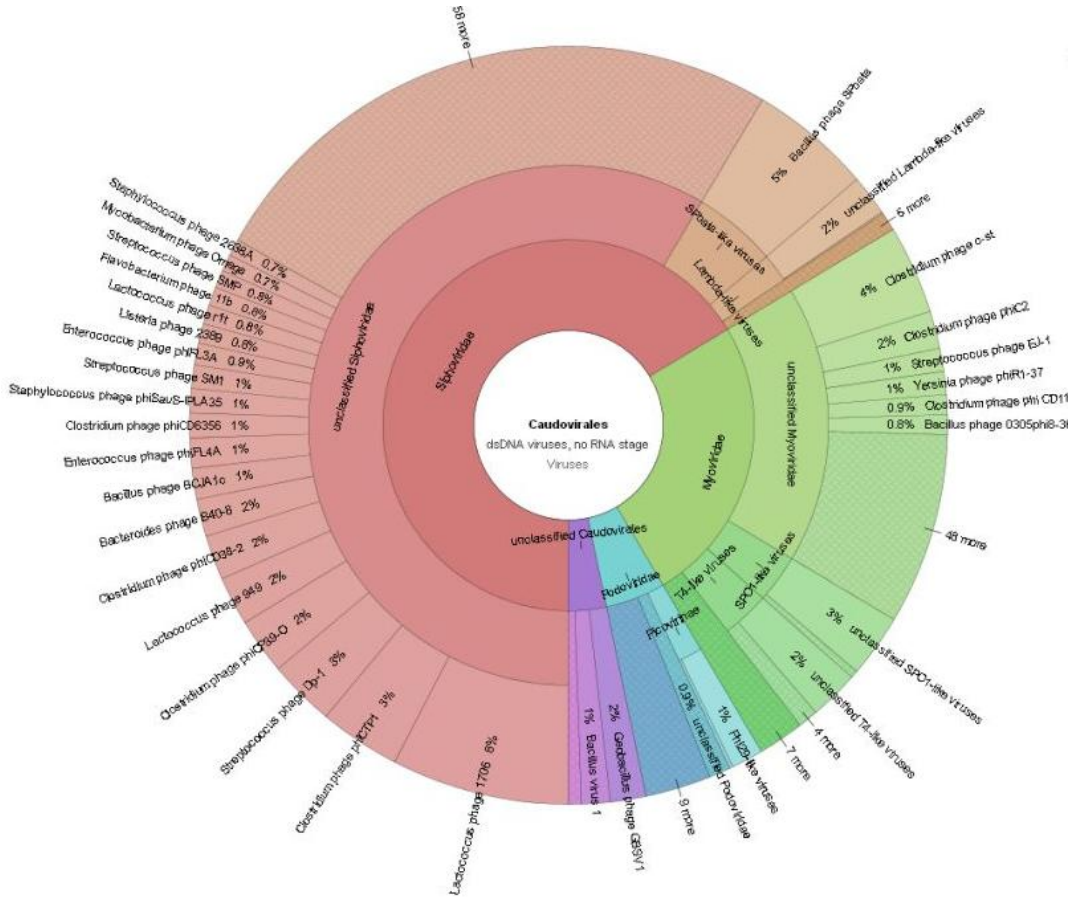

Figure 1. Taxonomic composition analysis of sequence data relating to the Caudovirales family, from the T4 virome (GAAS analysis, total sequences and threshold of 10^{-5} on e-value). Abundance for each phage calculated as a percentage of the total number of sequences designated as Caudovirales. Figure generated using Krona.

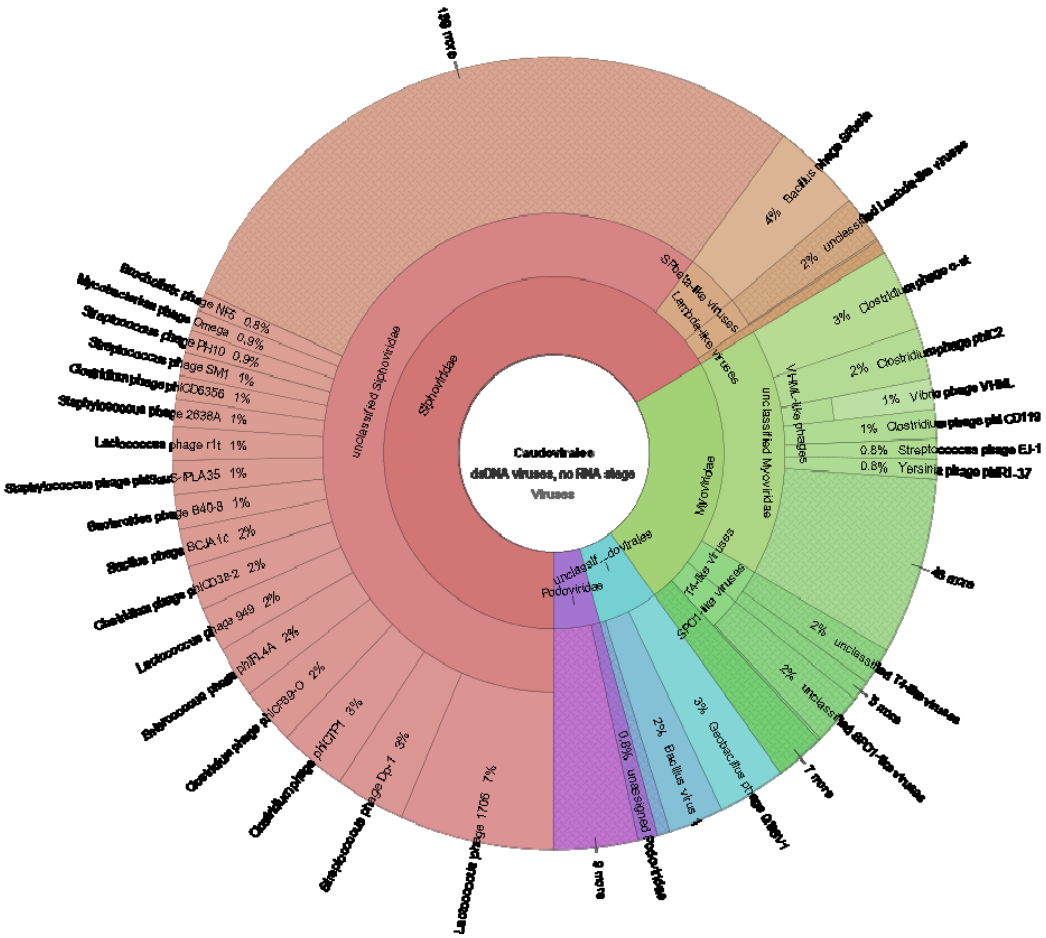

Figure 2. Taxonomic composition analysis of sequence data relating to the Caudovirales family, from the T12 virome (GAAS analysis, total sequences and threshold of 10^{-5} on e-value). Abundance for each phage calculated as a percentage of the total number of sequences designated as Caudovirales. Figure generated using Krona.

Table 9. Most abundant designated viral sequences (top 25 taxon) detected within the two virome datasets, corresponding to the 4 and 12h sample time points (T4 and T12). Data obtained following taxonomic classification of virome sequences using GAAS (threshold e-value 10^{-5}, Metavir).

T4 virome		T12 virome	
Virus taxon name	**% ***	**Virus taxon name**	**%***
Lactococcus phage 1706	6.12	*Lactococcus* phage 1706	5.01
Clostridium phage phiCTP1	2.67	*Geobacillus* phage GBSV1	3.54
Clostridium phage 39-O	2.61	*Bacillus* virus 1	2.80
Streptococcus phage Dp-1	2.12	*Streptococcus pyogenes* phage 315.3	2.70
Clostridium phage phiCD38-2	2.03	*Clostridium* phage 39-O	2.60
Geobacillus phage GBSV1	1.93	*Enterococcus* phage phiFL4A	2.41
Bacillus phage SPBc2	1.74	*Streptococcus* phage Dp-1	2.19
Enterococcus phage phiFL4A	1.59	*Clostridium* phage phiCTP1	2.05
Bacillus virus 1	1.51	*Clostridium* phage phiCD38-2	1.82
Bacillus phage BCJA1c	1.45	*Lactococcus* phage r1t	1.78
Clostridium phage phiC2	1.39	*Bacillus* phage BCJA1c	1.63
Geobacillus virus E2	1.36	*Clostridium* phage phiC2	1.61
Bacillus phage B103	1.36	*Clostridium* phage phiSM101	1.58
Lactococcus phage r1t	1.27	*Bacteroides* phage B40-8	1.38
Bacteroides phage B40-8	1.27	*Enterococcus* phage EFRM31	1.37
Streptococcus pyogenes phage 315.3	1.26	*Staphylococcus* phage phiSauS-IPLA35	1.35
Clostridium phage phiCD6356	1.23	*Vibrio* phage VHML	1.23
Streptococcus phage SM1	1.20	*Streptococcus* phage SM1	1.21
Enterococcus phage phiFL3A	1.03	*Bacillus* phage SPBc2	1.20
Streptococcus phage EJ-1	1.03	*Streptococcus* phage PH10	1.15
Bacillus phage phi29	1.02	*Clostridium* phage phiCD6356	1.10
Streptococcus phage SMP	1.01	*Staphylococcus* phage 2638A	1.08
Listeria phage 2389	0.98	*Geobacillus* virus E2	1.07
Flavobacterium phage 11b	0.97	*Clostridium* phage phi CD119	0.95
Staphylococcus phage tp310-2	0.89	*Streptococcus* phage EJ-1	0.88

*Abundance as percentage of total virus population based on GAAS analysis of sequences (without normalisation), e-value 10-5 threshold (Metavir).

Table 10. Archaeaphage detected within the two virome datasets following taxonomic classification of virome sequences using GAAS (Metavir threshold e-value 10^{-5} using the virus reference sequence database) with abundance presented as a percentage of total identified viruses

T4 virome		T12 virome	
Virus taxon name	**%***	**Virus taxon name**	**%***
Methanothermobacter phage psiM100	0.11	*Natrialba* phage PhiCh1	0.29
Natrialba phage PhiCh1	0.11	*Halorubrum* phage HF2	0.08
Acidianus rod-shaped virus 1	0.07	*Methanobacterium* phage psiM2	0.07
Acidianus two-tailed virus	0.05	*Methanothermobacter* phage psiM100	0.06
Sulfolobus virus STSV1	0.04	*Sulfolobus* virus STSV1	0.02
Acidianus filamentous virus 7	0.02		
Halorubrum phage HF2	0.02		
Archaeal BJ1 virus	0.02		

* Abundance as percentage of total virus population based on GAAS analysis of sequences (without normalisation), e-value 10^{-5} threshold (Metavir).

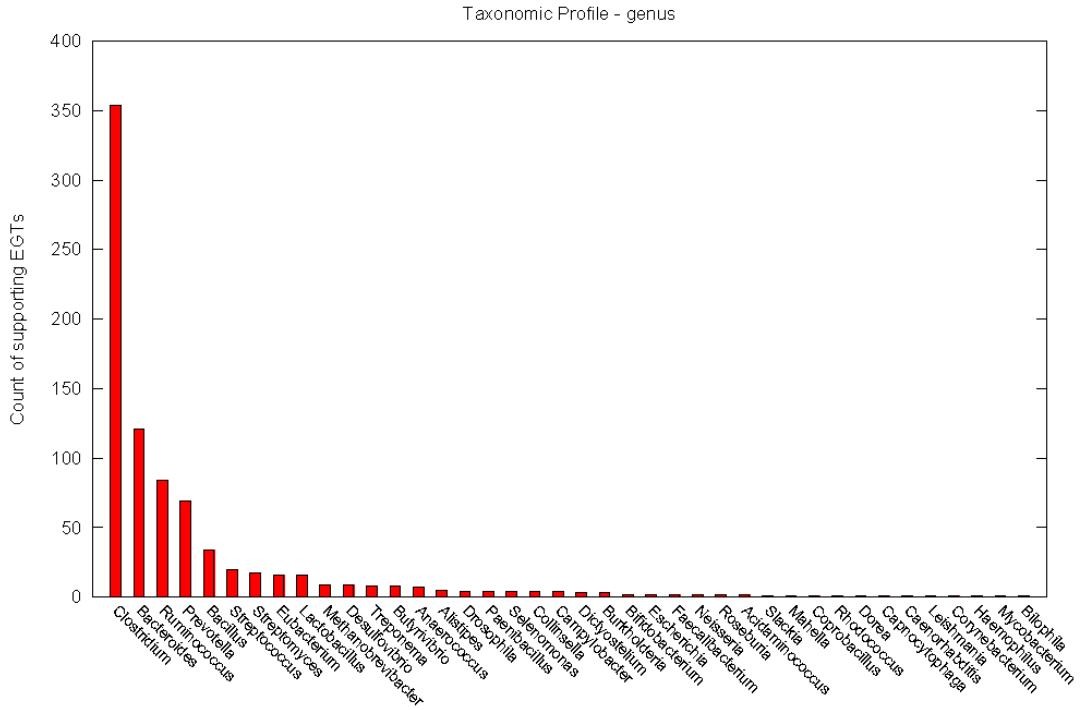

Figure 3. Taxonomic results for the T4 virome, top 40 most abundant taxon (abundance presented as a count of supporting EGTs) classified to genus level (WebCARMA).

