Genomic investigation of the faecal RNA virome in children from Oukasie clinic, North West Province, South Africa

By

Milton Tshidiso Mogotsi

Submitted in fulfilment with the requirements for the degree

Magister Scientiae

In the

Department of Microbial, Biochemical and Food Biotechnology
Faculty of Natural and Agricultural Sciences
University of the Free State
Bloemfontein
South Africa

And

Division of Virology
Faculty of Health Sciences
University of the Free State
Bloemfontein
South Africa
31 January 2019

Supervisor: Dr. Martin M. Nyaga

Co-supervisor: Prof. Hester G. O'Neill

UNIVERSITY OF THE **FREE STATE**
UNIVERSITEIT VAN DIE **VRYSTAAT**
YUNIVESITHI YA **FREISTATA**

This dissertation is dedicated to my family, especially my grandmother, Miss Malekgetho Ruth Mogotsi.

Declaration

"I, Milton Tshidiso Mogotsi, declare that the dissertation hereby submitted for the qualification *Magister Scientiae* (Microbiology) at the University of the Free State is my own independent work and has not been previously submitted by me for a qualification at another university/faculty. Furthermore, I concede copyright of the dissertation in favour of the University of the Free State."

Signature: 

Milton Tshidiso Mogotsi

Acknowledgements

I would like to extend my heartfelt gratitude and thanks to:

- **God**, for giving me the strength and inspiration to complete this M.Sc.
- **Dr Martin M. Nyaga**, my supervisor, for his guidance, constructive criticism and allowing me to grow as a scientist and for all the lessons that extended beyond the laboratory.
- **Prof Hester G. O'Neill** for her invaluable assistance, insightful suggestions, guidance and support throughout this study as my co-supervisor.
- **Medical Research Council/Diarrhoeal Pathogens Research Unit, Sefako Makgatho Health Sciences University** for providing me with stool samples to do this project and for offering me training on rotavirus detection by ELISA and PAGE, as well as rotavirus genotyping.
- **Mr Armand Bester** for his assistance with metagenomic data analysis.
- **Dr Benjamin Kumwenda (University of Malawi)** for his expert advice and guidance with bioinformatics.
- **Next Generation Sequencing Unit** colleagues for their continued support and assistance.
- **Molecular Virology and Clinical Biochemistry** lab members for their inputs, support and encouragement.
- The staff and students of the **Department of Microbial, Biochemical and Food Biotechnology** for their support and guidance.
- **My family** and **friends** for their undying love, support and encouragement throughout my studies and never doubting the decisions I made.
- The financial assistance of the **National Research Foundation (NRF)** towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF. (Grant no. **SFH160720180180**).
- The financial assistance of the **Poliomyelitis Research Foundation (PRF)** towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the PRF. (Grant no**. 17/62**).
- The financial assistance of the **South African Medical Research Council (SAMRC)** through the **Self-Initiated Research Grant (SIR)** awarded to Dr. Martin Nyaga for this virome project.
- The **University of the Free State Postgraduate School** for financial assistance.

Contents

List of figures

Chapter 2

Chapter 3

Chapter 4

List of tables

Chapter 2

Chapter 3

Chapter 4

List of abbreviations

(+)ssRNA:	Positive-sense ssRNA
(-)ssRNA:	Negative-sense ssRNA
βME:	Beta mercaptoethanol
AGE:	Acute gastroenteritis
ATM:	Amplicon tagment mix
BLAST:	Basic local alignment search tool
bp:	basepair
$CaCl_2$:	Calcium chloride
CD:	Crohn's Disease
cDNA:	Complementary deoxyribonucleic acid
CRISPR:	Clustered regularly interspaced palindromic repeats
DNA:	Deoxyribonucleic acid
DNase:	Deoxyribonuclease
ds:	Double-stranded
EB:	Elution buffer
EDTA:	Ethylenediaminetetra acetic acid
EM:	Electron Microscope
gb:	Gigabase
gDNA:	Genomic deoxyribonucleic acid
GIT:	Gastrointestinal tract
HS:	High sensitivity
HSREC:	Health Sciences Research Ethics Committee
HTS:	High-throughput sequencing
IBD:	Inflammatory Bowel Disease
IRS:	Inhibitor removal solution
K/mm^2:	Thousand per square millimetre
kb:	Kilobase

$MgCl_2$:	Magnesium chloride
MRC-DPRU:	Medical Research Council-Diarrhoeal Pathogens Research Unit
NaOH:	Sodium hydroxide
NCBI:	National Center for Biotechnology Information
NetoVIR:	Novel enrich techniques of viromes
NGS:	Next generation sequencing
NPM:	Nextera PCR mastermix
ORF:	Open reading frame
PAMPs:	Pathogen-associated molecular patterns
PBS:	Phosphate buffered saline
PCR:	Polymerase chain reaction
QC:	Quality control
QIIME:	Quantitative insights into microbial ecology, a bioinformatics software
qPCR:	Quantitative polymerase chain reaction
RefSeq:	Reference sequence database
RNA:	Ribonucleic acid
RNase:	Ribonuclease
rpm:	Rotations per minute
rRNA:	Ribosomal ribonucleic acid
RSB:	Resuspension buffer
RT-PCR:	Reverse transcription PCR
SDA:	Strand displacement amplification
SIA:	Sequence independent amplification
SMU:	Sefako Makgatho Health Sciences University
ss:	Single-stranded
TD buffer:	Tagment DNA buffer
UniProt:	Universal protein database
WHO:	World Health Organisation
WTA:	Whole transcriptome amplification

Conference presentation(s)

Milton T. Mogotsi, Peter N. Mwangi, Philip A. Bester, Hester G. O'Neill, Martin M. Nyaga. Genomic investigation of the Faecal RNA virome in children from Oukasie clinic, North West Province, South Africa.

13th International double-stranded RNA virus symposium, Houffalize, Belgium, 24-28 September 2018. (Poster presentation).

Abstract

The advancements in high-throughput sequencing (HTS) and improvements in bioinformatics tools have enabled partial description of the human gut microbiome and continue to receive increasing attention. Novel enteric eukaryotic viruses have been associated with severe childhood diarrhoea in low-income areas worldwide. New virome data has shown that childhood diarrhoea contains higher abundance of viruses most of which were previously non-pathogenic such as those within the families *Adenoviridae*, *Picornaviridae* and *Reoviridae*. Nevertheless, a huge knowledge gap exists about the composition and diversity of the viruses colonizing the gastrointestinal tract of asymptomatic humans, which may be of clinical importance, especially in low-income countries. A major drawback for this poor characterization of the human gut virome has been attributed to lack of optimised methods to conduct such studies. However, an effective virome enrichment method called NetoVIR, developed by Conceiação-Neto and co-workers in 2015 for preparation of viral metagenomics samples has bridged the gap.

In this study, viral metagenomics was employed to characterize the gut RNA virome of children under one-year old from the Oukasie clinic in the North West Province of South Africa. Faecal samples (n=12) were collected from four healthy infants at three time intervals (on average 7, 13 and 25 weeks old), to enable comparison of the changes in virome composition from baseline throughout the collection period. The samples were enriched for viral particles, followed by RNA extraction and RT-PCR. Library construction was done using a Nextera XT library preparation kit. The prepared libraries were sequenced on a MiSeq instrument to generate 251 bp paired-end reads. Using an in-house analysis pipeline, quality control of the generated reads was performed with FASTQc and Prinseq programmes. Quality-filtered reads were *de novo* assembled using metaSPAdes. Contigs were analysed by BLASTX searched against the NCBI database using DIAMOND, by aligning protein sequences against the NCBI protein database. Lastly, contigs that mapped to viruses were extracted for further statistical analysis.

Numerous human enteric viruses were detected in all faecal samples. *Reoviridae* and in particular rotaviruses were detected in all 12 samples (100 %). However, majority of the viral contigs belonged to *Picornaviridae* family including viruses such parechoviruses, echoviruses, coxsackieviruses, enteroviruses and polioviruses, making it the most abundant. *Astroviridae* such as astroviruses and *Caliciviridae* such as noroviruses were detected at low abundance. Additionally, few sequences matched to plant viruses (pepper mild mottle virus), which was likely introduced through diet. Several viruses of animal origin were also present in gut of two of the participants. This study has proved that viral metagenomics can be an appropriate method in characterization of the human virome, providing insight into viral community structure and diversity of human enteric viruses. Although the faecal

samples used in this study were negative by rotavirus screening using ELISA, it is interesting to observe that such high abundance of rotavirus sequences were still detected in the gut of asymptomatic individuals. The detection of polioviruses in one of the participants is a matter of public health concern since polioviruses have been eradicated by vaccination from many countries with only a handful still reporting sporadic cases. However, further analysis revealed that these were oral poliovirus vaccine sequences.

It is evident that the infant's gastrointestinal tract is colonized by different viral populations, irrespective of their health status. Despite the small sample size, this metagenomic study has provided some insight into the composition and diversity of viruses present in the gut of children. Lastly, the obtained data could be useful in the development of prevention strategies, as it provides information on virus species circulating in particular geographic areas, and to some extent can also suggest potential zoonotic transmissions.

Keywords: Virome, gastrointestinal tract, polioviruses, metagenomics, enteric viruses, *Reoviridae,* next generation sequencing, RNA, NetoVIR.

Chapter 1: Introduction to the study

1.1. Introduction

The human intestinal microbiome is important for human health, behaviour and disease (O'Hara and Shanahan, 2006; Bäumler and Sperandio, 2016). It is essential for human development by improving functions of the immune system and for enhancing metabolism and the digestion of food (Hooper *et al*., 2001; Macpherson and Harris, 2004; Bäumler and Sperandio, 2016; Kho and Lal, 2018). This human gut microbiome consists of complex populations of diverse and dynamic microorganisms, including bacteria, archaea, protists, fungi, and viruses (Virgin *et al*., 2014; Lim *et al*., 2015). Furthermore, numerous commensal bacteria within the microbiome offer additional benefits to the host such as conferring protection against pathogenic microorganisms (Bäumler and Sperandio, 2016). Nevertheless, microbiome researchers have established that every individual's microbiome has the potential to induce their susceptibility to infections that can lead to chronic enteric diseases (Langdon *et al*., 2016). In fact, reports have shown that a wide range of diseases including diabetes, inflammatory bowel disease and rheumatoid arthritis and occur due to disturbances within the microbiome (Littman and Pamer, 2011; Cho and Blaser, 2012; Qin *et al*., 2014).

Enteric microbial colonization occurs at birth with exposure to microorganisms from the immediate environment, and establishing a diverse enteric microbiota in early stages is imperative to prevent diseases later on. It is well-known that bacterial species increase in diversity during the first years of life due to several factors, namely, birth mode, diet, antibiotic usage, genetics, geographic area as well as lifestyle (Palmer *et al*., 2007; Tamburini *et al*., 2016; Milani *et al*., 2017). However, the viral populations within the infant gut (gut virome) remains less studied during this developmental stage (McCann *et al*., 2018). Despite an existing knowledge gap, the human gut virome is beginning to be understood and this is primarily due to new developments in next generation sequencing (NGS) (Haynes and Rohwer, 2011; Minot *et al*., 2013). Particularly, studies have reported alterations of the gut virome in both healthy and diseased states. Nevertheless, much remains unknown about the virome composition of healthy individuals, with a significant fraction of the sequences from human virome studies representing unknown viruses that are probably not present in current databases (Krishnamurthy and Wang, 2017).

The human virome is essentially a collection of all the viruses that are found in or on human beings. Continuously being updated, the human virome consists of eukaryotic viruses, prokaryotic viruses and endogenous viral elements integrated in the human genome (Virgin, 2014). The NGS-based metagenomic research has highlighted that the human gut virome plays a crucial role in the intestinal immunity and homeostasis (Focà *et al*., 2015). Although the gut virome has significant effects on human health, both in healthy and immunocompromised subjects, causing illnesses such as acute gastroenteritis (AGE) (Clark and McKendrick, 2004; Glass *et al*., 2009;

Eckardt and Baumgart, 2011; Kapusinszky *et al*., 2012), the composition of the human enteric virome is still poorly understood (Focà *et al*., 2015).

Acute gastroenteritis (AGE) is responsible for childhood illnesses and deaths across the world (Liu *et al*., 2016). Despite improvements in hygiene and prevention strategies, which significantly reduced the mortality rate due to diarrhoea from 15 % to 9 % between 2008 and 2015 among children below the age of five years, infectious diarrhoea is still a serious public health issue all over the world (Black *et al*., 2010; Liu *et al*., 2016). Globally, most of these diarrhoeal diseases in children are mainly due to viral infections. Interestingly, it has been established that viruses are prevalent in the gastrointestinal tract even in asymptomatic cases (Focà *et al*., 2015). Moreover, findings from previous studies suggested that most enteric infections by viral pathogens are observed mostly early in life rather than in adulthood. This is said to be occurring due to changes in the virome based on factors such as age of an individual and the surrounding environment (Breitbart *et al*., 2008; Reyes *et al*., 2010).

Group A rotaviruses within the *Reoviridae* family are the major aetiological agents associated with severe diarrhoeal disease in children below the age of 5 years, in both resource-poor and industrialized countries (Tate *et al*., 2012; Tate *et al*., 2016; Troeger *et al*., 2018). Despite a considerable drop following the introduction of rotavirus vaccines over a decade ago, infant hospitalizations due to viral diarrhoea continued to be reported (Spina *et al*., 2015; Kim *et al*., 2017; Thongprachum *et al*., 2017; Vizzi *et al*., 2017). Other known viral agents that have been involved in cases of childhood diarrhoea include members of the family *Astroviridae* such as human astroviruses (Sdiri-Loulizi *et al*., 2008; De Benedictis *et al*., 2011; Jiang *et al*., 2013), and human caliciviruses from the family *Caliciviridae* (Glass *et al*., 2001; Simpson *et al*., 2003; Kim *et al*., 2017). In addition, previous studies have revealed the presence of non-human viruses in the stool samples of children, suggesting potential interspecies transmission (Li *et al*., 2010; Li *et al*., 2011; Phan *et al*., 2012). Among the RNA viruses found in the gut, a prevalence of plant viruses has been demonstrated, presumably introduced through diet (Sdiri-Loulizi *et al*., 2008; Valentini *et al*., 2013).

Application of metagenomic sequencing has become very useful in virome studies, as viruses lack a universal marker such as the conserved bacterial 16S ribosomal ribonucleic acid (rRNA) gene. Despite the capacity of NGS-based metagenomics to analyse all microbial genomes, the relatively bigger genome size of bacteria tends to complicate detailed analysis of the virome. Furthermore, these approaches tend to overlook viral RNA genomes present in the microbiome despite the fact that RNA viruses are implicated in most cases of gastroenteritis (Zhang *et al*., 2006; Breitbart *et al*., 2008). The complexity of human virome analysis is further intensified by the dynamic interactions of bacteriophages present in abundance with their human hosts.

1.2. Problem statement

Despite increasing research interest and new information from viral metagenomics studies, limited knowledge exists about the total viral communities from human stool samples. Moreover, children and infants are at a greater risk of enteric illnesses due to viral infections than adults for a number of reasons. Mainly the immune system, which controls the infection processes, is at a developmental level in children. This difference can lead to more severe infections than in adults who have a fully developed immune system. Furthermore, much less is understood about factors that lead to the increase in prevalence in virus related gastroenteritis in both children and adults. Although these microorganisms play essential roles in metabolism, immunity and absorption of nutrients by host species (Hooper and Gordon, 2001), their composition and abundance change in different stages of life depending on several factors such as diet and environment (Reyes *et al*., 2012; Minot *et al*., 2013). In most parts of Africa, the complete composition of viruses in the gut (human gut virome) has only been described partially and remains mostly unknown, with the RNA virome being largely unexplored (Virgin *et al*., 2014; Rascovan *et al*., 2016). Furthermore, there are possible medical and financial implications that potentially virulent novel viruses may pose to human health. For instance, the virome and analysis of its conformation is of great concern because it distinguishes viruses which can induce clinical diarrhoeal disease, sub-clinical growth impedance and also increase knowledge of viruses that are part of the normal flora. However, in South Africa, little is known regarding investigations characterizing the human total gut virome and the roles these viruses play in health and disease. The current study was therefore undertaken to explore and characterize the faecal RNA viruses in children from South Africa, which could worsen the severity of diarrhoeal disease and/or contribute to underlying disease agents that may alter normal host immunity, thereby increasing cases of gastroenteritis and/or diarrhoea-related hospitalizations and deaths in the country.

1.3. Significance of study

This study seeks to determine and analyse the faecal RNA viruses present in children that form the normal flora and those with potential to intensify and lead to more serious diarrhoeal diseases. The data obtained from this study would give insight into the composition and the diversity of the gut RNA virome in children. Information from this study can guide research in the development of prevention strategies of viral infections leading to diarrhoea, thus minimizing the high mortalities and morbidities in children, not only in South Africa but also in other parts of the world. The use of NGS has led to virus discovery and the study of the entire virome composition of various stool samples. Utilizing NGS techniques and existing bioinformatics tools to perform this study provides a bigger picture of the composition and diversity of viruses colonizing the gut of children by assisting in classification into families, genus and species. In addition, reference sequences will be submitted to the public

domain database and will provide insights for future virome studies. The data to be obtained from this study could be vital in providing a description of the viruses identified and should also represent a baseline for future studies investigating viral populations in healthy infants in South Africa.

1.4. Research aim

The aim of this study was to apply a viral metagenomics approach to determine and characterize the total RNA viruses colonizing the gastrointestinal tract (enteric RNA virome) of apparently healthy children under one-year old in South Africa using an IIlumina MiSeq platform.

