Metagenomic Characterisation of the Gastrointestinal Virome of Neonatal Pigs

Oskar Karlsson Lindsjö
Faculty of Veterinary Medicine and Animal Science
Department of Biomedical Sciences and Veterinary Public Health
Uppsala

Doctoral Thesis
Swedish University of Agricultural Sciences
Uppsala 2016

Acta Universitatis agriculturae Sueciae
2016:84

Cover: Host-virome interaction
(photo: A. Lindsjö)

ISSN 1652-6880
ISBN (print version) 978- 91-576-8670-1
ISBN (electronic version) 978- 91-576-8671-8

Print: SLU Service/Repro, Uppsala 2016

Metagenomic Characterisation of the Gastrointestinal Virome of Neonatal Pigs

Abstract

Microorganisms that colonise the gastrointestinal tract are responsible for a large portion of the genetic diversity of the body. These microorganisms are of bacterial, archaeal and viral origin. The living space of these microorganisms, the microbiome, holds numerous interactions both between each other and the host. The viral part of the microbiome, the virome, consists of a multitude of virus species. These viruses infect and modulate cells from all three domains of life. Even though viruses have been acknowledged for their abilities to induce disease in its host, knowledge about the total diversity of viruses within the virome, and the role it plays in health and disease, is so far scarce. It is thought that the virome co-evolved with the host and that its establishment in mammals occurs early in life.

The virome can be studied by the use of viral metagenomics, the study of all viral genetic material within a sample. Viral metagenomics was used in this thesis to generate datasets for comparative metagenomics. These datasets were then used for disease investigation and to compare similarities in the viromes of two mammalian species, pigs and humans.

This thesis establishes a methodological framework for studying the virome in mammals, by use of viral metagenomics. A methodology for amplifying the metagenome prior to sequencing was assessed and a software for bioinformatics analysis of viral metagenomes was developed. With the methodologies developed herein, the eukaryotic virome of neonatal piglets suffering from diarrhoea was investigated. Several known enteric viruses were detected using viral metagenomics on healthy and diarrhoeic neonatal piglets. However, no virus was present exclusively within sick or healthy piglets and no virologial cause could be established for the neonatal diarrhoea. Comparative viral metagenomics was also used to establish if similarities existed between neonates of porcine and human origin, as well as between adults and neonates. Similarities were detected between adults of both species, who seems to be sharing a considerable part of their virome. There was also a notable difference between neonatal viromes and adult viromes, further supporting established theories about diversification over time of the virome.

Keywords: Metagenomics, Pig, Bioinformatics, Neonatal, Virome, Microbiome, Neonatal diarrhoea, Comparative metagenomics

Author's address: Oskar Karlsson Lindsjö, SLU, Department of Biomedical Sciences and Veterinary Public Health, P.O. Box 7028, 750 07 Uppsala, Sweden
E-mail: oskar.e.karlsson@.slu.se

Dedication

To friends and family, for keeping me sane…

Sapere Aude

Quintus Horatius Flaccus

Contents

List of Publications

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

I Karlsson, O. E., Belák, S., & Granberg, F. (2013). The effect of preprocessing by sequence-independent, single-primer amplification (SISPA) on metagenomic detection of viruses. *Biosecurity and bioterrorism: biodefense strategy, practice, and science*, *11*(S1), S227-S234.

II Norling M[*], Karlsson-Lindsjö OE[*], Gourlé H, Bongcam-Rudloff E, Hayer J (2016) MetLab: An *In Silico* experimental design, simulation and analysis tool for viral metagenomics studies. *PLoS ONE 11*(8): e0160334. doi:10.1371/journal.pone.0160334

III Karlsson OE[§], Larsson J[§], Hayer J, Berg M, Jacobson M (2016) The intestinal eukaryotic virome in healthy and diarrhoeic neonatal piglets. *PLoS ONE 11*(3): e0151481. doi:10.1371/journal.pone.0151481

IV Karlsson Lindsjö OE, Larsson J, Berg M, Jacobson M, Hayer J. (2016) Comparative viral metagenomics of the gastrointestinal virome in neonate and adult humans and domesticated pigs. (manuscript).

* Contributed equally to this publication

§ Contributed equally to this publication

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

I Designed the experiment together with co-authors. Performed part of the lab work. Performed part of the analysis. Drafted the manuscript and handled correspondence with the journal.

II Designed the experiment. Performed testing of the pipeline. Drafted the manuscript together with the shared main author. Contributed to correspondence with the journal together with the corresponding author.

III Designed the experiment together with the shared main author. Performed the practical work together with the shared main author. Performed most of the analysis together with the shared main author and JH. Drafted the manuscript with the shared main author and handled correspondence with journal.

IV Designed the experiment. Performed the practical work together with the JL and JH. Performed most of the analysis together with JH and JL. Drafted the manuscript.

Abbreviations

Ad2	Human Adenovirus type 2
AIV	Avian influenza virus
ASFV	African swine fever virus
BLAST	Basic local alignment search tool
BTV	Bluetongue virus
BWA	Burrows wheeler aligner
Contigs	Contiguous sequences
DNA	Deoxyribonucleic acid
GPU	Graphics processing unit
GUI	Graphical user interface
HIMG	Human infant metagenome
HMM	Hidden Markov model
HTS	High-throughput sequencing
LCA	Lowest common ancestor
Mb/h	Megabases per hour
MDA	Multiple displacement amplification
MS2	Bacteriophage MS2
NCBI	National center for biotechnology information
NGI	National Genomics Infrastructure
OTU	Operational taxonomic units
PCR	Polymerase chain reaction
Phages	Bacteriophages
Phi 6	Pseudomonas phage phi6
PhiX 174	Enterobacteria phage phiX174
PMG	Pig metagenome
PNMG	Pig neonatal metagenome
PPV	Porcine parvovirus
RNA	Ribonucleic acid

rRNA	Ribosomal RNA
SISPA	Sequence-independent, single-primer amplification
SRA	Short read archive
WGA	Whole genome amplification
WNV	West Nile virus

1 Introduction

Viruses are obligate intracellular parasites that rely on their hosts for replication; historically they have been investigated as a cause of disease in animals and humans (Mokili *et al.*, 2012). Some classical examples are the viruses that cause rabies, foot-and-mouth disease, rinderpest, bluetongue, and more recently, Zika and Ebola. Several viruses have been discovered that fall outside the classical definitions of pathogens, so-called orphaned viruses (Li & Delwart, 2011). Together with the viral community infecting prokaryotic organisms, the pathogenic viruses and the orphaned viruses form the virome (Virgin *et al.*, 2009). The advances in the last decades within high-throughput sequencing have enabled the virome to be investigated on a larger scale and with higher resolution than before (Delwart, 2007; Edwards & Rohwer, 2005). This allowed researchers to investigate the virome of several animal tissues and fluids as well as environmental locations of interest (Belak *et al.*, 2013; Hurwitz & Sullivan, 2013; Lipkin, 2013; Mokili *et al.*, 2012; Blomstrom, 2011; Daniel, 2005). As a result, a complex community of viral agents has been discovered (Cadwell, 2015; Virgin, 2014). These communities within the host may serve many functions, both as beneficial symbionts with the host and in interaction within the larger community of microorganisms inhabiting the organism (Tremaroli & Bäckhed, 2012; Kinross *et al.*, 2011). It has now become apparent that the microbial community is much more diverse than initially estimated. This raises questions about the role of viruses within the virome and their interaction with the host in disease and health.

This thesis addresses two parts of studying the virome — sample preparation and bioinformatics data analysis. The intestinal virome of healthy and sick neonatal pigs was investigated using comparative viral metagenomics. Finally, the virome of neonates of human and porcine origin was compared with that of adult humans and pigs to investigate the possible presence of a core virome shared between species.

1.1 The Microbiome

The understanding of the microbiome and its relevance for health and disease in its host has increased immensely over the last decade (Virgin, 2014; Virgin *et al.*, 2009). The microbiome was first defined by Whipps and Cooke (1988) as follows:

> A convenient ecological framework in which to examine biocontrol systems is that of the microbiome. This may be defined as a characteristic microbial community occupying a reasonably well defined habitat which has distinct physio-chemical properties. The term thus not only refers to the microorganisms involved but also encompasses their theatre of activity.

The word microbiome is interchangeable with the term microbiota in most of the literature. However, literature often excludes viruses in the term microbiota whereas the term microbiome often includes them (Marchesi & Ravel, 2015). The microbiota deals primarily with the composition of a microbial community whereas the microbiome includes not only organism composition but also genetic makeup and environmental conditions. In this thesis, the term microbiome will be used exclusively, and will include the composition of the community of microorganisms, including viruses, as well as interactions, host factors and temporal change. The microorganisms that form the microbiome encompass bacteria, archaea, fungi, parasites and viruses within a set environment e.g., the intestine, the skin, soil or a specific part of the ocean (Marchesi & Ravel, 2015). Of these microorganisms, bacteria and fungi are the most studied. This is due in part to the relatively low technological needs for descriptive studies using 16s rRNA in the case of bacteria or the internal transcribed spacer in fungi for classification of species diversity, so-called metabarcoding (Taberlet *et al.*, 2012; Group *et al.*, 2009).

Bacteria together with their intracellular parasites, the bacteriophages, are the two most abundant microorganisms on Earth (Edwards & Rohwer, 2005; Rohwer, 2003; Torsvik *et al.*, 2002). As such, bacteriophages and archaeal viruses are a major controlling factor of the prokaryotic communities (Rodriguez-Valera *et al.*, 2009; Prangishvili *et al.*, 2006; Wilhelm & Suttle, 1999). Bacteriophages have two life stages: the lytic where the bacteriophage lyse their host, and the integrative where it enters its host genome, forming lysogens. Bacteriophages are also known for horizontal gene transfer between strains of bacteria and even closely related bacterial species (Weinbauer, 2004). Phages can also interact directly with the host, thereby changing the environment in which the bacteria reside (Hodyra-Stefaniak *et al.*, 2015; Duerkop & Hooper, 2013).

1.1.1 The Virome

The virome is unique within the microbiome due to the systemic nature of its inhabitants, which often influence larger biological systems (Karst, 2016; Virgin, 2014; Minot *et al.*, 2012; Minot *et al.*, 2011). It is estimated that an individual healthy human harbours >10 chronic viral infections and sometimes even more (Virgin *et al.*, 2009). Due to their nature as parasitic entities, the effect of the viruses on the microbiome is often profound. Viruses modulate and control large part of the prokaryotic microbiome and can thereby affect both microbiome composition as well as host. These interactions affecting organisms from different domains of life, so called trans-domain interactions, are common between phage/bacteria but also between virus/host/bacteria (Virgin, 2014). Studies of the virome is a key component to study for understanding the microbiome's role in mammals.

The virome can be said to be divided into four different parts: i) the eukaryotic virome, i.e., viruses replicating within organisms belonging to the domain Eukarya, ii) the bacterial virome, i.e., viruses replicating within hosts of domain Bacteria, iii) the archaeal virome, viruses replicating within hosts of domain Archaea, and iv) the endogenous viral elements (Cadwell, 2015; Virgin, 2014; Rohwer & Thurber, 2009), see figure 1. These four components are present in all animal microbiomes investigated so far and they represent an important part of the whole microbiome.

An estimate of the diversity of the global virome is in the range of 10^{31} members, with a majority residing in the second group, bacteriophages, based on an estimation of a 10-fold diversity compared to the prokaryotic diversity (Suttle, 2005). This estimation is however contested, as ongoing metagenomics sampling of the oceans does not support it and puts the estimated considerably lower, nearly three orders of magnitude less than previous estimates (Hurwitz & Sullivan, 2013; Ignacio-Espinoza *et al.*, 2013). Nevertheless, the viromes are the most diverse biomes of the earth with an astonishing number of species and genetic diversity. The size of the mammalian virome is not known, but it is estimated that there are around 330,000 species of eukaryotic viruses. However, a large portion of the virome consists of viruses infecting prokaryotic hosts and, as such, the grand total would be much higher (Anthony *et al.*, 2013; Reyes *et al.*, 2012).

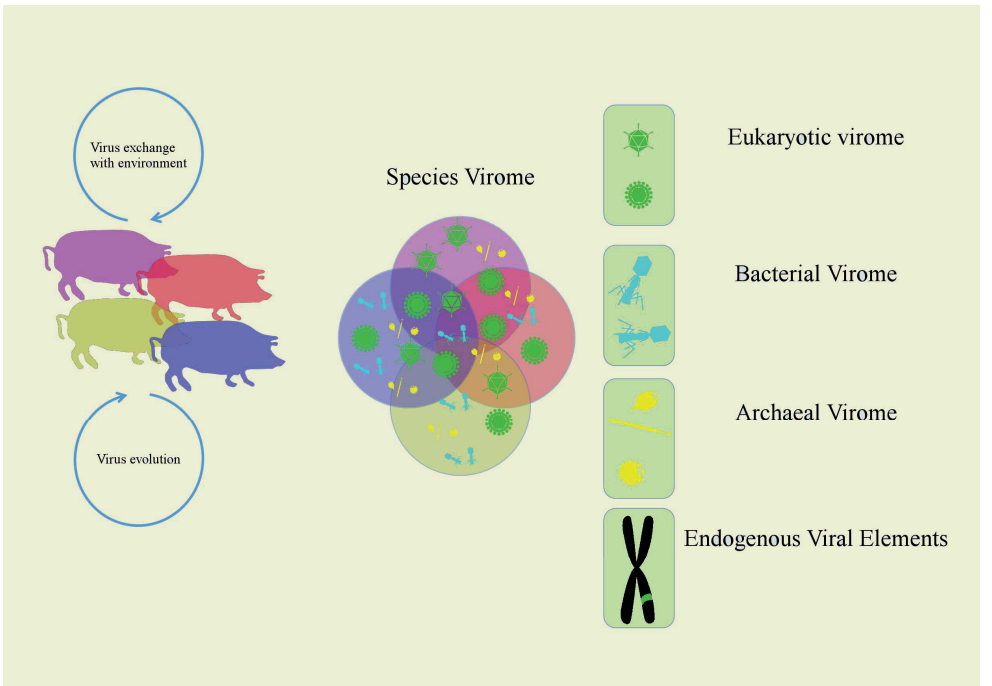

Figure 1. The virome is constantly changing within an individual due to outside influence and viral evolution. Within a host species certain characteristics of the virome is retained whereas some are unique to the individual. The virome in itself is composed of four groups of viruses: those infecting eukaryotes, those infecting bacteria, those infecting archaea and those residing within the host genome.

Of the eukaryotic viromes, the mammalian viromes have gained most attention as several studies have tried to grasp the relationship between the virome/microbiome/host in relation to potential diseases and virome/microbiome interactions (Mokili *et al.*, 2012). To fully encompass the virome and its role in shaping host response, as well as its interactions within the microbiome, it is necessary to consider its systemic effects (Handley, 2016), see figure 2. By allowing each species in a microbiome to be represented by a node within a network, Handley (2016) provides a systemic perspective of the virome. Viral species are thus represented by nodes with several edges, as their overall effect on the host and microbiome is often greater than that of other nodes (such as bacteria and host factors).

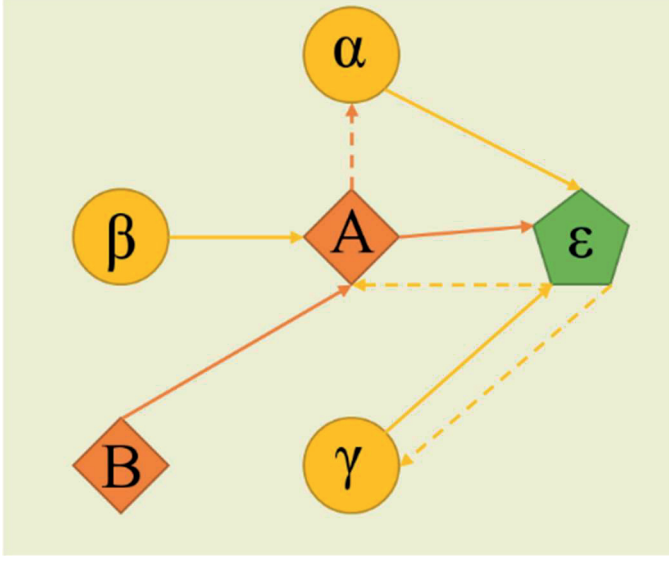

Figure 2. A simplified network of virus/bacteria/eukaryote interaction within the microbiome. Solid lines are direct modulation, such as infection. Dashed lines are indirect modulation, such as superinfection exclusion. A simple network of three viruses, yellow, and two bacteria, orange, can modulate itself and the host, green in several different ways. The six nodes have a total of eight possible edges in this simple network.

1.2 Methodologies for Studying the Microbiome

Before culture-independent methodologies were available, the microbiome was studied using traditional methodologies e.g., bacterial culturing, virus isolation and electron microscopy. Although providing several major breakthroughs, the methodologies were labour intensive and progress was limited (Cho & Blaser, 2012). Furthermore, most bacteria and viruses are not possible to cultivate.

With the introduction of high-throughput sequencing (HTS), first as parallel sequencing using Sanger sequencing and then with second or next-generation sequencers, the primary methodology shifted to nucleic-acid based detection through sequencing (Hugenholtz & Tyson, 2008). This has enabled a breakthrough in microbiome research allowing virome characterization to go from years down to just months or weeks of labour.

1.2.1 Virome Enrichment

The genomes constituting the virome are magnitudes smaller than the other genomes within the microbiome and commonly in the order of billions times smaller than the host genome. As such, unless the virome part is enriched, sequencing data from metagenomes are predominantly from the host and prokaryotic microbiome (Thurber *et al.*, 2009).

Virome enrichment is divided into two major parts: virome isolation and nucleic-acid based amplification. Virome isolation aims at removing host genetic material as well as the prokaryotic microbiome by centrifugation, filtration, rRNA depletion and nuclease treatment. Isolation can also be performed by bead-based capture of viral particles or genomes. Depending on the specific goals within the study, two or more techniques can be used in combination. Several in-depth studies have investigated the bias and the feasibility of single methods as well as combinations of single methods, see Hall *et al.* (2014) and Rosseel *et al.* (2015).

Nucleic-acid based amplification of the virome is generally based upon two methodological approaches: polymerase chain reaction (PCR) or multiple displacement amplification (MDA) (Rosseel *et al.*, 2015; Hall *et al.*, 2014). Both methodologies use random priming of the template, the virome, but the amplification strategy differs. The PCR approach aims at randomly priming the genomic elements of the virome and then amplifying the metagenome by PCR. This includes methodologies such as sequence-independent single-primer amplification (SISPA) (Allander *et al.*, 2001), see figure 3a. The second overreaching methodology MDA (figure 3b and 3c), uses a displacement DNA polymerase, e.g., Phi29 polymerase, to amplify DNA (Dean *et al.*, 2001). The reaction is primed by random hexamers on the template and the following amplification follows a pattern of branching and priming, as seen in figure 3b and 3c. Even though both methodologies have some inherent biases, both have been used with great success for amplification of whole genomes and metagenomes (Marine *et al.*, 2014; Rosseel *et al.*, 2013; Binga *et al.*, 2008).