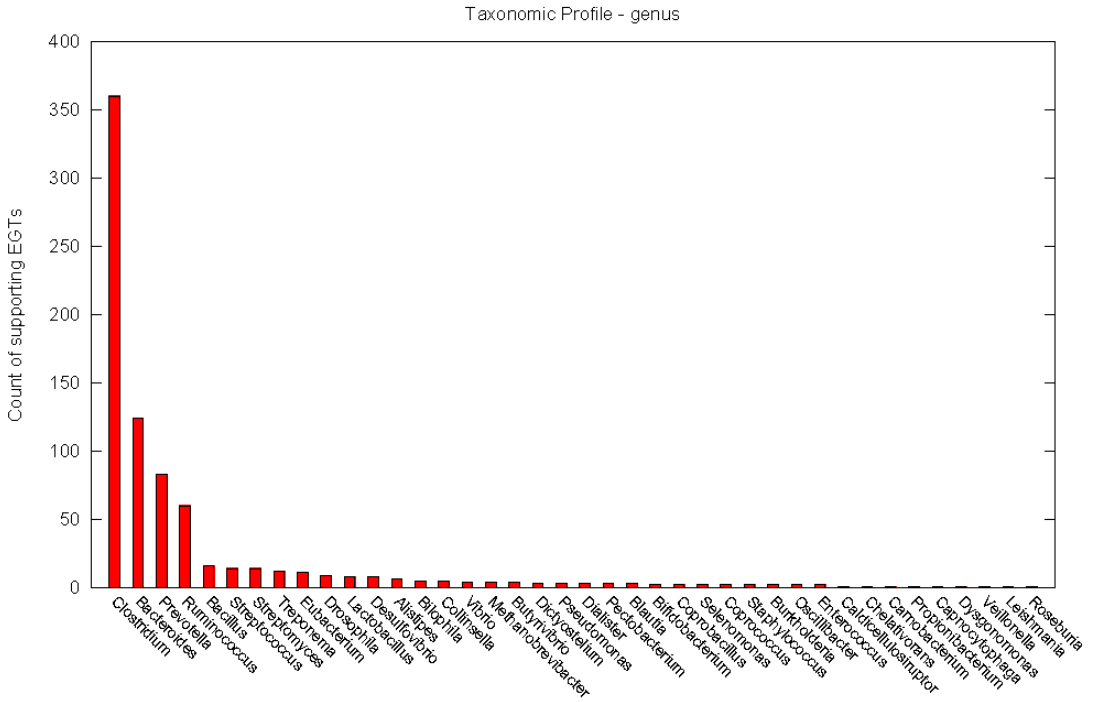

Figure 4. Taxonomic results for the T12 virome, top 40 most abundant taxon (abundance presented as a count of supporting EGTs) classified to genus level (WebCARMA).

Taxonomic analysis of the virome samples using metagenomic pipelines such as MG-RAST and WebCARMA (Figures 3. and 4.), indicated the presence of virome sequences associated with a broad range of microbial hosts taxons, with a small percentage found to be related to the non-microbial, Eukaryota superkingdom (approximately 2 % of total sequences for each virome (T4 and T12) as determined by WebCARMA taxonomic analysis). In addition, for each virome sequence dataset, a large proportion of sequences could not be assigned to a specific taxonomic grouping (74 % and 73 % of sequences unknown for the T4 and T12 virome datasets respectively, WebCARMA taxonomic analysis). This taxonomic analysis indicated that the rumen microbial communities associated with virome sequences were not static, with changes in the relative abundance of micro-associated sequences observed, including those associated with the dominant rumen methanogen genus, *Methanobrevibacter*, changing with time after feeding.

The virome datasets were also analysed using viral marker genes (Table 2.). These genes have been found to be genetically conserved amongst viral families, encoding for proteins often associated with core structural (for example, virus particle capsid (VP1) or tail fibre (T7gp17) proteins) or processing functions (for example, terminase large subunit protein (TerL) involved in viral DNA packaging). Both virome datasets were examined for the presence of viral marker genes accessed through the Metavir software pipeline and the occurrence of these marker genes is reported in Table 11., together with a previously unreported analysis using the same marker genes, for two other rumen virome datasets (Berg Miller *et al*. 2011). Interesting, analysis of these published datasets also indicated the presence of the same kinds of marker genes (G20, GP23, PolB, T7gp17 and TerL), with the exception of a single Rep marker gene (representing replication proteins found within the Circo-Nano-Geminiviridae) in the virome of a dry Holstein cow. The Holstein study however obtained a far larger number of total sequences for each cow virome (298 062 minimum number of raw sequences reported), therefore more hits to each of the respective marker genes, including multiple hits to the same marker sequence at the individual virus level may consequently be obtained using this analysis approach.

The terminase large subunit (TerL) gene was chosen as the basis for further phylogenetic analysis of representative virome sequences, resulting in the construction of phylogenetic trees. This analysis established the presence of specific viral marker genes within the virome datasets with homology to previously sequenced archaeaphage. Multiple phylogenetic trees were generated for the TerL gene to encompass all of the virus-specific TerL matches obtained (example tree for the T4 virome dataset, Figure A1.5). Given the complexity of these phylogenetic trees, a subset of results, showing a sequence from the T12 virome dataset showing the presence of a TerL sequence directly related to the archaeaphages *Methanobacterium* ψM2 and *Methanothermobacter* ψM100 is shown in Figure 5.

Further phylogenetic analysis with the DNA polymerase family B (PolB) marker gene, also indicated the presence of a T12 virome sequence (encoding for a 36 amino acid protein designated as a PolB gene (Metavir phylogenetic analysis tool). This sequence was most similar (98 % similarity on 35 bp) to the *Halorubrum* phage HF2. Interestingly, sequences also related to this phage, but not necessarily related to the PolB gene, were also detected following taxonomic classification of virome sequences using GAAS (Table 10).

Table 11. Viral marker gene analysis of all available rumen virome datasets, including the T4 and T12 samples from the current study and three additional published rumen virome datasets obtained for three Holstein cows at various stages of lactation (Berg Miller *et al*. 2011) and the number of sequences detected (98 % similarity on 35 bp, Metavir) within each virome dataset.

Virome dataset	Virus marker genes									
	G20	GP23	MCP	PolB	PsbA	RdRP	Rep	T7gp17	TerL	VP1
T4	1	1	0	0	0	0	0	0	89	0
T12	3	2	0	1	0	0	0	0	72	0
dry_7887	24	28	0	15	0	0	1	2	673	0
Lact_6993	65	48	0	13	0	0	0	28	972	0

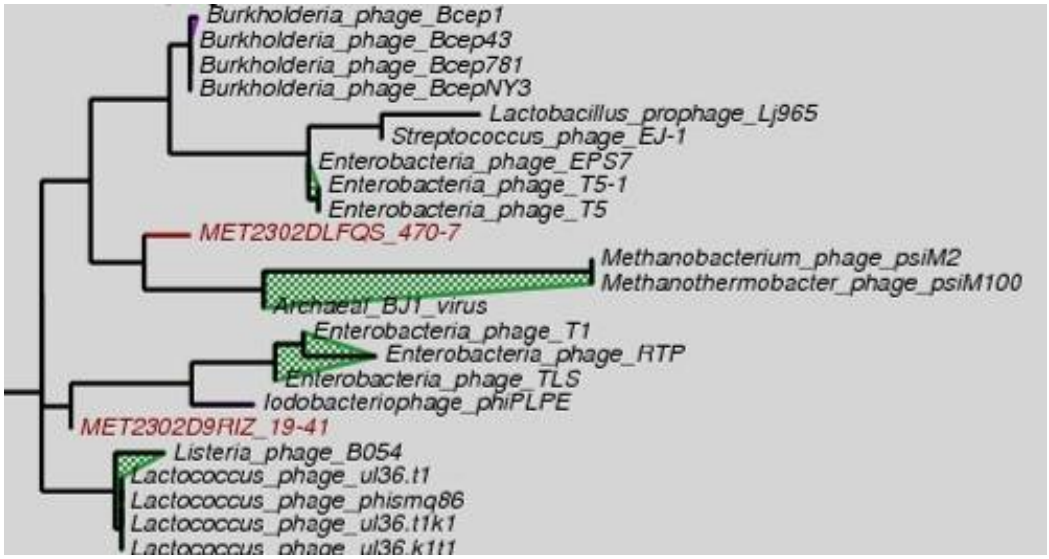

Figure 5. Subsection of Phylogenetic tree generated by Metavir showing a sequence related (98 % similarity on 35bp) to the known archaeaphages Methanabacterium ψM2 and Methanothermobacter ψM100. Analysis undertaken using BLASTx with terminase sequence data from reference sequences in the PFAM database (e-value $<10^{-3}$), alignment with HMMER

Table 12. Alpha diversity estimates for each virome sample (T4 and T12 corresponding to sample time (h) after feeding) determined by the Phage communities from Contig spectrum (PHACCS) program.

	T4	T12
Best scenario*	Lognormal	Lognormal
Richness (No. of genotypes)	2365	6440
Most abundant genotype (%)	1.615	3.145
Evenness**	0.905	0.818
Shannon-Weiner index†	7.030	7.176

* Best model used for computation within PHACCS based on smallest error.

**Complete evenness (measure of equitability) equal to 1.0.

† Shannon-Wiener index calculated as a measure of alpha diversity (the larger the value, the greater diversity within a sample).

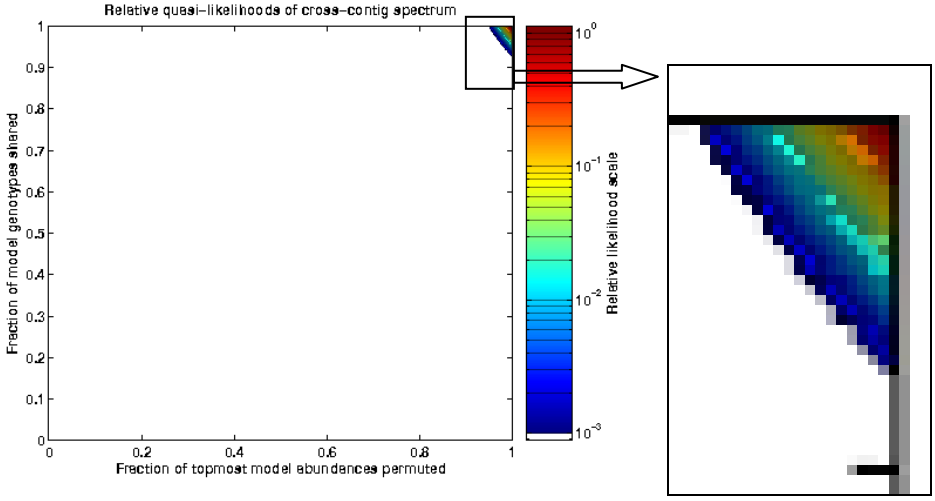

Figure 6. Comparison of the T4 and T12 sequence datasets (beta diversity measures) using contig spectrum generated from each dataset to model the fraction of genotypes shared between each dataset (Maxiphi program, Angly *et al.* 2006), with the T4 vs T12 cross-contig spectrum showing both samples tightly overlapping (graph upper right hand corner enlarged to the right), based on the best average distance of 0.715 and best average genome length of 74 000 bp. Beta diversity measures (T4 vs T12): best percent shared, 100 % and best percent permuted, 99.94 %.

When each virome dataset was analysed for alpha diversity on the basis of contig spectrum (Table 12.), the T12 virome was found to contain a greater total number of virus genotypes (6440 genotypes) than the T4 virome (2365 genotypes), with diversity slightly higher in the T12 virome (indicated by the Shannon-Weiner index). Interestingly, modelling of the between-sample variation (beta diversity) using Maxiphi (Angly *et al.* 2006) indicated that the best estimate of genotypes shared between T4 and T12 was 100 % (T4 vs T12 comparison) and the best percent permuted was 99.94 % (Figure 6.). These results indicated that the viruses present at the first sample time point (4 h) were also present at the second time point (12 h). The measures of alpha diversity (Table 12), indicated that virus population diversity considerably increased with time after feeding (as indicated by a decrease in the evenness estimate), therefore throughout the 8 h period, new virus genotypes either increased to detectable levels or were introduced into the rumen fluid (possibly through phage lytic events or blooms), effectively diluting the viruses present within the earlier time point.

The MG-RAST and WebCARMA pipelines also enabled a preliminary functional gene profile (virome genes related to enzymatic function) to be determined (Figures 9. and 10.) and facilitated comparison between the functional properties of the two rumen virome datasets (T4 and T12). For both viromes, the majority of functional enzymes were found to be associated molecular functioning (DNA binding, ATP binding and transcription), biological processes (DNA-dependant transcription regulation, DNA recombination and replication and DNA integration) and cellular component functions (associated with membrane, nucleus, intracellular and intracellular components). Other functional genes of potential interest include those associated with DNA methylation (common phage resistance mechanisms) and proteolysis (including phage-encoded genes involved in host-specific cell lysis activities).

A preliminary analysis was also undertaken (Blast-based comparison with seed subsystems, MG-RAST) to determine the proportion of virus sequences that could be attributed to functional genes present within (a) reference sequence phages; (b) phages known to integrate within bacterial or archaeal genomes (prophages); and (3) other transposable elements (Figures 7. and 8.). This analysis also showed differences in virome sequences obtained for each the two sample time points, with the T12 virome containing a slightly higher proportion of sequences which could be attributed to transposable elements. The most abundant phage-like elements identified using this approach included terminase proteins and other phage proteins associated with the rlt-like-Streptococcal phages.

Despite the functional profile of each rumen virome dataset being very similar, there were changes in the relative abundance of certain functional genes of interest (Table 13.). For example, genes related to (phage) DNA integration, were of similar proportions in each dataset (343 for T4 and 387 for T12) but in terms of the relative abundance position within each virome dataset, the DNA integration genes were more abundant in the T12 dataset (11th place) than the T4 dataset (16th place).

Figure 7. Sequences within the T4 virome attributed to phages and prophage elements identified using a blast-based comparison with seed subsystems (MG-RAST, figure constructed using krona).

Figure 8. Sequences within the T12 virome attributed to phages and prophage elements identified using a blast-based comparison with seed subsystems (MG-RAST, figure constructed using krona).