1.5. Research Objectives

In achieving the aforementioned aim, this study has been divided into several specific objectives:

i. To enrich and amplify the transcriptome of RNA viruses obtained from four participants and sequence on Illumina MiSeq platform.
ii. To assemble the data of enteric RNA viruses identified using MetaSpades and characterize the genomes using DIAMOND software.
iii. To determine the relative distribution of obtained gut RNA virome and compare to total microbial populations sequenced after enrichment.
iv. To evaluate the changes in virome composition per study participant over the three collection timeframes.

1.6. Dissertation organization

The dissertation consists of five chapters as outlined below. Each of the experimental chapter consists of introduction, methodology, results, discussions, and a list of references.

Chapter 1: Introduction

This chapter provides a general introduction of the project based on the research proposal.

Chapter 2: Literature review

This chapter is based on detailed review of literature on virome studies. This includes current theoretical knowledge and methodological contributions in human virome research.

Chapter 3: Viral enrichments, RNA Isolation and whole transcriptome amplification

This chapter describes the application of the recently developed NetoVIR protocol for the enrichment of viruses from faecal samples. Furthermore, amplification of viral transcriptome from enriched samples is also described in this chapter. The results discussed in this chapter highlight the applicability of this protocol in an attempt to minimize and eliminate the non-viral genomes.

Chapter 4: Human gut virome analysis and characterization by deep metagenomics sequencing

This chapter describes the genomic characterization of enteric virome in South African children, determination of prevalence and viral distribution of different microbial populations it describes the bioinformatics tools that were utilized to analyse the genomic data.

Chapter 5: General conclusions

This is the last chapter of the dissertation which includes general discussion and conclusions as well as future perspectives.

1.7. References

Acevedo, A and Andino, R. (2014). Library preparation for highly accurate population sequencing of RNA viruses**.** *Nature Protocols* **9(7)**: 1760-1769.**Bäumler, A. J. and Sperandio, V. (2016).** Interactions between the microbiota and pathogenic bacteria in the gut. *Nature* **535(7610)**: 85-93.

Bäumler, A. J. and Sperandio, V. (2016). Interactions between the microbiota and pathogenic bacteria in the gut**.** *Nature* **535(7610):** 85-93.

Black, R. E., Cousens, S., Johnson, H. L., Lawn, J. E., Rudan, I., Bassani, D. G., Jha, P., Campbell, H., Walker, C.F., Cibulskis, R. *et al*. (2010). Child Health Epidemiology Reference Group of WHO and UNICEF. Global, regional, and national causes of child mortality in 2008: a systematic analysis. *Lancet* **375**: 1969-1987.

Breitbart, M., Haynes, M., Kelley S., Angly, F., Edwards, R. A., Felts, B., Mahaffy, J. M., Mueller, J., Nulton, J., Rayhawk, S. *et al*. (2008). Viral diversity and dynamics in an infant gut. *Res Microbiol* **159(5)**: 367-373.

Cho, I. and Blaser, M. J. (2012). The human microbiome: at the interface of health and disease. *Nat Rev Genet* **13**: 260-270.

Clark, B. and McKendrick, M. (2004). A review of viral gastroenteritis. *Curr Opin Infect Dis* **17(5)**: 461-469.

De Benedictis, P., Schultz-Cherry, S., Burnham, A. and Cattoli, G. (2011). Astrovirus infections in humans and animals - molecular biology, genetic diversity, and interspecies transmissions. *Infect Genet Evol* **11**: 1529-1544.

Eckardt, A. J. and Baumgart, D. C. (2011). Viral gastroenteritis in adults. *Recent Pat Antiinfect Drug Discov* **6(1)**: 54-63.

Focà, A., Liberto, M. C., Quirino, A., Marascio, N., Zicca, E. and Pavia, G. (2015). Gut Inflammation and Immunity: What Is the Role of the Human Gut Virome? *Mediators Inflamm* **2015**: 326032.

Glass, R. I., Parashar, U. D. and Estes, M. K. (2009). Norovirus gastroenteritis. *N Engl J Med* **361(18)**: 1726-1785.

Glass, R. I., Bresee, J., Jiang, B., Gentsch, J., Ando, T., Fankhauser, R., Noel, J., Parashar, U., Rosen, B. and Monroe S. S. (2001). Gastroenteritis viruses. In *Novartis Foundation Symposium*, pp. 5-19. New York: Wiley.

Haynes, M. and Rohwer, F. (2011). The human virome. In *Metagenomics of the Human Body*, pp. 63-78. Edited by Nelson, K. E. New York: Springer:

Holtz, L., Finkbeiner, S., Zhao, G., Kirkwood, C., Girones, R., Pipas, J. and Wang, D. (2009). Klassevirus 1, a previously undescribed member of the family Picornaviridae, is globally widespread. *Virol J* **6**: 86.

Hooper, L. V., Wong, M. H., Thelin, A., Hansson, L., Falk, P. G., Gordon, G. I. (2001). Molecular Analysis of Commensal Host-Microbial Relationships in the Intestine. *Science* **291**: 881-884.

Hooper, L. V. and Gordon, J. I. (2001). Commensal host-bacterial relationships in the gut**.** *Science* **292**: 1115-1118.

Jiang, H., Holtz, L.R., Bauer, I., Franz, C. J., Zhao, G., Bodhidatta, L., Shrestha, S. K., Kang, G. and Wang, D. (2013). Comparison of novel MLB-clade, VA-clade and classic human astroviruses highlights constrained evolution of the classic human astrovirus nonstructural genes. *Virology* **436**: 8-14.

Kapusinszky, B., Minor, P. and Delwart, E. (2012). Nearly constant shedding of diverse enteric viruses by two healthy infants. *J Clinl Microbiol* **50(11)**: 3427-3434.

Kho, Z. Y. and Lal, S. K. (2018). The Human Gut Microbiome – A Potential Controller of Wellness and Disease. *Front Microbiol* **9**: 1835.

Kim, A., Chang, J. Y., Shin, S., Yi, H., Moon, J. S., Ko, J. S. and Oh, S. (2017). Epidemiology and factors related to clinical severity of acute gastroenteritis in hospitalized children after the introduction of rotavirus vaccination. *J Korean Med Sci* **32**: 465-474.

Koh, H., Baek, S. Y., Shin, J. I., Chung, K. S. and Jee, Y. M. (2008). Coinfection of viral agents in Korean children with acute watery diarrhea. *J Korean Med Sci* **23**: 937-940.

Krishnamurthy, S. R. and Wang, D. (2017). Origins and challenges of viral dark matter. *Virus Res* **239**: 136-142.

Langdon, A., Crook, N. and Dantas, G. (2016). The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation. *Genome Med* **8**: 39

Li, L., Shan, T., Soji, O. B., Alam, M. M., Kunz, T. H., Zaidi, S. Z. and Delwart, E. (2011). Possible cross-species transmission of circoviruses and cycloviruses among farm animals. *J Gen Virol* **92**: 768-772.

Li, L., Kapoor, A., Slikas, B., Bamidele, O. S., Wang, C., Shaukat, S., Masroor, M. A., Wilson, M. L., Ndjango, J. B., Peeters, M. *et al.* (2010). Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. *J Virol* **84**: 1674-1682.

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)**: 1228-1234.

Littman, D. R. and Pamer, E. G. (2011). Role of the commensal microbiota in normal and pathogenic host immune responses. *Cell Host Microbe* **10(4)**: 311-323.

Liu, L., Oza, S., Hogan, D., Chu, Y., Perin, J., Zhu, J., Lawn, J. E., Cousens, S., Mathers, C. and Black, R. E. (2016). Global, regional, and national causes of under-5 mortality in 2000-15: an updated systematic analysis with implications for the sustainable development goals. *Lancet* **388**: 3027-3035.

Liu, Y., Xu, Z. Q., Zhang, Q., Jin, M., Yu, J. M., Li, J. S., Liu, N., Cui, S. X., Kong, X. Y., Wang, H. *et al.* (2012). Simultaneous detection of seven enteric viruses associated with acute gastroenteritis by a multiplexed luminex based assay. *J Clin Microbiol* **50**: 2384–2389.

McCann, A., Ryan, F. J., Stockdale, S. R., Dalmasso, M., Blake, T., Ryan, C. A., Stanton, C., Mills, S., Ross, P. R. *et al.* (2018). Viromes of one year old infants reveal the impact of birth mode on microbiome diversity. *PeerJ* **6**: e4694.

Macpherson, A. J. and Harris, N. L. (2004). Interactions between commensal intestinal bacteria and the immune system. *Nat Rev Immunol* **4(6)**: 478-485.

Milani, C., Duranti, S., Bottacini, F., Casey, E., Turroni, F., Mahony, J., Belzer, C., Palacio, D. S., Arboleya, M. S. and Mancabelli, L. *et al.* (2017). The first microbial colonizers of the human gut: composition, activities, and health implications of the infant gut microbiota. *Microbiol Mol Biol Rev* **81(4)**: e00036-17.

Minot, S., Bryson, A., Chehoud, C., Wu, G. D., Lewis, J. D. and Bushman, F. D. (2013). Rapid evolution of the human gut virome. *Proc Natl Acad Sci U S A* **110**: 12450-12455.

O'Hara, A. M. and Shanahan, F. (2006). The gut flora as a forgotten organ. *EMBO Rep* **7**: 688-693.

Palmer, C., Bik, E.M., DiGiulio, D.B., Relman, D.A., Brown, P.O., (2007). Development of the human infant intestinal microbiota. *PLoS biol* **5(7)**: e177.

Phan, T. G., Li, L., O'Ryan, M. G., Cortes, H., Mamani, N., Bonkoungou, I. J., Wang, C., Leutenegger, C. M. and Delwart E. (2012). A third gyrovirus species in human faeces. *J Gen Virol* **93**: 1356-1361.

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., Guo, J., Le Chatelier, E., Yao, J., Wu, L. *et al*. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**: 59-64.

Rascovan, N., Duraisamy, R. and Desnues, C. (2016). Metagenomics and the Human Virome in Asymptomatic Individuals. *Annu Rev Microbiol* **70**: 125-41.

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

Sdiri-Loulizi, K., Gharbi-Khélifi, H., de Rougemont, A., Chouchane, S., Sakly, N., Ambert-Balay, K., Hassine, M., Guédiche, M. N., Aouni, M. and Pothier, P. (2008). Acute infantile gastroenteritis associated with human enteric viruses in Tunisia. *J Clin Microbiol* **46**: 1349-1355.

Simpson, R., Aliyu, S., Iturriza-Gómara, M., Desselberger, U. and Gray, J. (2003). Infantile viral gastroenteritis: on the way to closing the diagnostic gap. *J Med Virol* **70**: 258-262.

Spina, A., Kerr, K. G., Cormican, M., Barbut, F., Eigentler, A., Zerva, L., Tassios, P., Popescu, G. A., Rafila, A., Eerola, E. *et al.* (2015). Spectrum of enteropathogens detected by the FilmArray GI Panel in a multicentre study of community-acquired gastroenteritis. *Clin Microbiol Infect* **21**: 719-728.

Tamburini, S., Shen, N., Wu, H. C. and Clemente, J. C. (2016). The microbiome in early life: implications for health outcomes. *Nat Med* **22**, 713-717.

Tate, J. E., Burton, A. H., Boschi-Pinto, C., Steele, A. D., Duque, J., Parashar, U. D. and World Health Organisation-coordinated Global Rotavirus Surveillance Network. (2012). 2008 estimate of worldwide rotavirus-associated mortality in children younger than 5 years before the introduction of universal rotavirus vaccination programmes: a systematic review and meta-analysis. *Lancet Infect Dis* **12**: 136-141.

Tate, J. E., Burton, A. H., Boschi-Pint, C., Parashar, U. D. and World Health Organization-Coordinated Global Rotavirus Surveillance Network. (2016). Global, Regional, and National Estimates of Rotavirus Mortality in Children <5 Years of Age, 2000-2013. *Clin Infect Dis* **62(2)**: S96-S105

Thongprachum, A., Khamrin, P., Pham, N. T., Takanashi, S., Okitsu, S., Shimizu, H., Maneekarn, N., Hayakawa, S. and Ushijima, H. (2017). Multiplex RT-PCR for rapid detection of viruses commonly causing diarrhea in pediatric patients. *J Med Virol* **89**: 818-824.

Troeger, C., Khalil, I. A., Rao, P. C., Cao, S., Blacker, B. F., Ahmed, T., Armah, G., Bines, J. E., Brewer, T. G., Colombara, D. V. *et al.* (2018). Rotavirus vaccination and the global burden of rotavirus diarrhea among children younger than 5 years. *JAMA Pediatr* **72(10)**: 958-965.

Valentini, D., Vittucci, A. C., Grandin, A., Tozzi, A. E., Russo, C., Onori, M., Menichella, D., Bartuli, A. and Villani, A. (2013). Coinfection in acute gastroenteritis predicts a more severe clinical course in children. *Eur J Clin Microbiol Infect Dis* **32**: 909-915.

Virgin, H. W. **(2014).** The virome in mammalian physiology and disease. *Cell* **157(1)**: 142-150.

Vizzi, E., Piñeros, O. A., Oropeza, M. D., Naranjo, L., Suárez, J. A., Fernández, R., Zambrano, J. L., Celis, A. and Liprandi, F. (2017). Human rotavirus strains circulating in Venezuela after vaccine introduction: predominance of G2P[4] and reemergence of G1P[8]. *Virol J* **14**: 58.

Zhang, T., Breitbart, M., Lee, W. H., Run, J. Q., Wei, C. L., Soh, S. W., Hibberd, M. L., Liu, E. T., Rohwer, F. and Ruan, Y. (2006). RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol* **4**: e3.

Chapter 2: Literature review

2.1. General introduction

It is widely known that viruses inhabit almost every ecosystem, and they are the most abundant and diverse biological entities on earth, amounting to about 10^{31} virus particles (Breitbart and Rohwer, 2005). In fact, with new information coming from viral metagenomics studies, viruses have been shown to be more diverse than previously thought. Given this huge quantity of virus particles from various environments, determining how many of them correspond to infectious viruses is hardly possible. Analysis by electron microscopy (EM) has shown that most of these virus particles have morphological features similar to those of bacteriophages (Proctor *et al.,* 1993; Proctor, 1997). Until recently, viruses were only regarded as microbial agents that cause a broad spectrum of diseases or even deaths. However, the introduction of next generation sequencing techniques as well data from epidemiology studies have changed this perception, demonstrating that the human body harbours diverse populations of viruses even under non-pathological conditions. For example, Willner and colleagues detected numerous DNA sequences that belonged to *Poxviridae*, *Iridoviridae*, and *Mimiviridae* in a virome study (Willner *et al*., 2009). Moreover, some studies have reported the detection of giant viruses in the gut of both infants and adults (Breitbart *et al*., 2003; Reyes *et al*., 2010). Literature has reported that the human microbiome consists of close to 100 trillion cells, more or less the number of cells the human body is composed of (Turnbaugh *et al*., 2007). Moreover, faeces of healthy humans comprise nearly 10^{11} cells per gram, mostly dominated by bacteria (Wu *et al*., 2010; Human Microbiome Project, 2012).

The human gastrointestinal tract (GIT), specifically, is the natural habitat to complex microbial communities, including viruses. Improvements in sequencing methods did not only allow researchers to detect the presence of microbes, but it also shed light into how the gut microbiome influenced human health. Studies have indicated that the intestinal microbiome, through its interactions with the mucus layer, immune cells, and epithelial cells, can influence the health or disease state of host organisms (Virgin *et al*., 2014; Ursell *et al*., 2014). Research into the gut virome dynamics has established that the association between host and virome begins early in life, with compositional changes observed by the first year of life. Such occurrences are understood to coincide with changes in the diet and environmental exposure. Thus, individuals with a similar diet, appear to have comparable gut virome compositions (Minot *et al*., 2011; Minot *et al*., 2013).

Despite these studies, research on the viral component of the human microbiome and, in particular, the characterization of the healthy viral flora is still in its infancy. Nevertheless, previous studies have shown that the viral communities are vast and dynamic (Zhang *et al*., 2006; Reyes *et al*., 2010; Minot *et al*., 2011; Reyes *et al*., 2012; Minot *et al*., 2012; Reyes *et al*., 2013; Minot *et al*., 2013). Notably, pathogenic and zoonotic viruses are also

of great importance for human health. However, our knowledge about virus populations is very limited, and most of the viral communities are only partially characterized.

The advent of NGS technologies has indeed revolutionized the field of virus discovery (Lipkin, 2010 and 2013). Such technologies allow simultaneous characterization of the entire virus community, consisting of numerous species of viruses as well as the detection of novel viruses. These approaches, commonly referred to as viral metagenomics, offer an advantage of generating gigabases (gb) of genomic data with no prior knowledge of the microbes present in a given specimen. The identification of viruses is then achieved by comparing the produced sequence data to known sequence databases. In the past, cloning and Sanger sequencing were the primary techniques used in viral metagenomics (Breitbart *et al*., 2003; Djikeng *et al*., 2008), which unfortunately produced a small amount of genomic data. The application of these techniques to mammalian samples including humans can provide insight into how viruses interact with their host organisms and their environment. Therefore, it is anticipated that with the contribution and advancements in NGS-based methods, our knowledge and understanding of virus diversity and dynamics, and the impact thereof on human health will improve. In this chapter of the dissertation, the literature on human gut virome is reviewed.