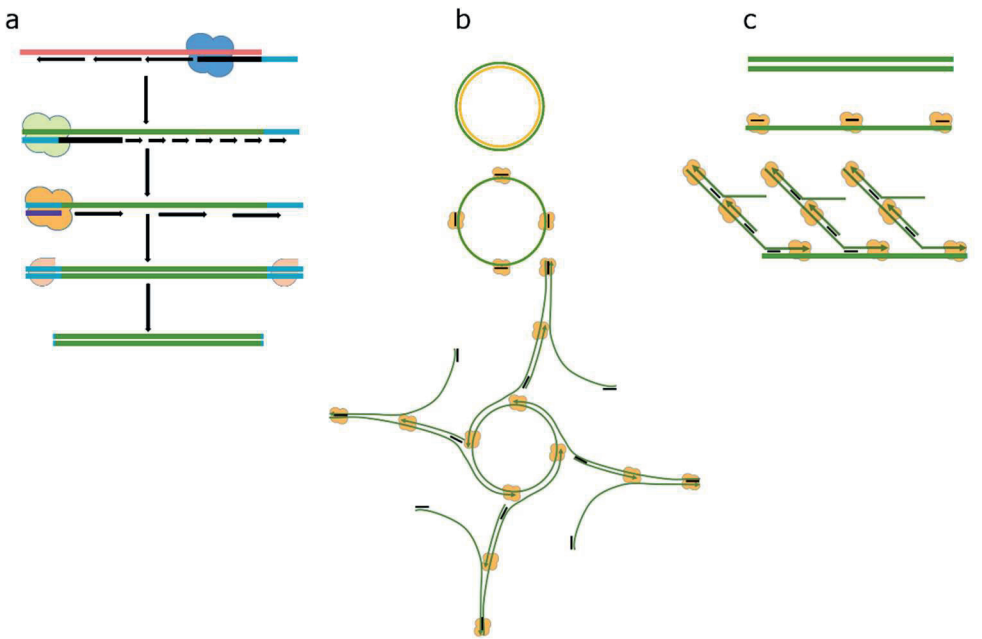

Figure 3. SISPA and MDA methodologies. a) SISPA methodology, SISPA-tag sequence, black and blue, is added to the template by reverse transcription, blue. The cDNA is converted dsDNA by Klenow polymerase, green. The template is then amplified using the PCR amplification primer, purple, and a polymerase, orange. After amplification the SISPA-tag is cleaved of by restriction endonuclease, pink. b) MDA methodology in which circular or c) linear DNA is randomly primed (black) and displacement DNA polymerase (yellow) then amplifies the DNA, creating branched DNA amplicons.

1.2.2 Sequencing

Over the last decade, the use of HTS for characterization of the virome has gone from a technology confined to sequencing centres to a technology widely used in both medium-sized and small labs (Belak *et al.*, 2013; Delwart, 2007). With the introduction of the second generation of sequencers, such as the 454 Life Sciences and the Solexa platform (later acquired by Illumina and will be referred to as "Illumina" throughout the thesis), large-scale virome characterisation moved from being an extremely labour extensive and costly procedure to a feasible methodology for pathogen detection and ecological studies (Bentley *et al.*, 2008; Lipkin, 2008; Wheeler *et al.*, 2008).

In 2011 the development accelerated with the introduction of affordable benchtop instruments for sequencing (Quail *et al.*, 2012; Rothberg *et al.*, 2011). The current rate of publication and citation within the field of viral metagenomics is extraordinary, with more than 350 published articles and over 8000 citations in 2015 (the most recent year for which full-year statistics are available).

Current HTS technologies are divided into two primary methodological fields: i) short read technologies (Illumina, IonTorrent, 454), and ii) single-molecule sequencing technologies (Oxford Nanopore and PacBio) (Goodwin *et al.*, 2016). Whereas the short read technologies work by sequencing several short fragments (75-400b) of the fractionated sample, single-molecule sequences can handle considerably longer sequences. Table 1 summarises the currently available technologies and their specifications, e.g., read length, Mb/h, data output.

Table 1. *Available sequencing technologies, their throughput, read length, Megabases per hour (Mb/h), and sequencing methodology. All data is collected from manufacturers homepage and represent specifications in optimal situations.*

Sequencing platform	Throughput	Read length	Mb/h	Methodology
Illumina MiniSeq	7.5 Gb	2*150	312	2-channel Sequencing by Synthesis
Illumina MiSeq	15 Gb	2*300	268	4-channel Sequencing by Synthesis
Illumina NextSeq	120 Gb	2*150	4,138	2-channel Sequencing by Synthesis
Illumina HiSeq 2500	1000 Gb	2*125	6,944	4-channel Sequencing by Synthesis
Illumina HiSeq 4000	1500 Gb	2*150	16,666	4-channel Sequencing by Synthesis
Ion Torrent	2 Gb	400	274	Semiconductor sequencing
Ion Proton	10 Gb	200	2,500	Semiconductor sequencing
IonTorrent S5	15 Gb	200	600	Semiconductor sequencing
PacBio	1 Gb	20,000	250	Single Molecule, Real Time Sequencing
PacBio Sequel	1.25 Gb	20,000	208	Single Molecule, Real Time Sequencing
Oxford Nanopore MinION	42 Gb	10,000	875	Nanopore DNA sequencing
Oxford Nanopore PromethION	256 Gb	10,000	5,333	Nanopore DNA sequencing

Three main sequencing platforms have produced most of the datasets for metagenomics: Illumina, IonTorrent and 454 Roche. The 454 was the first second-generation sequencer on the market. It was used up extensively until 2012 when it was overtaken by the launch of the benchtop sequencers in the IonTorrent and Illumina line-up. A considerable amount of data has been and still is produced by the 454 but the manufacturer no longer supports the platform.

1.2.3 Bioinformatics

Bioinformatics is the scientific field focused on developing methodologies for the efficient use and management of experimental data within biology (Hogeweg, 2011). Due to the extensive amount of data produced through HTS, bioinformatics is necessary for analysis and visualisation of metagenomes (Kunin *et al.*, 2008).

Bioinformatics for metagenomics studies requires several different tools for three main tasks: sequence quality control, sequence assembly, and sequence

classification (Kunin *et al.*, 2008). Sequence quality control tools was developed during the human genome project as a way to estimate and compare different sequences for increased validity of the assemblies. By providing a statistical framework for sequence quality based on the base calling from raw reads, a quality score Q is derived for each base in a sequence. Q is the probability that a base is miscalled at that specific location, such that Q20 predicts 1% error and Q30 predicts 0.001% error (Ewing & Green, 1998). An increased understanding of metagenome composition and the introduction of new sequencing platforms have pushed development for new quality control algorithms. Today there are several custom-made algorithms and software tools for determining a plethora of metrics, e.g., alien sequence contamination, low-complexity sample, adapter removal, mean Q-score filtering, length filtering etc. (Shrestha *et al.*, 2014; Patel & Jain, 2012; Schmieder & Edwards, 2011).

Sequence assembly is the process of forming long, contiguous sequences (contigs) from sequence reads (Pop, 2009). This process was initially based on simple algorithms, matching overlaps of sequence reads towards each other thereby forming a longer sequence. This was later iterated, performing the action on larger datasets thereby building larger contigs. As the number of sequences is increasing, so is the complexity of the operation and the execution time. With the HTS introduction, regular methodologies for sequence assembly became a bottleneck in the analysis (Pop, 2009). Development of new, algorithmically more optimised assemblers birthed the de bruijn assemblers, lowering execution time and enabling assembly of datasets several magnitudes more complex than before (Li *et al.*, 2012; Compeau *et al.*, 2011). Assembly of metagenomic data is however still hampered by the presence of several genomes within the dataset, which leads to computationally advanced problems. Given the large influx of metagenomic datasets during the last five years, development of specific tools for handling metagenomic assembly are on the horizon (Bankevich *et al.*, 2012; Boisvert *et al.*, 2012; Namiki *et al.*, 2012).

The final step of metagenomic bioinformatics analysis is the classification of sequence reads or contigs in their correct taxonomic group. This process was initially performed by sequence homology software, such as the basic local alignment search tool (BLAST) (Altschul *et al.*, 1990). This was refined with adding a lowest common ancestor (LCA) algorithm on top of the BLAST results, ordering the reads in taxonomical hierarchy (Huson *et al.*, 2007). As increased throughput of sequencing platforms eventually pushed the execution times from days to weeks' sequence read classification became the new bottleneck of metagenomics. Several tools have been developed to battle this,

ranging from execution time improvements on BLAST, e.g., GPU-BLAST, and DIAMOND to more optimised solutions such as k-mer based tools, e.g., Kraken and Kaiju (Menzel *et al.*, 2016; Buchfink *et al.*, 2015; Wood & Salzberg, 2014; Suzuki *et al.*, 2012; Vouzis & Sahinidis, 2011; Manavski & Valle, 2008). Methodologies using other theories for classification have been proposed; however, none of these have provided enough advantage to be widely used.

Efficient bioinformatic handling of metagenomic data is of outmost importance for microbiome studies. The increasing number of algorithms and software for handling different problems within the field has created a multitude of tools available to researchers. However, little benchmarking and compilation of tools into methodologies is performed and most of the development in the field is not available to general researchers without bioinformatics knowledge.

1.3 The Animal Virome – a Factor for Health and Disease

The composition of the virome in animals seems linked to both health and disease and can influence the host throughout a multitude of interactions (Stelekati & Wherry, 2012; Foxman & Iwasaki, 2011; Virgin *et al.*, 2009). Small changes to the virome, such as the introduction of a single pathogenic agent within a virome of several hundred species can quickly disrupt a multicellular organism, leading to degradation of biological function. This type of infection normally follows Koch's postulates, i.e., an infectious agent isolated from a diseased animal replicates the disease state, when introduced into a healthy animal. With the increasing knowledge of the role of the virome in infectious disease, this model is insufficient to explain certain disease states, like for example when several microorganisms together disturb the biological functions of the host (Mokili *et al.*, 2012). The paradigm shift from single pathogen detection to multi-microbe detection is changing the way we view infectious disease. With the increased knowledge from metagenomics studies of the virome, Mokili *et al* (2012) suggests the following:

> Alternatively, we propose the metagenomic Koch's postulates, which focus on the identification of metagenomic traits in disease subjects. The metagenomic traits are molecular markers such as sequence reads, assembled contigs, genes or full-genomes that can uniquely distinguish diseased metagenomes from those obtained from matched healthy control subjects.

In other words, they amend Koch's postulates to encompass not only single pathogens but also co-infections and virome composition of the host. This could in theory also be expanded to include host factors such as immunological maturation and genetic susceptibility, enabling a more holistic view of viral infection and pathogenesis.

Although the microbiome´s impact on disease and health has been established for nearly a decade (Gill *et al.*, 2006), investigations into viromes of healthy animals have been limited. The National Health Institute in the USA estimates that only 5% of the microbiome funding between 2012-2014 has been used to explore the virome (Stulberg *et al.*, 2016), even though estimates put the diversity of the virome equal to or higher than that of the prokaryotic components of the microbiome (Reyes *et al.*, 2012). Most studies on the virome are performed as part of pathogen discovery efforts, leaving a gap in the knowledge about the virome in healthy animals. Given the proposed plasticity of the prokaryotic microbiome in response to dietary changes as well as inflammatory states (Minot *et al.*, 2012; Reyes *et al.*, 2010), it is important to establish a baseline virome composition and its temporal changes.

The microbiome is mostly described in terms of commensal symbiosis. The separate species of the microbiome form a whole ecological community, which provides functionality to the host, thereby forming a symbiotic relationship between the two (Tremaroli & Bäckhed, 2012; Hooper *et al.*, 2001). Given the four defined parts of the virome, viruses can affect the host through trans-domain interactions in different ways:

- Eukaryotic viruses continuously infect higher organisms (Virgin *et al.*, 2009). This provides the organisms' immune response with a constant flux of activation thereby priming the immune system for coming infections (Foxman & Iwasaki, 2011). This constant interaction between the virome and its host is also trans-domain, as the constant flux of the immune system can provide a defence against microorganism invasion and colonisation, thereby not only affecting the host but also microorganisms populating the host (Handley, 2016; Norman *et al.*, 2015; Barton *et al.*, 2007).
- The bacterial virome infects and controls bacterial populations within the host microbiome. It also directly interacts with the host during its lytic cycle, during which phage particles directly interact with the host's immune receptors (De Vlaminck *et al.*, 2013; Duerkop & Hooper, 2013; Cuesta *et al.*, 2006).
- Less is known about the archaeal virome, but initial studies model it in the same way as the bacterial virome (Krupovic *et al.*, 2011; Rohwer & Thurber, 2009; Prangishvili *et al.*, 2006).

- Viral infection can also infer partial immunity by interaction with other viruses through superinfection exclusion, a state where a viral infection prevents further infection from other viral agents with close homology, e.g., an infection by Hepacivirus excludes infection by the other Hepacivurses to some degree (Tscherne *et al.*, 2007).
- The endogenous viral elements often confer beneficial functions to the host, either through direct function e.g., a gene is incorporated providing function for the host, or through indirect functionality, such as immunization (Feschotte & Gilbert, 2012).

1.4 Establishment of the Gastrointestinal Virome – Viral Interaction in Neonatal Hosts

The microbiome is established early, possibly already during gestation, and it matures into a stable state by the time the animal reaches young adulthood (Lim *et al.*, 2015; Aagaard *et al.*, 2014; Bergstrom *et al.*, 2014; Matamoros *et al.*, 2013; Penders *et al.*, 2006). The establishment of the microbiome in mammals is an important event in maturation. During the first months of life, the microbiome diversifies and as diet changes upon weaning, the microbiome enters a secondary expansive phase before entering its mature stage (Lim *et al.*, 2015; Schloss *et al.*, 2012; Kurokawa *et al.*, 2007). The microbiome composition of neonates is directly connected to that of the mothers and changes in the microbiome composition are affected by artificial birthing methods such as caesareans section (Mueller *et al.*, 2015; Goedert *et al.*, 2014).

Studies indicate that stressors to the microbiome during these stages of maturation can cause imbalances. One such stressor is antibiotic treatment, which kills bacteria and also induces pro-phage activation and thereby disturbs the prokaryotic microbiome with long-term effects after ended treatment (Allen *et al.*, 2011; Sommer & Dantas, 2011). Through trans-domain interactions, stressors on the microbiome can thereby affect the host long term, with possible inflammatory states and metabolic deficiencies (Norman *et al.*, 2015; Schwartz *et al.*, 2012; Tremaroli & Bäckhed, 2012; Kinross *et al.*, 2011).

Due to the extensive establishment of bacteria in the neonate microbiome, the virome is mostly dominated by prokaryotic viruses. However, recent advances also indicate that the eukaryotic virome is established during the first hours of life and becomes diversified much like the microbiome as a whole (Lim *et al.*, 2015). This eukaryotic virome includes both viruses with unknown host interactions as well as classical pathogenic viruses. It must be noted however that even though these viruses might be traditionally considered pathogens, the hosts remain asymptomatic.

1.5 The Core Virome

Most ecological systems comprise a core of species that are often resilient towards stressors and stable in diversity even though fluctuations might occur at the species composition level (Costello *et al.*, 2012). Earlier studies have indicated a core virome within different microbiomes in humans and animals (Cadwell, 2015; Stern *et al.*, 2012). However, knowledge is so far limited about the core virome and its functionality and interaction with the host and the microbiome, but its disruption seems to be damaging for the host (Karst, 2016). For example, virome disruption renders the host slower in recovery from inflammatory states (Yang *et al.*, 2016). It has also been shown that the presence of viral species can, to some extent, replace bacterial functions in the host (Kernbauer *et al.*, 2014). Kernbauer *et al* (2014) provided proof that enteric viruses, at least partially, provide the host with the same function as a commensal bacterium. The establishment of the prokaryotic microbiome is closely linked to that of the virome. These recent discoveries about the virome indicate that mammalian hosts retain a core virome. Additionally there are individual differences in virome composition, e.g. rare species who diversify the individual viromes but are not part of the core (Stern *et al.*, 2012).

1.6 The Implications of the Virome in Veterinary Medicine

Initial studies of the virome within mammals indicate a diverse population (Cadwell, 2015; Virgin, 2014; Reyes *et al.*, 2012). The gastrointestinal tract in particular is populated by a complex virome, retaining stability over time and developing throughout the host's life (Lim *et al.*, 2015; Ogilvie & Jones, 2015; Stern *et al.*, 2012). Determinants in host metabolic functions and stressors such as infection, inflammatory states and antibiotic treatment, influence the virome. These can lead to composition changes and sometime disturb the core functionality (Karst, 2016; Allen *et al.*, 2011; Sommer & Dantas, 2011). The virome interacts with the microbiome as well as with the host (Ogilvie & Jones, 2015). The role of the virome within the microbiome is not yet fully understood, even though initial studies indicate a profound effect on both the prokaryotic microbiome as well as on its eukaryotic host (Karst, 2016; Allen *et al.*, 2011).

Developing perspectives on the virome enable decades of old knowledge to be used beyond its initial intent, such as multi-factorial applications of Koch´s postulates and a re-emergence of phage-therapy (Mokili *et al.*, 2012; Lu & Koeris, 2011). Possible manipulation of the microbiome and its virome component relies on increased knowledge of the development, composition and functionality of the virome (Reyes *et al.*, 2012). With the increased

availability of HTS methodologies for exploration of the virome, the possibility to continuously monitor changes in the microbiome of relevant domesticated species is on the horizon (Lipkin, 2013; Delwart, 2007).

Knowledge about the microbiome has benefited immensely from HTS methodologies (Didelot *et al.*, 2012; Hall, 2007). Virome studies are still in their infancy, but are providing important perspectives on both infectious diseases and animal welfare. Given the rapid development of the technology, HTS is soon available even in resource-strained environments (Pennisi, 2016). Just as development of sequencing techniques has pushed the boundaries of knowledge within microbiomes and viromes, the coming years will result in more advanced experiments, clinical applications and possibly even precise manipulation of the virome.

2 Aims of the Thesis

The overall aim was to assess and develop molecular and bioinformatics methods that would allow the virome of neonatal animals to be characterized and investigated.