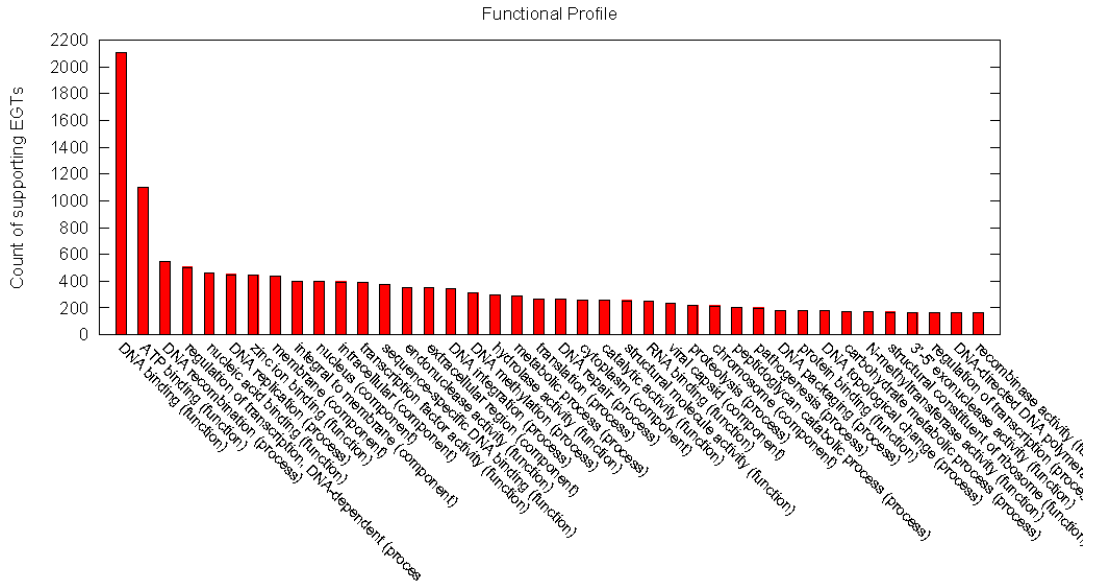

Figure 9. Functional profile for the 4 h time point (T4) rumen virome: 40 most abundant Gene Ontology terms (GO-terms, x-axis) and abundance (y-axis) as a total count of Environmental Gene Tags (EGTs) within the virome sequence dataset (WebCARMA).

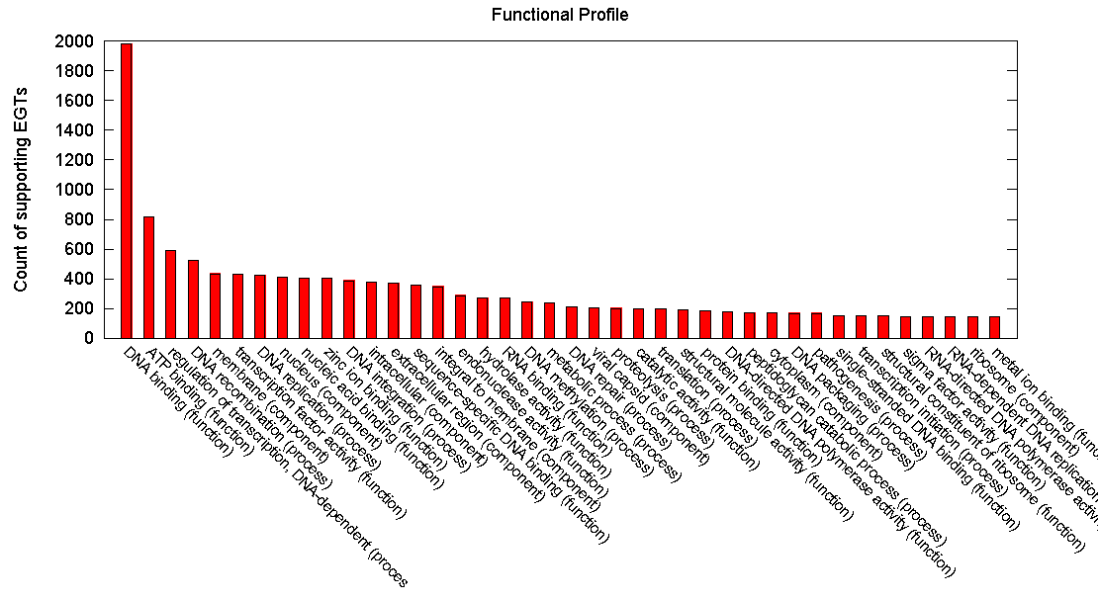

Figure 10. Functional profile for the 12 h time point (T12) rumen virome: 40 most abundant Gene Ontology terms (GO-terms, x-axis) and abundance (y-axis) as a total count of Environmental Gene Tags (EGTs) within the virome sequence dataset (WebCARMA).

Table 13. Top 20 most abundant functional gene groupings for each rumen virome dataset (4 h (T4) and 12 h (T12) sample times), identified the using WebCARMA pipeline for assignment of metagenomic sequence data to functional gene groupings.

T4		T12	
Functional gene grouping	Abundance*	Functional gene grouping	Abundance*
DNA binding (molecular_function)	2108	DNA binding (molecular_function)	1981
ATP binding (molecular_function)	1100	ATP binding (molecular_function)	819
DNA recombination (biological_process)	547	regulation of transcription, DNA-dependent (biological_process)	593
regulation of transcription, DNA-dependent (biological_process)	504	DNA recombination (biological_process)	523
nucleic acid binding (molecular_function)	461	membrane (cellular_component)	434
DNA replication (biological_process)	449	transcription factor activity (molecular_function)	431
zinc ion binding (molecular_function)	445	DNA replication (biological_process)	425
membrane (cellular_component)	436	nucleus (cellular_component)	408
integral to membrane (cellular_component)	398	nucleic acid binding (molecular_function)	405
nucleus (cellular_component)	396	zinc ion binding (molecular_function)	402
intracellular (cellular_component)	394	DNA integration (biological_process)	387
transcription factor activity (molecular_function)	393	intracellular (cellular_component)	379
sequence-specific DNA binding (molecular_function)	372	extracellular region (cellular_component)	369
endonuclease activity (molecular_function)	350	sequence-specific DNA binding (molecular_function)	359
extracellular region (cellular_component)	349	integral to membrane (cellular_component)	348
DNA integration (biological_process)	343	endonuclease activity (molecular_function)	288
DNA methylation (biological_process)	310	hydrolase activity (molecular_function)	274
hydrolase activity (molecular_function)	294	RNA binding (molecular_function)	269
metabolic process (biological_process)	286	DNA methylation (biological_process)	242
translation (biological_process)	265	metabolic process (biological_process)	237

*Abundance as a total count of supporting Environmental Gene Tags (EGTs)

Comparison of the T4 and T12 viromes with other virome datasets

Virome datasets corresponding to the 4 and 12 h sample time points were compared to virome datasets obtained for rumen fluid sampled from three Holstein cows (Berg Miller *et al*. 2011) and a range of other virome datasets from different environments. These additional environments included two Eukaryote ecosystems, representative of human faecal and gut samples, two ocean ecosystems, two freshwater lake ecosystems and two hypersaline ecosystems (described in Table 3.). A phylogenetic comparison (Figure 12.) and multidimensional scaling (MDS) plot comparison (Figure 11.) were undertaken using analysis tools within the Metavir software pipeline (Roux *et al*. 2011). Both of these comparison tools indicated that all five rumen fluid viromes were more closely related to each other than to viromes obtained from the other environments. The two rumen viromes from the current study (T4 and T12) were different in their phylogeny and species composition to each other, despite being sampled from the same animal, however these viromes were closer to each other than to the three previously published rumen viromes, perhaps as a consequence of either technical differences in sample preparation or differences associated with animal species, diet, age, stage of lactation or geography (previous study undertaken in Illinois, US). Interestingly, of the three Holstein cows examined in the previous study, the rumen virome from the lactating cow was found to be most phylogenetically different, with the dry and culled cow virome datasets having greater phylogenetic similarity (Figure 12.).

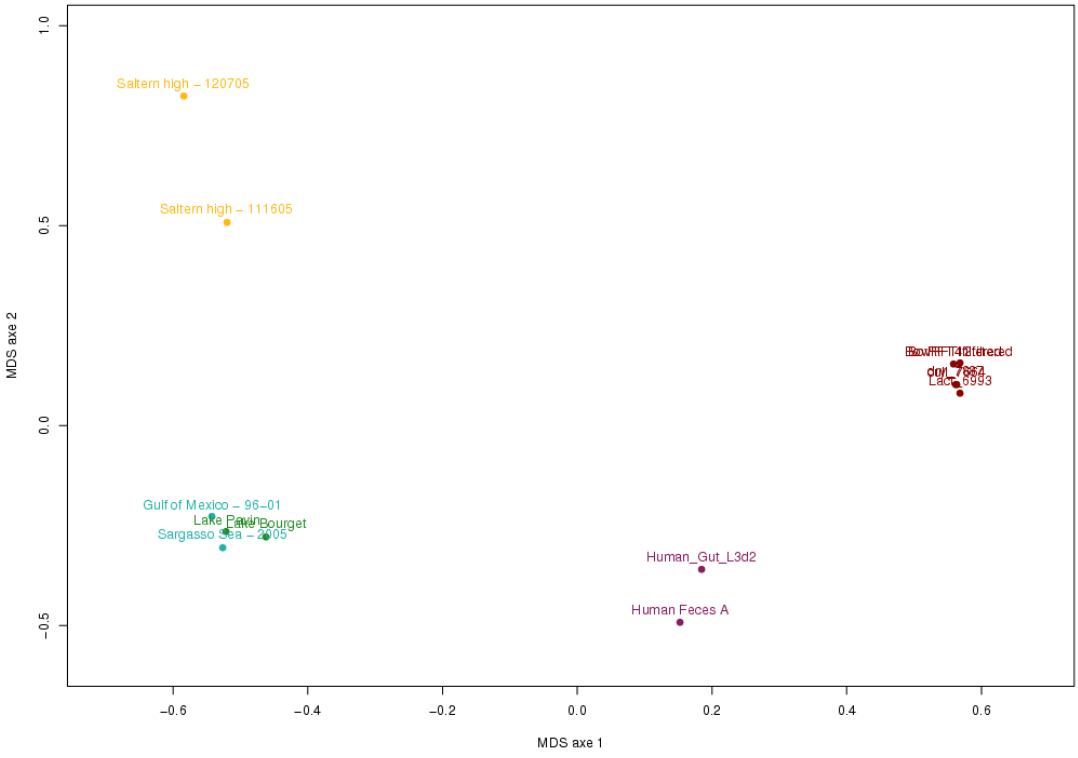

Figure 11. MDS plot of rumen viromes (red) and viromes associated with eukaryote (pink), marine (aqua), freshwater (green) and extreme (hypersaline) environments (yellow) (Metavir).

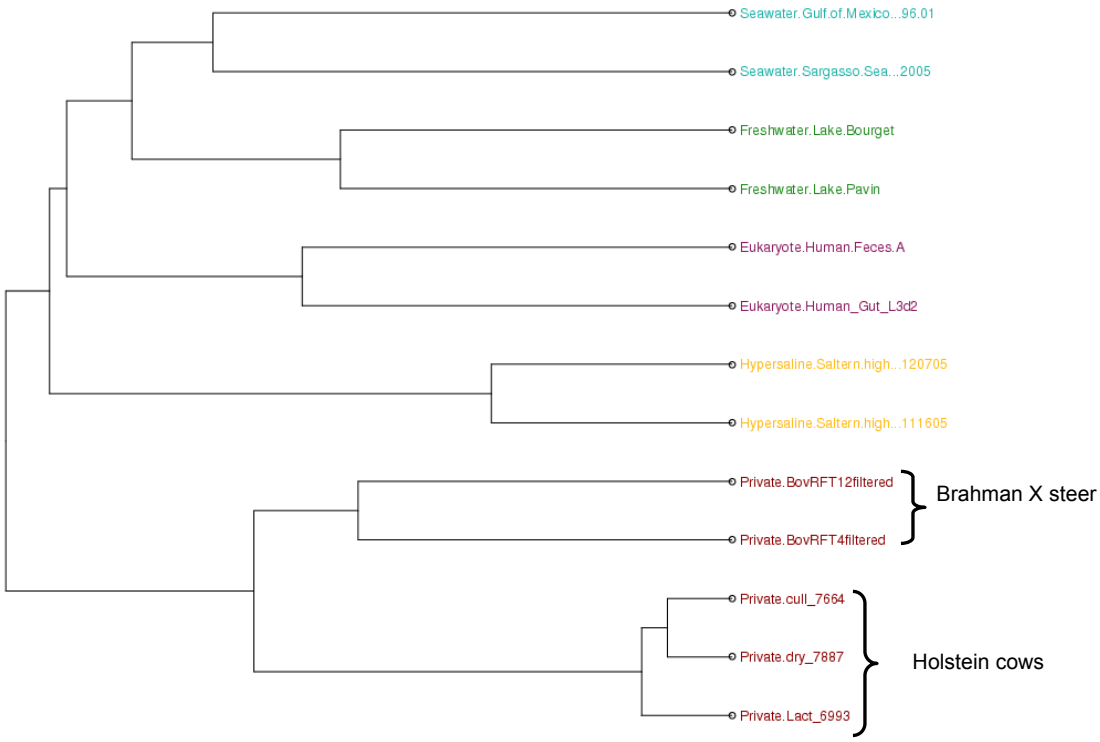

Figure 12. Phylogenetic comparison of five rumen viromes (designated as private analysis, red text) and viromes associated with eukaryote (pink), marine (light blue), freshwater (green) and extreme hypersaline environments (yellow) Samples compared using tBLASTx with score matrix used to cluster viromes using R software and pvclust, default parameters and 100 bootstraps (Metavir).