2.2. The human gut virome composition and characteristics

The human gut plays host to diverse microbial communities, including bacteria, viruses, fungi, and archaea (Mai and Draganov, 2009). Research has indicated that one of the factors that can influence the human health is the interaction of certain microbes with the immune system of the host (Virgin, 2014; Norman *et al*., 2014). The viral component of the human gut microbiome, otherwise the human gut virome, refers to a population of all viruses colonizing the human gut. This human gut virome comprises viruses infecting human cells (eukaryotic viruses), bacteria-infecting viruses, otherwise known as bacteriophages (prokaryotic viruses), as well as virus-derived genetic elements integrated into the host genomic material, which contribute to host gene expression (Virgin, 2014) (**Figure 2-1**). One gram (g) of human faeces is known to contain around 10^8 to 10^9 virus particles (Rohwer, 2003; Mokili *et al*., 2012). Research on viruses present in faecal samples has demonstrated that bacterial viruses are the most dominant viruses in the gut (Breitbart *et al*., 2008). Moreover, it has been established that a dynamic community structure exists within the gut, with the abundance of bacteriophages being nearly 10-fold more than the bacteria (Minot *et al*., 2013). Although the developments of new sequencing technologies have made a significant contribution in microbiome research, studies focusing on the non-bacterial components of the microbiome including the virome are very limited. The main obstruction for viruses includes, unlike the bacterial 16S gene, the absence of a universal marker for viruses, contaminating host genomic DNA, and the lack of suitable analytical tools and a robust genomic database hampers the progress for virome studies.

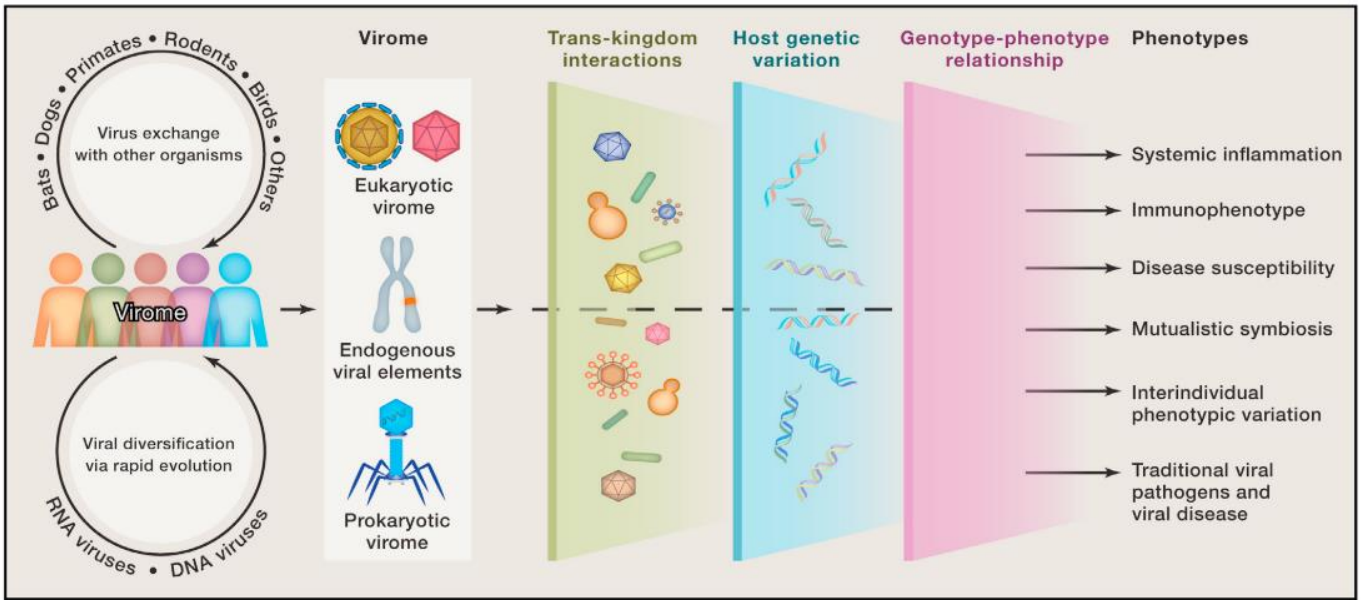

Figure 2-1: Virome components (eukaryotic virome, prokaryotic virome and endogenous viral elements), and their association to genotype/phenotype relationship. A rapid change in mammalian virome occurs continuously by exchange of virus particles between different organisms and as a result of individual virus evolution. Interactions between the virome and other members of the microbiome as well as variation in host genetics may influence a range of phenotypes important for health and disease (Taken from Virgin *et al*., 2014).

2.2.1. Bacteriophages

Bacteriophages are prokaryotic viruses that infect bacterial cells, and the total population of bacteriophages within a particular environment is called the phageome (Boyd, 2012). Based on evidence from metagenomic studies, the phageome is the most diverse component within the microbiome. Metagenomic studies have also revealed that most of the newly identified phage sequences do not share sequence similarities with those present in the current genomic databases, showing substantial diversity from known bacterial viruses (Minot *et al*., 2012). Despite the existing knowledge gaps in their classifications, research has demonstrated that dominant families of gut phages are dsDNA viruses within the families *Myoviridae*, *Siphoviridae* and *Podoviridae*, as well as the *Microviridae* family of ssDNA prokaryotic viruses (Ackermann, 2009).

The newly born infant's gut has been shown to be rapidly colonized by bacteria originating from the immediate environment and the mother (DiGiulio *et al*., 2008; Matamoros *et al*., 2013). As anticipated, this is usually followed by bacteriophage colonization, due to the presence of the host bacterial cells (Dalmasso *et al*., 2014). Furthermore, the infant's enteric phage population is less diverse, however there are frequent changes/variation occurring in this phageome, as compared to adults. As the infant's gut continues to mature, huge changes take

place within the phageome composition. However, the structure of gut phageome seems to stabilize with age (Dalmasso *et al*., 2014).

Although the pathways are still not fully understood, the effect of bacteriophages on bacterial community structure and function has been shown to influence the health of humans (Reyes *et al*., 2010; Cadwell *et al*., 2010; Minot *et al*., 2013). Metagenomic studies from the past few years have revealed that a wide variation exists within the gut bacterial virome among adults, and although a degree of similarity exists in the gut phages between individuals, the phageome appears to be unique to each person. The cause of such variabilities may be due to evolution of viruses within an individual and the body's response to the environment and the diet (Rodriguez-Valera *et al*., 2009; Reyes *et al*, 2010; Minot *et al*., 2011; Reyes *et al*., 2012; Minot *et al*., 2013). Furthermore, bacterial viruses can be regulators of the bacterial communities through the transfer of genes, eliminating competing bacteria, thereby allowing prophage-containing bacteria to colonize partially cleared areas. In addition viruses can encode toxins which may change the host's intestine to advance bacterial pathogenesis (Duerkop and Hooper, 2013).

2.2.2. Enteric viruses

Within the intestine, eukaryotic viruses represent a very small proportion compared to bacteriophages (Reyes *et al*., 2010; Minot *et al*., 2011; Zhang *et al*., 2006). The mammalian viruses colonizing the gastrointestinal tract are commonly referred to as enteric viruses. Nonetheless, metagenomic analysis of stool samples from healthy children has shown a complex community of viruses from a wide range of families including *Reoviridae*, *Astroviridae, Adenoviridae*, *Picornaviridae*, *Picobirnaviridae, Anelloviridae* and *Calicivirdae* (Kapusinszky *et al*., 2012). Enteric viruses have serious implications on human health, be it healthy or immune-compromised individuals, leading to illnesses such as severe gastroenteritis (Clark and McKendrick, 2004; Glass *et al*., 2009; Eckardt and Baumgart, 2011; Kapusinszky *et al*., 2012). For an example, picobirnaviruses, have been detected in faeces of humans with diarrhoea of unknown aetiology (Banyai *et al*., 2003; Finkbeiner *et al*., 2008; van Leeuwen *et al*., 2010), as well as in healthy subjects (Kapusinszky *et al*., 2012). The mode of transmission for most of the enteric viruses is via the faecal-oral route (Cliver, 1997). Norovirus, the common cause of non-bacterial induced acute gastroenteritis worldwide, is one of the first discovered enteric viruses (Kapikian *et al*., 1972). This was followed by the detection of other enteric viruses including astrovirus, sapovirus, adenovirus, enterovirus, and rotavirus (Glass *et al*., 2001).

Enteric viruses are the causative agents of viral gastroenteritis not only in humans, but also in other mammalian animals. Viral infection of farm animals has become of great concern and a threat to the economy in many countries due to loss of livestock (Halaihel *et al*., 2010). It has been reported that in human beings and non-human

mammals, an infection caused by the same viral agent may not result in a disease. On the contrary, both humans and porcine may suffer from severe or even fatal diarrhoea due to rotavirus infection (Desselberger, 2014).

Although viruses are highly host-specific, often infecting a limited range of host organisms, their capability to cross host species barrier must be taken into account for assessing their potential to human infections. Viruses that can be transmitted from animals to humans, known as zoonotic viruses, can also lead to diseases in humans. Example of zoonotic pathogens is avian influenza or Rabies (Christou, 2011; Abolnik, 2014). There are certain gut viruses known to be zoonotic (e.g., rotaviruses, noroviruses and astroviruses), and this zoonosis can occur either by direct transfer of viral agents from animals to human beings or through the ingestion of contaminated food product (Brugere-Picoux and Tessier, 2010; Machnowska *et al*., 2014). Among the RNA viruses that are frequently detected in the human gastrointestinal tract, studies have also reported the abundance of plant viruses, associated with diet (Minot *et al*., 2013). **Table 2-1** gives an overview of some of the studies that attempted to unravel the mammalian enteric virome within the past decade.

Table 2-1: Overview of faecal virome studies from 2008 to 2018.

Year	Topic	Material	Author
2008	Low diversity of known viruses in the infant gut <3 month old. Diversity increases with time.	Faeces	Breitbart *et al*., 2008
2009	Novel picornavirus: The Klassevirus.	Faeces	Greningeret *et al*., 2009
2010	Enteric viruses from monozygotic twins and their mothers.	Faeces	Reyes *et al*., 2010
2011	Inter-individual variation and dynamic response to diet.	Faeces	Minot *et al*., 2011
2012	Detection of novel circular ssDNA virus from bovine faecal sample.	Faeces	Kim *et al*., 2012
2013	Rapid evolution of the human gut virome.	Faeces	Minot *et al*., 2013
2014	Geographic variation in the human eukaryotic virome.	Faeces	Holtz *et al*., 2014
2015	The human gut virome: a multifaceted majority.	Faeces	Ogilvie *et al*., 2015
2016	Gut virome diversity in asymptomatic pigs in East Africa.	Faeces	Amimo *et al*., 2016
2017	Enteric virome of sympatric wild and domestic canids.	Faeces	Nádia Conceição-Neto *et al*., 2017
2018	Shared and distinct features of human milk and infant gut virome.	Faeces/ breast milk	Pannaraj *et al*., 2018

2.3. The gut virome in health and disease

Owing to the high prevalence of viruses in the gastrointestinal tract under non-pathological conditions, the gut mucosa clearly sustains numerous viral infections establishing a virome that can either benefit and/or harm the host (Barr *et al*., 2013). Thus, there's high probability that the enteric viral populations can influence the host phenotype in a healthy state, during inflammation and disease, through interactions with both other components of the gut microbiome and host genetics factors. In particular, bacteriophages could alter the interactions between the bacteria and host by infecting bacteria, and it is also possible that the gut bacterial microbiome can regulate the gut virome (Duerkop and Hooper, 2013). Intestinal phages may contribute to the shift from health to disease bringing about dysbiosis, an alteration in the gut microbial communities (de Paepe *et al*., 2014). A few models by which commensal bacterial viruses can affect the gut microbiome have been proposed, as listed below (de Paepe *et al*., 2014).

a. Kill the winner mechanism

This mechanism suggests that phages kill and reduce the population of only the dominant commensal bacteria (the "winning" microorganisms) in the gut microbiome. This phage predation mechanism is supported by the presence of clustered regularly interspaced short palindromic repeat (CRISPR) systems in human commensal bacteria. CRISPR spacers identifies and silence exogenous genetic elements such as bacteriophages, to confer some form of acquired immunity (Duerkop and Hooper, 2013).

b. Biological weapon model

In this mechanism, commensal bacteria used the bacteriophages to kill another competing bacteria for the enteric environment (Bossi *et al*., 2003; Brown *et al*., 2006). In this scenario, the phage would provide immune protection to its carrier bacteria against further infection (Barr *et al*., 2013). Acting as "biological weapons", phages would cause lysis of competing microorganisms, leading to dysbiosis and sometimes inflammatory response (de Paepe *et al*., 2014).

c. Community shuffling model

In this model (Mills *et al*., 2013), conditions such as inflammation, antibiotic therapy and oxidative stress trigger prophage induction in certain bacteria including *Escherichia coli* (Zhang *et al*., 2000) and *Clostridium difficile* (Meessen-Pinard *et al*., 2012). In this model, the prophage induction could lead to intestinal dysbiosis, which can in turn trigger inflammatory bowel disease (IBD), Crohn's Disease (CD) and colon cancer (Sun *et al*., 2011; Hanahan and Weinberg, 2011).

d. Role of other enteric viruses

Metagenomic studies have linked several enteric mammalian viruses to acute diarrhoea in children, showing high abundance of viruses from the families *Picornaviridae, Adenoviridae*, and *Reoviridae*. The genus *Enterovirus* was highly abundant within the *Picornaviridae* family (Holtz *et al*., 2014). **Table 2-2** summarizes different viruses found in faecal samples and their disease association.

Table 2-2: Some of the known viruses detected by viral metagenomics (Adapted from Scarpelleni *et al*., 2015).

Virus type	Virus genus/species	Genome type	Environment	Disease
Eukaryotic viruses:	Rotavirus Astrovirus Norovirus Enterovirus	RNA	Human faecal samples	Gastroenteritis
Plant viruses:	Pepper mild mottle virus Oat blue dwarf virus Tobacco Mosaic virus Maize chlorotic mottle virus	RNA	Human faecal samples and plants	Pathogenic only to plants, not humans

2.4. Interaction between the gut virome and the immune system

Modulation of the immune system by the existing interaction between the enteric immune system and components of the gut microbiome, including viruses can impact the host's health and disease (Duerkop and Hooper, 2013; Virgin, 2014).

The role of bacteriophages on the immune system

Evidence exists that bacteriophages, In addition to regulating the bacterial populations, could also interact with the human immune system, directly so. This can be supported by the observation in previous studies, wherein the bacteriophages that were orally administered translocate *in vivo* to systemic tissue, and induce innate and adaptive immunity (Duerr *et al*., 2004; Hamzeh-Mivehroud *et al*., 2008). Several other studies have also provided evidence of bacteriophage-induced humoral immune response (Uhr *et al*., 1962; Inchley and Howard, 1969). However, the process by which these phages stimulate innate immunity is not fully understood. In healthy humans, the cytokines secreted by immune cells plays a crucial role in regulating the balance between the virome and the immune system. Such cells can identify antigenic elements or pathogen-associated molecular patterns (PAMPs), including those produced by viruses (Virgin, 2014).

Furthermore, certain phages can also use commensal bacteria to drive their own genome, and in some circumstances, immunodeficiency among others, induce the expression of bacteriophage particles, which can elicit the immune response (Duerkop *et al*., 2012). In another study, bacteriophage proteins have been shown to enhance the potency of DNA vaccines (Cuesta *et al*., 2006).

The role of eukaryotic viruses on the immune system

Not much is known about the interaction between mammalian enteric viruses and the host immune system. However, findings from limited studies have shown that the eukaryotic virome can affect the host defense mechanisms against infections by viruses and bacteria. Moreover, certain viral pathogens chronically residing in tissues of healthy individuals, like herpesvirus may lead to underlying infections which can offer host protection against bacterial infections (Barton *et al*., 2007).

Conversely, other chronic viral infections can weaken host immune functioning and enhance susceptibility to infection. Specifically, immunodeficiency viruses have been associated with damage to the intestinal barrier, causing expansion of the gut virome (Duerkop and Hooper, 2013). Chronic immune suppression leads to total host immune deficiency, presenting the opportunity for some pathogens to damage the gut epithelial cells.

Consequently, these events facilitate the translocation of gut viruses and commensal bacteria across the epithelial surface, causing inflammation and systemic infection (Handley *et al*., 2012).

2.5. Tools for virus detection and virome characterization

In previous years, the detection of viruses in a clinical or environmental sample was very challenging and time-consuming due to limitations of the techniques available. It's only recently that several improvements were made as well as the introduction of new technologies for virus discovery and virome characterization. These may include DNA arrays and NGS techniques, which allow for fast and sensitive detection and characterization of viruses (Wang *et* al., 2003; Chiu, 2013; Pallen, 2014). These developments have led to a growing interest and a massive increase in the use of NGS technologies in several biological fields, particularly in pathogen discovery and virome studies (Lipkin, 2013). Numerous studies have applied these methods to explore the human virome and have successfully revealed the existence and composition of the human gut virome in both diseased and healthy subjects. Nevertheless, complete composition of the human gut virome and its impact on health are yet to be determined (Lipkin, 2010; Minot *et al*., 2011; Minot *et al*., 2013; Reyes *et al*., 2013; Lipkin 2013).

Several tools exist that can be used in virus identification and virome characterization. These methods can be classified into two, namely traditional and modern techniques. However, some traditional methods are only suitable for identification of individual pathogens, but not necessarily appropriate for characterization of whole viral populations as they cannot simultaneously identify multiple viral agents in a single specimen. For instance, electron microscopy, polymerase chain reaction (PCR) and Sanger sequencing have been used for virus identification but may not allow characterization of the entire virome. As a result, these traditional techniques have been substituted by modern techniques which are mainly molecular techniques, based on the detection of the virus nucleic material. They primarily involve amplification of nucleic acid followed by sequencing of the products, thereof. Several of these techniques and their applications are discussed below.