Specific objectives included:

- Assessment of bias introduced from sequence-independent, single-primer amplification (Paper I)
- Evaluation of bioinformatic methods for classification of viral reads within metagenomic datasets (Paper II)
- Development of a set of bioinformatic tools for use within viral metagenomics (Paper II)
- Characterisation of the eukaryotic virome within healthy and sick neonatal pigs, using viral metagenomics (Paper III)
- Comparison of viromes of healthy neonates of *Sus scroufa domesticus* and *Homo sapiens* to study similarities between the viromes (Manuscript IV)
- Investigation of stable parts of the virome in neonate and adult *Sus scroufa domesticus* and *Homo sapiens* (Manuscript IV)

3 Comments on Methodologies and Materials

This section describes in general terms the methods and the materials used for the studies in this thesis. The methods are described in greater detail within each paper (I-III) and manuscript IV. The studies can be divided into two main parts, methodological assessment (I, II) and application (III, IV). This section will cover the animal samples used for studies III and IV, the synthetic metagenome used in papers I and III, the data collections used within papers II and manuscript IV and the methodologies assessed within paper I and II.

When designing experiments, the choice of methodology must always reflect the hypothesis and aim of the study as well as the expected outcome. In papers III and manuscript IV, the goal was to determine whether neonatal animals have an initial establishment of a virome at birth and if a known viral agent were involved in emerging neonatal porcine diarrhoea in Sweden. This was tested with viral metagenomics on sick and healthy porcine neonates and with comparative viral metagenomics on datasets representing healthy neonates and adults from two species. To enable this, a molecular methodology for viral enrichment was evaluated in paper I and a framework for bioinformatics analysis was developed in paper II.

3.1 Construction of a Synthetic Metagenome (Paper I, Paper III)

In the experimental design of metagenomic studies, the use of positive controls is not as common as in other fields of molecular diagnosis and detection. This leads to a lack of comparability between methodologies and questions about the validity of the results (Knight *et al.*, 2012). Synthetic or model metagenomes — metagenomes produced from previously cultured viruses — can be used within virome studies as positive controls, as spikes for low diversity samples, and for development of methodological improvements

(Howe *et al.*, 2014; Marine *et al.*, 2011). Within this thesis, the term synthetic metagenome will be used exclusively.

3.1.1 Synthetic Metagenome Used for Investigating SISPA Bias (Paper I)

In paper I a synthetic metagenome was developed for testing an amplification strategy and as a positive control within future studies. Four previously genetically characterized viruses were included: human adenovirus type 2 (Ad2), african swine fever virus-E75 (ASFV), avian paramyxovirus type 1 (APV-1) and avian influenza virus H7N7 (AIV).

Ideally a synthetic metagenome should be constructed by spiking a sample with virons, thereby acting as a process control for both virome isolation and amplification. However, this can be technically challenging with pathogens such as ASFV, because it requires access to biosecurity facilities and trained personnel. Therefore, the synthetic metagenome used in Paper I was based on extracted nucleic acid of the four selected viruses. The viruses were grown and isolated independently of each other. The nucleic acids were then purified and combined after quantification (Nanodrop ND-1000) to represent a synthetic virome with one virus, Ad2, in excess and the other three at lower concentrations, as described in paper I. The synthetic metagenome was used in paper I to evaluate bias introduced by SISPA.

3.1.2 Improved Synthetic Metagenome (Paper III)

For paper III, the synthetic metagenome was redesigned and expanded. Ad2 and APV-1 were removed and but several other viruses were included in addition to ASFV and AIV: *Classical swine fever virus* (CSFV), *West Nile virus* (WNV), *Blue tongue virus* (BTV), *Porcine parvovirus* (PPV), *Enterobacteria phage phiX174* (PhiX 174), *Bacteriophage MS2* (MS2), *Pseudomonas phage phi6 (*Phi 6) and *Phage 28b*. This expanded synthetic metagenome covers the most common forms of genomic structures for eukaryotic viruses as well as bacteriophages and therefore provided a better point of reference. ASFV, CSFV, AIV, WNV, BTV and PPV were all kindly provided by the Department of Microbiology at the National Veterinary Institute (SVA), Uppsala, Sweden. Phi 6, MS2 and Phage 28b were provided by Eva Emmoth, also at the Department of Microbiology at SVA. Phi X174 virion DNA was purchased from New England Biolabs (UK), Hertfordshire, Hitchin, United Kingdom.

3.2 Whole Genome Amplification Strategies (Paper I, Paper III)

The study in paper I focused on SISPA and possible biases introduced when using it for whole genome amplification (WGA). SISPA or variants thereof are commonly used for metagenomic experiments. Alternatively, MDA can be used for amplification. MDA also has an inherent amplification bias due to a strong preference for circular genomes, increased amplification of shorter targets, and uneven amplification of larger targets (Marine *et al.*, 2014; Kim & Bae, 2011; Yilmaz *et al.*, 2010). Due to the abundance of circular DNA viruses in intestinal environments, the use of MDA thus makes it unsuitable for studies of the intestinal virome.

In paper I a synthetic metagenome was used to test the SISPA methodology to determine what biases were introduced by SISPA in a viral metagenomic dataset. Two versions of the same metagenome, one diluted and amplified and one undiluted, were compared for changes in genome coverage, abundance of reads mapping towards the viruses and the composition between viruses within the metagenome.

3.2.1 Improved SISPA Methodology (Paper III)

During 2013, an optimised SISPA methodology was published by Rosseel *et al.* (2013). Strong evidence was provided for the factors behind SISPA bias and a vastly improved methodology was presented. The optimised methodology combined several different primers and several SISPA products to improve the evenness of coverage for RNA viruses. The authors also included optimised SISPA-tag-sequence, increased the random region of the SISPA-tag, and optimised the priming part of the SISPA-tag to better suit the need of even amplification within the whole genome. Their method, however, has not been tested for whole metagenomes even though the original SISPA methodology and variations thereof have been used for amplification of viromes at several occasions (Li *et al.*, 2015; Granberg *et al.*, 2013; Mokili *et al.*, 2012; Blomstrom *et al.*, 2010; Allander *et al.*, 2001). In paper III, the optimised SISPA methodology proposed by Rosseel *et al* (2013) was used to generate amplified metagenomes.

3.3 Samples from Pigs Suffering from new Neonatal Piglet Diarrhoea Syndrome (Papers III, Manuscript IV)

Samples for papers III and manuscript IV were collected within the frame of the thesis work "Neonatal Porcine Diarrhoea" (Larsson, 2016). As described in paper III, recruitment of the herds was performed on a voluntary basis from farmers in the central parts of Sweden. Procedures for collecting samples were approved by the Ethics Committee for Animal Experimentation, Uppsala, Sweden (permission number: C120/11). Ten herds with reported piglet diarrhoea were chosen for sample collection and seven piglets were sampled within each herd. Of these seven piglets, two were healthy controls as judged by two veterinarians at collection and by two pathologists at necropsy. For full disclosure on the sampling procedures as well as the in-depth survey of new neonatal porcine diarrhoea syndrome in Sweden, see Larsson (2016).

Intestinal tissue (distal jejunum) was chosen for the metagenomics survey of known eukaryotic viruses in paper III. Material was collected and stored at -70°C before use. The samples were thawed and homogenised (Soft tissue Omni Tip™ Plastic Homogenizer Probes) in combination with freeze-thawing on dry ice (Daly *et al.*, 2011). The homogenate was then centrifuged after which the supernatant was filtered through a 0.45µm syringe filter and treated with nuclease. The use of nuclease treatment in metagenomics studies has often been raised as a possible bias, because it might destroy parts of the virome while still not removing all non-virome RNA and DNA (Li *et al.*, 2015; Rosseel *et al.*, 2015; Hall *et al.*, 2014). However, the data is not conclusive and nuclease treatments has been shown to decrease the complexity of the sample, thereby decreasing the execution time of bioinformatics analysis by several magnitudes. Therefore, nuclease treatment was chosen for sample preparation. DNA and RNA were extracted according to Blomström *et al.* (2010). DNA and RNA were then prepared for sequencing, see sections 3.2 and 3.7.

3.4 Theoretical Estimation of Required Sequencing Depth (Papers II, III)

The methodology adapted and implemented in the software in paper II, originally proposed by Wendl *et al.* (2013), provides an estimate for the necessary sequencing depth based on the stochastic size of genomes present, the number of species present, and the abundance of the smallest genome in the metagenome (Wendl *et al.*, 2013). However, knowledge of virus composition in most metagenomes is lacking. In metagenomes for which no prior knowledge exists, it is challenging to estimate the number of species and the stochastic size of genomes present. Therefore, estimates have to be performed

based on similar, previously characterised, datasets and caution must be taken to make sure enough sequencing depth is reached. If possible, a pilot experiment should be run to provide an estimate. It should also be noted that several other tools exist for estimation of sequencing depth in metagenomics experiments (Rodriguez & Konstantinidis, 2014b; Hooper *et al.*, 2010; Stanhope, 2010). The use of two or more methods for the estimation increases the accuracy (Rodriguez & Konstantinidis, 2014a). The adaption in paper II relied on the user supplying two metrics, the genome size of the lowest abundance species, and the estimated abundance of the species within the metagenome. Accuracy is based upon the user's ability to estimate these metrics, either from previous experiments or literature. Given the user input, a list is supplied with a range of different sequencing platforms where the probability of reaching full coverage of the type species is presented in number of sequence runs needed.

3.5 Sequencing (Paper II)

A primary concern for detection of viruses using HTS is the needed sequencing depth to accurately classify the virome and avoid false negatives resulting from under-sequencing and failure to detect part of the virome. In paper II a methodology for estimating the needed sequence depth was implemented as an experimental design software. This methodology was used for designing the sequencing needs within paper III.

Several platforms are available for sequencing. Given the theoretical sequence depth needed for discovery of viruses in the gastrointestinal virome, as calculated by the application presented in paper II, both the IonTorrent and Illumina MiSeq platforms were able to provide enough sequence data. The PacBio system theoretically provides enough depth for discovery of viruses, given the calculations from paper II, but few studies have been able so far to prepare samples for metagenomics sequencing on the PacBio. The longer read length of the PacBio platform should enable simpler classification as well as improve greatly the detection of new and highly divergent viruses (Wommack *et al.*, 2008). So far a single study using the PacBio technology for metagenomics has been published. It investigated the human skin microbiome, where the resolution on detection increased with the longer read length (Tsai *et al.*, 2016).

3.5.1 Design of Sequencing Experiment for Detection of Eukaryotic Viruses by use of Viral Metagenomics (Paper III)

In paper III a study of the eukaryotic virome was performed on healthy and sick neonatal piglets. During the experimental design, rigorous work was performed to determine optimal sequencing of the samples as well as sequencing technology to use for data generation.

As the Illumina and Ion Torrent technologies were deemed suitable, pilot studies were performed on both platforms. The pilot study was performed on seven pigs, two healthy and five sick. Sequencing was performed on the IonTorrent and the MiSeq sequencing platforms. All metagenomes were sequenced to a depth of 1.5 Gb (unpublished data). The sequencing data were classified using the methodology developed in paper II. The optimal sequencing depth for the complete sequencing study within paper III was calculated using the experimental design module from paper II. The final evaluation of sequencing platform was performed with several factors in mind: cost, speed, quality and support. Based on the pilot studies and extensive method evaluation, the samples within study III were sent to the National Genomics Infrastructure (NGI), in Uppsala, Sweden for sequencing on the Ion Proton platform. In line with the theoretical calculations, the obtained data provided a sequence depth of 1.5Gb per sample and were well within our detection range for the virome, calculated to around 300 Mbp.

3.6 Taxonomic Classification (Papers I, II, III and Manuscript IV)

Taxonomic classification was performed within all studies of this thesis. The study in paper I used a basic approach applying direct genomic mapping by Burrows-Wheeler aligner (BWA) combined with a BLAST-LCA (Li & Durbin, 2009). Although suitable on smaller datasets, larger datasets with a more complex virome are too technically challenging to analyse this way.

3.6.1 Database for Viral Classification (Papers II, III and Manuscript IV)

Paper II describes a software package for experimental planning, simulation, and analysis of metagenomic experiments. The software uses a curated viral database embedded with the k-mer based classification tool Kraken (Wood & Salzberg, 2014). The database was built from the refseq archaeal and bacterial database combined with the viral and phage divisions of the NCBI nt-database (Pruitt *et al.*, 2005). The database represents all currently known and publicly available viruses and enabled Kraken to classify divergent strains of already existing viruses as well as already known viruses. To allow for detection of highly divergent viruses, this was coupled with an

implementation of Hmmer3 and FragGeneScan (FGS) software using a previously published database with hidden markov models (HMM) for protein viral families, vFam (Skewes-Cox *et al.*, 2014; Rho *et al.*, 2010).

3.6.2 K-mer Based Classification of Sequence Reads (Papers II, III)

Within the bioinformatic pipeline suggested in paper II, the k-mer based homology classifier Kraken was chosen as a tool for taxonomic classification. This was done based on both accuracy, in *in silico* simulated samples, and execution speed of the software. Kraken outperformed the other tools both in accuracy and execution speed and did so consistently when scaling up the experiments. It should be noted that the area of taxonomic sequence classification is rapidly developing and that several k-mer based tools were released during this thesis work. As such, there is no guarantee that Kraken would be the tool of choice in the coming six months, something that will be taken into consideration for the next version of the software. The methodology also suffers from its strict method for comparing query sequence towards the database; this might lead to false negatives because it misclassifies reads as unknown instead of assigning them to viral homologues.

3.6.3 Hidden Markov Model Approach for Classification of Viral Reads

The alternative approach, HMM detection of viral protein families based on Hmmer3, FGS and the vFam database, does not suffer from the same limitation. Instead, its broad classification scheme only allows classification at the protein family level. Profiles built upon sequence homology within viral proteins and applied for homology searches such as in Hmmer3/FGS/vFam are useful for detecting highly divergent viruses. Coupled with an exact method such as BLAST or Kraken, it should provide a comprehensive picture of the virome, as concluded by Skewes-Cox *et al.* (2014):

> When we compared the vFams to BLAST in real metagenomic datasets, the vFams demonstrated an improved detection accuracy when viruses in the dataset were more divergent or when the metagenomic reads acquired through massively parallel sequencing were derived from less conserved regions of the viral genome.

They further recommended:

> A straightforward implementation leveraging both search methods could entail 1) a nucleotide BLAST search to a curated set of known non-viral genome sequences (including the host genome, if available) likely to appear in the metagenomic sequence data; 2) a BLAST search to a viral database to capture and taxonomically assign higher identity matches; and 3) a search of the vFams, extending the search space into more divergent territory.

The final methodology proposed in paper II included the following implementation: Filtering of sequence reads towards the host genome, an improved Kraken database for detection of viral reads and an implementation of Hmmer3/FGS/vFam for detection of highly divergent viral reads.

3.6.4 Taxonomic Analysis Performed on the Eukaryotic Virome in Healthy and Diarrhoeic Piglets (Paper III)

The methodology from paper II was applied to the datasets generated for the study in paper III. Results in paper III were validated using Diamond, a BLAST homologue combined with the LCA approach in Megan. No notable differences were found between the Kraken/Hmmer3/FGS/vFAM classification and Diamond (Buchfink *et al.*, 2015; Huson *et al.*, 2007). Both methodologies, Kraken/Hmmer3/FGS/vFam and Diamond, also work independently of each other and independently of assembly, something that normally strains the standard methodology of BLAST/LCA. With the Kraken/Hmmer3/FGS/vFam methodology, the classification of large datasets is done within hours and smaller datasets can be handled even on a standard laptop.

3.6.5 Taxonomic Classification of the Comparative Metagenomics Dataset (Manuscript IV)

In manuscript IV, the classification tool was exchanged to Kaiju with the nr-database (based on the non-redundant protein database at NCBI) released from University of Copenhagen (Menzel *et al.*, 2016). Kaiju allows for classification at a protein homology level rather than Krakens nucleotide level. This partially remedies the problem with highly divergent viruses and provides greater accuracy in classification. Classification data for the HIMG datasets were acquired directly from the original publication, see Lim *et al.* (2015).

3.7 Comparative Metagenomics and Diversity Calculations (Manuscript IV)

In study IV, a comparative metagenomics approach was applied to investigate the possible presence of a core virome within the gastrointestinal tract of pigs and humans. Classification data from four datasets were used: adult porcine material (PMG), adult human material (HMG), neonatal porcine material (PNMG) and neonatal human material (HIMG). Classification was performed at the viral family level and diversity data were calculated using vegan in R 3.3.1 (Team, 2016).

3.7.1 Public Datasets Used for Analysis and Comparative Studies (Manuscript IV)

In study IV, three public datasets were compared with the healthy neonatal piglets investigated in paper III, abbreviated PNMG. These were: "An integrated catalogue of reference genes in the human gut microbiome" project (HMG) (PRJEB5224), "STL Infant Twins Microbiome" (HIMG) (PRJNA284162), and "A catalogue of the pig gut metagenome" (PMG) (PRJEB11755) (Lim *et al.*, 2015; Li *et al.*, 2014; Le Chatelier *et al.*, 2013). See table 2 for full disclosure of datasets and their accession numbers. All data was retrieved from the NCBI short read archive (SRA).

Table 2. *Samples and accession numbers for the data used within study IV*

Dataset	Study	Biosample	SRA	Sample Key
PNMG	PRJEB11519	SAMEA3637942	ERS945091	A1
		SAMEA3637943	ERS945092	A2
		SAMEA3637945	ERS945094	B1
		SAMEA3637946	ERS945095	B2
		SAMEA3637948	ERS945097	C1
		SAMEA3637949	ERS945098	C2
		SAMEA3637951	ERS945100	D1
		SAMEA3637952	ERS945101	D2
		SAMEA3637954	ERS945103	E1
		SAMEA3637955	ERS945104	E2
		SAMEA3637957	ERS945106	F1
		SAMEA3637958	ERS945107	F2
		SAMEA3637960	ERS945109	G1
		SAMEA3637961	ERS945110	G2
		SAMEA3637963	ERS945112	H1
		SAMEA3637964	ERS945113	H2
		SAMEA3637966	ERS945115	I1

		SAMEA3637969	ERS945118	J1
		SAMEA3637970	ERS945119	J2
HIMG	SRP058399	SAMN03659353	SRS938306	HIMG_1
		SAMN03659353	SRS938306	HIMG_1
		SAMN03659359	SRS938389	HIMG_2
		SAMN03659359	SRS938389	HIMG_2
		SAMN03659365	SRS938395	HIMG_3
		SAMN03659365	SRS938395	HIMG_3
		SAMN03659371	SRS938402	HIMG_4
		SAMN03659371	SRS938402	HIMG_4
		SAMN03659377	SRS938407	HIMG_5
		SAMN03659377	SRS938407	HIMG_5
		SAMN03659383	SRS938414	HIMG_6
		SAMN03659389	SRS938421	HIMG_7
		SAMN03659389	SRS938421	HIMG_7
		SAMN03659395	SRS938426	HIMG_8
		SAMN03659395	SRS938426	HIMG_8
PMG	PRJEB11755	SAMEA3663204	ERS970353	PMG_1
		SAMEA3663203	ERS970352	PMG_2
		SAMEA3663202	ERS970351	PMG_3
		SAMEA3663201	ERS970350	PMG_4
		SAMEA3663200	ERS970349	PMG_5
		SAMEA3663199	ERS970348	PMG_6
		SAMEA3663198	ERS970347	PMG_7
		SAMEA3663197	ERS970346	PMG_8
		SAMEA3663196	ERS970345	PMG_9
		SAMEA3663195	ERS970344	PMG_10
HMG	PRJEB4336	SAMEA2153005	ERR321573	HMG_1
		SAMEA2156032	ERR321574	HMG_2
		SAMEA2146552	ERR321575	HMG_3
		SAMEA2153852	ERR321576	HMG_4
		SAMEA1965858	ERR321577	HMG_5
		SAMEA2144396	ERR321578	HMG_6
		SAMEA2151383	ERR321579	HMG_7
		SAMEA2158769	ERR321580	HMG_8
		SAMEA2156714	ERR321581	HMG_9
		SAMEA2152591	ERR321582	HMG_10

All datasets were published and publicly available at the time of analysis. Pigs from the PMG metagenome study were chosen depending on breed

((Landrace × Yorkshire) × Duroc) and geographical location (Denmark), as those factors are known to influence the microbiome. The piglets investigated in paper III were of similar breed and collected within the same geographical region, Scandinavia. The adult humans were selected to be from the Danish sub-group within the dataset, enabling a better geographical correlation to the pigs used in the study. Unfortunately, there were no neonatal human viromes available from the region at the time of the study.