4.2.2 Diurnal fluctuations in rumen bacterial populations

From an input of 293,074 raw sequences containing both bacterial and archaeal sequences, a total of 74,827 bacterial sequences were obtained that met the parameters outlined in Table A2.1 (Appendix 2). A total of 2073 OTUs were obtained from the bacterial sequences and removing the identified chimeric OTUs reduced the total number of sequences by 6% and the number of OTUs to 1885. Filtering the sequences to remove OTUs not present in all four sets of sequences and then pooling the four sets into one set for each time point reduced the total number of OTUs to 1423 and the total number of sequences was reduced to 59,217. A rarefaction curve of the bacterial 16S rRNA gene sequences prepared to estimate species richness for each time point, showed the curves flattening as the number of sequences sampled increased (Figure A1.4, Appendix 1), indicating that more intensive sampling would be likely to yield only few additional new OTUs.

Utilising the 16S rRNA gene data obtained for each of the four rumen fluid collection time-points, the different bacterial taxa present were identified to Genus level. The rumen bacterial population of this steer was dominated by Gram positive Firmicutes and Gram negative Bacteroidetes at all sample time points. Sub-populations of mainly Gram negative Proteobacteria, Spirochaetes, Tenericutes (encompassing mycoplasma-like bacteria), Fibrobacteres, Actinobacteria and other unclassified bacteria were also present at in lower proportions.

An OTU heatmap to represent the predominant bacterial populations was prepared and the most abundant 30 OTUs presented in Figure 13., with colours (dark blue to red) used to highlight the relative levels of OTU abundance in each rumen fluid sample. Several time-related differences in the relative abundance of OTUS were apparent, for example, when comparing the relative abundance of OTUs present between the 4 h and 12 h samples, there were 151 OTUs, representing 6.5 % of sequences within Clostridiales and 8.2 % within Bateriodetes, that were present in the 4 h sample, yet absent from the 12 h sample. Notable increases in relative abundance (increasing to 2.5 % of sequences), were also recorded over time for specific genera such as *Acetivibrio, Prevotella*, and individual unclassified Firmicutes and Clostridia (Figure 14.). An individual Clostridiales however, remained the most dominant population within the overall rumen bacterial population (specific OTU representing over 10 % of total sequences), over the 12 h period. Further analysis of this Clostridiales OTU sequence indicated a 100 % homology over 422 bp to a number of uncultured clones, including (1) a clone from a rumen fluid fibre adherent microbiome study (EU845513, steer 71, rumen fluid collection 1 h after feeding); and (2) a clone from a rumen fluid diet study, (GQ327558 dry Holstein cow, rumen fluid collection immediately prior to feeding). Using the RDP II database (Cole *et al*. 2009) the OTU also had a 99 % match over 422 bp to the unclassified Clostridiales rumen bacterium R-7 (AB239481), isolated from sheep rumen fluid, attaching to cellulose powder. These findings suggest that this dominant OTU corresponds to a commonly found yet largely uncharacterised, cellulose-degrading and possibly fibre adherent, Gram positive, spore-forming anaerobic bacterium constituting part of the normal rumen microflora when animals are consuming high fibre diets.

Differences in bacterial community structure over time were also illustrated using an OTU network (Figure 15.). In this figure, bacterial populations present within all four rumen fluid samples, representative of core rumen bacterial genera, are grouped together in the centre of the network. Bacterial populations not shared between

samples are depicted as stretching out beyond the central, core populations. In this way, distinct, time-dependant changes can be observed for a relatively small proportion of the bacterial population, with each rumen fluid sample containing bacterial genera which were not shared between all time points.

To further assess the time-related differences in diversity occurring within the bacterial community structure, the bacterial communities detected corresponding to each of the four collection time points were analysed using phylogenetically aware measures of diversity (both weighted and unweighted unifrac) and a jackknifed replicate UPGMA phylogenetic tree of the total bacterial OTU sequences was produced (Figure 16.). This analysis indicated that the bacterial communities at the 0 h and 4 h time points were most related, clustering together with 20-50 % support, whilst the bacterial communities found at the 8 h and 12 h time points clustered together but with < 25 % support. These results indicated a shift in rumen bacterial populations, occurring between 4 and 8 hours after feeding. Additional measures of population alpha diversity, including total species observed, Evenness and the Shannon-Wiener index (Table 14.) also indicated that rumen bacterial populations decreasing in species diversity over the 12 h time period after feeding.

Figure 13. Heat map of the top 30 OTUs from the bacterial partial 16S rRNA gene data set with the number of sequences within the OTUs for each sample time point. OTUs representing over 10 % of the sequences within the sample are coloured red; 5 to 10 % orange; 2.5 to 5 % light blue; 1 to 2.5 % mid blue and less than 1 % dark blue.

Consensus Lineage	0 hours	4 hours	8 hours	12 hours	#OTU ID
Clostridiales	2538	1613	2402	1014	1689
Clostridiales	968	648	630	298	1006
Oscillibacter	764	508	805	359	397
Ruminococcaceae	504	389	380	175	1536
Ruminococcaceae	386	284	312	185	137
Ruminococcaceae	285	182	250	127	16
Bacteroidetes	274	242	308	159	881
Pseudobutyrivibrio	253	115	175	118	1850
Bacteroidetes	217	157	133	56	479
Ruminococcus	194	161	188	129	622
Clostridiales	188	161	206	111	1631
Acetivibrio	175	107	193	94	1797
Clostridia	166	136	198	100	433
Prevotella	142	124	156	99	409
Ruminococcaceae	139	69	107	46	444
Firmicutes	136	101	166	91	2035
Butyrivibrio	136	56	117	84	1829
Clostridiales	134	72	86	45	1076
Lachnospiraceae	130	130	146	75	1956
Bacteroidetes	124	94	129	35	892
Ruminococcaceae	124	81	104	35	410
Prevotella	121	105	158	79	203
Ruminococcus	113	47	87	58	1683
Clostridiales	110	56	52		461
Bacteroidetes	108	119	133	80	1255
Flavobacteriaceae	108	64	71	24	1909
Ruminococcaceae	107	67	96	46	1895
Bacteroidetes	104	87	65	54	908
Clostridiales	103	119	132	55	1342
Prevotella	102	68	78	60	1934
Clostridia	102	93	124	59	1955

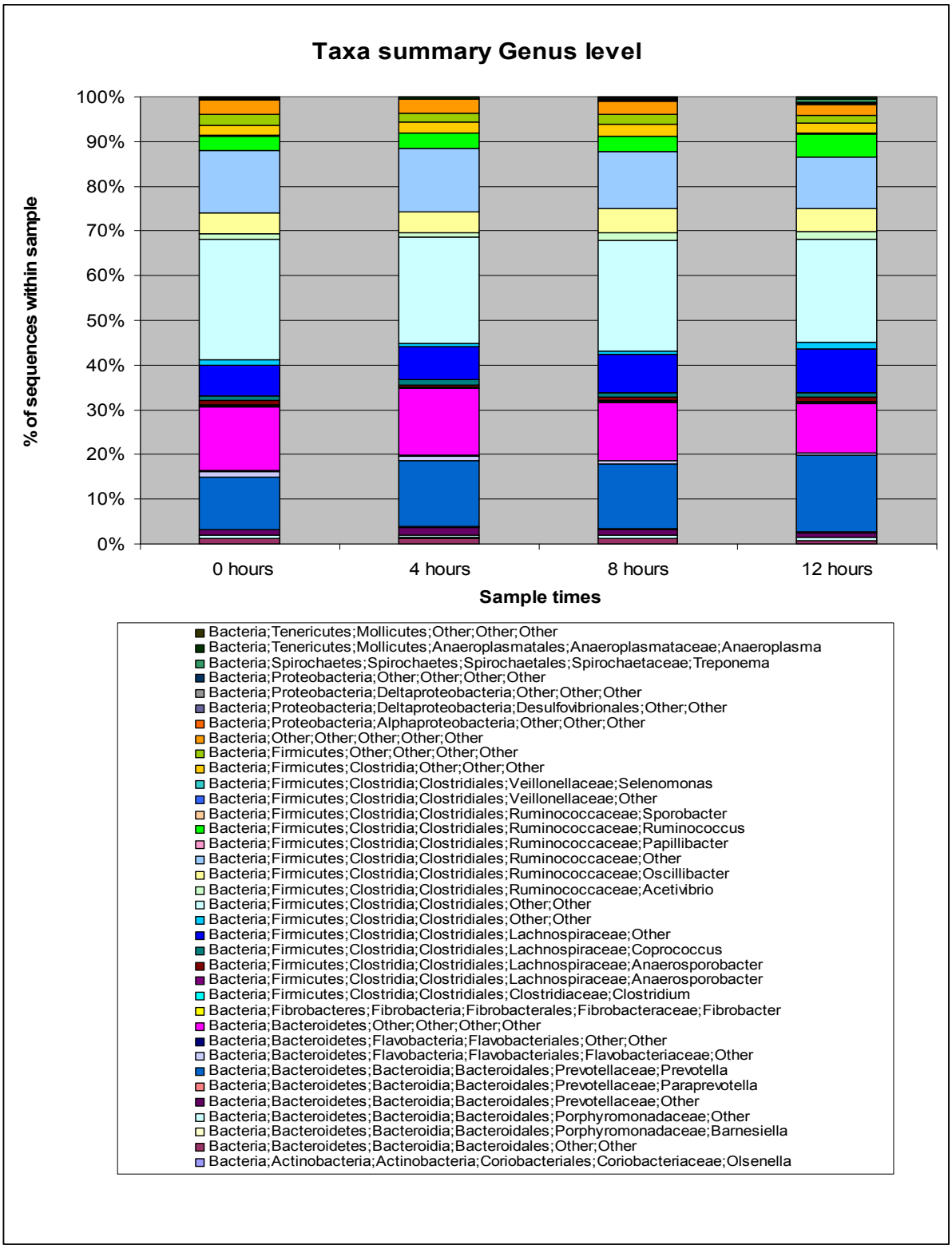

Figure 14. Taxonomic summary of bacterial OTUs at the Genus level presented as a percentage of the total number of sequences obtained for the respective sample collection time point.

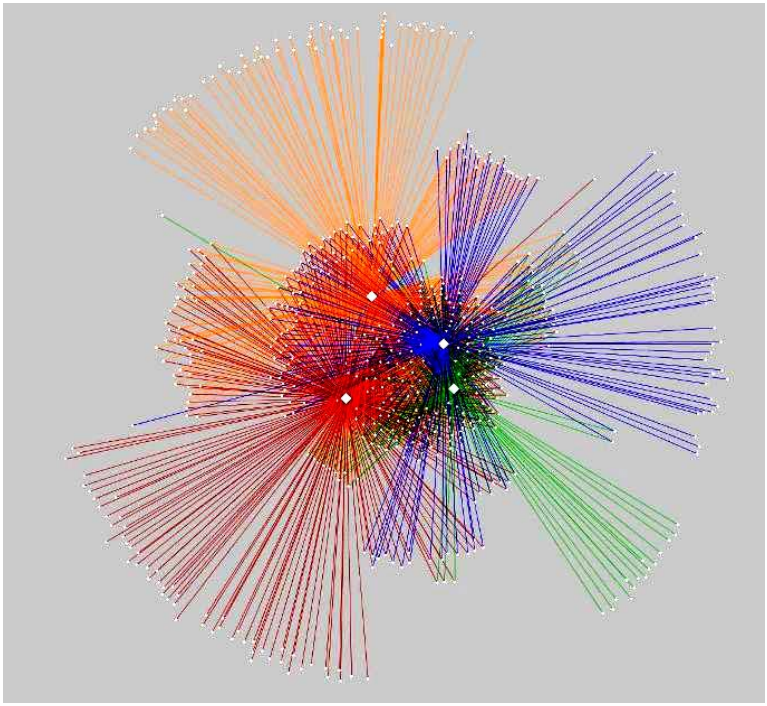

Figure 15. An OTU network of bacterial OTUs present at different time points with large white diamonds representing the samples and OTUs represented by small white diamonds. The edges belonging to each sample were coloured separately with 0 h (red) before feeding and 4 h (blue), 8 h (orange) and 12 h (green) post-feeding.

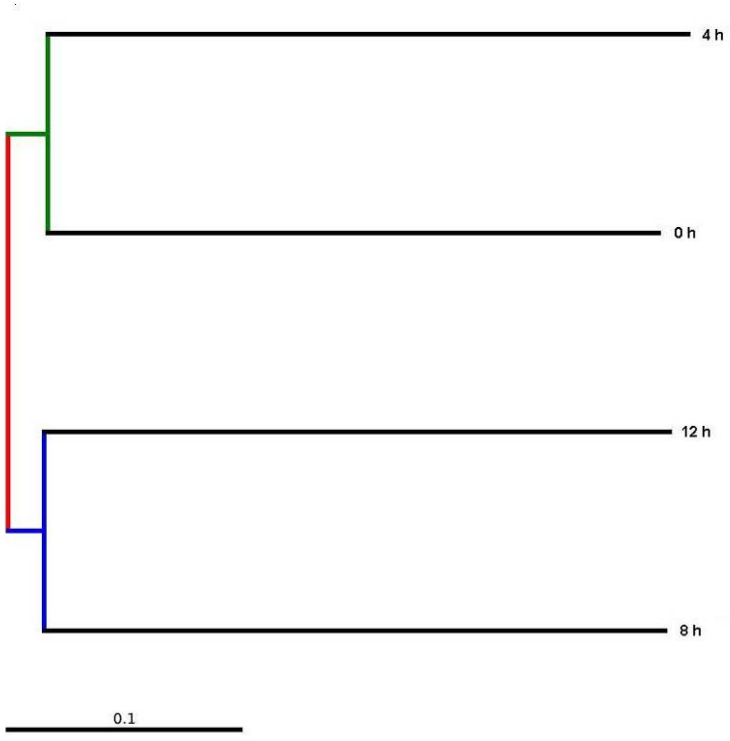

Figure 16. Measure of beta diversity, simplified distance tree for the bacterial communities (Jack knifed UPGMA) in rumen fluid sampled at four time points – 0 h, 4 h, 8 h and 12 h. Internal nodes are coloured to indicate support – red for 75-100% support, green for 20-50% and blue for < 25% support. Scale bar = 1 substitution per 10 nucleotides, corresponding to 10 % sequence divergence.