2.5.1. Genome amplification methods

Probably the most broadly applied method in molecular biology is the PCR. This method, in which a DNA sequence is exponentially amplified to produce numerous copies of the template DNA, was developed in 1983 by Mullis and colleagues (Mullis *et al*., 1986). After a decade, another method for genome amplification, known as strand displacement amplification (SDA), was developed by Walker and colleagues (Walker *et al*., 1992). The DNA is also exponentially amplified whereby strands displaced from a sense reaction serve as a target for an antisense reaction and vice versa (Walker *et al*., 1992). However, the main disadvantage of the PCR methods is that only known viruses can be detected since they use designed primers that will target a specific sequence. As a

consequence, these techniques may not be applicable for the detection of novel viruses. On the contrary, a method that allows non-specific amplification exists. This technique, referred to as the sequence-independent amplification (SIA) involves the use of random primers and can also be used in metagenomic studies, as it can provide sufficient nucleic material for sequencing (Delwart, 2007; Potgieter *et al*., 2009).

2.5.2. DNA sequencing

DNA sequencing is a molecular biology method that enables researchers to determine the order of the four nucleotides that make up a DNA molecule. Improvements in DNA sequencing methods continue to transform all branches of biology and medicine, from its role in determining the cause of diseases to virus discovery and drug development (Woollard *et al*., 2011; Datta *et al*., 2012). In the next few pages, progress made in sequencing technologies and the applications thereof, is discussed.

a. First generation sequencing

The development of DNA sequencing techniques started with two different groups led by Frederick Sanger in 1975 and by Allan Maxam and Walter Gilbert in 1977. Sanger's method was based on selective incorporation of a chain-terminating dideoxynucleotides using a DNA polymerase (Sanger and Coulson, 1975; Sanger *et al*., 1977). The Maxam and Gilbert method involved chemical modification of DNA and subsequent base-specific cleavage of the DNA (Maxam and Gilbert, 1977). Gel electrophoresis and radioactive labeling were used for separation and visualization of fragments in the two techniques. Sanger sequencing was their easiest and did not require many harmful chemicals. The two most important steps in the further development were automatization of the sequencing reactions and advanced base detection methods. Automated laser-based fluorescence dye detection and capillary electrophoresis were then later introduced (Smith *et al*., 1986). However, Sanger technologies could be too time-consuming and costly to sequence of virome or huge mammalian genomes.

b. Second generation sequencing

Interestingly, combining Sanger technologies with fluorescence detection led to the next generation sequencing (NGS), which allowed numerous sequence reactions to take place simultaneously. The principle of these NGS technologies was still based on Sanger sequencing, however, they follow different procedures with regards to fragmentation of genomic material, amplification of fragments and base detection methods. **Table 2-3** summarizes the most common sequencing instruments for first to third generation sequencing technologies. In particular, Illumina technology is one of the most popular second generation or NGS platform. It is based on the sequencing by synthesis and reversible termination chemistry (Bentley *et al*., 2008). Ion Torrent sequencing, which also follows sequencing by synthesis chemistry is based on ion semiconductor technology, whereby fluctuations in pH (a measure of how acidic or basic a solution is) are detected (Grada and Weinbrecht, 2013). These new

technologies come with the advantages of high speed, high accuracy, coverage and throughput and are also less time-consuming. Most relevant, they have made viral metagenomics research possible due to their ability to sequence samples without cultivation or prior knowledge of the sample composition.

Table 2-3: Common sequencing platforms available.

	First generation sequencing	**Second generation sequencing**		**Third generation sequencing**
Instrument	Sanger sequencing (capillary-based)	Illumina MiSeq	Ion Torrent	PacBio
Chemistry	Fluorescent dideoxy terminator	Fluorescent emission from ligated dye-labelled nucleotides	Proton detection	Fluorescent, single molecule sequencing
Read length	750bp	≤600bp	>200bp	≤10kbp
Applications	Small nucleic acid fragments	Whole genomes, metagenomics, targeted sequencing	Whole genomes, metagenomics, targeted sequencing	*De-novo* assembly of genomes including long repetitive sequences
Remarks	Low throughput, high accuracy	High coverage and output, high accuracy, costly	High coverage, longer reads	High coverage, very long reads, costly

2.5.3. Virome enrichment

The main challenge in performing detailed virome analysis is the relatively low abundance of viral genetic material as compared to bacterial and background host nucleic acids (Yang et al., 2011; Mokili et al., 2013). The enrichment of samples for viromes is a very critical step for a successful characterization. However, loss of viruses and introduction bias must be minimized. Few studies have evaluated the efficiency of the various enrichment methods (Hall *et al*., 2014; Conceição-Neto *et al*., 2015).

Hall and colleagues used three viruses (adenovirus, influenza A and human enterovirus) to test combinations of procedures such as centrifugation, filtration and nuclease digestion (Hall *et al*., 2014) to demonstrate that various methods have significant impact on virus recovery. Rosseel and colleagues determined the efficiency of filtration,

DNase treatment and rRNA removal in serum and tissue samples spiked with Newcastle disease virus. This method was evaluated only with one virus and is exclusively applicable for RNA viruses (Rosseel *et al*., 2015).

For a more detailed study, Conceição-Neto and colleagues developed a protocol for virome enrichment that aimed to substantially remove non-viral nucleic acids. Based on their findings, they concluded that a thorough homogenization of most biological or environmental samples is important for optimal enrichment of virus particles. For faecal samples in particular, it was discovered that the use of beads in homogenization of faecal samples can lead to significant virus losses, and as a result should be omitted (Conceição-Neto *et al*., 2015). Here, homogenization was followed by centrifugation to pellet the larger microbial particles and cellular debris. They found that centrifugation of the homogenate removed majority of the bacteria and also resulted in 99 % decrease of ribosomal RNA. The recovered supernatant from centrifugation was then subjected to filtration to separate viral particles from larger prokaryotic and host cells using either 0.45 µm or 0.22 µm filters (Conceição-Neto *et al*., 2015). Although filtration steps resulted in the efficient removal of over 99 % of bacterial cells, an equal quantity of the bigger molecular weight mimivirus particles were lost thus the 0.45 µm was more appropriate for enrichment of total virome (Conceição-Neto *et al*., 2015). Nuclease treatment of the resulting filtrate is one of the key steps in virome enrichment as it digest the free-floating nucleic acids. Lastly, in order to have enough starting material for sequencing, random amplification of viral genome was carried out (Conceição-Neto *et al*., 2015). Although enrichment protocols should aim to increase the quantity of viral nucleic material, whilst reducing background genome, some of the proposed amplification protocols have been shown to introduce bias (Kim *et al*., 2011; Li *et al*., 2015).

2.6. Bioinformatics approaches and tools for virome analysis

Unlike 16S rRNA sequencing data analysis, for which analysis packages like QIIME (Caporaso *et al*., 2010) exists, such tools do not exist for virome analysis. Available tools can perform functional annotation of viromes and estimate viral diversity **(Table 2-4).** Efforts are made to implement pipelines that can take raw reads, filter out host and bacterial genomes and assign taxonomy and functionality to viruses within samples. Nevertheless, some methods have been used to analyze viral metagenomic data to determine the composition, to detect known and new virus species. These methods entail firstly the removal of adapter sequences added during the library preparation stage, and this is followed by removal of low quality reads. BLAST (Altschul *et al*., 1990) alignment is then used to filter out host genomes. The sequence reads can be analyzed as short individual fragments or as larger contiguous sequences, or contigs in short, generated by *de novo* assembly (Namiki *et al*., 2012).

Analysis of larger contigs allows for better screening of viruses from various sampling points. Although there are currently available tools for building of contigs, there are a number of challenges that necessitate further optimization of these methods. For instance, relative abundance of genomes and sequence heterogeneity may influence the results. BLAST is one of the methods that can be used downstream to identify sequence homology of reads and contigs to reference genomes in the viral database, and therefore determine the composition and abundance (Langmead and Salzberg, 2012). Open reading frames (ORFs) can also be called on contigs to find a particular gene (Wang *et al.*, 2010). More than 7000 reference whole genome sequences are available in the NCBI genome database. In addition to Refseq and UniProt (Hulo *et al.*, 2011; O'Leary *et al.*, 2016), there are custom viral protein databases for samples from a range of environments and humans (Norman *et al.*, 2015).

Table 2-4: Different tools and tools for bioinformatic analysis of virome data.

Name	Description
MetaVir 2 (Roux *et al*., 2014)	Web-based tool for annotating viral reads or contigs
VIROME (Womack *et al*., 2012)	Web-based tool for classification of predicted of predicted ORFs from viral metagenomic data
VirusDetect (Zheng *et al*., 2017)	A web-based tool for detection of viruses from small RNA sequence data.
VirusSeeker (Zhao *et al*., 2017)	BLAST-based analytical method for virus identification and virome composition.
VirFinder (Ren *et al*., 2017)	K-mer frequency-based, metagenomic identification.

2.7. Conclusions

Despite our limited understanding of the human virome, the rapid progress and improvements in high throughput sequencing and analysis tools continue to expand our knowledge of the human virome. It is now known that the human virome is considerably huge and diverse, and it is continuously evolving at a faster rate. Although characterization of virome composition and diversity is important, one of the most key areas of research is the interaction of viruses with other members of the microbiome. Importantly, the interaction of viruses with bacterial cells can modulate viral infectivity and pathogenesis. Moreover, modulation of the immune system by certain chronic viral infections can affect pathogenicity of other microbes. Research has shown the value of NGS approaches in understanding these complexities within the human gut virome community and their role on human health.

Although the application of NGS in virology remains challenging due to the absence of a universal viral sequence signature, with decreasing sequencing costs, development of new analysis tools, and improved annotation of virome databases, the capacity to determine the virome composition is widening. However, none of the available tools has addressed the issue of viral dark matter. This is not surprising, considering the estimated vast populations of viruses on earth, and the limited viral sequence databases currently available. However, this may not an issue for identification human pathogens, most of which have been well characterized due their medical relevance. Despite the obvious challenges in viral metagenomic studies, current solutions and approaches in the analysis of metagenomes are broadening our views on human gut ecology at a fast pace.

NGS can effectively provide a broad picture of viruses and bacteriophages colonizing the gastrointestinal tract of humans. The impact of virome on human health calls for efforts to comprehensively characterize the human virome and annotate the viral agents inhabiting humans. It is also important to note that the virome is only one component of our metagenome, therefore it should not be analyzed separately from other members of the microbiome.

2.8. References

Abolnik, C. (2014). A current review of avian influenza in pigeons and doves (*Columbidae*). *Vet Microbiol* **170(3-4)**: 181-96.

Ackermann, **H. W. (2009).** Phage classification and characterization. *Methods Mol Biol* **501**: 127-140.

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

Amimo, J. O., Okoth, E., Junga, J. O., Ogara, W. O., Njahira, M. N., Wang, Q., Vlasova, A. N., Saif, L. J. and Djikeng, A. (2014). Molecular detection and genetic characterization of kobuviruses and astroviruses in asymptomatic local pigs in East Africa. *Arch Virol* **159**: 1313-1319.

Banyai, K., Jakab, F., Reuter, G., Bene, J., Uj, M., Melegh, B. and Szücs, G. (2003). Sequence heterogeneity among human picobirnaviruses detected in a gastroenteritis outbreak. *Arch Virol* **148(12)**: 2281-2291.

Barr, J. J., Auro, R., Furlan M., Whiteson, K. L., Erb, M. L., Pogliano, J., Stotland, A., Wolkowicz, R., Cutting, A. S., Doran, K. S. *et al*. (2013). Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proc Natl Acad Sci U S A* **110(26)**: 10771-10776.

Barton, E. S., White, D. W., Cathelyn, J. S., Brett-McClellan, K, A., Engle, M., Diamond, M. S., Miller, V. L. and Virgin, H. W. (2007). Herpesvirus latency confers symbiotic protection from bacterial infection. *Nature* **447(7142)**: 326-329.

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. and Bignell, H. R. *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53-59.

Bossi, L., Fuentes, J. A., Mora, G. and Figueroa-Bossi, N. (2003). Prophage contribution to bacterial population dynamics. *J Bacteriol* **185(21)**: 6467-6471.

Boyd, E. F. (2012). Bacteriophage-Encoded Bacterial Virulence Factors and Phage–Pathogenicity Island Interactions. *Adv Virus Res* **82**: 91-118.

Breitbart, M. and Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends Microbiol* **13**: 278-284.

Breitbart, M., Haynes, M., Kelley S., Angly, F., Edwards, R. A., Felts, B., Mahaffy, J. M., Mueller, J., Nulton, J., Rayhawk, S. *et al*. (2008). Viral diversity and dynamics in an infant gut. *Res Microbiol* **159(5)**: 367-373.

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J.M., Nulton, J., Salamon, P. and Rohwer, F. (2003). Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* **185**: 6220-6223.

Brown, S. P., Le Chat, L., de Paepe, M. and Taddei, F. (2006). Ecology of microbial invasions: amplification allows virus carriers to invade more rapidly when rare. *Curr Biol* **16(20)**: 2048-2052.

Brugere-Picoux, J. and Tessier, P. (2010). Viral gastroenteritis in domestic animals and zoonoses. *Bull Acad Natl Med* **194(8)**: 1439-1449.

Cadwell, K., Patel, K. K., Maloney, N. S., Liu, T. C., Ng, A. C., Storer, C. E., Head, R. D., Xavier, R., Stappenbeck, T. S. and Virgin, H. W. (2010). Virus-plus- susceptibility gene interaction determines Crohn's disease gene Atg16L1 phenotypes in intestine. *Cell* **141(7)**: 1135-1145.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I. *et al*. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335-336.

Chiu, C. Y. (2013). Viral pathogen discovery. *Curr Opin Microbiol* **16**: 468-478.

Christou L. (2011). The global burden of bacterial and viral zoonotic infections. *Clin Microbiol Infect* **17(3)**: 326-30.

Clark, B. and McKendrick, M. (2004). A review of viral gastroenteritis. *Curr Opin Infect Dis* **17(5)**: 461-469.

Cliver, D. O. (1997). Virus transmission via food. *Food Technol* **51(4)**: 71-78.

Conceição-Neto, N., Zeller, M., Lefrère, H., De Bruyn, P., Beller, L., Deboutte, W., Yinda, C. K., Lavigne, R., Maes, P., van Ranst, M. *et al*. (2015). Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci Rep* **5**: 16532.

Conceição-Neto, N., Godinho, R., Álvares, F., Yinda, C. K., Deboutte, W., Zeller, M., Laenen, L., Heylen, E., Roque, S., Petrucci-Fonseca, F. *et al.* (2017). Viral gut metagenomics of sympatric wild and domestic canids, and monitoring of viruses: Insights from an endangered wolf population. *Ecol Evol* **7(12)**: 4135-4146.

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

Dalmasso, M., Hill, C. and Ross, R. P. (2014). Exploiting gut bacteriophages for human health. *Trends Microbiol* **22**: 399-405.

Datta, S., Budhauliya, R., Das, B., Chatterjee, S., Vanlalhmuaka. and Veer, V. (2012). Next-generation sequencing in clinical virology: Discovery of new viruses. *World J Virol* **4(3)**: 265-76.

de Paepe, M., Leclerc, M., Tinsley, C. R. and Petit, M. A. (2014). Bacteriophages: an underestimated role in human and animal health? *Front Cell Infect Microbiol* **5(39)**.

Delwart, E. L. (2007). Viral metagenomics. *Rev Med Virol* **17(2)**: 115-131

Desselberger, U. (2014). Global issues related to enteric viral infections. *VirusDisease* **25(2)**: 147-149.

DiGiulio, D. B., Romero. R., Amogan, H. P., Kusanovic, J. P., Bik, E. M., Gotsch. F, et al. (2008). Microbial prevalence, diversity and abundance in amniotic fluid during preterm labor: A molecular and culture-based investigation. *PLoS One* **3(8)**: e3056.

Djikeng, A., Halpin, R., Kuzmickas, R., Depasse, J., Feldblyum, J., Sengamalay, N., Afonso, C., Zhang, X., Anderson, N.G., Ghedin, E. *et al.* (2008). Viral genome sequencing by random priming methods. *BMC Genomics* **9**: 5.

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

Duerkop, B. A., Clements, C. V., Rollins, D., Rodrigues, J. L. M. and Hooper, L. V. (2012). A composite bacteriophage alters colonization by an intestinal commensal bacterium. *Proc Nat Acad Sci of the U S A* **109 (43)**: 17621-17626.

Duerr, D. M., White, S. J. and Schluesener, H. J. (2004). Identification of peptide sequences that induce the transport of phage across the gastrointestinal mucosal barrier. *J Virol Methods* **116(2)**: 177-180.

Eckardt, A. J. and Baumgart, D. C. (2011). Viral gastroenteritis in adults. *Recent Pat Antiinfect Drug Discov* **6(1)**: 54-63.

Finkbeiner, S. R., Allred, A. F., Tarr, P. I., Klein, E. J., Kirkwood, C. D. and Wang, D. (2008). Metagenomic analysis of human diarrhea: viral detection and discovery. *PLoS Pathog* **4(2)**: e1000011.

Glass, R. I., Bresee, J., Jiang, B., Gentsch, J., Ando, T., Fankhauser, R., Noel, J., Parashar, U., Rosen, B. and Monroe S. S. (2001). Gastroenteritis viruses. In *Novartis Foundation Symposium*, pp. 5-19. New York: Wiley.