3.7.2 Estimations of Viral Diversity

Diversity was measured as gamma diversity, i.e., total species diversity, for sample groups. This was performed within sample groups and within super groups, i.e., the whole population of neonates, adults or all samples. Beta diversity/similarity between samples and between groups of samples was estimated using a qualitative measurement of presence/absence calculated with Jaccard-index. Beta-diversity was calculated on curated data, removing samples with no viral reads detected. Based on beta-diversity, dissimilarity was calculated and used to produce dendrograms and to cluster samples and viral families within heat maps. All diversity indexes were calculated using the vegan package (Dixon, 2003).

Heat maps were generated from curated data, removing viral families with less than 100 reads over the whole sample population and samples with no viral reads. Heat maps were produced using the heatmap.2 function in the gplots library in R 3.3.1 (Warnes *et al.*, 2016). Clustering was performed at sample and family level using the hclust function with unweighted pair group method with arithmetic mean. Heat maps and dendrograms for the virome composition were also calculated on a heavily curated dataset, increasing the family cutoff to 1000 viral reads per viral family, and thereby removing a large fraction of the outliers visible in the original data. This enabled a clearer picture of the eukaryotic core virome.

4 Results and Discussion

The increasingly rapid evolution of our understanding of the virome and its role in disease and health is driven by technology and the accessibility of that technology. Even though the existence of the virome has long been known and partially understood, the full implications of its interaction with the host and the hosts' microorganisms have not been acknowledged. With the development of high-throughput sequencing during the human genome project and then the introduction of parallel sequencing platforms and second-generation sequencers, biologists started addressing questions about the microbiome and the virome component. As the technology advanced, so did the questions one could answer. Today microbiome studies have evolved even within the limited timespan of this thesis, with currently over in total 400 publications studying the whole or part of the virome.

4.1 Methodological Evaluation of Sequence-Independent Single Primer Amplification (Paper I)

As described in the introduction, initial enrichment of the virome is often performed before sequence analysis to decrease background DNA from other parts of the microbiome and the host. This process normally depletes the nucleic acid material to a level where normal sequencing library methodologies are inadequate. To prepare sequencing libraries, amplification of the depleted genomic material is therefore necessary.

In paper I the SISPA methodology was investigated for amplification of a viral metagenomic sample. To enable a controlled study, a synthetic metagenome was constructed using four known viruses: ASFV, APMV-1, Ad2 and AIV. Two samples from the synthetic metagenome were sequenced, one at 1000× dilution with SISPA amplification and one unamplified. Sequencing was performed on the IonTorrent sequencing platform, at NGI in Uppsala. This

was followed by direct mapping of sequence reads towards reference genomes using BWA and metagenomics classification of sample content by BLAST/LCA (Li & Durbin, 2009). In both samples, all four viruses were detected. However, the number of reads identified as viral differed between the two samples, where the amplified sample consisted of twice as many viral reads. The reads were not, however, evenly spread over the viral genomes, where Ad2 accounted for ~99% of the classified viral reads in the amplified sample. This is in contrast to the unamplified sample where ~95% of the viral reads were classified as Ad2. In addition to this, the low abundant viruses had on average only half the coverage in the SISPA amplified samples compared to the unamplified. The change in average coverage seen between SISPA-amplified and unamplified samples is due to the PCR approach of amplification inherent in SISPA, where the initial template of Ad2 is prominent in the reaction and thus gets amplified to a much larger extent. Looking closer at the retrieval of the Ad2 genome, it was directly evident that the SISPA amplification that did occur was not evenly distributed over the genome. The regions at the start and end of the genome were much more amplified than the more central regions, see figure 4. In paper I, SISPA was deemed to bias the sample by inducing sequence specific amplification and by favouring abundant targets for amplification. Although this was a clear methodological disadvantage, paper I concluded that SISPA did not prevent positive detection of viruses within the viral metagenome despite the bias.

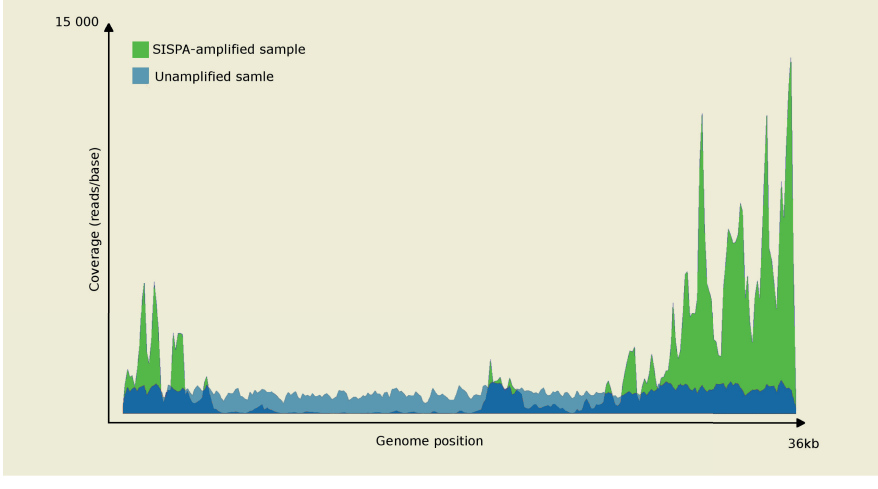

Figure 4. Genome coverage of Ad2 before and after amplification by SISPA. Modified from paper I, The effect of preprocessing by sequence-independent, single-primer amplification (SISPA) on metagenomic detection of viruses.

4.1.1 Implementation of an Improved SISPA Protocol

The coverage bias noted in study I was investigated in a parallel study by Rosseel *et al.* (2013). They found that the tag-sequence induced areas of extreme amplification. This occurs when the tag-sequence and not the random primer binds to the template, i.e., when the amplification part has template specificity. To accommodate these known biases, we used the optimised protocol suggested by Rosseel *et al.* (2013) including an adaptation of the primers by adding longer stretches of random nucleotides (dodecamers instead of hexamers) and two optimised tag-sequences in paper III. Given the overall picture, and together with the results within papers I and paper III, SISPA provides a rapid and inexpensive way to amplify metagenomes. Even though this comes with a set of biases, the improvements made by Rosseel *et al.* (2013) — in combination with using an expanded synthetic metagenome as positive control — means SISPA can be used as a rapid methodology for amplifying metagenomes. In comparison to MDA, it does not favour circular genomes, which are often in high abundance in the gastrointestinal tract. The main bias is instead based on template abundance, which possibly favours pathogenic viruses in scenarios where an acute infection is present in the sampled metagenome. However, both SISPA and MDA limit the range of applications for viral metagenomics, as both of them are random amplifications that induce biases. This prevents them from being used in a quantitative setting and it complicates of whole viral genomes from the sequence data. Emerging methodologies such as the linear amplification for deep sequence and optimised linker amplified shotgun library might result in better suited methodologies for amplification in viral metagenomics, even though these are still untested in large-scale studies (Duhaime *et al.*, 2012; Shankaranarayanan *et al.*, 2012).

4.2 Assessment and Compilation of Bioinformatic Tools (Paper II)

Paper II describes the development of the software MetLab. MetLab encompasses three modules for design and analysis of viral metagenomics experiments. The first module enables support for designing experiments, providing the user with an estimated amount of sequencing for detection of viral species within a sample. The second module produces simulated data for viral metagenomic experiments, based on the user's choice of sequencing methodology as well as species distribution of the sample. The third module is an analytical pipeline, providing the user with a predefined set of tools and a modular execution order for analysis of viral metagenomes.

4.2.1 Experimental Design Module

The first module was the experimental design module. Replication of experiments within metagenomics is complicated by several factors, experimental design being one of them. Tools enabling researchers to standardise the needed sequencing depth as well as choice of sequencing technology could be of great help for designing experiments. The implementation of Stevens theorem in the experimental design module allows the user to estimate the needed sequencing depth based on abundance of a species within the metagenome and genome size of the species (Wendl *et al.*, 2013). The experimental design module is shown in figure 5.

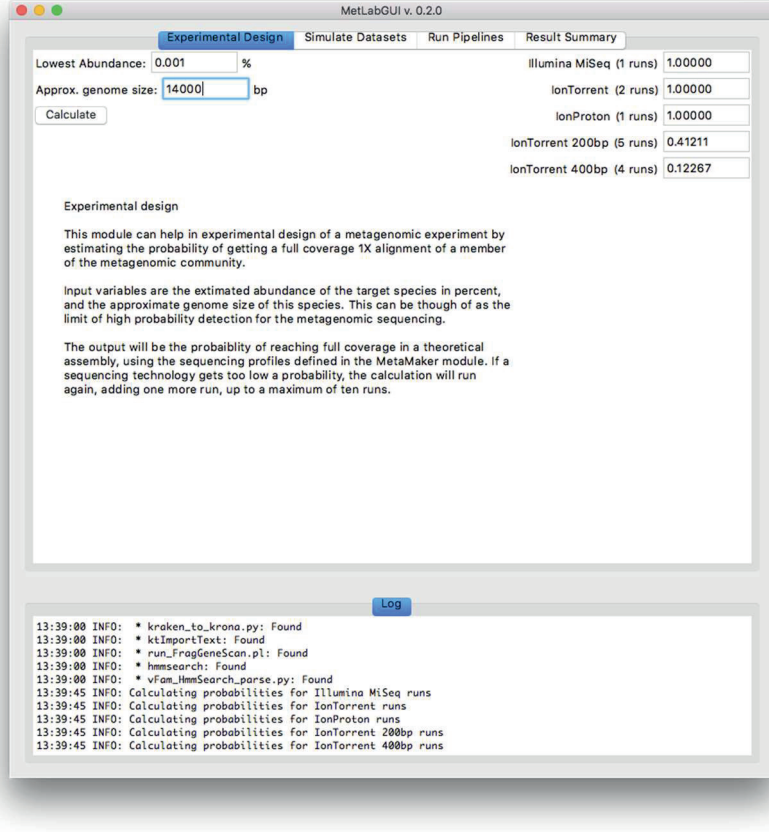

Figure 5. Experimental design module of MetLab. User input consists of genome size and expected lowest abundance of the species. Output consists of the confidence values based on sequencing runs, i.e. number of runs needed to reach coverage of all included genomes.

The module was built upon the work of Wendl *et al.* (2013), who formalized a generalisation of Steven's theorem to cover a particular domain, e.g., a species, from a population of domains, e.g., a metagenome. In the experimental design module, this was used to estimate confidence values for different sequencing techniques to detect a species of abundance, α, in a dataset of size R×L, where L is read length and R number of reads. The graphical user interface (GUI) implementation was designed with simplicity in mind, allowing the user to only modify two variables, abundance of a species and size of the species genome. The software then calculates confidence values and present them as runs of sequencing on different platforms.

4.2.2 Sequence Simulator

The second module implemented the metamaker software. Metamaker takes previously produced sequencing data, analyses the errors introduced by the sequencing technology and the read length of the sequences. Then it extrapolates error profiles from the dataset, specific for the technology analysed. This is then applied as a model for simulating new datasets built on the same error profile as the test-data. The GUI seen in figure 6 enables the user to modify the simulation by variable species count (sample diversity), species distribution, species taxa and sequencing variables, e.g., sequencing technology, the use of mate pair sequencing data, and the insert size. The metamaker sequence simulator module can be used to produce simulated datasets for testing new software, benchmarking execution speed, e.g., running time of software or pipelines, and for estimating sequencing runs for the experimental design module.

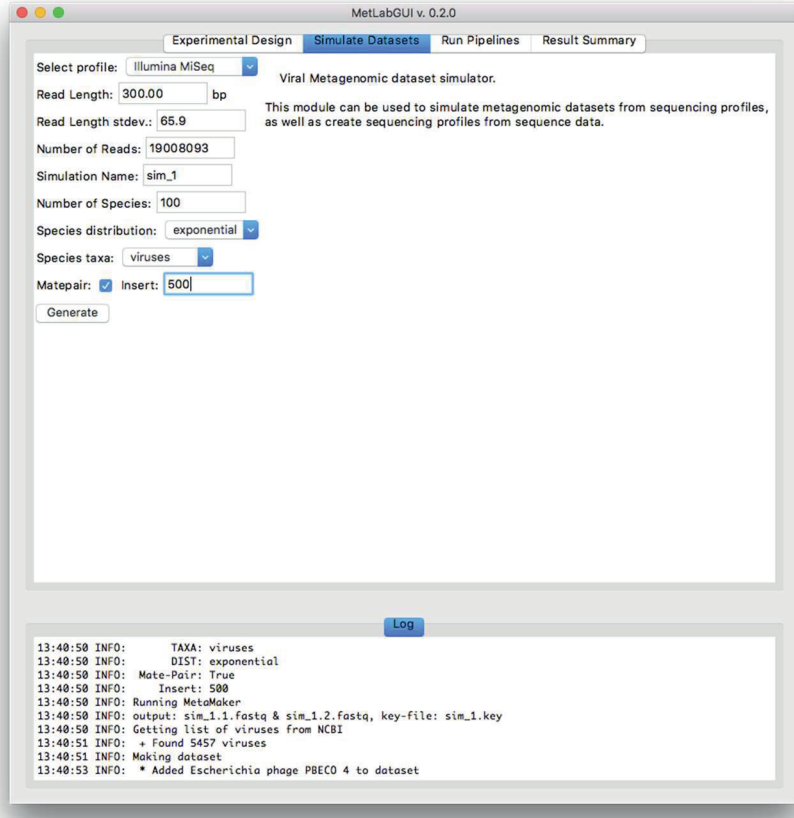

Figure 6. Simulation module of MetLab. User input is limited to sequencing technology, number of species in the simulated metagenome, species taxa and distribution model, exponential or linear.

4.2.3 Analytical Pipeline

The third module of MetLab is a customisable pipeline for analysis of metagenomes. This module enables the user to analyse small-to-moderate sized metagenomics datasets on a standard laptop. This is achieved by the use of a reduced database (removing redundancy) based on the archaeal and bacterial refseq databases as well as the viral and bacteriophage nt-database. The software can also run with an extended database on high-end computers. The GUI for the analysis module is shown in figure 7.

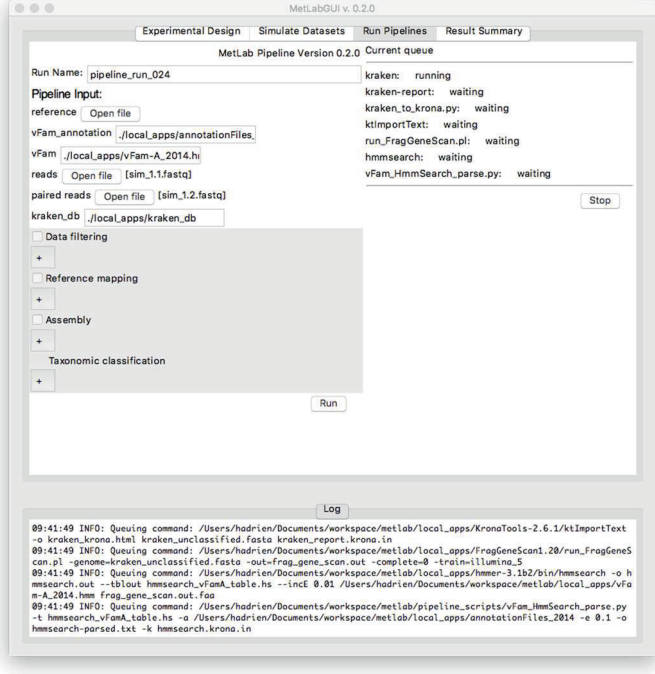

Figure 7. Metagenome analysis pipeline of MetLab. Pipeline input is query data, databases used for Kraken and Hmmer3/FGS/vFam. The analytical pipeline is customisable allowing the users to exclude different parts of the analysis.

The user can specify the host reference sequences, vFam parameters (if a customised vFam database is used), query reads, and Kraken database. All different parts of the module are customisable, enabling the user to skip or add assembly steps to the analysis as needed. Output from the module is presented as krona charts as well as tsv files for easy import into R and other statistical tools (Ondov *et al.*, 2011). This is provided together with a log with execution statistics. Analytical pipelines tend to differ between labs and between studies. MetLab provides the research community with an easy to use standardised pipeline that can be replicated between labs for comparative studies. As studies of viromes become more commonplace, standardisation of both the designs and analysis will be needed for increased usability of the data.

4.3 The Eukaryotic Virome of Diarrhoeic and Healthy Neonatal Piglets (Paper III)

With the methodologies developed in papers I and II, a framework for investigating the virome was developed. This was implemented in paper III which investigated the presence of eukaryotic viruses within the intestinal microbiome of neonatal piglets. Sixty-nine piglets were investigated, 50 suffering from new neonatal porcine diarrhoea and 19 healthy controls. The controls were sequenced individually and the diarrhoeic samples were sequenced in pools of five.

After sequencing, samples were analysed using the methodology developed in Paper II, resulting in 32 datasets: 19 for the healthy controls, 10 for the pools with diseased piglets, one positive control, one process control, and one non template control. All three controls behaved according to expectations. An average of 220 Mb of data was generated per sample with a mean read length of 185 bases, on the IonProton platform at NGI in Uppsala.