Table 14. Statistical analysis summary, alpha diversity measures for the bacterial and archaeal populations corresponding to each sample time point.

Microbial Population	Time (h)	Total number of observed species (OTUs)*	Equitability**	Shannon-Wiener Index†
Bacteria	0	402	0.828	7.160
	4	360	0.853	7.240
	8	411	0.839	7.284
	12	244	0.871	6.909
Archaea	0	17	0.357	1.444
	4	19	0.453	1.921
	8	24	0.373	1.712
	12	27	0.528	2.517

*Observed species (OTUs, measure of richness) normalised between samples with the minimum number of bacterial sequences for all samples estimated (from rarefaction analysis) to be 8060, minimum number of archaeal sequences estimated (from rarefaction analysis) to be 899.
** Equitability as a measure of species evenness (complete evenness equal to 1.0).
† Shannon-Wiener index calculated as a measure of alpha diversity (the larger the value, the greater diversity within a sample).

4.2.3 Diurnal fluctuations in rumen archaeal populations

From an input of 293,074 raw sequences containing both bacterial and archaeal sequences, a total of 5,348 archaeal sequences were obtained that met the parameters outlined in Table 4. A total of 69 OTUs were obtained from the archaeal sequences and removing chimeric OTUs reduced the total number of OTUs to 49. A rarefaction curve generated of the archaeal 16S rRNA gene sequences obtained for each rumen fluid collection time point, showed the curves flattening (Figure A.1.3, Appendix 1.) indicating that more intensive sampling would be likely to yield few additional new OTUs.

Similarly to the bacterial 16S rRNA gene data, for each of the four sample collection time points, the different archaeal taxa identified to genus level are presented as a stacked bar graph (Figure 18.). The archaeal population of the steer was dominated by several populations belonging to the genus *Methanobrevibacter,* contributing 80 % of the overall sequences for each sample time point. Archaeal populations with lesser abundance included representatives of the genus *Methanomicrobium,* the order Thermoplasmatales (related to thermophilic acidiphiles) and the family Methanobacteriaceae. Unclassified representatives of the phylum Crenarchaeota, which usually encompasses archaea only isolated from extreme habitats, were also detected in very low abundance within the archaeal datasets.

An OTU heatmap to represent the predominant archaeal populations was prepared and the most abundant 20 OTUs presented (Figure 17.) with colours (dark blue to red) used to highlight the relative levels of OTU abundance in each rumen fluid sample. Several time-related differences in the relative abundance of OTUs were observed, for example, one of the most dominant archaeal populations unrelated to the genus *Methanobrevibacter*, the OTU designated as *Methanomicrobium*, increased slowly in relative sequence abundance between 0 and 8 h (< 1 % of sequences), to being most abundant (10 to 20% of sequences) 12 h after feeding. Other non-*Methanobrevibacter* OTUs belonging to the *Thermogymnomonas*, *Methanimicrococcus* and Thermoplasmatales, all increased in relative sequence

abundance with time after feeding. Several OTUs corresponding to *Methanobrevibacter* sp. however, were always dominant within the rumen archaeal population. Further analysis of one OTU (#66) which represented 50 to 72 % of sequences within each of the four time points, was found to have a 99 % match over 479 bp to the *Methanobrevibacter millerae* strain ZA-10, a formate-utilising methanogen isolated from bovine rumen contents (Rea *et. al*, 2007).

Differences occurring within the archaeal community structure over time were also illustrated using an OTU network (Figure 19.). In this figure, archaeal populations present within all four rumen fluid samples, representative of core rumen archaea genera, are grouped together in the centre of the network (represented by small diamond shapes). Archaeal populations not shared between all samples are depicted as stretching out beyond the central, core populations. Where outlying archaeal OTUs are also present, for example, in another sample time-point, the coloured lines intersect. In this way, similarly to the bacterial OTU network (Figure 15.), distinct time-dependant changes can be observed for a relatively small proportion of the archaeal population. The bacterial OTU network however, encompasses many more OTUs, representing a much more diverse and populous community. The archaeal OTU network therefore more clearly depicts time-dependant changes in community structure. For example, the rumen fluid sample collection taken immediately prior to feeding (t = 0 h) contains only one OTU completely unique to that collection time point, the number of unique OTUs then increases in relation to time after feeding, with the 4, 8 and 12 h samples having 2, 7 and 8 unique archaeal OTUs respectively. The greatest shift in archaeal diversity therefore appears to occur between 4 and 8 h after feeding, with a marked increase in archaeal community diversity observed within the 12 h time period.

To further assess the time-related differences in archaeal populations, the archaeal communities identified for each of the four collection time points were analysed using phylogenetically aware measures of beta diversity (both weighted and unweighted unifrac) and a jackknifed replicate UPGMA phylogenetic tree of the total archaeal OTU sequences prepared (Figure 20). This analysis indicated that the archaeal communities at the 0 h and 4 h time points were most related, clustering together with 20-50 % support and the archaeal communities found at the 8 h and 12 h time points also clustered together at 20-50 % support. These results indicated a shift in rumen archaeal populations, occurring sometime between 4 and 8 hours after feeding. Additional statistical measures of population alpha diversity, including total species observed, Evenness and the Shannon-Wiener index (Table 14.) also indicated that rumen archaeal populations increased in overall species number and diversity during the 12 h period.

Figure 17. Heat map of the top 20 OTUs from the archaeal partial 16S rRNA gene sequence data set with the number of sequences within the OTUs for each sample time point. OTUs representing over 50 % of the sequences within the sample are coloured red; 20 to 25 % orange; 10 to 20 % sky blue; 5 to 10 % light blue; 1 to 5 % mid-blue and less than 1 % dark blue.

Consensus Lineage	0 hour	4 hour	8 hour	12 hour	#OTU ID
Methanobrevibacter	847	553	957	909	66
Methanomicrobium	12	37	63	289	11
Methanobrevibacter	289	205	130	174	4
Thermoplasmatales	14	27	70	141	59
Methanobacteriaceae	18	36	26	71	14
Thermogymnomonas	12	11	18	39	29
Methanimicrococcus		8	1	21	34
Thermoplasmatales			4	20	18
Methanobrevibacter	7	4	15	15	42
Methanobrevibacter	17	9	11	15	45
Thermoplasmatales	2	4	12	14	31
Thermoplasmatales	1	3	7	14	52
Thermoplasmatales	2	2	3	11	30
Methanosphaera	4	5	3	9	13
Methanobacteriaceae			1	5	27
Methanomicrobium	1		1	5	67
Methanobrevibacter		4	1	4	57
Methanobrevibacter	5	2	1	3	9
Methanobrevibacter	2	1		3	46
Thermoplasmatales	1		2	5	68

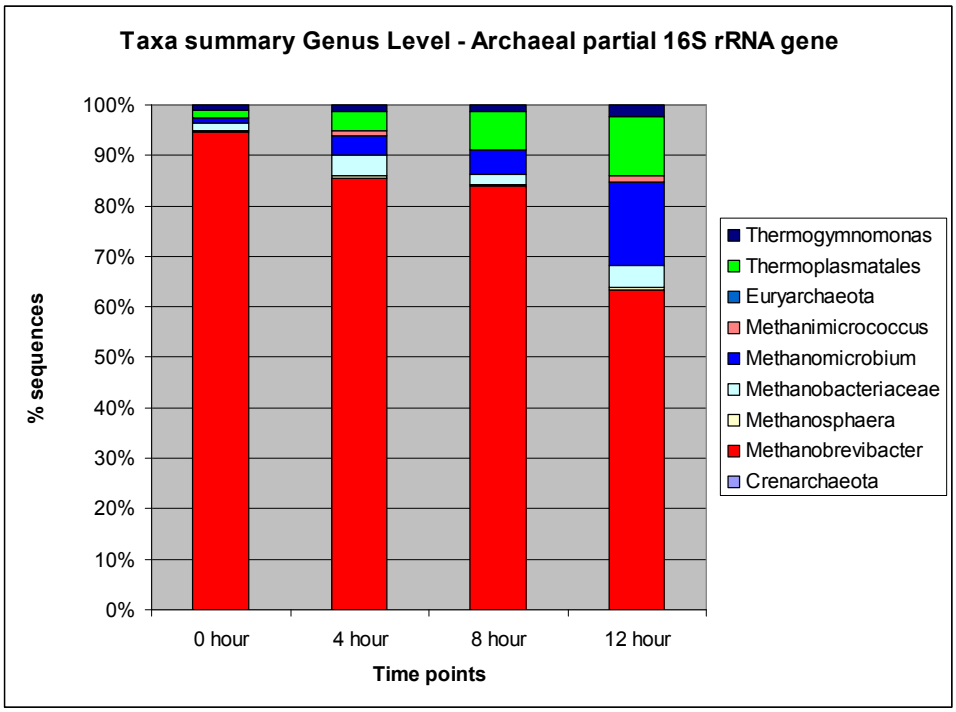

Figure 18. Taxonomic summary of archaeal OTUs at the Genus level presented as a percentage of the total number of sequences obtained for the respective sample time point.

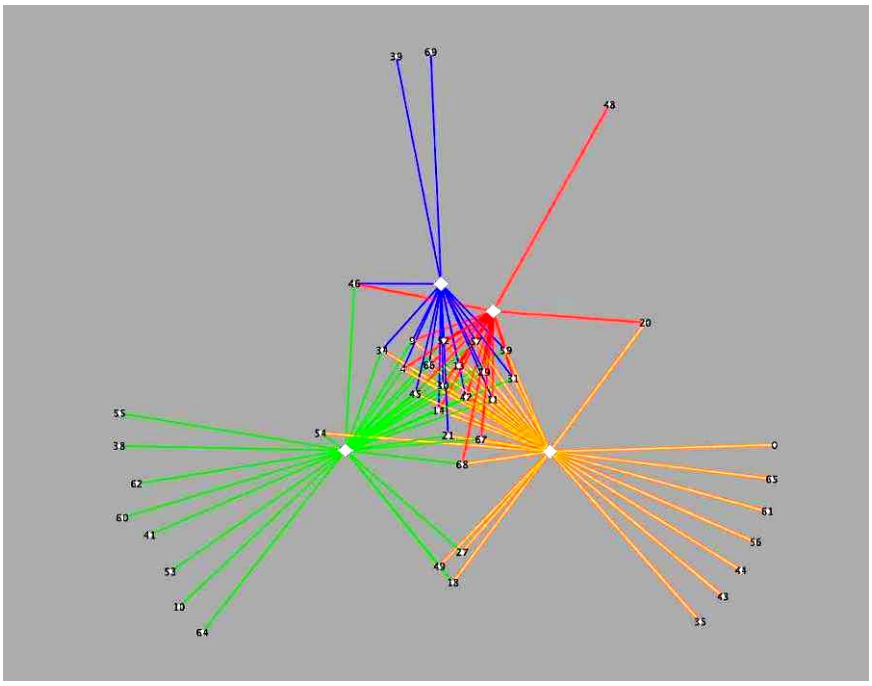

Figure 19. An OTU network of archaeal OTUs at different time points, with large white diamonds representing the samples and OTUs represented by smaller white diamonds. The edges belonging to each sample are coloured separately with 0 h (red) before feeding and 4 h (blue), 8 h (orange) and 12 h (green) post feeding.

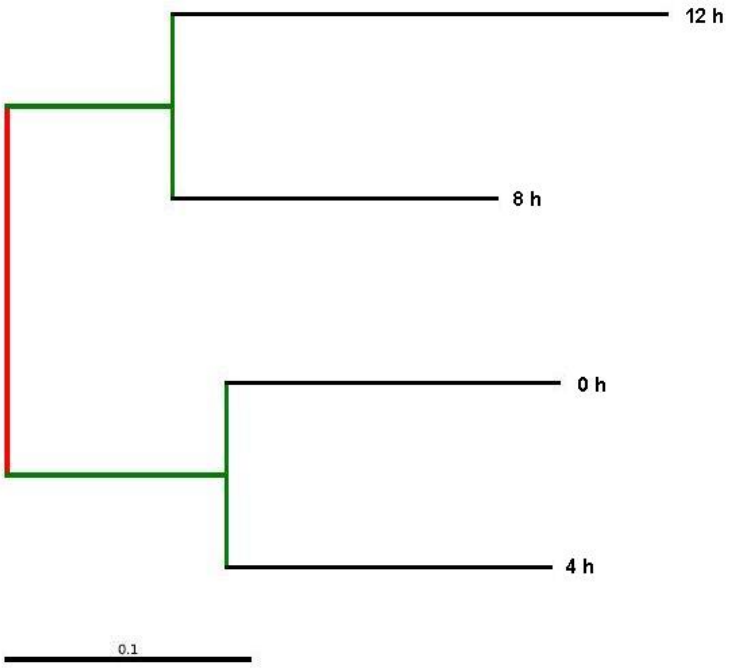

Figure 20. Measure of beta diversity, simplified distance tree for archaeal communities (Jack knifed UPGMA) in rumen fluid sampled at four time points (0, 4, 8 and 12 h). Internal nodes on the tree are coloured to indicate support – red for 75-100 % support and green for 20-50 % support. Scale bar = 1 substitution per 10 nucleotides, corresponding to 10 % sequence divergence.

4.3 Genome analysis of φAR29

The rumen phage φAR29 is normally present within the genome of the bacterial host *Prevotella ruminicola* AR29, producing intact phage particles when exposed to the phage inducing agent mitomycin C (Klieve *et al*. 1989). This phage has been the focus of previous scientific attention, with φAR29 genes being incorporated into vector systems used in the development of recombinant rumen bacteria (Gregg *et al*. 1994). The complete sequence of this phage and several phage genes were determined by Dr Shawn Seet, as part of his PhD studies (Seet, 2005). The expansion of virus sequence databases in recent years has enabled us to revisit this phage sequence and identify several additional phage genes within the draft sequence (Table 15.) although of the 55 coding domains (CDs) identified, only approximately one third of φAR29 genes clearly identified with genes with known function. Genes identified related to phage particle assembly and packaging (terminase gene), genes controlling phage genome integration into the bacterial host genome (integrase and integration protein), a single tRNA gene and other non-structural proteins (e.g. N-acetylmuramoyl-L-alanine amidase, which may function as a lysozyme). Interestingly, the φAR29 genome contains genes also found in the genome of bacteria other than the known host strain Prevotella ruminicola AR29, including *Bacteroides* sp. D2, *Bacteroides ovatis* (ATCC 8483), *Bacteroides* sp. D1 and *Bacteroides* sp. 224.