Glass, R. I., Parashar, U. D. and Estes, M. K. (2009). Norovirus gastroenteritis. *New Engl J Med* **361(18)**: 1726-1785.

Grada, A. and Weinbrecht, K. (2013). Next-generation sequencing: methodology and application. *J Invest Dermatol* **133(8)**: e11.

Greninger, A. L., Runckel, C., Chiu, C. Y., Haggerty, T., Parsonnet, J., Ganem, D. and DeRisi, J. L. (2009). The complete genome of klassevirus - a novel picornavirus in pediatric stool. *Virol J* **16**: 82.

Halaihel, N., Masía, R. M., Fernández-Jiménez, M., Ribes, J. M., Montava, R., De Blas, I., Gironés, O., Alonso, J. L. and Buesa, J. (2010). Enteric calicivirus and rotavirus infections in domestic pigs. *Epidemiol Infect* **138(4)**: 542-548.

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

Hamzeh-Mivehroud, M., Mahmoudpour, A., Rezazadeh, H. and Dastmalchi, S. (2008). Non-specific translocation of peptide-displaying bacteriophage particles across the gastrointestinal barrier. *Europ J Pharm Biopharm* **70(2)**: 577-581.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* **144(5)**: 646-674.

Handley, S. A., Thackray, L. B., Zhao, G., Presti, R., Miller, A. D., Droit, L., Abbink, P., Maxfield, L. F., Kambal, A., Duan, E. *et al.* (2012). Pathogenic simian immunodeficiency virus infection is associated with expansion of the enteric virome. *Cell* **151(2)**: 253-266.

Holtz, L. R., Cao, S., Zhao, G., Bauer, I. K., Denno, D. M., Klein, E. J., Antonio, M., Stine, O. C., Snelling, T. L., Kirkwood, C. D. and Wang, D. (2014). Geographic variation in the eukaryotic virome of human diarrhea. *Virology* **468-470**: 556-564.

Hulo, C., de Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I. and Le Mercier, P. (2011). ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res* **39**: D576-D582.

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486(7402)**: 207-214.

Inchley, C. J. and Howard, J. G. (1969). The immunogenicity of phagocytosed T4 bacteriophage: cell replacement studies with splenectomized and irradiated mice. *Clin Exp Immunol* **5(1)**: 189-198.

Kapikian, A. Z., Wyatt, R. G., Dolin, R., Thornhill, T. S., Kalica, A. R. and Chanock, R. M. (1972). Visualization by immune electron microscopy of a 27-nm particle associated with acute infectious nonbacterial gastroenteritis. *J Virol* **10(5)**: 1075-1081.

Kapusinszky, B., Minor, P. and Delwart, E. (2012). Nearly constant shedding of diverse enteric viruses by two healthy infants. *J Clin Microbiol* **50(11)**: 3427-3434.

Kim, H. K., Park, S. J., Nguyen, V. G., Song, D. S., Moon, H. J., Kang, B. K and Park, B. K. (2012). Identification of a novel single-stranded, circular DNA virus from bovine stool. *J Gen Virol* **93**: 635-639.

Kim, K. H. and Bae, J. W. (2011). Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl Environ Microbiol* **77**: 7663-7668.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357-359.

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

Lipkin, W.I. (2010). Microbe hunting. *Microbiol Mol Biol Rev* **74**: 363-377.

Lipkin, W.I. (2013). The changing face of pathogen discovery and surveillance. *Nat Rev Microbiol* **11**: 133-141.

Machnowska, P., Ellerbroek, L. and Johne, R. (2014). Detection and characterization of potentially zoonotic viruses in faeces of pigs at slaughter in Germany. *Vet Microbiol* **168(1)**: 60-68.

Matamoros, S., Gras-Leguen, C., Le Vacon, F., Potel, G. and de La Cochetiere, M. F. (2013). Development of intestinal microbiota in infants and its impact on health. *Trends Microbiol* **21(4)**: 167-173.

Maxam, A. M., and Gilbert, W. (1977). A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74(2)**: 560-564.

Meessen-Pinard, M., Sekulovic, O. and Fortier, L. C. (2012). Evidence of in vivo prophage induction during clostridium difficile infection. *Appl Environ Microbiol* **78(21)**: 7662-7670.

Mills, S., Shanahan, F., Stanton, C., Hill, C., Coffey, A. and Ross, R. P. (2013). Movers and shakers: influence of bacteriophages in shaping the mammalian gut microbiota. *Gut Microbes* **4(1)**: 4-16.

Minot, S., Bryson, A., Chehoud, C., Wu, G. D., Lewis, J. D. and Bushman, F. D. (2013). Rapid evolution of the human gut virome. *Proc Natl Acad Sci U S A* **110**: 12450-12455.

Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D. and Bushman, F. D. (2012). Hypervariable loci in the human gut virome. *Proc Natl Acad Sci U S A* **109**: 3962-3966.

Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S. A., Wu, G. D., Lewis, J. D. and Bushman, F. D. (2011). The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res* **21**: 1616-1625.

Mokili, J. L., Dutilh, B. E., Lim, Y. W., Schneider, B. S., Taylor, T., Haynes, M. R., Metzgar, D., Myers, C. A., Blair, P. J., Nosrat B. *et al*. (2013). Identification of a novel human papillomavirus

Mokili, J. L., Rohwer, F. and Dutilh, **B. E. (2012).** Metagenomics and future perspectives in virus discovery. *Curr Opin Virol* **2(1)**: 63-77.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. and Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol* **51(1)**: 263-273.

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

Norman, J. M., Handley, S. A. and Virgin, H. W. (2014). Kingdom-agnostic metagenomics and the importance of complete characterization of enteric microbial communities. *Gastroenterol* **146**: 1459-1469.

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. *et al*. (2015). Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160**: 447-460.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D. *et al*. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**: D733-D745

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

Pallen M. J. (2014). Diagnostic metagenomics: potential applications to bacterial, viral and parasitic infections. *Parasitology* **141**: 1856-1862.

Pannaraj, P. S., Ly, M., Cerini, C., Saavedra, M., Aldrovandi, G. M., Saboory, A. A., Johnson, K. M., Pride, D. T. (2018). Shared and Distinct Features of Human Milk and Infant Stool Viromes. *Front Microbiol* **9**: 1162.

Potgieter, A. C., Page, N. A., Liebenberg, J., Wright, I. M., Landt, O. and van Dijk, A. A. (2009). Improved strategies for sequence-independent amplification and sequencing of viral double-stranded RNA genomes. *J Gen Virol* **90(6)**: 1423-1432.

Proctor, L. M. (1997). Advances in the study of marine viruses. *Microsc Res Tech* **37**: 136-161.

Proctor, L. M., Okubo, A. and Fuhrman, J.A. (1993). Calibrating estimates of phage-induced mortality in marine bacteria: Ultrastructural studies of marine bacteriophage development from one-step growth experiments. *Microb Ecol* **25**: 161-182.

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. and Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**: 69.

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

Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F. and Gordon, J. I. (2012). Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat Rev Microbiol* **10**: 607-617.

Reyes, A., Wu, M., McNulty, N. P., Rohwer, F. L. and Gordon, J. I. (2013). Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut. *Proc Natl Acad Sci U S A* **110(50)**: 20236-20241.

Rodriguez-Valera, F., Martin-Cuadrado, A. B., Rodriguez-Brito, B., Pasić, L., Thingstad, T. F., Rohwer, F. and Mira, A. (2009). Explaining microbial population genomics through phage predation. *Nat Rev Microbiol* **7**: 828-836.

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

Roux, S., Tournayre, J., Mahul, A., Debroas, D. and Enault, F. (2014). Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **15**: 76.

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94(3)**: 441-448.

Sanger, F. and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Biotechnology* **24**:104-108.

Scarpellini, E., Ianiro, G., Attili, F., Bassanelli, C., De Santis, A. and Gasbarrini, A. (2015). The human gut microbiota and virome: Potential therapeutic implications. *Dig Liver Dis* **47(12)**: 1007-1012.

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B. and Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature* **321(6071)**: 674-679.

Sun, L., Nava, G. M. and Stappenbeck, T. S. (2011). Host genetic susceptibility, dysbiosis, and viral triggers in inflammatory bowel disease. *Curr Opin Gastroenterol* **27(4)**: 321-327.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R. and Gordon, J. I. (2007). The human microbiome project. *Nature* **449**: 804-810.

Uhr J. W., Dancis J., Franklin, E. C., Finkelstein, M. S. and Lewis E. W. (1962). The antibody response to bacteriophage phi-X 174 in newborn premature infants. *J Clinl Invest* **41**: 1509-1513.

Ursell, L. K., Haiser, H. J., Van Treuren, W., Garg, N., Reddivari, L., Vanamala, J., Dorrestein, P. C., Turnbaugh, P. J. and Knight, R. (2014). The intestinal metabolome: an intersection between microbiota and host. *Gastroenterol* **146(6)**: 1470-1476

van Leeuwen, M., Williams, M. M. W., Koraka, P., Simon, J. H., Smits, S. L. and Osterhaus, A. D. M. E. (2010). Human picobirnaviruses identified by molecular screening of diarrhea samples. *J Clin Microbiol* **48(5)**: 1787-1794.

Virgin, H. W. (2014). The virome in mammalian physiology and disease. *Cell* **157**: 142-150**.**

Walker, G. T., Fraiser, M. S., Schram, J. L., Little, M. C., Nadeau, J. G. and Malinowski, D. P. (1992). Strand displacement amplification--an isothermal, in vitro DNA amplification technique. *Nucleic Acids Res* **20(7)**: 1691-1696.

Wang, D., Urisman, A., Liu, Y-T., Springer, M., Ksiazek, T. G., Erdman, D. D., Mardis, E. R., Hickenbotham, M., Magrini, V., Eldred, J. J *et al*. (2003). Viral Discovery and Sequence Recovery Using DNA Microarrays. *PLoS Biol* **1(2):** e2.

Wang, S., Sundaram, J. P. and Spiro, D. (2010). VIGOR, an annotation program for small viral genomes. *BMC Bioinformatics* **11**: 451.

Willner, D., Furlan, M., Haynes, M., Schmieder, R., Angly, F.E., Silva, J., Tammadoni, S., Nosrat, B., Conrad, D. and Rohwer, F. (2009). Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PloS One* **4**: e7370.

Wommack, K. E., Bhavsar, J., Polson, S. W., Chen, J., Dumas, M., Srinivasiah, S., Furman, M., Jamindar, S. and Nasko, D. J. (2012). VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Stand Genomic Sci* **6**: 427-439.

Woollard, P. M., Mehta, N. A., Vamathevan, J. J., Van Horn, S., Bonde, B. K. and Dow, D. J. (2011). The application of next-generation sequencing technologies to drug discovery and development. *Drug Discov Today* **16(11-12)**: 512-9.

Wu, G. D., Lewis, J. D., Hoffmann, C., Chen, Y. Y., Knight, R., Bittinger, K., Hwang, J., Chen, J., Berkowsky, R., Nessel, L., Li, H. *et al*. (2010). Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiol* **10**: 206.

Yang, J., Yang, F., Ren, L., Xiong, Z., Wu, Z., Dong, J., Sun, L., Zhang, T., Hu, Y., Du, J. *et al*. (2011). Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. *J Clin Microbiol* **49(10)**: 3463-3469.

Zhang, T., Breitbart, M., Lee, W. H., Run, J. Q., Wei, C. L., Soh, S. W., Hibberd, M. L., Liu, E. T., Rohwer, F. and Ruan, Y. (2006). RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol* **4**: e3.

Zhang, X., McDaniel, A. D., Wolf, L. E., Keusch, G. T., Waldor, M. K. and Acheson, D. W. K. (2000). Quinolone antibiotics induce Shiga toxin-encoding bacteriophages, toxin production, and death in mice. *J Infect Dis* **181(2)**: 664-670.

Zhao, G., Wu, G., Lim, E. S., Wu, G., Lim, E. S., Droit, L., Krishnamurthy, S., Barouch, D. H., Virgin, H. W. and Wang, D. (2017). VirusSeeker, a computational pipeline for virus discovery and virome. *Virol* **503**: 21-30.

Zheng, Y., Gao, S., Padmanabhan, C., Li, R., Galvez, M., Gutierrez, D., Fuentes, S., Ling, K. S., Kreuze, J. and Fei, Z. (2017). VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. *Virol* **500**: 130-138.

Chapter 3: Virome enrichment, whole transcriptome amplification and Illumina sequencing

3.1. Introduction

A human body is colonized by numerous populations of viruses which are unique to each individual (Minot and Bryson, 2013). Viral metagenomics is an incredibly valuable method that can give a clearer picture about the structure and composition of the human virome and it continues to be a useful approach in the detection of new viral agents (Edwards and Rohwer, 2005). Earlier methods used in viral metagenomics studies included cloning as well as Sanger sequencing (Allander *et al*., 2005; Djikeng *et al*., 2008), which have recently been replaced by more advanced next generation sequencing (NGS) technologies. Despite improvements in viral discovery due to sequencing output and decrease in running costs (Barzon *et al*., 2011; Radford *et al*., 2012; Frey *et al*., 2014), the small amount of virus nucleic material compared to non-viral nucleic acids such as that of humans and bacteria, remains a major setback (Rosseel *et al*., 2015). An approach that can enhance the quantity of viral nucleic acids while reducing host and bacterial nucleic acids is needed for successful virome analysis. Although a metagenomics approach has made it possible to analyze all microbial communities from any sample, detailed analysis of the virome is usually made difficult by the larger size of bacterial genome (Conceição-Neto *et al.*, 2015). Additionally, such methods tend to show bias against viral RNA nucleic material present in microbial communities (Breitbart *et al*., 2008).

In order to understand how the human gut virome influences health and disease, optimized techniques for reproducible viral metagenomic analysis need to be developed. For efficient study of the virome, it is important that pre-enrichment procedures are performed for viruses from a sample. To increase the nucleic starting material prior to library preparations for NGS, random amplification is usually recommended (Reyes *et al*., 2012). When performing these steps, one can easily lose virus particles during sample preparation and introduce bias, therefore caution must be exercised (Conceição-Neto *et al.*, 2015).

Different enrichment protocols for viral metagenomics were developed and tested by a number of researchers. Recently, Kohl and colleagues used quantitative PCR (qPCR) to determine loss of viruses (orthoreovirus, vaccinia virus, sendai virus and influenza virus) from tissue after different purification and amplification procedures (Kohl *et al*., 2015). The same year, Rosseel and colleagues optimized a protocol consisting of three steps, filtration, DNase digestion and ribosomal RNA removal in serum and tissue samples spiked with Newcastle disease virus. However, this was applicable only for RNA viruses (Rosseel *et al*., 2015). The most recent study that, also mentioned in Chapter 2, aimed to validate the virome preparation protocol (NetoVIR) was done by Conceição-Neto and colleagues. This study took into account the bacterial component. They assessed the bias introduced by different steps of the protocol on a mock-virome, on 16S rRNA and also on mock-bacterial populations, to obtain the best virus-to-bacteria ratio (Conceição-Neto *et al.*, 2015). This sample preparation protocol, which can be

customized, involved steps such as homogenization, centrifugation, filtration, DNase digestion and sequence independent amplification. This protocol was shown to be efficient in minimizing bacterial cells, while enriching for virus particles, although some larger viruses like mimiviruses were lost (Conceição-Neto *et al*., 2015). Indeed, NGS-based metagenomics have transformed many fields of biological sciences. Examples include the application of next generation sequencing techniques in other aspects of virology such as investigation of quasispecies, viral evolution and analyses of antiviral resistance (Mardis, 2008; Barzon *et al*., 2011).

In this chapter, we describe the enrichment of viruses in stool samples with the NetoVIR protocol (Conceição-Neto *et al.*, 2015). We further discuss the application of metagenomics sequencing using the massively parallel Illumina MiSeq platform (https://www.illumina.com/systems/sequencing-platforms/MiSeq.html) to characterize the gut virome composition of children below the age of one year, focusing particularly on RNA viruses, as the majority of important (re)emerging viral infections are caused by RNA viruses (Cleaveland *et al*., 2001; Woolhouse and Gowtage-Sequeria, 2005; Djikeng *et al*., 2008; Jones *et al*., 2008).

3.2. Materials and methods

3.2.1. Ethics Statement and Sample Collection

This research study was conducted with the approval of the University of the Free State's Health Sciences Research Ethics Committee (HSREC) **(Ethics no.: HSREC 130/2016 B (UFS-HSD2017/0374)**. Faecal specimens were collected from infants in Oukasie clinic in the North West Province, South Africa and archived at -80 °C at the Medical Research Council, Diarrhoeal Pathogens Research Unit (MRC-DPRU), Sefako Makgatho Health Sciences University (SMU) based in Ga-Rankuwa Township, Pretoria, Gauteng Province, South Africa. The archived stool samples were then obtained from the MRC-DPRU and transferred to the Next Generation Sequencing Unit, University of the Free State, Bloemfontein, South Africa and stored at -80 °C upon arrival until they were ready for processing.

Table 3-1: Demographic data of the four study participants. A, B, C and D are different study participants.