Viruses from eight different families known to infect mammals were detected in the whole sample group, by classification with Kraken, Diamond and Hmmer3/FGS/vFam. These virus families were *Adenoviridae*, *Anelloviridae*, *Astroviridae*, *Caliciviridae*, *Circoviridae*, *Parvoviridae*, *Picornaviridae* and *Reoviridae*. Additionally, retroviral sequences were retrieved from the samples as well as a number of unclassified viruses. There were no viruses exclusively associated with the diarrhoeic animals. The most common finding was *Picornaviridae*, predominantly *Aichivirus C*. *Aichivirus C* is commonly thought to be part of the normal virome of pigs. Even so, the transmission route has been suggested to be faecal-oral (Kitajima *et al.*, 2011; Sdiri-Loulizi *et al.*, 2010). The finding in Paper III, of *Aichivirus C* as early as

24 hours after birth might indicate direct infection at birth or transplacental transmission. The findings within paper III were in agreement with previous findings in older animals within the same field of study (Sachsenroder *et al.*, 2014; Lager *et al.*, 2012; Shan *et al.*, 2011). Interestingly, the average number of viral families was somewhat fewer in our samples than in those of previous studies. This could be due to the younger age of the animals or a difference in sample material, which was tissue in study III and faecal in the other studies. The low diversity of the eukaryotic viruses within the virome is also supported by the study on humans presented by Lim *et al.* (2015), where the virome diversifies over the first 24 months of life.

No direct viral cause for the diarrhoeic syndrome was evident using metagenomics. Several viruses were present within hours of birth, representing a number of known enteric viruses. This indicates that the virome is established quickly, within hours of birth, and that its establishment is an individual process, even in herd living animals with large litters.

4.4 Comparative Metagenomics Between Neonatal and Adult Viromes (Manuscript IV)

In paper III, the viruses infecting eukaryotes were investigated as part of an ongoing search for the aetiological agent of neonatal porcine diarrhoea (Larsson, 2016). Within the porcine neonatal metagenome (PNMG) dataset, there were 19 healthy controls. Study IV built upon those data and combined it with three other datasets: two from humans, representing neonates (human infant metagenome, HIMG) and adults (human metagenome, HMG) and one from adult pigs (porcine metagenome, PMG). Study IV investigated similarities between neonatal and adult samples. This was performed both on porcine and human material. This was also investigated between the two species.

For study IV, the operational taxonomic units (OTU) were defined as viral families. Reads from the four datasets were classified using Kaiju; reads unclassified at the family level were excluded. The diversity index was calculated as: alpha for local (within sample) diversity, beta for between sample diversity (similarity between tested), and gamma for total diversity (total OTU within sample groups, such that every OTU was unique and counted once). Samples were split into viruses infecting prokaryotes and those infecting eukaryotes, after which heat maps were generated. Samples were bi-clustered based on beta-diversity, at the viral family and sample levels.

4.4.1 Prokaryotic Virome

The prokaryotic virome of the four datasets is dominated by four families, *Microviridae*, *Myoviridae*, *Podoviridae* and *Siphoviridae*, as seen in figure 8. Notable was the lower presence of *Microviridae* within the porcine neonates compared with the other sample groups. Cluster analysis of the prokaryotic virome (Figure 9) also showed a clear group with the previously mentioned group of viral families as well as a secondary cluster containing *Bicaudaviridae, Fuselloviridae, Inoviridae*, and *Lipothnixviridae*. *Corticoviridae* clustering alone represented a single sample.

Diversity in the prokaryotic virome did not appear to change much between neonatal and adult, with four families of viruses dominating. The gamma-diversity for the adult samples was 9 and for the neonates, 7, see table 3. The combined Gamma for neonates turned out the same as for the adult samples, indicating presence of 9 different viruses combined in the two datasets, PNMG and HIMG. This might indicate that different viruses populate the human neonates compare to the porcine. This is also seen in beta-diversity, see figure 8, where the neonates of porcine origin do not cluster with the neonates of human origin. Porcine neonates had fewer samples with *Microviridae* (10/19), compared to the human neonates (8/8). This might have been due to different sampling conditions, as faecal matter was sampled in the human samples and tissue in the pigs or it may reflect a genuine pattern in the establishment of the prokaryotic virome. However, adult samples did not show this dissimilarity.

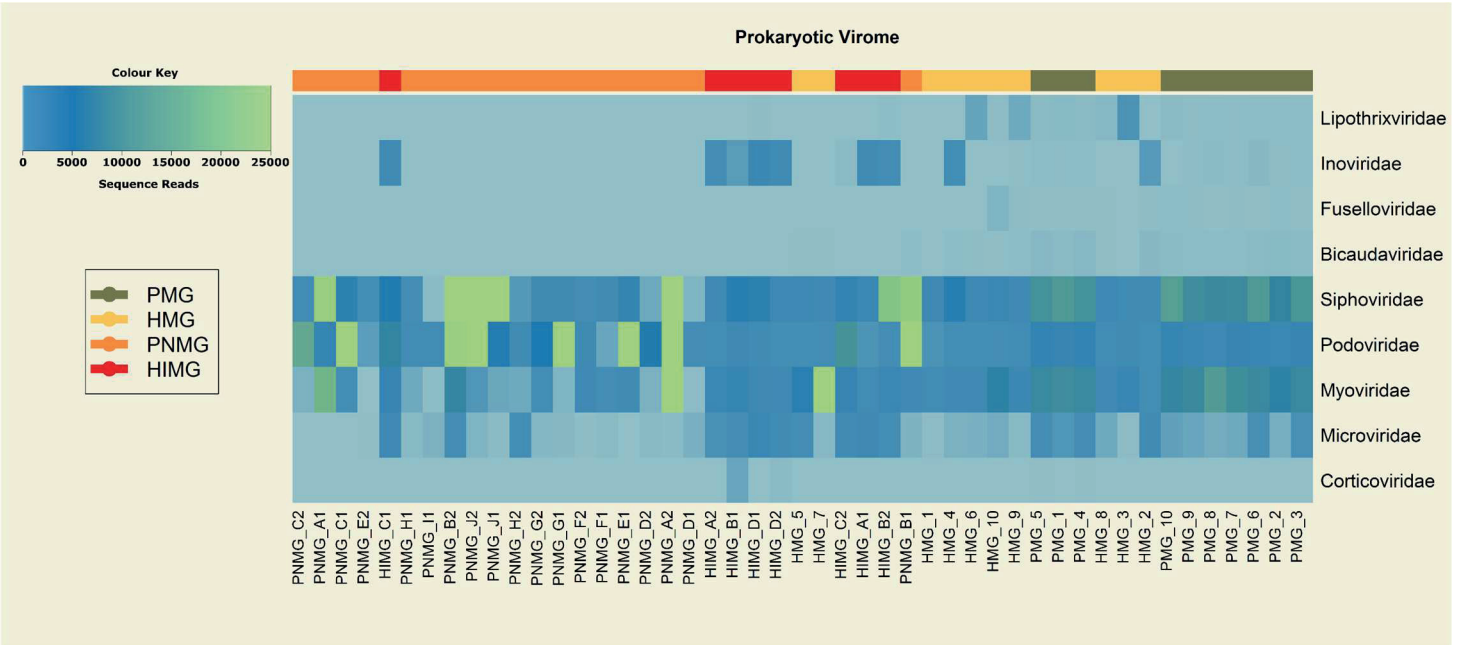

Figure 8. Heat map of the prokaryotic virome. The columns represent samples and the rows represent virus families. Column and row clustering are based on a beta-diversity, Jaccard-distance. Top row represents samples, by colour code as seen in legend to the left. Heat map intensity represents sequence reads for each entry as seen in legend top left.

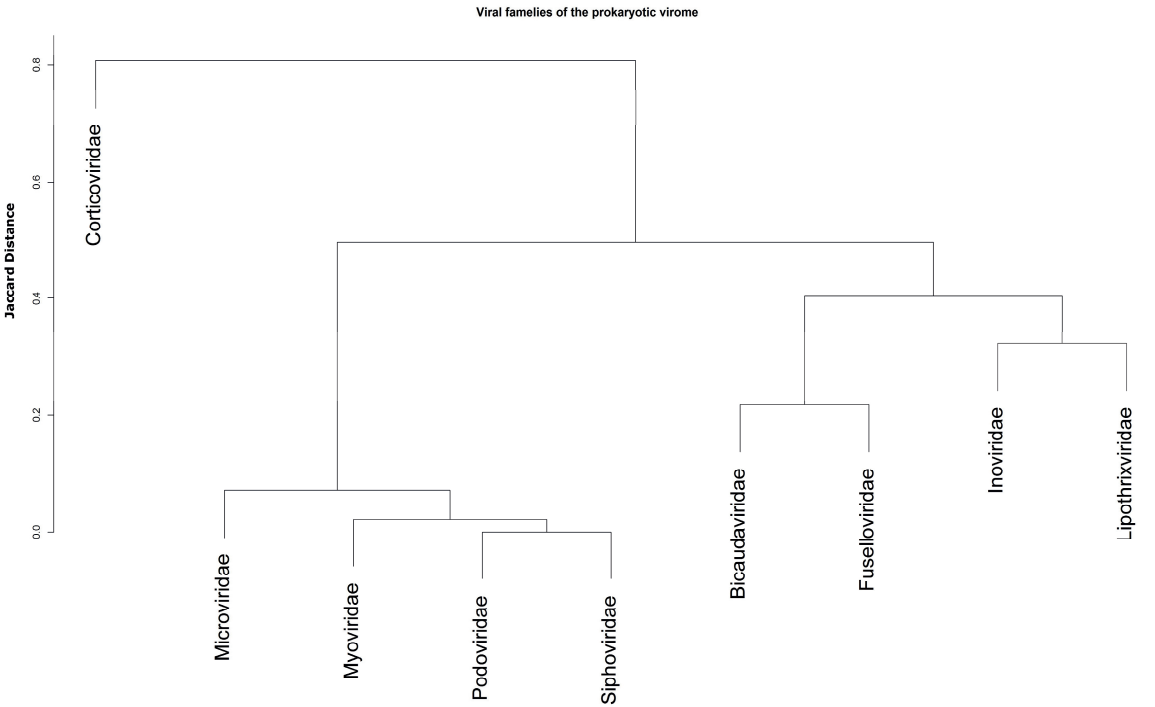

Figure 9. Clustering based on Jaccard-index for viral families within the prokaryotic virome. Distance indicate low similarity to the other. base on occurrence within the dataset. Two main clusters are formed with a single outlier, *Corticoviridae.*

Table 3. *Gamma diversity for the prokaryotic virome*

Dataset	Gamma
PMG	9
HMG	9
PNMG	7
HIMG	7
PMG+HMG	9
PNMG+HIMG	9

4.4.2 Eukaryotic Virome

The eukaryotic virome was considerably more diverse than the prokaryotic one and also showed a greater dissimilarity between neonates, and between the neonates and adults. The difference was greatest between the human neonates and the other datasets. The human and porcine neonates were born into considerably different environments so a difference in diversity was not unexpected. The eukaryotic viral diversity differs from the pattern seen in the prokaryotic viral diversity as HIMG have considerable less viral families present compared to the other three datasets. Notable is the difference between the PNMG dataset and the adult dataset that was considerably less than between the human neonates and the other datasets, as seen in figure 10 and 11. HIMG formed several small clusters within the analysis. PMG and HMG clustered together, indicating high similarity between the datasets (measured in diversity). PNMG clustered over several smaller clusters, forming three clusters and an outlier. That PMG and HMG clustered together was interesting, as it could indicate that pigs might be used as models for gastrointestinal virome composition in addition too bacteriological (Pang *et al.*, 2007).

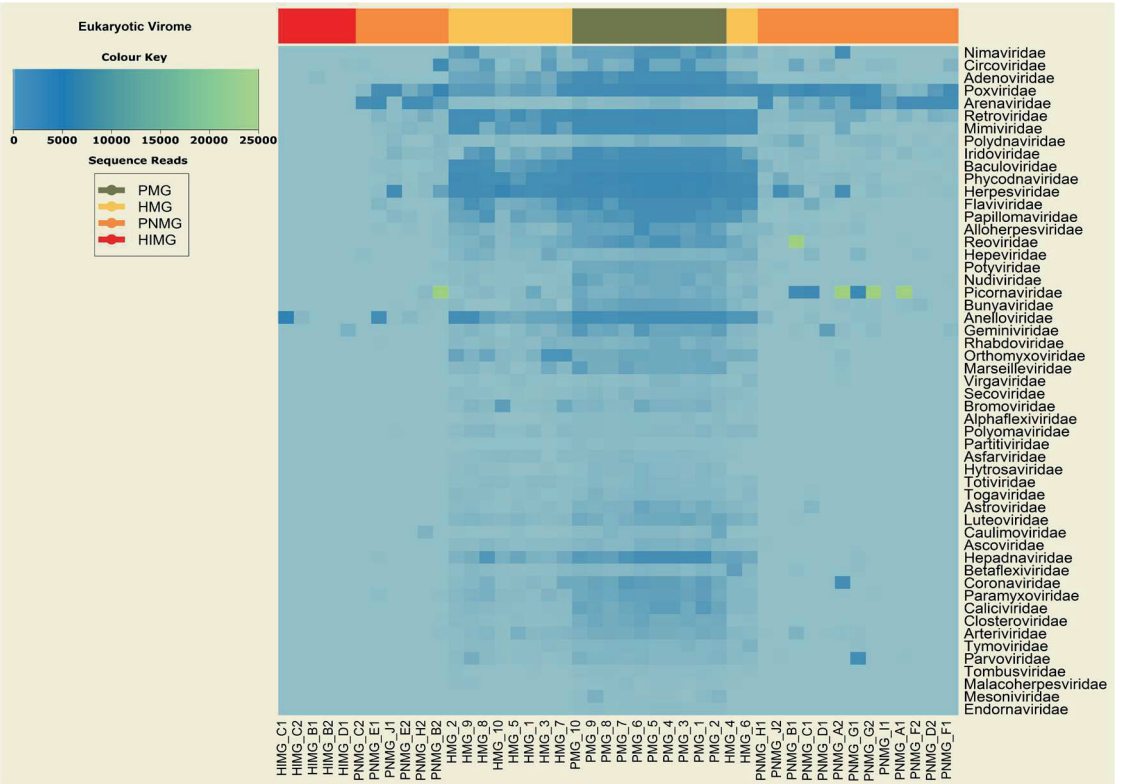

Figure 10. Heat map of the eukaryotic virome. Column and row clustering are based on a beta-diversity, Jaccard-distance. Top row represents samples, by colour code as seen in legend to the left. Heat map intensity represents sequence reads for each entry as seen in legend top left.

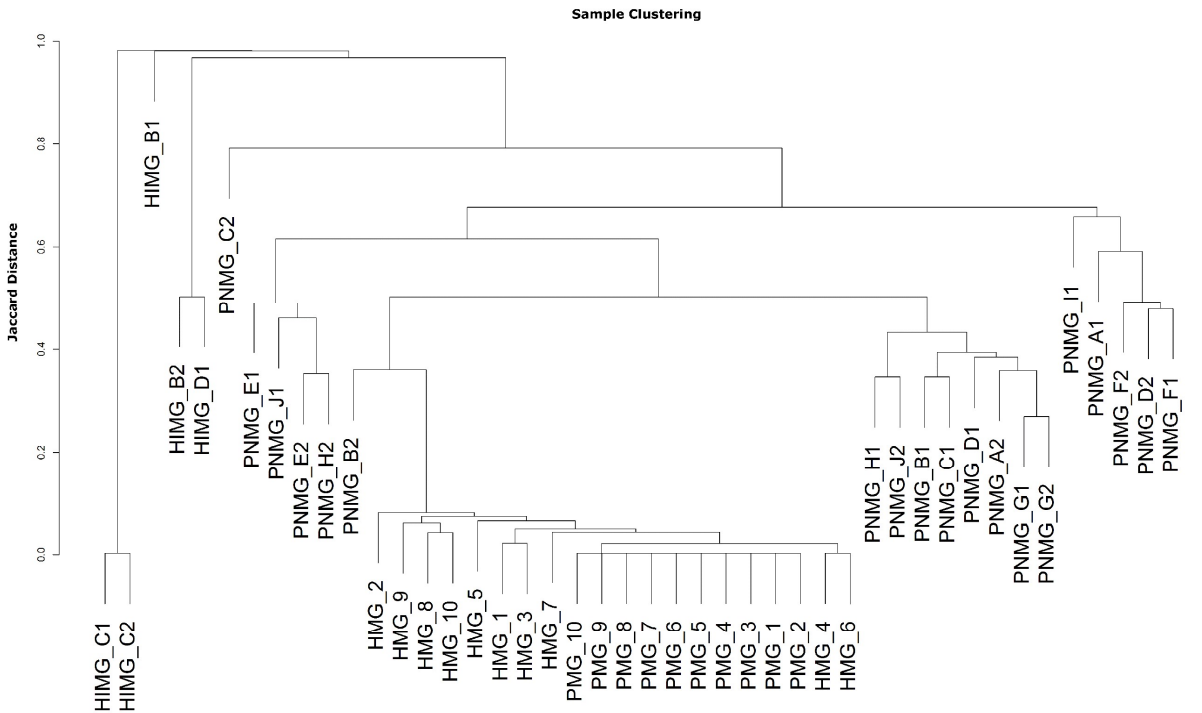

Figure 11. Clustering based on Jaccard-index for the eukaryotic virome on samples level. Notable is how the adult samples cluster together, indicating high similarity (similar diversity).

Another interesting finding was the low gamma diversity within the HIMG samples, see table 4. This might be due to several reasons. Viral diversity in the human neonates might be low due to differences in amplification methodology. This is however unlikely as they were amplified using two different methods. The lower diversity might also be a result of classification error, as the three other samples groups were classified using the Kaiju classification rather than the BLAST used by Lim *et al.* (2015). Finally, the sequence technology differed between datasets; the porcine neonates were sequenced with IonProton and the human neonates with Illumina. None of these explanations would account for all the differences seen in diversity. However, the method of delivery differed between the human neonates and the porcine neonates, where three out of four couple of twins were delivered by caesarean (Lim *et al.*, 2015). This is a known factor that will change the composition of the microbiome and might change the composition of the virome (Dominguez-Bello *et al.*, 2010). As mentioned before, differences in birth environment should also be considered, as the human neonates were born in a hospital whereas the porcine neonates are born in pig pens on a farm. The number of reads classified as *Picornaviridae* in the PNMG was also in stark contrast to the absence of such reads in the HIMG dataset. Detection of reads classified as *Picornaviridae* occurred in 28 out of 47 samples, ~60%, for the whole sample collection, but only in 3 out of the 18 human samples (including the neonates). However, viruses in the family *Picornaviridae* typically infect the gastrointestinal tract of pigs and finding them there in high prevalence was expected.