Undertaking this sequence analysis within the scope of the current project has enabled us to further develop the bio-informatics skills required to analyse this kind of phage-based genomic data.

Table 15. Preliminary analysis of φAR29, identification of genes with known function and relative sequence position.

φAR29 Genes with known function	**Genome position (nt)**
Phage terminase, large subunit	4472-5599
tRNA gene	17028-16914
Putative filamentous protein	16295-16675
Integration protein	19246-20007
Excisionase (Xis)	20019-20549
Integrase IntN1	20870-22249
Putative two-component sensor histidine kinase (relication/transcriptor proteins)	23231-23746
N-acetylmuramoyl-L-alanine amidase	24243-23743
Glutamate synthase [NADPH] large chain	30553-29564
MAEBL, putative	33209-30660

5.0 Discussion / Conclusion

The investigations undertaken during the course of this one-year project involved developing methodologies for the purification and concentration of viral particles from rumen fluid and the extraction of viral genomic material (DNA and RNA) of sufficient quality and quantity for further metagenomic analysis, using high through-put 454-pyrosequencing (Genome Sequencer, GS FLX Titanium platform, Roche). The methodology developed was based on the methodology originally developed by Klieve and Swain (1993) and later utilised by the only published rumen virome study (Berg Miller *et al.* 2011). In the current project however, the methodology was markedly modified in several areas. The revised methodology involved (a) the introduction of a PEG/NaCl precipitation step to aid the extraction of virus particles from rumen fluid samples; (b) DNase treatment steps to remove contaminating bacterial DNA; and (c) omitted the process of embedding the virus particles in low melting temperature agarose (as would be required for PFGE analysis). The methodology developed also managed to avoid the use of additional PCR amplification steps which have previous been found to introduce sequencing bias (Henn *et al*. 2010).

The methodology developed for purifying viral RNA from rumen fluid is completely novel, with the RNA virus populations of the rumen currently undescribed in the scientific literature. Unfortunately due to delays incurred by the external sequencing provider, sequence data for the RNA virus population of the rumen is not reported. The current study successful in obtaining sequence data for two rumen virome samples encompassing the DNA virus populations, four rumen bacterial population samples and four rumen archaeal population samples, the latter two microbial population studies being based on 16S rRNA gene amplicon sequencing, using multiple PCR primers sets specific for each microbial population.

All samples prepared for high through-put sequencing arose from a short animal trial conducted at UQ Gatton, with four rumen fluid samples collected from a once-a-day fed Brahman cross steer over the course of a 12 hour period, with the first sample collected immediately before feeding. This experiment was undertaken in order to verify if single collections of rumen fluid would adequately reflect the full extent of the virus population, given that particularly the DNA phage population of the rumen, has been previously shown by PFGE to fluctuate in response to the time after feeding (Swain *et al*. 1996). The previously published rumen virome study (Berg Miller *et a*l. 2011) only predicted the rumen microbial population using the taxonomic profiles of the three virome datasets described and comparison to completely different bovine rumen fluid metagenomic studies (Brulc *et al*. 2009 and Hess *et al*. 2011). In the current study, examining the bacterial and archaeal populations of the rumen samples also prepared for virome analysis, contributed to the depth of the study, facilitating the identification of (a) microbial groups which may be directly contributing to the rumen fluid virome (blooms of lysogenic phage); and (b) rumen microbial communities which may be susceptible to phage infection. Interestingly, examining the rumen microbial communities (bacteria and archaea) over the course of a 12 h period, provided evidence that these microbial communities are not static in response to feeding, with distinct patterns of diurnal fluctuation emerging. While changes in the rumen bacterial populations over time, in response to feeding, have been previously reported (Leedle *et al*. 1982), to our knowledge, this is the first examination of these changes to be investigated using a high through-put sequencing approach. Changes in the rumen archaeal (methanogen) populations were particularly obvious, given the lesser abundance and complexity of this microbial domain within the rumen, compared to the bacterial population. Despite

populations of *Methanobrevibacter* being dominant at all sample time-points, the methanogen population increased in diversity over time, with increasing abundance of populations such as *Methanomicrobium*, *Thermogymnomonas* and *Methanimicrococcus*. These time-dependant changes in rumen microbial populations are most likely to be in direct response to the changing availability of substrates for microbial growth, as suggested by the earlier studies (Leedle *et al*. 1982).

The rumen virus population also changed in response to time after feeding. Predominant phage types represented within the sequence data, included phages for host species identified as being highly dominant within the respective rumen bacterial populations (Results 4.2.2). As a consequence of the Lucerne and Rhodes Grass diet being fed to the steer employed in the current study, rumen bacterial populations were dominated by those Firmicutes belonging to the order Clostridiales. In the 12 h (T12) virome dataset, for example, viral sequences corresponding to 10 different *Clostridium* phages were detected, six of which had an abundance of > 1% of the total virome population. Interestingly for the 12 h (T12) virome, viral sequences corresponding to 21 different *Lactoccocus* phages were also detected, with only 2 having >1 % abundance, yet isolates of this bacterial genus were under-represented within the matching bacterial dataset. A combination of two attributing factors may explain this discrepancy. Firstly the rumen bacterial population was found to be dominated by gram positive organisms closely related to *Lactococcus* (also belonging to the Division Firmicutes, class Bacilli, family Streptococcaceae), so it is not unlikely that phages able to infect *Lactococcus* may have similar structural and functional genes to phages infecting closely related rumen Firmicutes. Secondly, the *Lactococcus* genus includes isolates which are very well characterised, being heavily utilised in the dairy (cheese-making) industry and phages for this genus have been extensively studied with many *Lactococcus* phage sequences deposited in the virus reference sequence databases. Rumen-derived phages however are very poorly represented within the virus reference sequence databases, therefore sequence analysis of rumen virome datasets may tend towards to the identification of genes only currently designated to the most closely related viruses present in the database. The limitations of currently available viral reference sequence databases also contributed to the relatively small proportion of sequences within each virome dataset (approximately 9 % of sequences when using an e-value threshold of 10^{-5}) that could be directly ascribed to existing virus taxons.

Despite these limitations, a considerable number of virus sequences were identified within the rumen virome datasets, with an estimated total of 6112 and 5546 viruses identified within the 4 h (T4) and 12 h (T12) virome sequence datasets respectively (GAAS estimate with e-value threshold of 10^{-5}). In agreement with the previous rumen virome study (Berg Miller *et al.* 2011) and large scale sequencing, rumen metagenomic studies which also picked up some virus-related sequences (Brulc *et al.* 2009; Hess *et al*. 2011), the vast majority of identified rumen virome sequences were associated with the virus order Caudovirales (representing 94 % of identified virus sequences). This order includes ds DNA, tailed bacteriophages, traditionally differentiated into families on the basis of phage particle morphology, specifically tail structure. The virome dataset obtained for each time-point of rumen fluid sample collection had similar proportions of each virus family, with sequences attributed to the Siphoviridae being the most abundant (63 % of sequences designated as Caudovirales), with only small differences (< 2 % of sequences) occurring between the relative proportion of Myoviridae and Podoviridae present in each respective virome dataset. Time-related differences occurring between each virome dataset were more obvious when examining the sequences at the level of individual

identified virus taxon. This approach even distinguished differences in the relative abundance and variety of sequences attributed to known archaeaphage present in each virome dataset (Table 10).

Previous studies have suggested that the rumen phage population is sustained at high concentrations of approximately 10^{10} phage particles per ml rumen fluid (Klieve and Swain, 1993), as a consequence of continual phage replication events, resulting in blooms of phage particles and associated microbial host lysis. It has also been thought that a very high proportion of rumen bacteria contain lysogenic phage integrated within their genome (Klieve *et al*. 1989). As more rumen bacteria and archaea have their genome fully sequenced, the predicted high prevalence of lysogenic phage is being confirmed (Leahy *et al*. 2010, Berg Miller *et al.* 2011). Taxonomic analysis of the two rumen virome sequence datasets confirmed the very high abundance of lysogenic phages, with a very large proportion of sequences obtained from the virus preparations (which were essentially free of microbial DNA contamination), directly attributed to genes also found within microbial genomes included in microbial sequence databases. For example, for the 4 h sample time point virome, containing a total of 69,133 sequences, 25.25 % of sequences were related to bacterial genes and 0.15 % of sequences were related to archaeal genes. This confirms the findings of the only other rumen virome study (Berg Miller *et al.* 2011) and supports previous culture-based studies which suggested that the activity of lysogenic phages may play a major role in sustaining rumen phage populations and may actively contribute to horizontal gene transfer within the rumen microbial ecosystem (Klieve *et al*. 1989).

The PHACCS program (Angly *et al.* 2005) was also used to model the virus populations present within each virome (alpha diversity measures) and the Maxiphi program (Angly *et al.* 2006) was used to determine compare differences in virus populations between the viromes (beta diversity measures). A more mathematical approach is being widely adopted in virome population studies, where the majority of viruses present are unclassified (Angly *et al*. 2006), for estimating virus species richness and diversity, without necessarily determining virus taxonomy. The program Maxiphi was also used to determine time-related changes in the virus populations. This analysis approach indicated that the viruses present at the first sample time point (4 h) were also present at the second time point (12 h), with measures of alpha diversity indicated that virus population diversity considerably increased with time after feeding (as indicated by a decrease in the evenness estimate). Therefore throughout the 8 h period, new virus genotypes either increased to detectable levels or were introduced into the rumen fluid (possibly through phage lytic events or blooms), effectively diluting the viruses present within the earlier time point. While these results are interesting and correspond with the changes in virus population diversity observed in PFGE-based studies (Swain *et al*. 1996), further determination of the additional two sample time-points obtained for prior to feeding and 8 h after feeding may be required to affirm this initial finding.

The analysis of virome sequence data is widely recognised as being relatively difficult due to the lack of highly conserved genes (such as the species-specific 16 S rRNA gene of bacteria and archaea) which could be utilised to genetically differentiate virus families. This area of virus research is rapidly evolving and several virus marker genes have been recognised as being conserved enough between specific groups of viruses to enable these genes to be utilised for phylogenetic analysis. The Metavir virome analysis web site (Roux *et al*. 2011) has incorporated several of these genes into it's analysis toolbox and the two rumen virome sequence datasets and three other published rumen virome datasets (derived from Holstein cow rumen fluid, Berg Miller *et al*. 2011) were analysed in this way. Of the marker

genes available, the Caudovirales-specific terminase large subunit gene (TerL) and DNA polymerase family B gene (PolB) were found to be the most useful for rumen-derived viruses. Use of these marker genes also facilitated the detection of virome sequences directly related to the reference sequence archaeaphages, *Methanobacterium* phage psiM2, *Methanothermobacter* phage psiM100 and *Halorubrum* phage HF2. The *Methanobrevibacter ruminantium* M1 integrated prophage φmru (Leahy *et al*. 2010) is not yet included in the reference sequence database utilised for this analysis, so further analysis would be required to determine if marker gene sequences related to this methanogen prophage are included within the rumen virome sequence datasets. It is anticipated that the PolB marker gene may be particularly useful for the identification of archaeaphage-related sequences within larger virome datasets, as PolB-like genes are found in all of the main archaeal lineages (Nanoarchaeota, Crenarchaeota and Euryarchaeota) and archaeal viruses such as Halovirus HF1 (Monier *et al*. 2008).

In conclusion, the laboratory-based and bioinformatics methodology developed during the course of this project may be applied to the examination of virus populations in rumen fluid samples obtained from larger scale experiments and has greatly advanced the current understanding of the bovine rumen virome and associated microbial populations. This preliminary study successfully utilised metagenomic analysis tools to establish the presence of viruses related to known archaeaphages within bovine rumen fluid samples, through taxonomic analysis and the identification of key viral marker genes. It is anticipated that using a metagenomic approach as presented in this study will greatly facilitate the identification of environmental samples rich in archaeaphage (for targeted archaeaphage isolation experiments) and also identify archaeaphage functional genes, such as those involved in host-specific cell lysis (proteolysis genes such as endoisopeptidases) which may be utilised in the future as archaeocins.

6.0 Bibliography

Adams, J.C., Gazaway, M.D., Brailsford, M.D., Hartman, P.A. and Jacobson, N.L. (1966) Isolation of bacteriophages from the bovine rumen. *Experientia* **22,** 717-718.

Angly, F., Rodriguez-Brito, B., Bangor, D., McNairnie, P., Breitbart, M., Salamon, P., Felts, B., Nulton, J., Mahaffy, J. and Rohwer, F. (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BCM Bioinformatics* **6**, 41.

Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., Chan, A.M., Haynes, M., Kelley, S., Liu, H., Mahaffy, J.M., Mueller, J.E., Nulton, J., Olsen, R., Parsons, R., Rayhawk, S., Suttle, C.A. and Rohwer, F. (2006) The marine viromes of four oceanic regions. *PLOS Biology*. **4,** e362-e368.

Angly FE, Willner D, Prieto-Davó A, Edwards RA, Schmieder R, Vega-Thurber R, Antonopoulos DA, Barott K, Cottrell MT, Desnues C, Dinsdale EA, Furlan M, Haynes M, Henn MR, Hu Y, Kirchman DL, McDole T, McPherson JD, Meyer F, Miller RM, Mundt E, Naviaux RK, Rodriguez-Mueller B, Stevens R, Wegley L, Zhang L, Zhu B, Rohwer F. (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLOS Computational Biology*. **5**, e1000593.