Participant	Sample ID	Collection time/age	Stool collection date	Gender	Clinical status
A1	**4450**	6 weeks	02/07/2015	Female	Asymptomatic
A2	**6618**	10 weeks	07/08/2015	Female	Diarrhoeal
A3	**8908**	36 weeks	01/02/2016	Female	Asymptomatic
B1	**8903**	10 weeks	27/01/2016	Female	Asymptomatic
B2	**8941**	14 weeks	15/02/2016	Female	Asymptomatic
B3	**10287**	24 weeks	25/05/2016	Female	Asymptomatic
C1	**8824**	6 weeks	05/01/2016	Male	Asymptomatic
C2	**10127**	14 weeks	05/04/2016	Male	Asymptomatic
C3	**10233**	20 weeks	09/05/2016	Male	Asymptomatic
D1	**8355**	6 weeks	25/11/2015	Female	Asymptomatic
D2	**8910**	14 weeks	02/02/2016	Female	Asymptomatic
D3	**10098**	20 weeks	15/03/2016	Female	Asymptomatic

3.2.2. Sample Preparation and Viral Metagenomics Enrichment Procedure

Twelve (12) faecal samples were collected from four (4) children below the age of one year at the time between July 2015 and May 2016 (**Table 3-1**). The faecal samples were collected from each study participant (A, B, C and D) at three different time points (**Figure 3-1**; **Table 3-1**), with the mean time intervals being seven (7) weeks as a baseline, thirteen (13) weeks and twenty four (24) weeks. This type of collection was done to enable us to monitor the changes in the gut virome composition and diversity from baseline to the final collection for each of the four participants. Due to challenges pertaining to the parents being unable to bring children to the clinic on the scheduled appointment dates, a slight variation or inconsistency existed in the time intervals for sample collection among the four participants. **Figure 3-1** below summarizes the demographic data of the four participants under investigation.

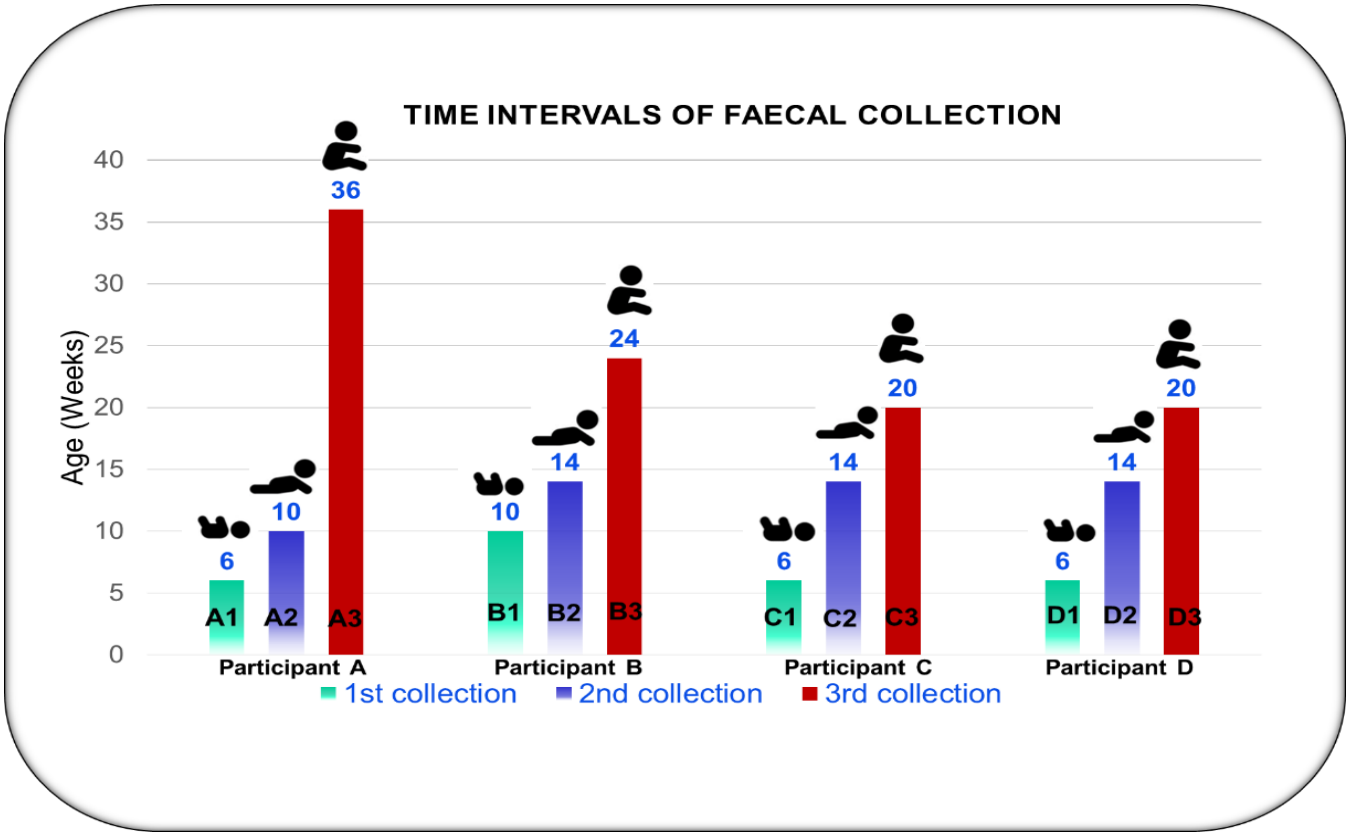

Figure 3-1: Graph depicting age and time intervals of faecal sample collection for the four participants**.**

In preparation for nucleic material extraction, all 12 faecal samples were removed from -80 °C and incubated on a sterile bench to stabilize to room temperature. The 12 samples, three from each participant were subjected to pre-treatment virome enrichment procedure to enrich for viral particles by following the **N**ovel **e**nrichment **t**echnique **o**f **VIR**omes (NetoVIR) protocol developed by the Laboratory for Viral Metagenomics, KU Leuven, Belgium (Conceição-Neto *et al*., 2015). Therefore, a 10 % faecal suspension was formulated by weighing 50 mg of stool specimen on an analytical balance (Radwag, Radom, Poland), the solid stool was placed into a 2 ml O-ring screw cap microcentrifuge tube (QSP, New Hampshire, United States) and 500 µl of freshly prepared and filter-sterilized phosphate buffered saline (PBS) (Sigma Aldrich, Missouri, United States) at pH 7.5 was added to the tube containing the stool. In the absence of an automated tissue/stool homogenizer, stool suspensions were homogenized by vigorous vortexing at maximum speed (3 400 rpm) for 5 minutes using a vortex mixer (Labnet international, Inc., New Jersey, United States) to obtain a homogenous mixture. The resulting homogenate was centrifuged at 15 000 *x g* for 5 minutes at room temperature on a benchtop micro-centrifuge (Labnet international, Inc., New Jersey, United States) to pellet cellular debris. In order to remove larger prokaryotic and eukaryotic cells, 300 µl of supernatant was collected and subsequently filtered through a 0.45 µm syringe filter

(Avacare, Johannesburg, South Africa). The filtrate (130 µl) was recovered and immediately subjected to nuclease treatment by adding to each sample 2 µl of Benzonase Nuclease (Sigma Aldrich, Missouri, United States), 1 µl of Micrococcal Nuclease (Thermo Scientific, Massachusetts, United States) and 7 µl of home-made buffer consisting of 1 M Tris (Merck, Darmstadt, Germany), 100 mM $CaCl_2$ (Sigma Aldrich, Missouri, United States) and 30 mM $MgCl_2$ (Invitrogen, California, United States), pH 8. The sample/enzyme mixture was incubated for 2 hours at 37 °C on AccuBlock Digital Dry Bath (Labnet international, Inc., New Jersey, United States) to digest unprotected nucleic acid that are expected to be floating freely in the sample following stool homogenization. Such nucleic material may originate from the host (human), bacteria, plants (diet) and other organisms. Following the 2 hour incubation, 7 µl of 10 nM EDTA (Invitrogen, California, United States) was added to each sample mix to inactivate the nuclease enzymes from further digestion of nucleic acids. The viral enriched samples were temporarily stored at -20 °C awaiting RNA extraction. **Figure 3-2** illustrates the virome enrichment procedure.

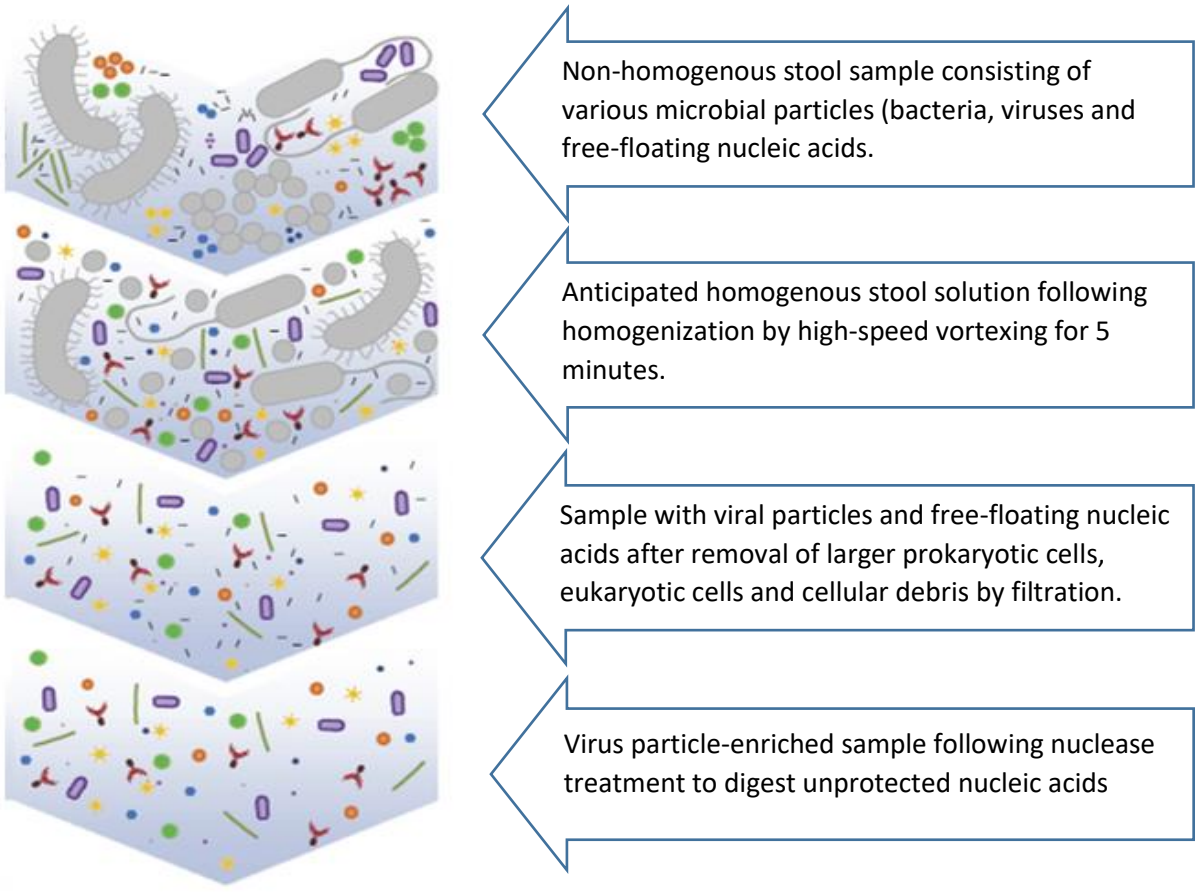

Figure 3-2: Schematic representation of the NetoVIR enrichment protocol (adpated from conceicao-Neto *et al.*, 2015).

3.2.3. Nucleic Acid Extraction

Extraction of viral RNA was performed on viral particle enriched samples using the RNeasy PowerMicrobiome RNA isolation kit (Qiagen, Hilden, Germany) with slight modifications to the manufacturer's protocol. Before starting with the experiment, a PM1/beta-mercaptoethanol (βME) solution was prepared by adding 10 µl of beta-mercaptoethanol (Merck, Darmstadt, Germany) to 990 µl of solution PM1 pre-warmed at 55 °C on AccuBlock Digital Dry Bath (Labnet international, Inc., New Jersey, United States). Solution PM1 is a lysis buffer and βME is a toxic agent that aids in the inactivation of RNases that are released during RNA isolation procedures by reducing disulfide bonds and changing the native conformation that is needed for normal functioning of the enzyme (Mommaerts *et al*., 2015). Since the starting material was a clear virus enriched liquid sample and not a solid stool sample, the first step of the original RNA extraction protocol which according to the protocol was to homogenize the stool sample in 0.1 mm glass beads suspended in a buffer, was not performed and 650 µl of PM1/βME was instead added to 147 µl of the enriched sample. The mixture was briefly vortexed (Labnet international, Inc., New Jersey, United States) to mix and incubated at room temperature for 5 minutes.

Another modification to the kits protocol was omission of homogenizing the faecal samples in the presence of beads for a virome study because bead-beating results in huge viral losses (Conceição-Neto *et al*., 2015) as the beads disrupt virus particles, thereby exposing their nucleic acids for digestion by nuclease enzymes.

Following 5 minutes incubation of the samples and PM1/ βME mixture at room temperature, 150 µl of Inhibitor Removal Solution (IRS) was added to 797 µl of the above lysed sample and mixed by brief vortexing followed by incubation at 4 °C for 5 minutes. Solution IRS was used to remove inhibitory substances that are associated with stool, such as digested food and heme from lysed red blood cells in stool, which may interfere with PCR and other downstream processes (www.qiagen.com). The tubes were centrifuged at 13,000 *x g* for 1 minute. Without disturbing the pellet, approximately 947 µl of the supernatant was removed from each sample tube and transferred to a clean 2 ml collection tubes. Solution PM3 at a volume of 650 µl and another 650 µl of solution PM4 were added to the tubes containing supernatant and vortexed briefly to mix. Solution PM3 contains the binding salts for total nucleic acid purification on-column and solution PM4 is 100 % ethanol. These solutions provide conditions needed for RNA and DNA binding to the spin filter. Thereafter, 650 µl of the supernatant was loaded onto a spin filter and centrifuged at 13 000 *x g* for 1 minute. Flow-through was removed and discarded and this was loaded three times until all the supernatant had been used up. At this step the total nucleic acids were bound to the spin filter.

Subsequently, 650 µl of a well-mixed solution PM5 which is an isopropanol combined with a wash buffer was added to the spin filter and centrifuged at 13 000 *x g* for 1 minute to remove salts from the membrane for optimal

performance of the on-column DNase step. Flow-through was discarded and the tubes were centrifuged again at 13 000 *x g* for 1 minute to remove any residual isopropanol. The spin filter baskets were placed into clean 2 ml collection tubes, after which 50 µl of DNase I solution (prepared as described above) was added. The DNase-treated spin filters were incubated at room temperature for 15 minutes to allow the DNase to soak into the membrane and digest the genomic DNA on the columns. Solution PM7 (400 µl) was added to the columns and centrifuged at 13 000 *x g* for 1 minute to inactivate the DNase enzyme and remove it from the column membrane along with digested DNA. The flow through was discarded and 650 µl of solution PM5 (isopropanol in wash buffer) was again added and centrifuged at 13 000 *x g* for 1 minute. The flow-through was discarded once more followed by addition of 650 µl solution PM4 (100 % ethanol) and centrifuged at 13 000 *x g* for 1 minute for final washing and removal of salts from the column before the RNA elution step. Flow-through was discarded and centrifuged again at 13 000 *x g* for 2 minutes to remove excess wash solution. This final dry spin ensured that all the ethanol was cleared from the membrane for efficient elution of high quality RNA. The spin filter basket was placed into a clean 2 ml collection tube. Lastly, the RNA was eluted by adding 50 µl of RNase-free water to the center of the spin filter followed by centrifugation at 13 000 *x g* for 1 minute. According to the protocol, the RNA solubilized from the spin filter membrane into RNase-free water required no further purification and was ready for use in downstream enzymatic applications. For quantification, 2 µl of the eluted RNA was used to determine the RNA concentration and also to assess its quality by an absorbance-based assay on a µlite Biodrop spectrophotometer (Biodrop, Cambridge, United Kingdom), using RNase-free water as a blank.

3.2.4. Host Ribosomal RNA (rRNA) Removal

The eluted, purified and quantified RNA (as described in section 3.2.3 above) was subjected to host ribosomal RNA (rRNA) removal to enrich for viral RNA. To achieve this, NEBNext ribosomal RNA (rRNA) depletion (Human/Rat/Mouse) kit (New England Biolabs, Massachusetts, United States) was used following the manufacturer's instructions. To describe this procedure in detail, using the RNA concentration readings obtained from Biodrop spectrophotometer (Biodrop, Cambridge, United Kingdom), a starting material ranging between 5 ng – 1 µg total RNA (DNA-free) was prepared in a 12 µl total volume. Firstly, RNA/DNA-probe reaction mix was prepared in a 200 µl tube on ice by adding 12 µl of purified RNA to a mixture consisting of 1 µl NEBNext rRNA Depletion Solution and 2 µl of Probe Hybridization Buffer. After a brief centrifugation to mix the components, the tubes were immediately placed in a thermocycler for a PCR using the following program: a heated lid set to 105 °C, denaturation at 95 °C for 2 minutes, cooling from 95 °C to 22 °C at a ramping rate of 0.1 °C per second and a final hold step at 22 °C for 5 minutes.