Table 4. *Gamma diversity for the eukaryotic virome*

Dataset	Gamma
PMG	53
HMG	53
PNMG	48
HIMG	4

4.4.3 The Virome

When the combined eukaryotic and prokaryotic virome was analysed, the four sample groups formed distinctive clusters as seen in figure 12, only showing viral families with more than 1000 reads assigned. Adult samples clustered together, indicating high similarity, and neonates clustered in two separate clusters. It is interesting that the neonates were so different from each other, i.e. not clustering together, whereas the adults had a similar diversity. This might indicate that the initial virome composition of humans differs from that of pigs.

However, several factors should be taken into account: virome isolation (performed on neonatal pigs and not on adult pigs or humans), amplification strategy (two methods on neonatal humans, one on neonatal pigs and no amplification on adult samples), and sequencing technology (Illumina on humans, IonProton on neonatal pigs and Illumina on adult samples). Even though these factors confound a direct comparison, they hardly account for all the differences in similarity between the sample sets. The sample population is however too limited to draw firm conclusions.

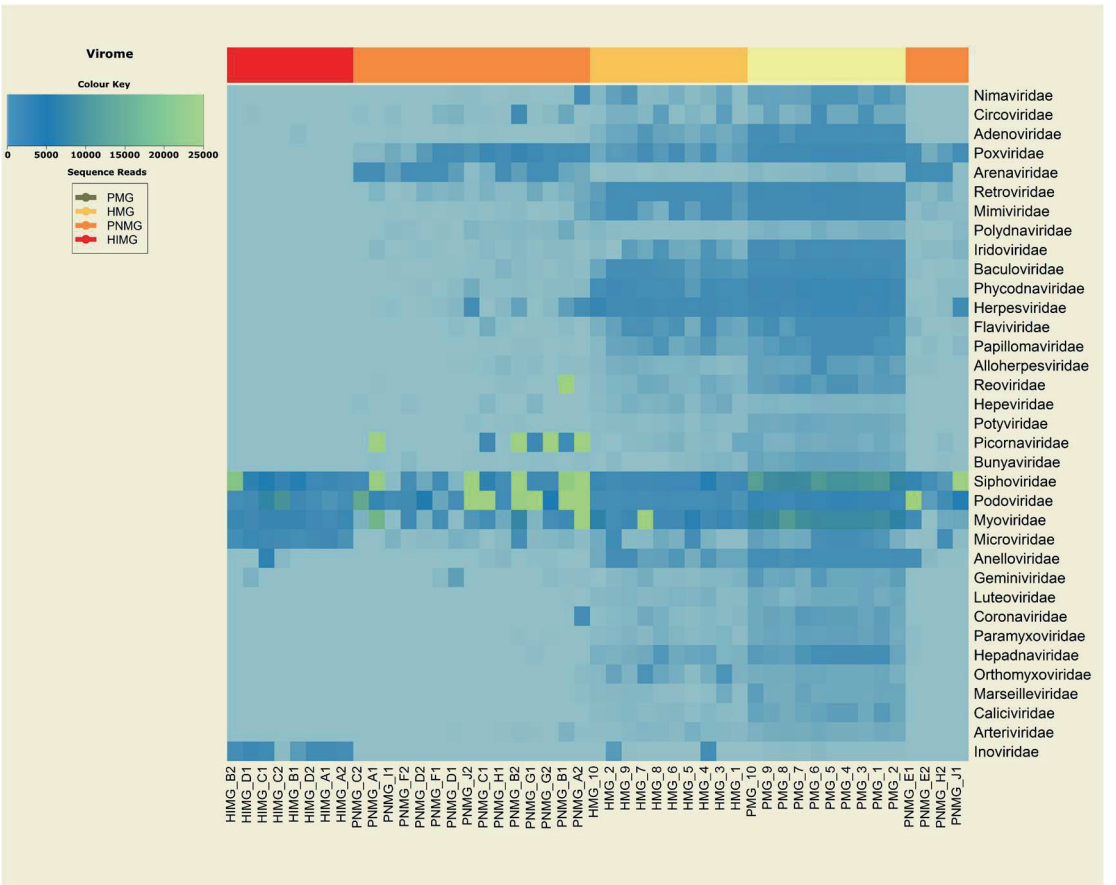

Figure 12. Heat map of virome composition. Column and row clustering are based on a beta-diversity, Jaccard-distance. Top row represents samples, by colour code as seen in legend to the left. Heat map intensity represents sequence reads for each entry as seen in legend top left. Note the neonatal samples clustering together and the adult samples mixing. Clearly visible is the higher diversity in the adult samples as well as the four viral families representing the most common viral families in the prokaryotic virome.

Four families were identified as a core within the prokaryotic virome in pigs and humans: *Microviridae*, *Myoviridae*, *Podoviridae* and *Siphoviridae*. Even though *Microviridae* viruses were somewhat fewer in the neonatal pigs, 16/19, then in the HIMG, 8/8, similarity clustering indicates that *Microviridae* belong in the core group. Stern *et al.* (2012) found that:

> Our results suggest the existence of a non-negligible common reservoir of phages that is spread among unrelated individuals residing in distant geographical locations.

Given the pattern suggested by Stern *et al.* (2012) in humans and the one observed within our study IV, it is possible that a common core of phages are shared between mammals. If this holds up under scrutiny in larger studies, it could enable a roadmap for manipulation of the microbiome and provide indicators for overall health of the animal.

The eukaryotic virome had no similar pattern of viral families present over all sample groups. However, the adult samples, HMG and PMG, shared several viral families. If total diversity is limited to only viral families where the total number of viral reads exceeds 1000, thereby removing outliers and low abundance viruses, the pattern changes, see figure 13.

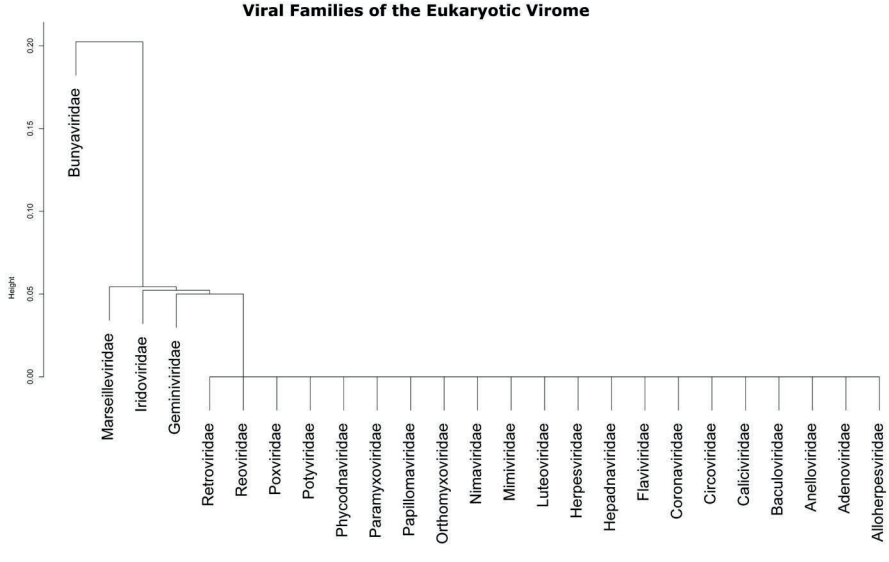

Figure 13. Clustering based on Jaccard-index for the eukaryotic virome on viral family level. Notable is the high similarity between the baseline of viruses, indicating a common presence within the dataset. Five smaller cluster represent the outliers in the eukaryotic virome.

Clustering of the eukaryotic virome, figure 13, now show a baseline of viral families with some viral families clustering alone or in small groups. Given the limited number of samples used, this pattern would have to be studied in a larger group of individuals before conclusive proof can be presented, but the pattern indicates that mammals of different species could be sharing a large portion of their eukaryotic virome.

4.4.4 Implications of Difference in Similarity Between the Neonatal and Adult Viromes

In study IV, a comparison was performed between the neonatal and adult viromes of two species. The preliminary results indicated that there is an in-between-species similarity within the prokaryotic virome. Differences in the eukaryotic virome between neonates and adults was profound, but between adults there was a discernible similarity and several viral families were shared.
Viral species are an important factor modulating the microbiome and therefore animal health. The prokaryotic virome reacts to antibiotic stressors, inflammatory states, diet changes and metabolic states such as obesity (Yang *et al.*, 2016; Norman *et al.*, 2015; Abeles & Pride, 2014; Allen *et al.*, 2011). The eukaryotic virome has not been explored to the same degree, but it does provide immunogenic benefits to the host and possibly protection through superinfection exclusions (Beperet *et al.*, 2014; Schaller *et al.*, 2007).

If the results in study IV are reproducible in larger study groups and between other species, it would indicate not only a shared genetic ancestry but also a shared microbiome ancestry. Given the theory that the microbiome and the host evolved as a unit, the possibility of a shared core in the microbiome is not that different from our shared genetic similarities. If applied to medicine, this could mean a breakthrough for microbiome research, as crucial basic science for manipulating the microbiome could be performed in animals with high virome similarity and then extrapolated to humans. This could provide medical practitioners in all fields with a novel and powerful tool to battle the many metabolic and chronic inflammatory diseases associated with a disturbed microbiome as well as provide better tools for predicting outcome of phage-therapy.

5 Concluding Remarks and Future Perspectives

Viral metagenomics provide an unprecedented ability to characterise the viromes in various host species. Within this thesis methodological assessment and development was performed to improve the ability to amplify samples before sequencing as well as analyse samples after sequencing. These methodologies were then applied on samples from neonatal piglets with diarrhoea of unknown aetiology to discern possible differences between healthy and sick neonates. Lastly, comparative viral metagenomics was performed on four datasets, two from neonates and two from adults of porcine and human origin. Diversity was compared between the datasets to discern possible similarity between neonates and adults as well as between the two species.

In conclusion the results of this thesis were:

- Sequence-independent, single-primer amplification can be used to amplify viral metagenomes before sequencing. The methodology does however introduce biases in the dataset, e.g. abundance change in the composition of the metagenome and uneven amplification of genomes within the sample.
- An implementation of Stevens's theorem was developed together with a sequencing simulation tool to provide researchers with reliable estimates on needed sequencing depth as well as *in silico* generated datasets for method assessment and development.
- Two previously developed methodologies for taxonomic classification, Kraken and Hmmer3/FGS/vFam, were combined into an analytical pipeline for analysis of viral metagenomes. The implementation is quick, less computationally extensive than previous methodology and can be executed on a normal office computer.

- A complete analytical pipeline was developed for quality control, assembly, classification and visualisation of viral metagenomes. This pipeline was combined with the experimental design and dataset simulator into a comprehensive software for researchers working with viral metagenomics.
- A possible aetiological background of neonatal porcine diarrhoea was investigated. There was no conclusive proof for a viral cause to the disease state in these particular cases.
- Comparative metagenomics was employed on four datasets to discern changes between neonatal and adult viromes as well as possible similarities between pig and human viromes.
  - A core of bacteriophage families could be detected, showing high similarity between neonates as well as between species.
  - There was little similarity between neonates of human and porcine origin in the eukaryotic virome.
  - Higher diversity in adults was attributed to a higher diversity of eukaryotic viruses. Limiting the eukaryotic virome in adults to exclude outliers provided indications that a core virome can be found in adults of two different species.

5.1 Future perspectives

Viral metagenomics is providing crucial data for understanding viruses within the microbiome, the virome. Introduction of HTS has provided researchers with technological possibilities for characterisation of viruses by use of metagenomics. This also raises several points to be addressed in the future.

- Virome isolation and enrichment should be developed to enable bias free, linear representation of viromes prior to sequencing.
- Sequencing methodologies employed for characterisation of viromes should aim at single molecule sequencing with long reads. This would enable a more accurate classification of viral reads within a sample. Current single molecule sequencing technologies are also amplification free, providing a linear representation of the sample.
- Experimental design for characterisation of viromes should aim at following best practise approach. This would include sequencing depth, metadata collection, replicates and controls. If employed properly, this would enable comparison of metagenomic datasets with greater confidence.

- Tools for analysis should be standardised to have a common point of reference attached to experiments as metadata. This would greatly improve datamining and use of big data methodologies in metagenomics.
- A mammalian virome initiative should be considered. Given results within this thesis as well as in literature the virome of humans and other mammals could share considerable characteristics. If that is the case careful manipulation of the virome could lead to improved health and disease management.

6 Populärvetenskaplig Sammanfattning

Alla högre organismer, så som människor och djur, är koloniserade av ett komplext samhälle av mikroorganismer. Dessa mikroorganismer kallas med ett ord för mikrobiomet. Mikrobiomet innehåller tusentals organismer och är i konstant förändring. Det anpassar sig efter värdens hälsa och dess födointag likväl som efter invasion av skadliga mikroorganismer, infektioner. Till mikrobiomet räknas bakterier, parasiter, arkéer, virus samt de faktorer som påverkar dem.

De virus som uppehåller sig inom mikrobiomet kallas med ett ord för viromet. Viromet påverkar både mikrobiomet och värddjuret. Detta sker både genom direkt påverkan, så som ett virus som infekterar värddjuret och genom indirekt verkan, som ett virus som infekterar en bakterie och därmed överför sjukdomsalstrande egenskaper till den bakterien. Viromet är därmed viktigt att studera då dess egenskaper kan vara avgörande för hälsa såväl som sjukdom.

Tack vare det senaste årtiondets utveckling av sekvenseringsteknologin, den molekylärbiologiska metod med vilken man läser av arvsmassa hos organismer, kan forskare nu läsa arvsmassan i viromet, dessa dataset kallas metagenom. Med hjälp av den informationen kan man härleda vilka olika typer av virus som är närvarande i viromet och därmed få insikter i vilka egenskaper som ett djurs individuella virom har. För att genomföra detta måste man först isolera viromet, vilket sker genom att man reducerar närvaron av värddjurets och andra mikroorganismers arvsmassa. Detta sker genom filtrering, centrifugering och behandling med nukleas, ett enzym som bryter ner fri arvsmassa så att endast skyddad arvsmassa består. Det sista steget är möjligt då viruspartiklar är byggda med en skyddande kapsel, en så kallad virion, som skyddar virusets arvsmassa från att brytas ner av nukleasbehandlingen. Efter behandlingen återstår väldigt lite arvsmassa. För att man ska kunna sekvensera den måste man amplifiera arvsmassan, vilket sker genom två huvudtyper av metoder.

Efter amplifiering genomförs sekvensering och arvsmassan i provet läses av. Den resulterande datamängden uppgår ofta till flera gånger den mänskliga arvsmassan, motsvarande 3000 böcker med runt 600 sidor, och är otroligt komplex. På grund av komplexiteten måste man använda datorer för att kartlägga informationen inuti arvsmassan från viromet. Detta kallas för bioinformatik och är de metoder och program som hanterar biologisk information.

I den här avhandlingen har både den molekylärbiologiska metodbiten och bioinformatiken utvärderats och förbättrats. Ett program har utvecklats för att analysera viromet och visualisera dess innehåll. Metoderna har sedan testats på två fall, ett fall där en möjlig virusinfektion undersöktes i smågrisar och ett fall där man studerade likheter mellan viromet i vuxna djur och späddjur hos människa och gris.

I fallet med spädgrisarna hittades inte något virus som förklarade sjukdomstillståndet. Dock hittades ett antal tidigare kända virus utspridda i de individuella viromen. Tack vara data vid den här studien kunde man nu med säkerhet säga att virus inte orsakade sjukdomen hos smågrisarna. Detta föranledda till det sista fallet, där likheter mellan späddjur och vuxna djur jämfördes. För att göra jämförelsen använde man sig av tre grupper av data förutom det tidigare nämnda: humana spädbarn, vuxna grisar och vuxna människor.

I det sista fallet hittade man gemensamma nämnare för viromet över samtliga grupper i den del av viromet som infekterar bakterier. Även i den del av viromet som infekterar värddjuren hittade man likheter, främst mellan vuxna individer av de två djurslagen. Studien ger en första inblick i hur två arter som lever och utvecklats tillsammans delar egenskaper i sitt virom. Om det här stämmer även vid större studier är det en framtida möjlighet för att modifiera viromet för ökad hälsa och välbefinnande bos både djur och människor.

References

Aagaard, K., Ma, J., Antony, K.M., Ganu, R., Petrosino, J. & Versalovic, J. (2014). The Placenta Harbors a Unique Microbiome. *Science Translational Medicine,* 6(237).

Abeles, S.R. & Pride, D.T. (2014). Molecular Bases and Role of Viruses in the Human Microbiome. *Journal of Molecular Biology,* 426(23), pp. 3892-3906.

Allander, T., Emerson, S.U., Engle, R.E., Purcell, R.H. & Bukh, J. (2001). A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. *Proceedings of the National Academy of Sciences of the United States of America,* 98(20), pp. 11609-11614.

Allen, H.K., Looft, T., Bayles, D.O., Humphrey, S., Levine, U.Y., Alt, D. & Stanton, T.B. (2011). Antibiotics in Feed Induce Prophages in Swine Fecal Microbiomes. *Mbio,* 2(6).

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology,* 215(3), pp. 403-410.

Anthony, S.J., Epstein, J.H., Murray, K.A., Navarrete-Macias, I., Zambrana-Torrelio, C.M., Solovyov, A., Ojeda-Flores, R., Arrigo, N.C., Islam, A., Khan, S.A., Hosseini, P., Bogich, T.L., Olival, K.J., Sanchez-Leon, M.D., Karesh, W.B., Goldstein, T., Luby, S.P., Morse, S.S., Mazet, J.A.K., Daszak, P. & Lipkin, W.I. (2013). A Strategy To Estimate Unknown Viral Diversity in Mammals. *Mbio,* 4(5).

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A. & Pevzner, P.A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology,* 19(5), pp. 455-477.

Barton, E.S., White, D.W., Cathelyn, J.S., Brett-McClellan, K.A., Engle, M., Diamond, M.S., Miller, V.L. & Virgin, H.W. (2007). Herpesvirus latency confers symbiotic protection from bacterial infection. *Nature,* 447(7142), pp. 326-U7.