Berg-Miller, M., Yeoman, C.J., Chia, N., Tringe, S.G., Angly, F.E., Edwards, R.A., Flint, H.J., Lamed, R., Bayer, E.A. and White, B.A. (2011) Phage-bacteria relationships and CRISPR elements revealed by a metagenomic survey of the rumen microbiome. *Environ Microbiol* **14***,* 207-227*.*

Brulc, J.M., Antonopoulos, D.A., Miller, M.E., Wilson, M.K., Yannarell, A.C., Dinsdale, E.A., Edwards, R.E., Frank, E.D., Emerson, J.B., Wacklin, P., Coutinho, P.M., Henrissat, B., Nelson, K.E. and White, B.A. (2009) Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside ydrolases *Proc Natl Acad Sci U.S.A.* **106**, 1948-1953.

Caporaso J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J. and Knight, R. (2010a) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*. **7**, 335-336.

Caporaso, J.G., Bittinger, K., Bushman, F.D., DeSantis, T.Z., Andersen, G.L. and Knight, R. (2010b) PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics.* **26**, 266–267.

Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., Marsh, T., Garrity, G.M. and Tiedje, J.M. (2009) The ribosomal database project: improved alignments and new tools for rRNA analysis. *Nucl. Acids Res*. **36**, D141-D145.

Comeau AM, Chan AM & Suttle CA (2006) Genetic richness of vibriophages isolated in a coastal environment. *Environ Microbiol* **8**, 1164–1176.

Culley, A. and Steward, G. (2007). New genera of RNA viruses in subtropical seawater, inferred from polymerase gene sequences. *Appl Environ Microbiol* **73**, 5937–5944.

DasSarma, S. and Fleischmann, E.F. (1995) In '*Archaea: A Laboratory Manual – Halophiles.* pp. 269-72. Cold Spring Harbour Laboratory Press: New York.

Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M *et al*. (2008). Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* **452**, 340–343.

Dinsdale, E.A., Edwards, R.A., Hall, D., Angly, F., Breitbart, M., Brulc, J.M., Furlan, M., Desnues, C., Haynes, M., Li, L. et al. (2008) Functional metagenomic profiling of nine biomes. *Nature,* **452**, 629-632.

Dorigo, U., Jacquet, S., Humbert, J.F. (2004) Cyanophage diversity, inferred from g20 gene analyses, in the largest natural lake in France, Lake Bourget. Appl Environ Microbiol 70, 1017-1022.

Gantner, S., Andersson, A.F., Alonso-Saez, L. and Bertilsson, S. (2011) Novel primers for 16S rRNA-based archaeal community analyses in environmental samples. *J Microbiol Methods* **84**, 12-18.

Gerlach, W. and Stoye, J. (2011) Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic acids Res* **39**, e91.

Hemmerich, C., Buechlein, A., Podicheti, R., Revanna, K.V. and Dong, Q. (2010) An Ergatis-based prokaryotic genome annotation web server. *Bioinformatics* **26**,1122-1124.

Hess M, Sczyrba A, Egan R, Kim T-W, Chokhawala H, Schroth G *et al*. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**, 463-467.

Hoogenraad, N.J., Hird, F.J.R., Holmes, I. and Millis, N.F. (1967) Bacteriophages in rumen contents of sheep. *J Gen Virol* **1**:575-576.

Huang, Y., Beifang, N., Goa, Y., Fu, L. and Li, W. (2010) CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680-682.

Iverson, W.G. and Millis, N. F. (1976) Characterisation of *Streptococcus bovis* bacteriophage. *Can J Microbiol* **22**, 847-852.

Kim, M.S., Park, E.J., Roh, S.W. and Bae, J.W. (2011) Diversity and abundance of single-stranded DNA viruses in human feces. *Appl Environ Microbiol* **77**, 8062-8070.

Klieve, A.V. and Bauchop, T. (1988) Morphological diversity of ruminal bacteriophages from sheep and cattle. *Appl Environ Microbiol* **54**, 1637-1641.

Klieve, A.V., Hudman, J.F. and Bauchop, T. (1989) Inducible bacteriophages from ruminal bacteria. *Appl Environ Microbiol* **55**, 1630-1634.

Klieve, A.V. and Bauchop, T. (1991) Phage resistance and altered growth habit in a strain of *Streptococcus bovis. FEMS microbiol lett* **80**,155-160.

Klieve, A.V., Gregg, K. and Bauchop, T. (1991) Isolation and characteristics of lytic phages from *Bacteroides ruminicola* ss *brevis. Current Microbiol* **23**,183-187.

Klieve, A.V. and Swain, R.A. (1993) Estimation of ruminal bacteriophage numbers by pulsed-field gel electrophoresis and laser densitometry. *Appl Environ Microbiol* **59**, 2299-2303.

Klieve, A.V., Heck, G.L., Prance, M., and Shu, Q. (1999) Genetic homogeneity and phage susceptibility of ruminal strains of *Streptococcus bovis* isolated in Australia. *Lett Appl Microbiol* **29**,108-112.

Klieve, A.V., Bain, P.A., Yokoyama, M.T., Ouwerkerk, D., Forster, R.J. and Turner, A.F. (2004) Bacteriophages that infect the cellulolytic ruminal bacterium *Ruminococcus albus* AR67. *Lett Appl Microbiol* **38**, 333-338.

Klieve, A.V. (2005) 2.2. Bacteriophages. In *'Methods in gut microbial ecology for ruminants*'. Eds H.P.S. Makkar, C.S. McSweeney pp. 39-46. International Atomic Energy Agency. Springer Academic Press: Netherlands.

Klieve, A.V. and Gilbert, R.A. (2005) 4.2 Bacteriophage populations. In *'Methods in gut microbial ecology for ruminants*'. Eds H.P.S. Makkar, C.S. McSweeney pp. 129-137. International Atomic Energy Agency. Springer Academic Press: Netherlands.

Lane, D.J. (1991) 16S/23S rRNA sequencing. In '*Nucleic acid techniques in bacterial systematics*.' Eds. Stackebrant, E., Goodfellow, M. pp. 115-175. Academic Press: United Kingdom.

Leahy, S.C., Kelly, W.J., Altermann, E., Ronimus, R.S., Yeoman, C.J., Pacheco, D.M., Li, D., Kong, Z., McTavish, S., Sang, C., Lambie, S.C., Janssen, P.H., Dey, D. and Attwood, G.T. (2010) The genome sequence of the rumen methanogens *Methanobrevibacter ruminantium* reveals new possibilities for controlling ruminant methane emissions. *PLoS One* **5**, e8926.

Leedle, J.A.Z., Bryant, M.P. and Hespell, R.B. (1982) Diurnal variations in bacterial numbers and fluid parameters in ruminal contents of animals fed low- and high-forage diets. *Appl Environ Microbiol* **44**, 402-412.

Lopes, C.T., Franz, M., Kazi, F., Donaldson, S.L., Morris, Q., and Bader, G.D. (2010) Cytoscape Web: an interactive web-based network browser. Bioinformatics,

Monnier, A., Claverie, H., Ogata, H. (2008) Taxonomic distribution of large DNA viruses in the sea. *Genome Biol* **9**, R106.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C. Nielsen, T. *et al*. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65.

Rea S., Bowman J. P., Popovski S., Pimm, C., and Wright, A. D. (2007) Methanobrevibacter millerae sp nov and Methanobrevibacter olleyae sp nov., methanogens from the ovine and bovine rumen that can utilize formate for growth. *Int J Syst Evol Microbiol* **57**, 450-456.

Rosario, K., Duffy, S. and Breitbart, M. (2009) Diverse circovirus-like genome architectures revealed by environmental metagenomics. *J Gen Virol* **90**, 2418-2424.

Roux, S., Faubladier, M., Mahul, A., Paulhe, N., Bernard, A., Debroas, D. and Enault, F. (2011) Metavir: a web server dedicated to virome analysis. *Bioinformatics* **27**, 3074-3075.

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., Colombet, J., Sime-Ngando, T., and Deboas, D. (2012) Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* **7**, e33641.

Rowe, J.M., Fabre, M.F., Gobena, D., Wilson, W., and Wilhelm, S.W. (2011) Application of the major capsid protein as a marker of the phylogenetic diversity of Emiliania huxleyiviruses. *FEMS Micobiol Ecol* **76**, 373-380.

Seet, S.G.M. (2005) Genome sequence of bacteriophage φAR29: a basis for integrative plasmid vectors. PhD thesis, Murdoch University.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G. Van Horn, D.J. and Weber, C.F. (2009) Introducing mother: Open-source, platform independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**, 7537-7541.

Schattner, P., Brooks, A.N. and Lowe, T.M. (2005) *Nucleic Acids Res*, **33**, W686-689.

Schoenfeld, T., Patterson, M., Richardson, P.M., Wommack, K.E., Young, M. and Mead, D. (2008) Assembly of viral metagenomes from Yellowstone hot springs. *Appl Environ Microbiol* **74**, 4164-4174.

Smoot, M.E., Ono, K., Ruschienski, J., Wang, P., and Ideker, T. (2011) Cytoscape 2.8: new features for data integration and network visualization. 27, 431-432.

Sullivan, M.B., Krastins, B., Hughs, J.L., Kelly, L., Chase, M., Sarracino, D. and Chisholm, S.W. (2009) The genome and structural proteome of an ocean siphovirus: a new window into the cyanobacetrial 'mobilome'. *Environ Microbiol* **11**, 2935-2951.

Swain, R.A., Nolan, J.V. and Klieve, A.V. (1996) Natural variability and diurnal fluctuations within the bacteriophage population of the rumen. *Appl Environ Microbiol* **62**, 994-997.

Watanabe, K., Kodama, Y. and Harayama, S. (2001) Design and evaluation of PCR primers to amplify bacterial 16S ribosomal DNA fragments used for community fingerprinting *J Microbiol Methods* **44**, 253-62.

Yu Z. and Forster R.J. (2005) Nucleic acid extraction, oligonucleotide probes and PCR methods. In *'Methods in gut microbial ecology for ruminants*'. Eds H.P.S. Makkar, C.S. McSweeney pp. 81-104. International Atomic Energy Agency. Springer Academic Press: Netherlands.

7.1 Appendix 1.

Sequence data quality summary

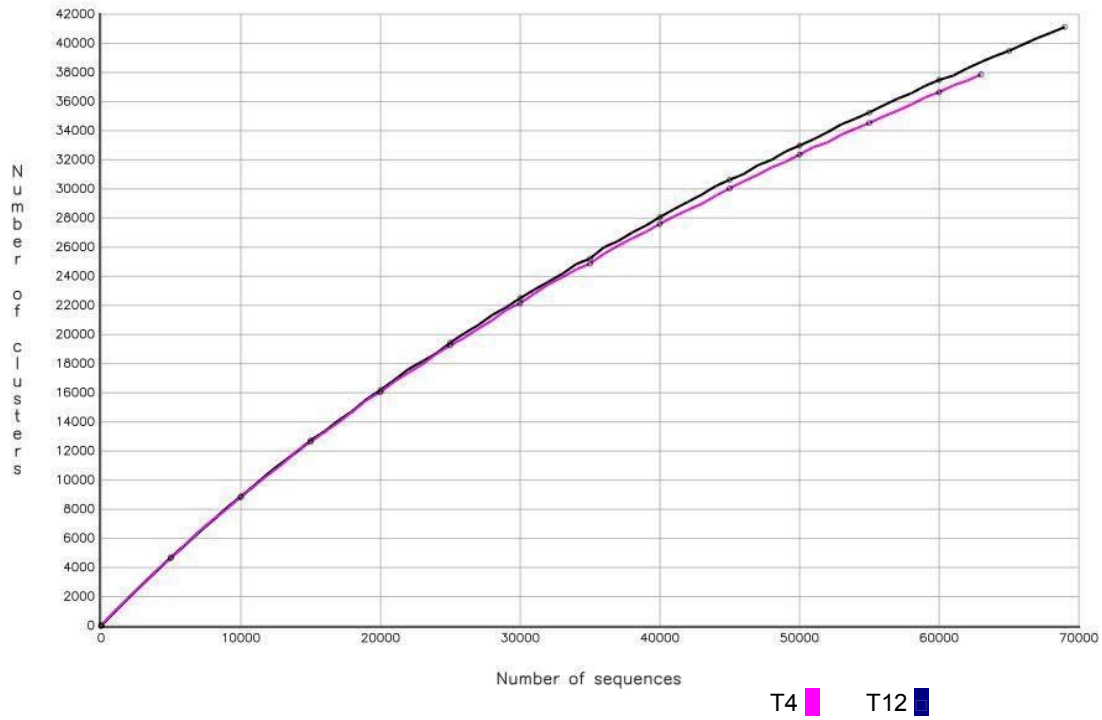

Figure A1.1 Rarefaction curves of complete (whole) virome datasets generated with the Metavir pipeline based on 98 % identity for virome samples corresponding to 4 h (T4) and 12h (T12) after feeding.