In this particular reaction, the rRNA-specific DNA probes contained in the NEBNext rRNA Depletion Solution hybridize to the host rRNA in the sample. Upon completion of the above hybridization reaction, the tubes were removed from the thermocycler, spun down in a tabletop centrifuge and immediately placed on ice. Following probe hybridization step, ribonuclease digestion was done whereby the RNA/DNA-probe mix were treated with RNase H enzyme which is an endoribonuclease enzyme that specifically hydrolyzes the phosphodiester bonds in the RNA that has formed a hybrid with the DNA probe molecules, and obviously this RNase H enzyme does not digest single- or double-stranded DNA (Donis-Keller, 1979; Gubbler and Hoffman, 1983; Goodwin and Rottman, 1992). For this RNase H step, the enzyme master mix consisting of 2 µl NEBNext RNase H, 2 µl RNase H reaction buffer and 1 µl nuclease-free water per reaction was prepared on ice and mixed by pipetting up and down for 10 times and added to the RNA/DNA-probe sample above, followed by pipette-mixing and brief centrifugation. The samples were placed in a thermocycler with lid heated at 40 °C and incubated at 37 °C for 30 minutes. Upon completion, the tubes were removed from the thermocycler, spun down in a tabletop centrifuge and placed on ice. In that step, the host rRNA that had hybridized to DNA probes was digested by the RNase H enzyme. The final reaction step in this rRNA depletion procedure entailed a second phase DNase I treatment to digest the DNA probes and this was also carried also out on ice as described below. DNase I digestion master mix was prepared by combining the following reagents: 5 µl DNase I reaction buffer, 2.5 µl DNase I (RNase-free) and 22.5 µl nuclease-free water. From the above mix, 30 µl was added into the RNase H treated sample in the previous step, mixed and centrifuged. The samples were incubated in a thermocycler at 37 °C for 30 minutes with lid heated at 40 °C. After incubation, the samples were spun down in a tabletop centrifuge and placed on ice.

Following the enzymatic rRNA depletion for enrichment of viral RNA, the treated RNA was purified using Agencourt RNAClean XP beads (Beckman Coulter, Indiana, United States) for removal of host rRNA and other unwanted impurities such probes and excessive enzymes by size-selection and magnetic separation, followed by an ethanol wash step. Before beginning with purification, the samples were transferred from PCR tubes to a 96-well skirted PCR plate and the Agencourt RNAClean XP beads (Beckman Coulter, Indiana, United States) were briefly vortexed to evenly re-suspend the beads in solution and 110 µl of these re-suspended beads was added to 50 µl of the above RNA sample. The solution was mixed well by pipetting up and down, making sure all the liquid was expelled out of the tip in the final mix, and incubated on ice for 15 minutes. The plate was placed on a PCR magnetic stand (PerkinElmer, Massachusetts, United States), in order to separate the beads from the solution. Once the solution was clear (after 4 minutes of incubation), the supernatant which is expected to contain the mentioned impurities was removed and discarded while making sure not to disturb the pelleted beads containing bound RNA. After all supernatant was removed, 200 µl of freshly prepared 80 % ethanol was added into each well containing the beads while in the magnetic stand and incubated at room temperature for 30 seconds to wash off

impurities, excess enzymes and digested nucleic material. After 30 seconds of incubation, the supernatant was removed and discarded making sure not to disturb the pelleted beads with bound RNA. This wash step was repeated twice, followed by a brief centrifugation and final removal of all traces of ethanol to prevent carry-over of ethanol which might interfere with downstream processes such as reverse transcription polymerase chain reaction (RT-PCR). With the plate still on the magnetic stand, the beads were air dried on the bench at room temperature for five minutes. Caution was taken not to over-dry the beads as that could result in reduced recovery of the targeted RNA. Once all the ethanol had evaporated by air-drying, the RNA was eluted by re-suspending the beads in 8 µl of nuclease free water away from the magnetic stand. The solution was carefully mixed by pipetting up and down ten times followed by two minutes incubation at room temperature. The plate was placed into the magnetic stand for five minutes to elute the target RNA molecule by magnetic separation of the beads. Once the solution was clear enough, 6 µl of the supernatant containing the eluted RNA was removed and transferred to a new 96-well PCR plate. The enriched and purified rRNA-depleted RNA samples were temporarily stored at -20 °C until ready for the synthesis of complementary DNA (cDNA).

3.2.5. Reverse Transcription and Whole Transcriptome Amplification

For library preparations purposes, the starting material was first converted to DNA. Therefore, the enriched, isolated viral RNA was reverse transcribed to generate cDNA. For synthesis of cDNA, a QIASeq FX Single Cell RNA Library Preparation Kit (Qiagen, Hilden, Germany) was used. Purified ribosomal RNA (rRNA) depleted viral RNA (input RNA amount of 60 - 100 ng) at a volume of 8 µl was transferred into a 2 ml sterile microcentrifuge tubes (all 12 samples processed in parallel), followed by addition of 3 µl NA Denaturation Buffer, the two reagents were mixed by vortexing and briefly centrifuged. The solution was incubated at 95 °C for 3 minutes, then cooling to 4 °C in a thermocycler to denature the RNA. Once the denaturation solution had cooled, 2 µl of genomic DNA (gDNA) Wipeout Buffer was added, mixed by pipetting, centrifuged briefly and incubated at 42 °C for 10 minutes to remove any genomic DNA that may be present in the RNA sample. During the 10 minutes incubation, a Quantiscript RT mix was prepared by combining the following reagents and mixed by pipetting: 4 µl RT/Polymerase Buffer, 1 µl oligo dT primer, 1 µl random primer and 1 µl Quantiscript RT enzyme mix.

After 10 minutes incubation, 7 µl of the freshly prepared Quantiscript RT mix was added to the RNA sample, mixed, centrifuged and incubated at 42 °C for 60 minutes in a thermocycler. This is the reverse transcription reaction that converted the RNA to cDNA. The reaction was stopped by subsequent incubation at 95 °C for 3 minutes and lastly cooled on ice. In preparation for whole transcriptome amplification (WTA), adapter ligation of the generated cDNA was done by adding a ligation mix consisting of 8 µl ligase buffer and 2 µl ligase mix to the Quantiscript RT mix in a PCR tube. In this reaction, the ends of cDNA strands were ligated with short oligos to enable the second PCR

reaction of cDNA amplification. The reaction was stopped by incubation at 95 °C for 5 minutes. For WTA, 30 µl of the REPLI-g SensiPhi amplification mix comprising of 29 µl REPLI-g sc Reaction and 1 µl Buffer REPLI-g SensiPhi DNA Polymerase was added to the ligation reaction from the previous step, mixed, centrifuged briefly and incubated at 30 °C for 2 hours in a thermocycler. The reaction was stopped by incubating at 65 °C for 5 minutes.

3.2.6. Quantification and Quality Control of Amplified Complementary DNA (cDNA)

3.2.6.1. Quality control of nucleic material – Spectrophotometric assay

Before proceeding to library preparation, the quality and purity of starting material was determined in order to generate good quality libraries. Therefore, the amplified transcriptome from previous steps was assessed for purity using the ratio of absorbance at 260 nm and 280 nm (A260/A280 ratio). This was done by measuring 2 µl of each of the cDNA samples on µlite Biodrop spectrophotometer (Biodrop, Cambridge, United Kingdom). A DNA sample was regarded as pure if the absorbance ratio is approximately between or near 1.8 to 2.2. All 12 samples were measured and the readings were recorded (**Table 3-5**).

3.2.6.2. Quantification of nucleic material – Fluorimetric assay

It was a prerequisite to determine the concentration of the cDNA samples before proceeding with library preparations, to be able to use the correct starting concentration. The amplified transcriptome samples were therefore quantified on Qubit fluorometer (Life Technologies, California, United States) using dsDNA High Sensitivity assay (described, herein) which is highly selective for double-stranded DNA (dsDNA).

The Qubit fluorometer (Life Technologies, California, United States) is a benchtop instrument that measures the concentration of DNA, RNA, and protein in a sample (Acar *et al*., 2009; Hamza *et al*., 2009). In contrast to UV-absorbance method using a spectrophotometer to determine absorbance of light of at 260 nm for nucleic acids, Qubit fluorometer (Life Technologies, California, United States) makes use of fluorescent dyes to quantify nucleic acids in a sample and this principle makes it more accurate. Furthermore, absorbance methods cannot distinguish between DNA, RNA, protein or free nucleotides, making them non-specific for NGS downstream processes (Manchester, 1995; Huberman, 1995; Glasel, 1995; Manchester, 1996).

In Qubit HS DNA assay, a fluorescent dye specifically binds to a DNA molecule and as soon as it binds to the target molecule, the dye becomes highly fluorescent (Schweitzer and Scaiano, 2003; McKnight *et al*., 2006). The fluorescence signal is then detected and converted into DNA concentration by the Qubit fluorometer using DNA standards with pre-determined concentration.

Figure 3-3 is a schematic representation of Qubit assay procedure. In our assay, a Qubit working solution was prepared by combining 1 µl of Qubit reagent which contains fluorescent dyes and 199 µl of Qubit buffer for each of the 12 samples and two Qubit DNA standards. A working solution was prepared for 12 samples and two Qubit standards. In total, 28 µl of Qubit reagent was added to 2 786 µl of Qubit buffer, prepared in a 15 ml conical tube. After vortexing for 30 seconds to ensure thorough mixing of the reagents, 198 µl of the freshly prepared working solution, it was transferred to each of the twelve sterile Qubit tubes (for cDNA samples) and 190 µl of the working solution was also added to each of the two sterile Qubit tubes (for Qubit standards). Furthermore, 2 µl was taken from each of the twelve cDNA samples and added to 198 µl of the freshly prepared working solution in a Qubit tubes and 10 µl of each of the Qubit DNA standards was added to 190 µl of the working solution to a total of 200 µl, respectively. All tubes were vortexed for 3 seconds and incubated at room temperature for 2 minutes, after which the readings were measured on Qubit 3.0 fluorometer. Selecting a high sensitivity dsDNA assay on the Qubit fluorometer instrument, the two standards were measured on the instrument to generate a standard graph that was used to convert the DNA fluorescence to DNA concentrations. After measuring the DNA standards, concentrations of the twelve cDNA sample were determined and recorded in nanogram per microlitre (ng/µl) units.

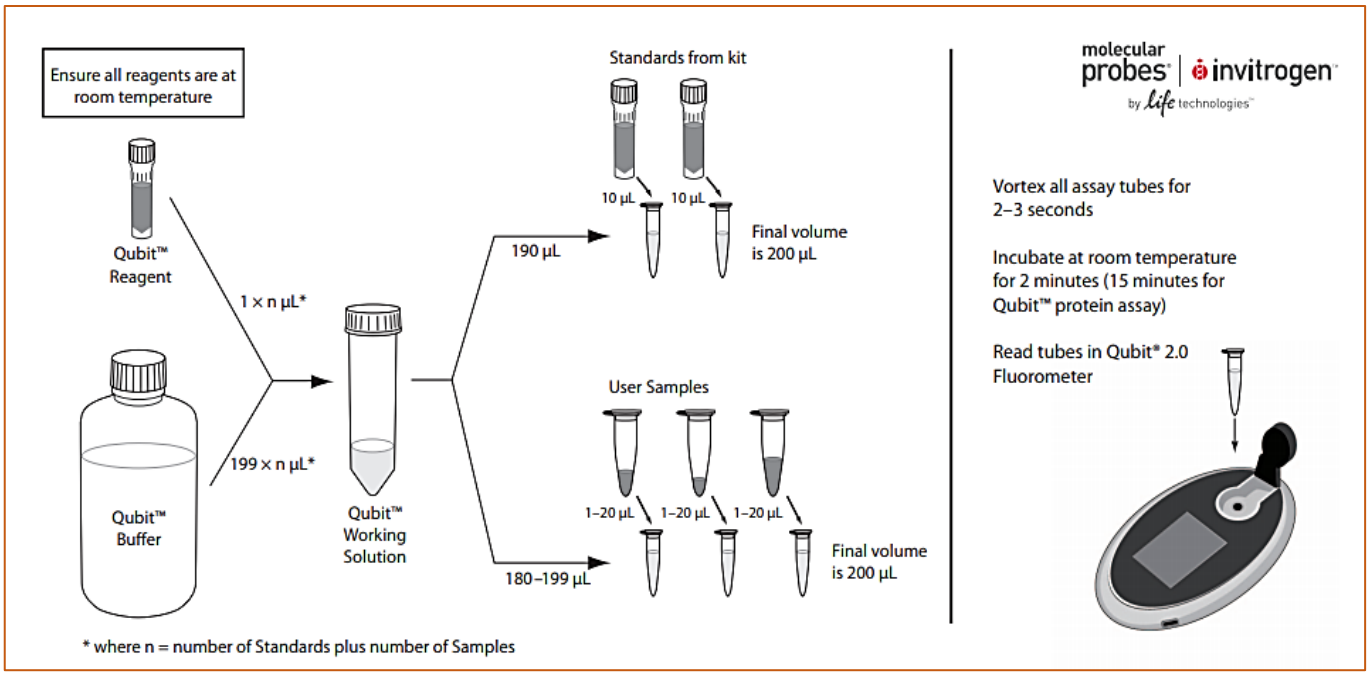

Figure 3-3: Qubit Assay procedure for dsDNA quantification (available at www.invitrogen.com/qubit)

3.2.7. Library Preparations

3.2.7.1. Overview

Library construction is the processing of starting DNA material in preparation for cluster generation and high-throughput sequencing. The procedure varies depending on the kit specification for the kind of library needed. Generally, library preparations entail fragmentation of nucleic material or sizing of the target DNA material to a desired length and this is often coupled with tagging of the incised fragments with adapters. The next step is the addition of universal sequencing primers that are complementary to the flow-cell adaptors and the incorporation of unique dual indices to both ends of the DNA fragments, a process referred to as barcoding. Moreover, purification of the barcoded libraries using magnetic beads and 80 % ethanol is done to remove PCR contaminants such as unincorporated primers/adapters, enzymes and salts. Once this is done, the libraries are quality assessed and quantified prior to being normalization to equal concentrations and pooling.

3.2.7.2. Normalization of starting DNA material

Library preparations for next generation sequencing were performed from the 12 quality assessed cDNA samples that were synthesized as described above. Nextera XT DNA Library preparation kit (Illumina, Inc., California, United States) was used for construction of NGS libraries with several modifications as per the NetoVIR protocol (Conceição-Neto *et al*., 2015). Working with Qubit readings of synthesized cDNA obtained, appropriate dilution calculations were done to adjust the obtained Qubit concentrations to an input DNA concentration of 1.2 ng/µl for all 12 samples. After dilutions, Qubit assay was performed as described above on the diluted samples to confirm the normalized library concentrations.

3.2.7.3. Fragmentation of starting DNA material

As mentioned in section 3.2.7.2 above, several modifications were made to the Nextera XT DNA library preparation protocol (Illumina, Inc., California, United States). It is important to mention that the size of DNA fragments resulting from the enzymatic fragmentation depends on the input amount of DNA, the fragmentation time and the PCR extension step. Therefore, to generate longer fragments, the DNA material used in this protocol was normalized to a concentration of 1.2 ng/µl instead of 0.2 ng/µl as recommended in the original protocol. In addition, fragmentation incubation time was reduced from 5 to 4 minutes and the PCR extension step was increased to 45 seconds. The first step of the library constructions was DNA fragmentation/tagmentation, which involved cleaving of the dsDNA while simultaneously tagging each DNA fragment with an adapter using transposome enzyme. Firstly, the reagents were removed from -20 °C and thawed on ice. Once fully thawed, each tube was inverted 5 times and centrifuged for 5 seconds. Working on a 96-well PCR plate (Soreson Bioscience Inc.,

Utah, United States), 5 µl of Tagment DNA Buffer (TD) was added to the first 12 wells the 96 well plate, followed by addition of 2.5 µl each of the 12 cDNA samples at a concentration of approximately 1.2 ng/µl. Each sample was mixed by pipetting up and down ten times. Thereafter, 2.5 µl of Amplicon Tagment Mix (ATM) was then added to each well of the sample and mixed by pipetting. The PCR plate was sealed with a Microseal 'A' film (Biorad, California, United States) and centrifuged at 280 *x g* at 20 °C for 1 minute (Hermle, Labnet international, Inc., New Jersey, United States) and immediately transferred to a thermal cycler (MULTIGENE OPTIMAX, Labnet international, Inc., New Jersey, United States) to start the tagmentation reaction by incubation for 4 minutes at 55 °C minutes with a lid pre-heated at 60 °C, and cooled at 10 °C. Once completed, the PCR plate was removed from the thermocycler and tagmentation reaction was stopped by adding 2.5 µl of Neutralize Buffer (NT) which neutralizes the enzyme that fragmented the template DNA. The PCR plate was again sealed with a Microseal 'A' film (Biorad, California, United States) again centrifuged at 280 *x g* at 20 °C for 1 minute followed by 5 minutes incubation at room temperature.

3.2.7.4. Index PCR: Barcoding

This step of the library preparation procedure introduced dual indices or barcodes which were 8 base long DNA oligonucleotides, to uniquely mark each sample prior to multiplexing. In addition to indexing, sequencing primers that were compatible with Illumina platforms were also introduced to both ends of the DNA fragments.

Prior to indexing, a sample sheet was created for all 12 samples using Illumina Experiment Manager software version 1.15 (Illumina, Inc., California, United States). This was to enable the identification and separation of pooled samples using their unique index sequences and also included information regarding the workflow that was executed by the sequencing instrument and the read length. To the tagmented DNA above, 2.5 µl of index primer 1 (N70X), 2.5 µl of index primer 2 (S50X) and 7.5µl Nextera PCR master mix (NPM) were added to 12.5 µl of each of the samples based on the unique combinations created for each sample as tabulated in **Table 3-2**. To avoid index cross-contamination, the old caps of the index tubes were discarded and replaced with new sterile ones. The reactions were mixed by pipetting up and down 10 times, the PCR plate was sealed with a Microseal 'A' film (Biorad, California, United States) and centrifuged at 280 *x g* at 20 °C for one minute.

Table 3-2: Sample sheet with unique Nextera XT index combination of the 12 samples.