Belak, S., Karlsson, O.E., Blomstrom, A.L., Berg, M. & Granberg, F. (2013). New viruses in veterinary medicine, detected by metagenomic approaches. *Veterinary Microbiology,* 165(1-2), pp. 95-101.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Cheetham, R.K., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J.,

Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X.H., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M.J., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X.L., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M.D., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Catenazzi, M.C.E., Chang, S., Cooley, R.N., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fajardo, K.V.F., Furey, W.S., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoschler, K., Hurwitz, S., Ivanov, D.V., Johnson, M.Q., James, T., Jones, T.A.H., Kang, G.D., Kerelska, T.H., Kersey, A.D., Khrebtukova, I., Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee, X., Liao, A.K., Loch, J.A., Lok, M., Luo, S.J., Mammen, R.M., Martin, J.W., McCauley, P.G., McNitt, P., Mehta, P., Moon, K.W., Mullens, J.W., Newington, T., Ning, Z.M., Ng, B.L., Novo, S.M., O'Neill, M.J., Osborne, M.A., Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, A.C., Pike, A.C., Pinkard, D.C., Pliskin, D.P., Podhasky, J., Quijano, V.J., Raczy, C., Rae, V.H., Rawlings, S.R., Rodriguez, A.C., Roe, P.M., Rogers, J., Bacigalupo, M.C.R., Romanov, N., Romieu, A., Roth, R.K., Rourke, N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short, S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Sohna, J.E.S., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., vandeVondele, S., Verhovsky, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J.W., Worsley, G.J., Yan, J.Y., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J.C., Hurles, M.E., McCooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R. & Smith, A.J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature,* 456(7218), pp. 53-59.

Beperet, I., Irons, S.L., Simon, O., King, L.A., Williams, T., Possee, R.D., Lopez-Ferber, M. & Caballero, P. (2014). Superinfection Exclusion in Alphabaculovirus Infections Is Concomitant with Actin Reorganization. *Journal of Virology,* 88(6), pp. 3548-3556.

Bergstrom, A., Skov, T.H., Bahl, M.I., Roager, H.M., Christensen, L.B., Ejlerskov, K.T., Molgaard, C., Michaelsen, K.F. & Licht, T.R. (2014). Establishment of Intestinal Microbiota during Early Life: a Longitudinal, Explorative Study of a Large Cohort of Danish Infants. *Applied and Environmental Microbiology,* 80(9), pp. 2889-2900.

Binga, E.K., Lasken, R.S. & Neufeld, J.D. (2008). Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *Isme Journal,* 2(3), pp. 233-241.

Blomstrom, A.L. (2011). Viral metagenomics as an emerging and powerful tool in veterinary medicine. *Veterinary Quarterly,* 31(3), pp. 107-114.

Blomstrom, A.L., Widen, F., Hammer, A.S., Belak, S. & Berg, M. (2010). Detection of a Novel Astrovirus in Brain Tissue of Mink Suffering from Shaking Mink Syndrome by Use of Viral Metagenomics. *Journal of Clinical Microbiology,* 48(12), pp. 4392-4396.

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F. & Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biology,* 13(12).

Buchfink, B., Xie, C. & Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods,* 12(1), pp. 59-60.

Cadwell, K. (2015). The Virome in Host Health and Disease. *Immunity,* 42(5), pp. 805-813.

Cho, I. & Blaser, M.J. (2012). APPLICATIONS OF NEXT-GENERATION SEQUENCING The human microbiome: at the interface of health and disease. *Nature Reviews Genetics,* 13(4), pp. 260-270.

Compeau, P.E.C., Pevzner, P.A. & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology,* 29(11), pp. 987-991.

Costello, E.K., Stagaman, K., Dethlefsen, L., Bohannan, B.J.M. & Relman, D.A. (2012). The Application of Ecological Theory Toward an Understanding of the Human Microbiome. *Science,* 336(6086), pp. 1255-1262.

Cuesta, A.M., Suarez, E., Larsen, M., Jensen, K.B., Sanz, L., Compte, M., Kristensen, P. & Alvarez-Vallina, L. (2006). Enhancement of DNA vaccine potency through linkage of antigen to filamentous bacteriophage coat protein III domain I. *Immunology,* 117(4), pp. 502-506.

Daly, G.M., Bexfield, N., Heaney, J., Stubbs, S., Mayer, A.P., Palser, A., Kellam, P., Drou, N., Caccamo, M., Tiley, L., Alexander, G.J.M., Bernal, W. & Heeney, J.L. (2011). A Viral Discovery Methodology for Clinical Biopsy Samples Utilising Massively Parallel Next Generation Sequencing. *Plos One,* 6(12).

Daniel, R. (2005). The metagenomics of soil. *Nature Reviews Microbiology,* 3(6), pp. 470-8.

De Vlaminck, I., Khush, K.K., Strehl, C., Kohli, B., Luikart, H., Neff, N.F., Okamoto, J., Snyder, T.M., Cornfield, D.N., Nicolls, M.R., Weill, D., Bernstein, D., Valantine, H.A. & Quake, S.R. (2013). Temporal Response of the Human Virome to Immunosuppression and Antiviral Therapy. *Cell,* 155(5), pp. 1178-1187.

Dean, F.B., Nelson, J.R., Giesler, T.L. & Lasken, R.S. (2001). Rapid amplification of plasmid and phage DNA using phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Research,* 11(6), pp. 1095-1099.

Delwart, E.L. (2007). Viral metagenomics. *Reviews in Medical Virology,* 17(2), pp. 115-131.

Didelot, X., Bowden, R., Wilson, D.J., Peto, T.E.A. & Crook, D.W. (2012). Transforming clinical microbiology with bacterial genome sequencing. *Nature Reviews Genetics,* 13(9), pp. 601-612.

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science,* 14(6), pp. 927-930.

Dominguez-Bello, M.G., Costello, E.K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N. & Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences of the United States of America,* 107(26), pp. 11971-11975.

Duerkop, B.A. & Hooper, L.V. (2013). Resident viruses and their interactions with the immune system. *Nature Immunology,* 14(7), pp. 654-659.

Duhaime, M.B., Deng, L., Poulos, B.T. & Sullivan, M.B. (2012). Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous

assessment and optimization of the linker amplification method. *Environmental Microbiology,* 14(9), pp. 2526-2537.

Edwards, R.A. & Rohwer, F. (2005). Viral metagenomics. *Nature Reviews Microbiology,* 3(6), pp. 504-510.

Ewing, B. & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research,* 8(3), pp. 186-194.

Feschotte, C. & Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. *Nature Reviews Genetics,* 13(4), pp. 283-U88.

Foxman, E.F. & Iwasaki, A. (2011). Genome-virome interactions: examining the role of common viral infections in complex disease. *Nature Reviews Microbiology,* 9(4), pp. 254-264.

Gill, S.R., Pop, M., DeBoy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., Gordon, J.I., Relman, D.A., Fraser-Liggett, C.M. & Nelson, K.E. (2006). Metagenomic analysis of the human distal gut microbiome. *Science,* 312(5778), pp. 1355-1359.

Goedert, J.J., Hua, X., Yu, G. & Shi, J. (2014). Diversity and composition of the adult fecal microbiome associated with history of cesarean birth or appendectomy: analysis of the American Gut Project. *EBioMedicine,* 1(2), pp. 167-172.

Goodwin, S., McPherson, J.D. & McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics,* 17(6), pp. 333-351.

Granberg, F., Vicente-Rubiano, M., Rubio-Guerri, C., Karlsson, O.E., Kukielka, D., Belak, S. & Sanchez-Vizcaino, J.M. (2013). Metagenomic Detection of Viral Pathogens in Spanish Honeybees: Co- Infection by Aphid Lethal Paralysis, Israel Acute Paralysis and Lake Sinai Viruses. *Plos One,* 8(2).

Group, N.H.W., Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J.A., Bonazzi, V., McEwen, J.E., Wetterstrand, K.A., Deal, C., Baker, C.C., Di Francesco, V., Howcroft, T.K., Karp, R.W., Lunsford, R.D., Wellington, C.R., Belachew, T., Wright, M., Giblin, C., David, H., Mills, M., Salomon, R., Mullins, C., Akolkar, B., Begg, L., Davis, C., Grandison, L., Humble, M., Khalsa, J., Little, A.R., Peavy, H., Pontzer, C., Portnoy, M., Sayre, M.H., Starke-Reed, P., Zakhari, S., Read, J., Watson, B. & Guyer, M. (2009). The NIH Human Microbiome Project. *Genome Research,* 19(12), pp. 2317-23.

Hall, N. (2007). Advanced sequencing technologies and their wider impact in microbiology. *Journal of Experimental Biology,* 210(9), pp. 1518-1525.

Hall, R.J., Wang, J., Todd, A.K., Bissielo, A.B., Yen, S.H., Strydom, H., Moore, N.E., Ren, X.Y., Huang, Q.S., Carter, P.E. & Peacey, M. (2014). Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *Journal of Virological Methods,* 195, pp. 194-204.

Handley, S.A. (2016). The virome: a missing component of biological interaction networks in health and disease. *Genome Medicine,* 8.

Hodyra-Stefaniak, K., Miernikiewicz, P., Drapala, J., Drab, M., Jonczyk-Matysiak, E., Lecion, D., Kazmierczak, Z., Beta, W., Majewska, J., Harhala, M., Bubak, B., Klopot, A., Gorski, A. & Dabrowska, K. (2015). Mammalian Host-Versus-Phage immune response determines phage fate in vivo. *Scientific Reports,* 5.

Hogeweg, P. (2011). The Roots of Bioinformatics in Theoretical Biology. *Plos Computational Biology,* 7(3).

Hooper, L.V., Wong, M.H., Thelin, A., Hansson, L., Falk, P.G. & Gordon, J.I. (2001). Molecular analysis of commensal host-microbial relationships in the intestine. *Science,* 291(5505), pp. 881-4.

Hooper, S.D., Dalevi, D., Pati, A., Mavromatis, K., Ivanova, N.N. & Kyrpides, N.C. (2010). Estimating DNA coverage and abundance in metagenomes using a gamma approximation. *Bioinformatics,* 26(3), pp. 295-301.

Howe, A.C., Jansson, J.K., Malfatti, S.A., Tringe, S.G., Tiedje, J.M. & Brown, C.T. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *Proc Natl Acad Sci U S A,* 111(13), pp. 4904-9.

Hugenholtz, P. & Tyson, G.W. (2008). Microbiology: metagenomics. *Nature,* 455(7212), pp. 481-3.

Hurwitz, B.L. & Sullivan, M.B. (2013). The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *Plos One,* 8(2), p. e57355.

Huson, D.H., Auch, A.F., Qi, J. & Schuster, S.C. (2007). MEGAN analysis of metagenomic data. *Genome Research,* 17(3), pp. 377-86.

Ignacio-Espinoza, J.C., Solonenko, S.A. & Sullivan, M.B. (2013). The global virome: not as big as we thought? *Current Opinion in Virology,* 3(5), pp. 566-571.

Karst, S.M. (2016). Viral Safeguard: The Enteric Virome Protects against Gut Inflammation. *Immunity,* 44(4), pp. 715-8.

Kernbauer, E., Ding, Y. & Cadwell, K. (2014). An enteric virus can replace the beneficial function of commensal bacteria. *Nature,* 516(7529), pp. 94-98.

Kim, K.H. & Bae, J.W. (2011). Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl Environ Microbiol,* 77(21), pp. 7663-8.

Kinross, J.M., Darzi, A.W. & Nicholson, J.K. (2011). Gut microbiome-host interactions in health and disease. *Genome Medicine,* 3(3), p. 14.

Kitajima, M., Haramoto, E., Phanuwan, C. & Katayama, H. (2011). Prevalence and genetic diversity of Aichi viruses in wastewater and river water in Japan. *Appl Environ Microbiol,* 77(6), pp. 2184-7.

Knight, R., Jansson, J., Field, D., Fierer, N., Desai, N., Fuhrman, J.A., Hugenholtz, P., van der Lelie, D., Meyer, F. & Stevens, R. (2012). Unlocking the potential of metagenomics through replicated experimental design. *Nature Biotechnology,* 30(6), pp. 513-520.

Krupovic, M., Prangishvili, D., Hendrix, R.W. & Bamford, D.H. (2011). Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. *Microbiology and Molecular Biology Reviews,* 75(4), pp. 610-635.

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K. & Hugenholtz, P. (2008). A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev,* 72(4), pp. 557-78, Table of Contents.

Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A., Takami, H., Morita, H., Sharma, V.K., Srivastava, T.P., Taylor, T.D., Noguchi, H., Mori, H., Ogura, Y., Ehrlich, D.S., Itoh, K., Takagi, T., Sakaki, Y., Hayashi, T. & Hattori, M. (2007). Comparative

metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res,* 14(4), pp. 169-81.

Lager, K.M., Ng, T.F., Bayles, D.O., Alt, D.P., Delwart, E.L. & Cheung, A.K. (2012). Diversity of viruses detected by deep sequencing in pigs from a common background. *J Vet Diagn Invest,* 24(6), pp. 1177-9.

Larsson, J. (2016). *Neonatal porcine diarrhoea.* (Acta Universitatis agriculturae Sueciae, 2016). Uppsala, Sweden: SLU Service/Repro, Uppsala 2016.

Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J.M., Kennedy, S., Leonard, P., Li, J., Burgdorf, K., Grarup, N., Jorgensen, T., Brandslund, I., Nielsen, H.B., Juncker, A.S., Bertalan, M., Levenez, F., Pons, N., Rasmussen, S., Sunagawa, S., Tap, J., Tims, S., Zoetendal, E.G., Brunak, S., Clement, K., Dore, J., Kleerebezem, M., Kristiansen, K., Renault, P., Sicheritz-Ponten, T., de Vos, W.M., Zucker, J.D., Raes, J., Hansen, T., Meta, H.I.T.c., Bork, P., Wang, J., Ehrlich, S.D. & Pedersen, O. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature,* 500(7464), pp. 541-6.

Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25(14), pp. 1754-60.

Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J.R., Prifti, E., Nielsen, T., Juncker, A.S., Manichanh, C., Chen, B., Zhang, W., Levenez, F., Wang, J., Xu, X., Xiao, L., Liang, S., Zhang, D., Zhang, Z., Chen, W., Zhao, H., Al-Aama, J.Y., Edris, S., Yang, H., Wang, J., Hansen, T., Nielsen, H.B., Brunak, S., Kristiansen, K., Guarner, F., Pedersen, O., Dore, J., Ehrlich, S.D., Meta, H.I.T.C., Bork, P., Wang, J. & Meta, H.I.T.C. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nature Biotechnology,* 32(8), pp. 834-41.

Li, L., Deng, X., Mee, E.T., Collot-Teixeira, S., Anderson, R., Schepelmann, S., Minor, P.D. & Delwart, E. (2015). Comparing viral metagenomics methods using a highly multiplexed human viral pathogens reagent. *Journal of Virological Methods,* 213, pp. 139-146.

Li, L.L. & Delwart, E. (2011). From orphan virus to pathogen: the path to the clinical lab. *Current Opinion in Virology,* 1(4), pp. 282-288.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B. & Fan, W. (2012). Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief Funct Genomics,* 11(1), pp. 25-37.

Lim, E.S., Zhou, Y., Zhao, G., Bauer, I.K., Droit, L., Ndao, I.M., Warner, B.B., Tarr, P.I., Wang, D. & Holtz, L.R. (2015). Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat Med,* 21(10), pp. 1228-34.

Lipkin, W.I. (2008). Pathogen discovery. *PLoS Pathog,* 4(4), p. e1000002.

Lipkin, W.I. (2013). The changing face of pathogen discovery and surveillance. *Nature Reviews Microbiology,* 11(2), pp. 133-41.

Lu, T.K. & Koeris, M.S. (2011). The next generation of bacteriophage therapy. *Curr Opin Microbiol,* 14(5), pp. 524-31.

Manavski, S.A. & Valle, G. (2008). CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment. *BMC Bioinformatics,* 9 Suppl 2, p. S10.

Marchesi, J.R. & Ravel, J. (2015). The vocabulary of microbiome research: a proposal. *Microbiome,* 3, p. 31.

Marine, R., McCarren, C., Vorrasane, V., Nasko, D., Crowgey, E., Polson, S.W. & Wommack, K.E. (2014). Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome. *Microbiome,* 2(1), p. 3.

Marine, R., Polson, S.W., Ravel, J., Hatfull, G., Russell, D., Sullivan, M., Syed, F., Dumas, M. & Wommack, K.E. (2011). Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. *Appl Environ Microbiol,* 77(22), pp. 8071-9.

Matamoros, S., Gras-Leguen, C., Le Vacon, F., Potel, G. & de La Cochetiere, M.F. (2013). Development of intestinal microbiota in infants and its impact on health. *Trends Microbiol,* 21(4), pp. 167-73.

Menzel, P., Ng, K.L. & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun,* 7, p. 11257.

Minot, S., Grunberg, S., Wu, G.D., Lewis, J.D. & Bushman, F.D. (2012). Hypervariable loci in the human gut virome. *Proc Natl Acad Sci U S A,* 109(10), pp. 3962-6.

Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S.A., Wu, G.D., Lewis, J.D. & Bushman, F.D. (2011). The human gut virome: inter-individual variation and dynamic response to diet. *Genome Research,* 21(10), pp. 1616-25.

Mokili, J.L., Rohwer, F. & Dutilh, B.E. (2012). Metagenomics and future perspectives in virus discovery. *Current Opinion in Virology,* 2(1), pp. 63-77.

Mueller, N.T., Bakacs, E., Combellick, J., Grigoryan, Z. & Dominguez-Bello, M.G. (2015). The infant microbiome development: mom matters. *Trends Mol Med,* 21(2), pp. 109-17.

Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res,* 40(20), p. e155.

Norman, J.M., Handley, S.A., Baldridge, M.T., Droit, L., Liu, C.Y., Keller, B.C., Kambal, A., Monaco, C.L., Zhao, G., Fleshner, P., Stappenbeck, T.S., McGovern, D.P., Keshavarzian, A., Mutlu, E.A., Sauk, J., Gevers, D., Xavier, R.J., Wang, D., Parkes, M. & Virgin, H.W. (2015). Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell,* 160(3), pp. 447-60.

Ogilvie, L.A. & Jones, B.V. (2015). The human gut virome: a multifaceted majority. *Front Microbiol,* 6, p. 918.

Ondov, B.D., Bergman, N.H. & Phillippy, A.M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics,* 12, p. 385.

Pang, X., Hua, X., Yang, Q., Ding, D., Che, C., Cui, L., Jia, W., Bucheli, P. & Zhao, L. (2007). Inter-species transplantation of gut microbiota from human to pigs. *Isme Journal,* 1(2), pp. 156-62.

Patel, R.K. & Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *Plos One,* 7(2), p. e30619.

Penders, J., Thijs, C., Vink, C., Stelma, F.F., Snijders, B., Kummeling, I., van den Brandt, P.A. & Stobberingh, E.E. (2006). Factors influencing the composition of the intestinal microbiota in early infancy. *Pediatrics,* 118(2), pp. 511-21.

Pennisi, E. (2016). Pocket DNA sequencers make real-time diagnostics a reality. *Science,* 351(6275), pp. 800-801.

Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Brief Bioinform,* 10(4), pp. 354-66.

Prangishvili, D., Forterre, P. & Garrett, R.A. (2006). Viruses of the Archaea: a unifying view. *Nature Reviews Microbiology,* 4(11), pp. 837-48.