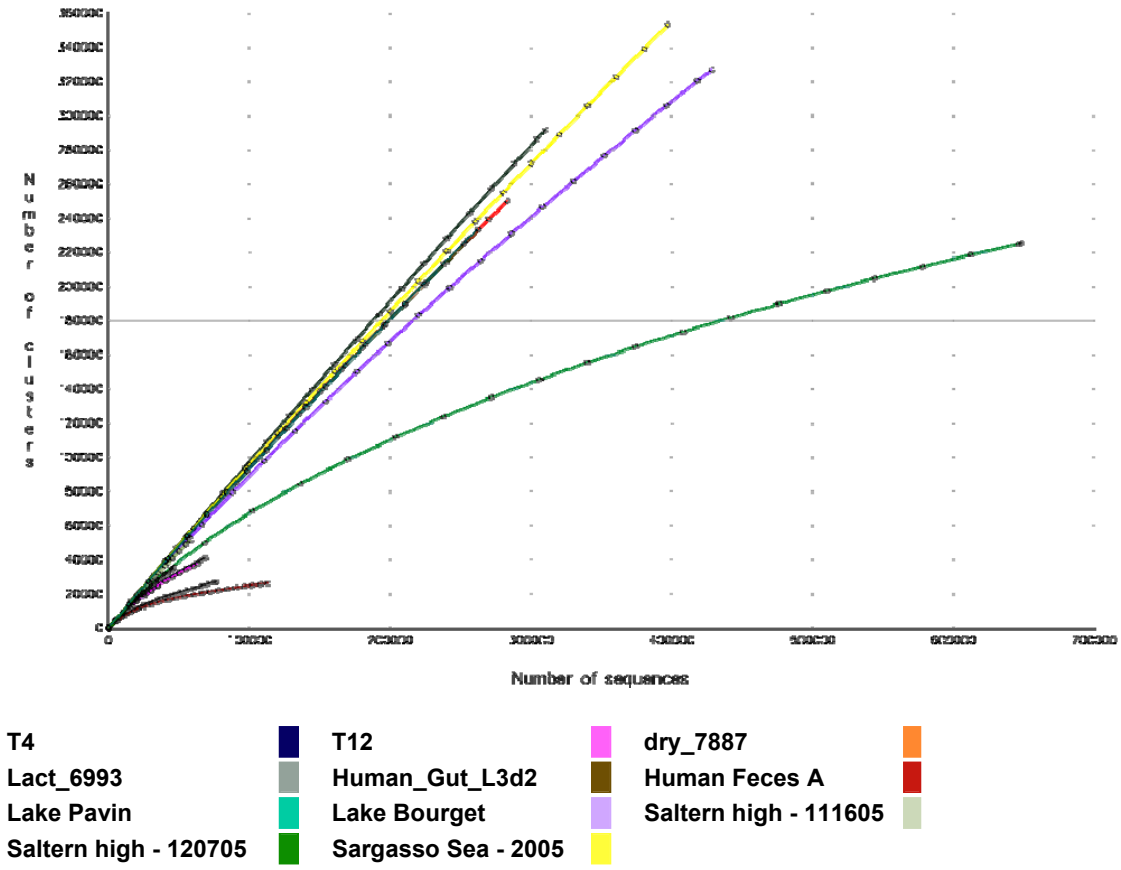

Figure A1.2 Rarefaction curves of complete (whole) virome datasets generated with the Metavir pipeline based on 98 % identity for virome samples corresponding to 4 h (T4) and 12h (T12) after feeding with virome data from several other published virome studies including bovine rumen (Lact_6993, dry_7887), freshwater lake (Lake Bourget, Lake Parvin), an ocean (Sargasso Sea – 2005), hypersaline environments (Saltern high- 120705 and 111605) and human digestive studies (human_gut_L3d2 and human faeces A).

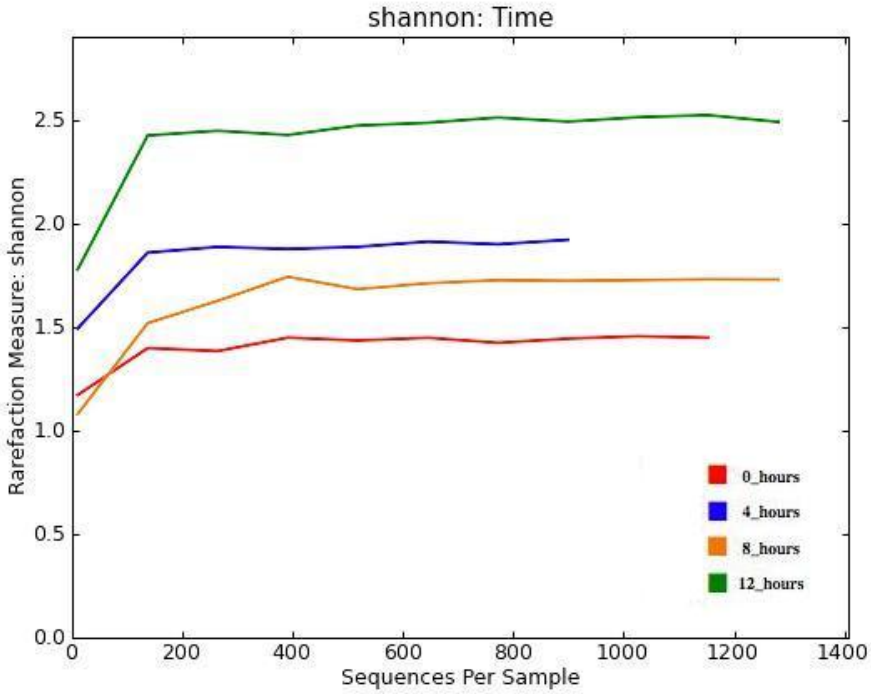

Figure A1.3 Rarefaction curves based on the Shannon index for the 16S rRNA gene sequences corresponding to the archaeal populations from rumen fluid collected at each time-point (h).

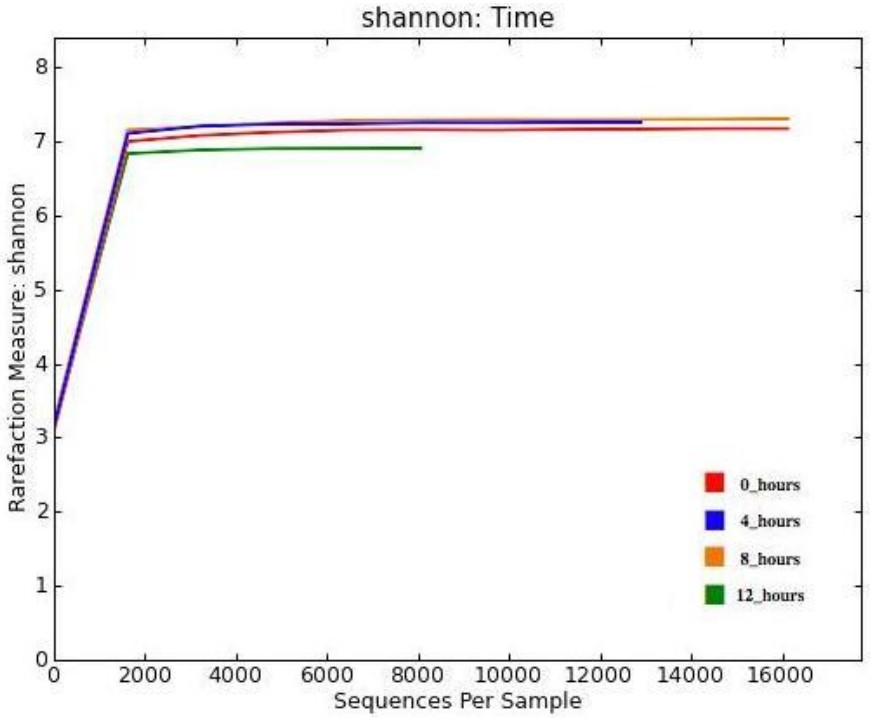

Figure A1.4 Rarefaction curves based on the Shannon index for the 16S rRNA gene sequences corresponding to the bacterial populations from rumen fluid collected at each time-point (h).

Figure A1.5 Example phylogenetic tree showing T4 virome sequences (in red) mapped in relation to their similarity (98 % similarity on 35 bp) to known TerL sequences, with the corresponding virus family listed in colour on the right. Analysis undertaken using BLASTx with terminase seq from reference sequences in PFAM database (e-value $<10^{-3}$), alignment with HMMER, Metavir.

7.2 Appendix 2.

Table A2.1 Sequence of primer barcodes used for specific amplification of the bacterial and archaeal 16S rRNA sequences respectively, with primers named according to the microbial community target (archaea or bacteria), rumen fluid collection time-point (0, 4, 8 and 12 h after feeding) and barcode number.

Primer name	Barcode Sequence
Arch-0-3	AGACGCACTC
Arch-4-2	ACGCTCGACA
Arch-8-4	AGCACTGTAG
Arch-12-5	ATCAGACACG
Bact-0-17	CGTCTAGTAC
Bact-0-2	ACGCTCGACA
Bact-0-3	AGACGCACTC
Bact-0-4	AGCACTGTAG
Bact-4-5	ATCAGACACG
Bact-4-6	ATATCGCGAG
Bact-4-7	CGTGTCTCTA
Bact-4-8	CTCGCGTGTC
Bact-8-9	TAGTATCAGC
Bact-8-10	TCTCTATGCG
Bact-8-11	TGATACGTCT
Bact-8-12	TACTGAGCTA
Bact-12-13	CATAGTAGTG
Bact-12-14	CGAGAGATAC
Bact-12-15	ATACGACGTA
Bact-12-16	TCACGTACTA

Table A2.2 Mapping files for QIIME analysis of bacterial and archaeal sequences

#SampleID	BarcodeSequence	LinkerPrimerSequence	Type	Time	Description
Bact.0.1	CGTCTAGTAC	TACGGGAGGCAGCAG	Bacteria	0_hours	Bact_0_17
Bact.0.2	ACGCTCGACA	TACGGGAGGCAGCAG	Bacteria	0_hours	Bact_0_2
Bact.0.3	AGACGCACTC	TACGGGAGGCAGCAG	Bacteria	0_hours	Bact_0_3
Bact.0.4	AGCACTGTAG	TACGGGAGGCAGCAG	Bacteria	0_hours	Bact_0_4
Bact.4.5	ATCAGACACG	TACGGGAGGCAGCAG	Bacteria	4_hours	Bact_4_5
Bact.4.6	ATATCGCGAG	TACGGGAGGCAGCAG	Bacteria	4_hours	Bact_4_6
Bact.4.7	CGTGTCTCTA	TACGGGAGGCAGCAG	Bacteria	4_hours	Bact_4_7
Bact.4.8	CTCGCGTGTC	TACGGGAGGCAGCAG	Bacteria	4_hours	Bact_4_8
Bact.8.9	TAGTATCAGC	TACGGGAGGCAGCAG	Bacteria	8_hours	Bact_8_9
Bact.8.10	TCTCTATGCG	TACGGGAGGCAGCAG	Bacteria	8_hours	Bact_8_10
Bact.8.11	TGATACGTCT	TACGGGAGGCAGCAG	Bacteria	8_hours	Bact_8_11
Bact.8.12	TACTGAGCTA	TACGGGAGGCAGCAG	Bacteria	8_hours	Bact_8_12
Bact.12.13	CATAGTAGTG	TACGGGAGGCAGCAG	Bacteria	12_hour	Bact_12_13
Bact.12.14	CGAGAGATAC	TACGGGAGGCAGCAG	Bacteria	12_hour	Bact_12_14
Bact.12.15	ATACGACGTA	TACGGGAGGCAGCAG	Bacteria	12_hour	Bact_12_15
Bact.12.16	TCACGTACTA	TACGGGAGGCAGCAG	Bacteria	12_hour	Bact_12_16

#SampleID	BarcodeSequence	LinkerPrimerSequence	Type	Time	Description
Arch.0.3	AGACGCACTC	CCCTAYGGGGYGCASCAG	Archaea	0_hours	Arch_0_3
Arch.4.2	ACGCTCGACA	CCCTAYGGGGYGCASCAG	Archaea	4_hours	Arch_4_2
Arch.8.4	AGCACTGTAG	CCCTAYGGGGYGCASCAG	Archaea	8_hours	Arch_8_4
Arch.12.5	ATCAGACACG	CCCTAYGGGGYGCASCAG	Archaea	12_hour	Arch_12_5

7.3 Appendix 3.

Letter from external sequencing provider describing delays experienced in obtaining sequencing data.

agrf
Australian Genome Research Facility Ltd

REGISTERED OFFICE
Gehrmann Laboratories, Research Road
The University of Queensland BRISBANE QLD 4072 Australia
Ph: 617 3365 4448 Fax: 617 3365 1823
Email: agrf@agrf.org.au
www.agrf.org.au

ABN 63 097 086 292

20th February 2012,

Dear Ros,

Due to issues encountered in late 2011 with AGRF's GS FLX Next Generation Sequencing Platform resulting in a failed sequencing run, there have been delays to the generation of quality data for the samples you submitted on the 28th of November 2011, and our advised turnaround time of 8 weeks has not been able to be achieved.

Further recent issues encountered with the servers at our Brisbane Node, have caused further delays to the supply of your data. The sequencing has been completed (GS FLX run completed 15/02/2012), and our Bioinformatics team is now in the process of retrieving the data and performing the required QC analysis, prior to the supply of your data.

On behalf of AGRF, I would like to sincerely apologise for these delays.

It is anticipated that your data will be available for you later this week, and is being processed as a matter of priority.

Should you have any additional queries or concerns in relation to this matter, please do not hesitate to contact me.

Yours sincerely

Rachel Kliese

Account Manager
Australian Genome Research Facility Ltd
Level 5 Gehrmann Laboratories
Research Road
University of Queensland
St Lucia QLD
Phone: +61 7 3346 3244
Fax: +61 7 3365 1823
Mobile: 0400 175 432

7.4 Appendix 4.

Publications arising from B.CCH.1025

Conference presentations

1. Gilbert, R.A., Ogg, C., Vo, J., Gulino, L.M., Ouwerkerk, D. and Klieve, A.V. (2012) Phages of rumen bacteria, whole genome sequencing and annotation. Abstract submitted for presentation, ASM 2012, The Australian Society for Microbiology 2012 Annual Scientific Meeting 1-4 July Brisbane, Australia. (Submission No. 59600153)

2. Gilbert, R.A., Ogg, C., Vo, J., Gulino, L.M., Ouwerkerk, D. and Klieve, A.V. (2012) A genomic strategy to identify archaeal viruses in the rumen. Abstract submitted for presentation, 8th INRA-RRI Symposium on Gut Microbiology. Gut Microbiota: Friend or Foe? June 17-20, Clermont-Ferrand, France.