Pruitt, K.D., Tatusova, T. & Maglott, D.R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res,* 33(Database issue), pp. D501-4.

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

Reyes, A., Haynes, M., Hanson, N., Angly, F.E., Heath, A.C., Rohwer, F. & Gordon, J.I. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature,* 466(7304), pp. 334-338.

Reyes, A., Semenkovich, N.P., Whiteson, K., Rohwer, F. & Gordon, J.I. (2012). Going viral: next-generation sequencing applied to phage populations in the human gut. *Nature Reviews Microbiology,* 10(9), pp. 607-17.

Rho, M., Tang, H. & Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res,* 38(20), p. e191.

Rodriguez-Valera, F., Martin-Cuadrado, A.-B., Rodriguez-Brito, B., Pašić, L., Thingstad, T.F., Rohwer, F. & Mira, A. (2009). Explaining microbial population genomics through phage predation. *Nature Reviews Microbiology,* 7(11), pp. 828-836.

Rodriguez, R.L. & Konstantinidis, K.T. (2014a). Estimating coverage in metagenomic data sets and why it matters. *Isme Journal,* 8(11), pp. 2349-51.

Rodriguez, R.L. & Konstantinidis, K.T. (2014b). Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics,* 30(5), pp. 629-35.

Rohwer, F. (2003). Global phage diversity. *Cell,* 113(2), p. 141.

Rohwer, F. & Thurber, R.V. (2009). Viruses manipulate the marine environment. *Nature,* 459(7244), pp. 207-12.

Rosseel, T., Ozhelvaci, O., Freimanis, G. & Van Borm, S. (2015). Evaluation of convenient pretreatment protocols for RNA virus metagenomics in serum and tissue samples. *Journal of Virological Methods,* 222, pp. 72-80.

Rosseel, T., Van Borm, S., Vandenbussche, F., Hoffmann, B., van den Berg, T., Beer, M. & Hoper, D. (2013). The origin of biased sequence depth in sequence-independent nucleic acid amplification and optimization for efficient massive parallel sequencing. *Plos One,* 8(9), p. e76144.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., Hoon, J., Simons, J.F., Marran, D., Myers, J.W.,

Davidson, J.F., Branting, A., Nobile, J.R., Puc, B.P., Light, D., Clark, T.A., Huber, M., Branciforte, J.T., Stoner, I.B., Cawley, S.E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J.A., Namsaraev, E., McKernan, K.J., Williams, A., Roth, G.T. & Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature,* 475(7356), pp. 348-52.

Sachsenroder, J., Twardziok, S.O., Scheuch, M. & Johne, R. (2014). The general composition of the faecal virome of pigs depends on age, but not on feeding with a probiotic bacterium. *Plos One,* 9(2), p. e88888.

Schaller, T., Appel, N., Koutsoudakis, G., Kallis, S., Lohmann, V., Pietschmann, T. & Bartenschlager, R. (2007). Analysis of hepatitis C virus superinfection exclusion by using novel fluorochrome gene-tagged viral genomes. *Journal of Virology,* 81(9), pp. 4591-603.

Schloss, P.D., Schubert, A.M., Zackular, J.P., Iverson, K.D., Young, V.B. & Petrosino, J.F. (2012). Stabilization of the murine gut microbiome following weaning. *Gut Microbes,* 3(4), pp. 383-93.

Schmieder, R. & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics,* 27(6), pp. 863-4.

Schwartz, S., Friedberg, I., Ivanov, I.V., Davidson, L.A., Goldsby, J.S., Dahl, D.B., Herman, D., Wang, M., Donovan, S.M. & Chapkin, R.S. (2012). A metagenomic study of diet-dependent interaction between gut microbiota and host in infants reveals differences in immune response. *Genome Biology,* 13(4), p. r32.

Sdiri-Loulizi, K., Hassine, M., Bour, J.B., Ambert-Balay, K., Mastouri, M., Aho, L.S., Gharbi-Khelifi, H., Aouni, Z., Sakly, N., Chouchane, S., Neji-Guediche, M., Pothier, P. & Aouni, M. (2010). Aichi virus IgG seroprevalence in Tunisia parallels genomic detection and clinical presentation in children with gastroenteritis. *Clin Vaccine Immunol,* 17(7), pp. 1111-6.

Shan, T., Li, L., Simmonds, P., Wang, C., Moeser, A. & Delwart, E. (2011). The fecal virome of pigs on a high-density farm. *Journal of Virology,* 85(22), pp. 11697-708.

Shankaranarayanan, P., Mendoza-Parra, M.A., van Gool, W., Trindade, L.M. & Gronemeyer, H. (2012). Single-tube linear DNA amplification for genome-wide studies using a few thousand cells. *Nat Protoc,* 7(2), pp. 328-38.

Shrestha, R.K., Lubinsky, B., Bansode, V.B., Moinz, M.B., McCormack, G.P. & Travers, S.A. (2014). QTrim: a novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform. *BMC Bioinformatics,* 15, p. 33.

Skewes-Cox, P., Sharpton, T.J., Pollard, K.S. & DeRisi, J.L. (2014). Profile hidden Markov models for the detection of viruses within metagenomic sequence data. *Plos One,* 9(8), p. e105067.

Sommer, M.O. & Dantas, G. (2011). Antibiotics and the resistant microbiome. *Curr Opin Microbiol,* 14(5), pp. 556-63.

Stanhope, S.A. (2010). Occupancy modeling, maximum contig size probabilities and designing metagenomics experiments. *Plos One,* 5(7), p. e11652.

Stelekati, E. & Wherry, E.J. (2012). Chronic bystander infections and immunity to unrelated antigens. *Cell Host Microbe,* 12(4), pp. 458-69.

Stern, A., Mick, E., Tirosh, I., Sagy, O. & Sorek, R. (2012). CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Research,* 22(10), pp. 1985-94.

Stulberg, E., Fravel, D., Proctor, L.M., Murray, D.M., LoTempio, J., Chrisey, L., Garland, J., Goodwin, K., Graber, J. & Harris, M.C. (2016). An assessment of US microbiome research. *Nature Microbiology,* 1, p. 15015.

Suttle, C.A. (2005). Viruses in the sea. *Nature,* 437(7057), pp. 356-61.

Suzuki, S., Ishida, T., Kurokawa, K. & Akiyama, Y. (2012). GHOSTM: a GPU-accelerated homology search tool for metagenomics. *Plos One,* 7(5), p. e36060.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol Ecol,* 21(8), pp. 2045-50.

Team, R.C. (2016). R: A Language and Environment for Statistical Computing. Vienna, Austria.

Thurber, R.V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. (2009). Laboratory procedures to generate viral metagenomes. *Nat Protoc,* 4(4), pp. 470-83.

Torsvik, V., Ovreas, L. & Thingstad, T.F. (2002). Prokaryotic diversity--magnitude, dynamics, and controlling factors. *Science,* 296(5570), pp. 1064-6.

Tremaroli, V. & Bäckhed, F. (2012). Functional interactions between the gut microbiota and host metabolism. *Nature,* 489(7415), pp. 242-249.

Tsai, Y.C., Conlan, S., Deming, C., Program, N.C.S., Segre, J.A., Kong, H.H., Korlach, J. & Oh, J. (2016). Resolving the Complexity of Human Skin Metagenomes Using Single-Molecule Sequencing. *Mbio,* 7(1), pp. e01948-15.

Tscherne, D.M., Evans, M.J., von Hahn, T., Jones, C.T., Stamataki, Z., McKeating, J.A., Lindenbach, B.D. & Rice, C.M. (2007). Superinfection exclusion in cells infected with hepatitis C virus. *Journal of Virology,* 81(8), pp. 3693-703.

Virgin, H.W. (2014). The virome in mammalian physiology and disease. *Cell,* 157(1), pp. 142-50.

Virgin, H.W., Wherry, E.J. & Ahmed, R. (2009). Redefining Chronic Viral Infection. *Cell,* 138(1), pp. 30-50.

Vouzis, P.D. & Sahinidis, N.V. (2011). GPU-BLAST: using graphics processors to accelerate protein sequence alignment. *Bioinformatics,* 27(2), pp. 182-8.

Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W.H.A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M. & Venables, B. (2016). gplots: Various R Programming Tools for Plotting Data.

Weinbauer, M.G. (2004). Ecology of prokaryotic viruses. *FEMS Microbiol Rev,* 28(2), pp. 127-81.

Wendl, M.C., Kota, K., Weinstock, G.M. & Mitreva, M. (2013). Coverage theories for metagenomic DNA sequencing based on a generalization of Stevens' theorem. *J Math Biol,* 67(5), pp. 1141-61.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C.L., Irzyk, G.P., Lupski, J.R., Chinault, C., Song, X.Z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D.M., Margulies, M., Weinstock, G.M., Gibbs, R.A. & Rothberg, J.M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature,* 452(7189), pp. 872-6.

Whipps, JM and RC Cooke. "Mycoparasitism And Plant Disease Control". Fungi In Biological Control Systems. M.N. Burge. 1st ed. Manchester: Manchester University Press, 1988. 176. Print.

Wilhelm, S.W. & Suttle, C.A. (1999). Viruses and nutrient cycles in the sea viruses play critical roles in the structure and function of aquatic food webs. *Bioscience,* 49(10), pp. 781-788.

Wommack, K.E., Bhavsar, J. & Ravel, J. (2008). Metagenomics: read length matters. *Appl Environ Microbiol,* 74(5), pp. 1453-63.

Wood, D.E. & Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology,* 15(3), p. R46.

Yang, J.Y., Kim, M.S., Kim, E., Cheon, J.H., Lee, Y.S., Kim, Y., Lee, S.H., Seo, S.U., Shin, S.H., Choi, S.S., Kim, B., Chang, S.Y., Ko, H.J., Bae, J.W. & Kweon, M.N. (2016). Enteric Viruses Ameliorate Gut Inflammation via Toll-like Receptor 3 and Toll-like Receptor 7-Mediated Interferon-beta Production. *Immunity,* 44(4), pp. 889-900.

Yilmaz, S., Allgaier, M. & Hugenholtz, P. (2010). Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nature Methods,* 7(12), pp. 943-4.

Acknowledgments

This work has been supported by grants from the framework of the EU project AniBioThreat (grant agreement: Home/2009/ISEC/AG/191) with financial support from the Prevention of and Fight against Crime Programme of the European Union, European Commission—Directorate General Home Affairs. This publication reflects views only of the authors, and the European Commission cannot be held responsible for any use that may be made of the information contained therein.
Financial support was obtained from The Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning, Formas (221-2012-586) awarded to Professor Mikael Berg.
The author would like to express special thanks to the EU FP7 Project reference: 612583: "Developing an European American NGS Network" (DEANN)
The author would like to thank the EU COST Action BM1006 Next Generation Sequencing Data Analysis Network SeqAhead for financial support to participate on workshops, training and discussions
This work was partially supported by the EU FP7 AllBio project [KBBE.2011.3.6-02]
Funding was also generously provided by HELGE AX:SON JOHNSONS foundation, used for analysis in paper II, III and manuscript IV.
The authors would like to acknowledge support of the National Genomics Infrastructure (NGI) / Uppsala Genome Center and UPPMAX for providing assistance in massive parallel sequencing and computational infrastructure. Work performed at NGI / Uppsala Genome Center has been funded by RFI/VR and Science for Life Laboratory, Sweden. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Science is, contrary to popular belief, a team sport and more people than I can mention here has provided and cared for my education throughout this work. There are however several that have done outstanding contributions to both the work herein as well as to my personal development.
I want to express my deep gratitude to my supervisor group, Mikael Berg, Sándor Belák, Mikael Leijon, Erik Bongcam-Rudloff, Anne-Lie Blomström, Fredrik Granberg and Maja Malmberg. Without your caring support this would never have been possible.

From the first meeting with **Mikael B** that snowy day in December 2010 until these last few days of writing the thesis our relationship has been one of trust and curiosity. I'm forever in your debt, you have believed in me and thereby nurtured a curios and competent researcher. You have guided me with those careful and considerate remarks from day one, through crazy times in the AniBio project, in travel all over Europe and finally in writing my thesis. You made this possible. **Sándor**, you have not only provided for my scientific education but you have also provided insight and knowledge far beyond what was expected. I have so enjoyed our company together and I hope that our discussions on scientific as daily matters will continue for long to come. **Mikael Leijon**, you have always had time to listen to me. From those first crazy ideas about the use of sequencers to these last few rounds sending the draft you have provided knowledge and commentary for me to grow from. **Erik**, from the whimsical first designs for my bioinformatics research, through workshops all over Europe and contacts all over the world you have been an extraordinary source of information. I hope that we can continue this collaboration for many more years! **Anne-Lie,** for being the first PhD student you supervised I must say the experience has been great. You always provide sensible suggestions at the same time as you encourage novel ideas and creativity. I hope you continue to take on PhD students for many years to come, it would be a boon to the scientific community. **Fredrik**, my friend, you have always been one of a kind, able to not only be extraordinary visionary but also back that up with rigorous work and skill. I have learned so much from you and I hope we can continue to do research and discuss kids and SciFi for many years to come. **Maja** was maybe not my first supervisor but she is one that have left a huge mark on my scientific and personal development. Throughout crazy discussions about the evolutionary role of viruses, to the possible use of CRISPR technology, to the discussions about life and feminism you have been a constant source of enlightenment to me, if I could wish for one thing it would be for you to supervise more PhD students. You have all done extraordinary well.

Jenny Larsson, my friend and co-worker for 5 years of this project. Who could have guessed what the one course would lead up to? What an amassing journey we have had. Few have the possibility to do one PhD project, even fever is crazy enough to try and do several at once. Strangely enough, crazy meet crazy and beautiful science was the result. All those late nights and failed experiments, all that laughter and shared pain and the sweet feeling of success! You are an amazing researcher and it has been an honour to work next to you, something I hope we will do more in the future. You are my friend for life.

Also, borrowing your superior for endless discussions about science as well as life has been wonderful, thank you **Magdalena**.

My, former, fellow PhD student **Martin Norling**, the probably best bioinformatican in the world (or so his father in law claim anyway). Being a mythical being with powers to fix any server in the world cannot be easy but you proved to be more than the myth. Thank you for your support and friendship these last years and may our friendship last long.

My friend, colleague and co-author **Juliette Hayer**. From our first rather confusing meeting in the hallway of animal genetics until this summer I have shared so many great moments with you. Thanks for all the support throughout articles and science. My friend, student and co-worker **Hadrien Gourlé**, you have the future before you! I have been amazed at your skill, both in bioinformatics as well as science, since your first period as a student at our group. You are an amazingly talented individual. My friend and co-worker **Tomas Klingström**, I remember that day you started and the discussion we had then. You have so much creativity in you, good luck with your thesis. And finally **Robert**, thanks for always keeping up the good work and being the most illustrious man on the face of the earth! Past and present personnel at the **SLU global bioinformatics centre**. There are so many, and so few at the same time. The amount of scientific publications coming out of this group never cease to amaze me. Keep on the good work!

And the personnel at HGEN. I have always felt at home at my second department, you are a great gang! Special thanks to HGEN lab who constantly handle problems with equipment, my mental state and my research, thanks **Sofia**, **Tomas**, **Hans-Henrik**, **Siw**, **Louise** and **Charlotte**!

All the SVA and AniBio personnel and co-workers who provided such valuable feedback and a fun an encouraging environment during my first years as a PhD student. **Rickard**, you truly are one of the best project leads ever! Thanks for all great discussions on science and biosecurity. **Camilla B**, I could not have wished for better support! Thanks for all the patience and help throughout the AniBio project. The PhD students, **Hanna**, **Joakim** and **Trine** who together with me formed our own little unofficial research school in bio-threat and biosecurity. Thanks for all the discussions, for co-authorship and an awesome stay in Lund! **Snus** and **Ullis**, who kept me in line (even though all that paperwork never got done). **Jean-Francois Valarcher** who have provided such interesting discussion as well as being my external reviewer at my half time. **Karin Ullman**, **Alia**, **Misha**, **Li Hong** and **Siamak** who have provided great discussions about science. And Finally **Helena B,** for all those great discussions about life and science! Thanks for your care and compassion about all things!

Past and present PhD students at the section of virology: **Jonas**, **Munir**, **Georgi**, **Sandra** and **Elin** for support when needed, for collaborations and for companionship throughout my studies. **Emeli**, since that sunny day (with my ugly scots plaid shorts) I knew you were special! Turns out I was right! Hang in there, its soon your turn and you will make a great doctor Torsson when you are done! **Harindranath Cholleti**, for funny and enlightening discussions about marriage, life and science! Good luck with you studies, you will make a great doctor soon! **Johanna L,** you have been such a support and a good friend throughout all this! Thank you so much for everything!
And to the personnel at BVF, thanks for all these years of fika, support and meetings! You have truly been a family away from family!

And all the PhD student of SLU/VH. You have been with me the whole way and it has been an incredible journey. I'm so proud to be part of such a great team of young and brilliant scientists. **Marlene A**, **Camilla W and Karin O** for constant support, provider of babies when I miss my own, discussion partners, party gang, fellow scientists. It has been my pleasure to get to know you all, I'm in awe of your skill and dedication and I hope we will have many more years together! **Viktor, Maria, Pernille,** for support and encouragement during this last hectic year! You made it (or in Viktor's case you are soon there)! And **Monika**, for being an awesome student and a true scientist!
Kristina, **Henrik**, **Stefan**, **Karin**, **Maria**, **Jan**, and **Sara** for all those FUN meetings. It was truly an amazing experience.

Special thanks to **Malin** and **Malin**, who always (and I mean always) provide good company for fika at any time during the working hours.

Olga Vinnere Pettersson and **Inger Jonasson** and the staff at NGI Uppsala for always taking on my crazy projects. Great science requires great data and you have provided throughout my whole study period.

Finally, my children, my wife and my family, you have been my sanity check, every day and provided me with laughter, love and encouragement. **Lucas**, **Nova** and **Stella**, you are an amazing gift that life has bestowed upon me. **Anna**, for always and forever, will you be in my heart. Thanks for all the laughter, all the comfort and all the love. To **Johanna** and **Fredrik (and the boys),** who always provide a second home for me and my family when needed. Thanks for all the great discussions throughout the years. **Kalle** and **Maria**, for being there, always and at any hour of the day for so many years now. Jonathan who is one of the most caring persons in the world. And to **Josephine**, I know I talk a lot but I actually did work hard for this! :D To my **parents** and **my sisters**, who nurtured their Professor Calculus throughout his childhood and always been there for encouragement and support. And to grandmother **Kerstin**, for all the care you have given throughout the years. And my loving family in Uppsala, **Tina**, **Thomas**, **Caroline** and **Charlotte** who have been such great support for me and my family these past five years. And finally **Kristian** and **Melker,** you might both be small now, but great things can come in small packages. I look forward to watching your life's unfold!