Sample number	Sample name	Index 1	Index 2
A1	**4450**	N709	S504
A2	**6618**	N709	S505
A3	**8908**	N709	S506
B1	**8903**	N709	S507
B2	**8941**	N709	S508
B3	**10287**	N709	S517
C1	**8824**	N710	S502
C2	**10127**	N710	S503
C3	**10233**	N710	S504
D1	**8355**	N710	S505
D2	**8910**	N710	S506
D3	**10098**	N710	S507

PCR reaction was then performed on a thermal cycler under the following reaction conditions: 72 °C for 3 minutes, 95 °C for 30 seconds, 15 cycles of 95 °C for 10 seconds, 55 °C for 30 seconds, 72 °C for 45 seconds and a hold at 4 °C. Immediately after the reaction had ended, a second clean-up of index PCR was performed using AMPure XP beads (Beckman Coulter, Indiana, United States) following the same procedure as described above.

3.2.7.5. PCR Clean Up

After indexing, a post-PCR clean-up step was necessary to remove any unwanted products or PCR contaminants such as residual enzymes and unincorporated index primers present in the samples. To do this, PCR clean-up was performed using AMPure XP magnetic beads (Beckman Coulter, Indiana, United States), which in addition to library purification, also size selected the desired fragment sizes by filtering low and high molecular weight fragments.

AMPure XP beads (Beckman Coulter, Indiana, United States) were removed from 4 °C and allowed 30 minutes to normalize to room temperature before use. Indexed libraries (50 µl) in a PCR plate were centrifuged at 280 *x g* for 1 minute at room temperature using Prism centrifuge (Labnet international, Inc., New Jersey, United States) so as to collect condensation. The AMPure XP beads (Beckman Coulter, Indiana, United States), at room temperature,

were vortexed for 30 seconds using a vortex mixer (Labnet international, Inc., New Jersey, United States) for even distribution of the beads in solution. Following this, a sufficient volume of beads was added to a trough (SPL Life Sciences, Naechon-Myeon, Korea), from which 30 µl of beads were collected using a P200 multichannel pipette (due to viscous solution) and added to each sample well of the PCR plate, resulting in a total volume of 80 µl. The plate was then sealed using Microseal 'A' film (Biorad, California, United States) followed by shaking at 1 800 rpm for 2 minutes using Micro-Multiple Genie (Scientific Industries Inc., New York, United States). This was incubated at room temperature for 5 minutes without shaking. This step allows the genomic material to bind to the magnetic beads through their negatively charged DNA molecules. Following incubation, the plate was placed on a magnetic stand until the supernatant became clear. It was very important that as soon as the plate was placed on the magnetic stand, the beads with bound genomic material were attracted to the side surface of the wells. The purpose of this was to isolate the genomic material from the rest of the solution with impurities. The supernatant which now contained the unwanted impurities was removed and discarded using a multichannel pipette set at 100 µl.

Ethanol wash step was performed by adding 200 µl of freshly prepared 80 % ethanol to the beads while still on the magnetic stand followed by 30 seconds incubation on the magnetic stand. The magnetic beads with bound DNA were washed with 80 % ethanol to remove impurities; DNA is not soluble in ethanol due to its chemical properties, therefore only other molecule apart from DNA was washed away from the beads. Thereafter, the supernatant was removed and discarded using a pipette set to 200 µl. This step was repeated twice followed by the subsequent removal of excessive ethanol. The beads were allowed to air-dry for 15 minutes while still on the magnetic stand after which the plate was removed from the stand. Using a P200 multichannel pipette, 52.5 µl of resuspension buffer (RSB) buffer collected from a trough was added to each sample well (DNA is soluble in RSB). The plate was sealed by Microseal 'A' film (Biorad, California, United States) and shaking at 1800 rpm was done for 2 minutes followed by incubation at room temperature for 2 minutes. The plate was then placed on a magnetic stand until the supernatant became clear. Finally, 50 µl of supernatant which contained the eluted DNA was removed from each of the sample wells and transferred to a new labelled 96-well PCR plate, sealed and stored in at -20 °C until the next step.

3.2.7.6. Quantification using HS dsDNA Assay by Qubit 3.0

Qubit quantification was performed to determine library concentrations after indexing and clean-up steps. This was done using a High Sensitivity (HS) dsDNA Assay kit on a Qubit 3.0 Fluorometer instrument.

3.2.7.7. Quality Validation using the Agilent 2100 Bioanalyzer

After attaching indices to uniquely mark each of the libraries, clean-up was done. However, this was not sufficient to conclude on the quality of the prepared libraries. Validation was required prior to sequencing to ensure good quality data is generated at the end. In addition, the validation also gave an idea of the library fragment size distribution. The library sizes wee important as they were used in calculations to normalize libraries to an equimolar concentration. Library validation and quality assessment was therefore done by running 1 µl of each of the 12 samples on an Agilent 2100 Bioanalyzer (Agilent technologies, California, United States).

A full report was generated for each run showing the electropherogram with fragment size distribution, gels with amplicon sizes or smear, and other important results.

3.2.7.8. Library Normalization to 4nM

Prior to sequencing, all indexed and cleaned libraries were pooled into one tube because sequencing bias could be created if pooled libraries had varying concentrations. Therefore, libraries were normalized to equimolar concentrations prior to pooling. A special formula, shown below, was used to firstly determine library concentrations in nanomolar (nM) units, and this formula took into consideration the average library size as determined on the Agilent Bioanalyzer as well as the Qubit concentrations. Further dilution to a desired concentration of 4 nM, or 2 nM for libraries that fell below the targeted 4nM was then performed.

$$\frac{\boldsymbol{Concentration\ in\ ng/\mu l}}{(\boldsymbol{660g/mol\ x\ Average\ library\ size})}\ \boldsymbol{X\ 10^6 = Concentration\ in\ nM}$$

All libraries were expected to have varying nanomolar concentrations and as a result, they were normalized to 4 nM before pooling.

3.2.7.9. Pooling of Libraries

In preparation for cluster generation and Illumina sequencing, all libraries which were in equal concentrations were pooled together into a single 1.5 ml microfuge tube by taking 5 µl of each library and transferring it into one tube. The pooled library was mixed by gently pipetting up and down several times. The pooled libraries could not contaminate because each had a unique index (barcode) which the MiSeq software would use to de-multiplex upon completion of the sequencing run. The single, pooled library was now ready for chemical denaturation and dilution using sodium hydroxide (NaOH) as discussed in 3.2.7.10.

3.2.7.10. Library Denaturing and Dilution

A fresh dilution of 0.2 N NaOH (Sigma Aldrich, Missouri, United States) was prepared and 5 µl of the 4 nM final pooled and well mixed library was added to 5 µl of freshly prepared 0.2 N NaOH (Sigma Aldrich, Missouri, United States) in a clean 1.5 ml microcentrifuge tube to result in a total volume of 10 µl. The mixture was vortexed briefly and centrifuged at 280 *x g* for 1 minute on a mini centrifuge (Labnet international, Inc., New Jersey, United States), followed by 5 minutes incubation at room temperature to denature DNA into single strands, while simultaneously diluting the DNA library to 2 nM. Once denatured, 990 µl pre-chilled hybridization buffer HT1 (Illumina, Inc., California, United States) was added to 10 µl of the 4 nM denatured DNA library to further dilute it to 20 pM. The 20 pM library was placed on ice until it was ready for dilution to final loading concentration of 10 pM. To get to the desired final concentration of 10 pM, a further dilution of the denatured library was done by combining 300 µl of the 20 pM library with 300 µl pre-chilled HT1 and mixed by inverting several times.

3.2.7.11. PhiX Control Denaturing and Dilution

PhiX (Illumina Inc., California, United States), which is a bacteriophage genome used as a positive control in Illumina sequencing runs was also diluted and denatured to a final loading concentration of 20 pM for optimal cluster density. This was done by mixing 2 µl of the 10 nM PhiX library and 3 µl of EB (elution) Buffer (Qiagen, Hilden, Germany) as an equivalent of 10 nM Tris-HCl, pH 8.5. This resulted in 5 µl of a 4 nM PhiX which was then denatured by adding 5 µl 0.2N NaOH, centrifuged at 280 *x g* for 1 minute and incubated at room temperature for 5 minutes. A further dilution to 20 pM was done by adding 990 µl pre-chilled HT1 buffer to 10 µl of 4 nM denatured PhiX library.

3.2.7.12. Combining Library and PhiX Control

In order to create diversity within the library, a PhiX control spike-in of 15 % was used by combining 90 µl of 20 pM denatured and diluted PhiX with 510 µl of 10 pM denatured and diluted library, giving rise to a final volume of 600 µl. This final library was incubated on ice until ready for final heat denaturation and loading.

3.2.8. Cluster Generation and Illumina Sequencing

This was the final step where the final library spiked with 15 % PhiX was loaded onto the sequencing cartridge. The V3 reagent cartridge, with a standard flow cell generating up to 25 million paired-end reads, which had been thawed in de-ionized water for 90 minutes to defrost the reagents was removed and dried completely by gently tapping it on a paper towel. A new flow cell cleaned with nuclease-free water, a PR2 buffer and a waste bottle were inserted into their respective compartments in the MiSeq instrument following the prompts from the MiSeq Control Software. The final library was heat denatured for two minutes at 96 °C on heating block and immediately

incubated on ice. The foil seal covering the reservoir labelled “Load Samples” was cleaned with a lint-free tissue and pierced with a 1 000 µl pipette tip prior to sample loading. The denatured pooled library (600 µl) consisting of 15 % PhiX control was transferred into the appropriate reservoir labelled “Load Samples” in the V3 sequencing cartridge. Upon loading the final library-PhiX mixture, the cartridge was then inserted into its compartment on a MiSeq instrument (Illumina Inc., California, United States). After reviewing and confirming all parameters, the run was started and allowed to proceed for 500 cycles to generate 2 x 250 bp paired-end reads.

Bridge amplification is the method used by Illumina platforms to generate millions of clusters prior to sequencing. In this process (illustrated in **Figure 3-4**), millions of DNA strands were hybridized to numerous complementary primers fixed on the flow cell surface (https://www.illumina.com). Using high-fidelity DNA polymerase, the strands were synthesized from the hybridized primers by 3’ extension. The initial strands were denatured, leaving the complementary strands fixed on the flow cell, which were then amplified by isothermal bridge amplification. The strands looped to hybridize to nearby primers.

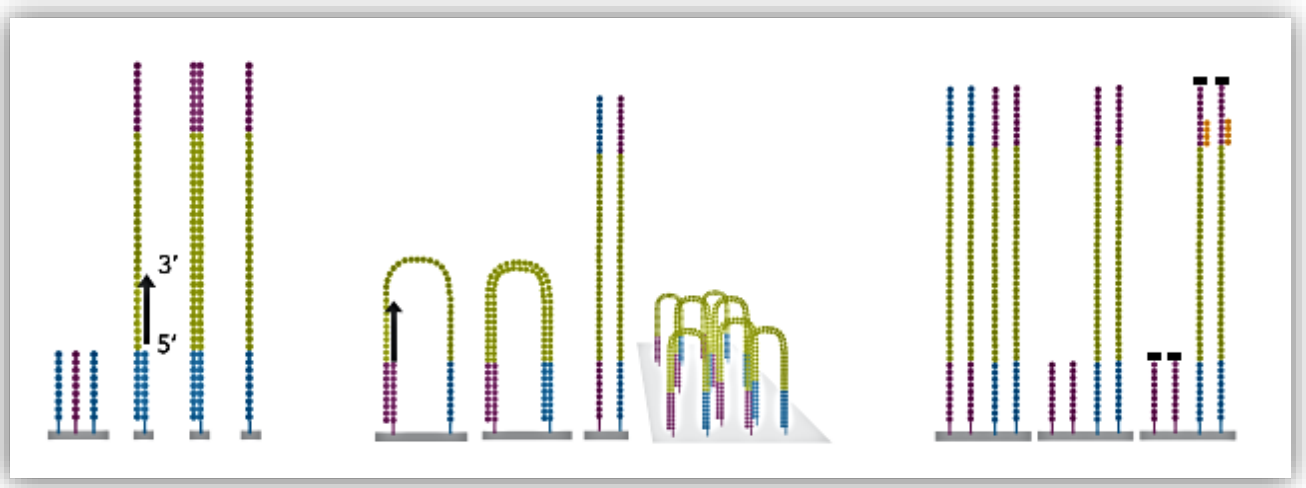

Figure 3-4: Cluster generation by bridge amplification on Illumina platforms. During clustering, each fragment molecule is isothermally amplified on a flow-cell, which is a glass slide with lanes coated with a lawn of oligos. DNA fragments with adapters binds to complementary flow cell oligos and are clonally amplified by bridge amplification. (Adapted from https://www.illumina.com/documents/products/datasheets/datasheet_cbot.pdf)

The process was repeated multiple times to create millions of dense clonal clusters. Ultimately, the flow cell contained millions of clusters ready for sequencing (https://www.illumina.com). Once clusters were generated, sequencing-by-synthesis commenced in a massively parallel way to produce millions of short (250 bp) sequence reads (paired-end) for the 500 cycles.

3.3. Results

3.3.1. Virome enrichments

Faecal specimens were treated by applying the NetoVIR procedure to enrich for virus particles in human faeces. The viral enriched samples were then used for RNA extraction. Host ribosomal RNA was removed from the extracted RNA using a commercial kit. Concentration and the integrity of extracted RNA were determined by spectrophotometry.

3.3.2. Viral RNA extraction

The virome enriched samples were used to extract RNA using the RNeasy PowerMicrobiome Kit (Qiagen, Hilden, Germany) with several modifications as described in the methodology. **Table 3-3** and **Table 3-4** show the quantification and quality assessment results of the extracted RNA prior and post rRNA depletion.

Table 3-3: RNA concentration readings and A260/A280 ratio determined on Biodrop (Biodrop, Cambridge, United Kingdom) before rRNA depletion for all twelve samples.

Sample ID	Participant ID	RNA Concentration (ng/µl)	A260/280 Ratio
A1	**4450**	9.41	1.74
A2	**6618**	348.80	2.10
A3	**8908**	10.35	2.77
B1	**8903**	563.30	2.12
B2	**8941**	10.77	2.40
B3	**10287**	68.68	2.10
C1	**8824**	2.02	1.93
C2	**10127**	7.86	2.12
C3	**10233**	9.93	1.95
D1	**8355**	35.10	2.09
D2	**8910**	49.03	2.10
D3	**10098**	85.41	2.30

RNA could be extracted from all 12 samples although some had very low quantity such as sample C1 which was the lowest with RNA concentration of 2.02 ng/µl and the highest was sample B1 with a concentration of 563 ng/µl. For quality control using the ratio of absorbance at wavelengths at 260 nm and 280 nm, the ratio of all 12 samples ranged from a minimum of 1.74 to 2.77, however the average was 2.14 which implies that majority of the RNA samples were of good quality with only three samples having a ratio higher than 2.12. Based on the readings obtained, the integrity of RNA was acceptable to proceed to the next step.

3.3.3. Ribosomal RNA depletion

Ribosomal RNA was removed from the extracted total RNA to enrich for viral RNA using NEBNext rRNA depletion kit (Human/Mouse/Rat) (New England Biolabs, Massachusetts, United States). Following this step, quantification of the treated samples was performed to determine the concentration and quality of the RNA. **Table 3-4** shows the concentration readings and absorbance ratio obtained at 260 nm and 280 nm wavelengths.

Table 3-4: RNA concentration readings and A260/A280 ratio determined on Biodrop spectrophotometer (Biodrop, Cambridge, United Kingdom) post rRNA depletion for all twelve samples.

Sample ID	Participant ID	RNA Concentration (ng/µl)	A260/280 Ratio
A1	**4450**	5.81	1.73
A2	**6618**	105.40	2.02
A3	**8908**	3.73	2.8
B1	**8903**	598.0	2.14
B2	**8941**	2.19	2.90
B3	**10287**	78.24	2.17
C1	**8824**	1.26	1.94
C2	**10127**	2.63	2.45
C3	**10233**	1.49	2.38
D1	**8355**	31.5	2.01
D2	**8910**	34.27	2.2
D3	**10098**	1.29	2.23

As expected the RNA concentration decreased after removal of host rRNA from the RNA samples, except for samples B1 and B3 where an increase in RNA concentration was observed. Furthermore, more than a 50 % reduction in RNA concentration was observed in six out of the twelve samples. This is indicative of high abundance of rRNA in the stool samples. The reason for a slight increase in RNA concentration for the two samples could the due to overestimation by absorbance-based method during the first quantification, handling errors or inefficient removal of impurities during the clean-up after rRNA depletion using magnetic beads. These impurities might have

been quantified since the absorbance-based method measures nucleic acids non-specifically. Nevertheless, the quality of the RNA was good for all samples.

3.3.4. Reverse Transcription and Whole Transcriptome Amplification

In order to prepare libraries for next generation sequencing, the extracted RNA needs to be converted to cDNA. Therefore reverse transcription was performed to synthesize complementary DNA, from which the whole transcriptome was subsequently amplified. Concentrations of the amplified cDNA products and the quality thereof were determined on Biodrop spectrophotometer (Biodrop, Cambridge, United Kingdom) and Qubit (Life Technologies, California, United States). The obtained results are provided in **Table 3-5 and Table 3-6**.

Table 3-5: Quantification and quality assessment of amplified cDNA on Biodrop spectrophotometer (Biodrop, Cambridge, United Kingdom).

Sample ID	Participant ID	cDNA Concentration (ng/µl)	A260/280 Ratio
A1	**4450**	3000.2	1.78
A2	**6618**	2755.3	1.92
A3	**8908**	3052.7	1.78
B1	**8903**	2977.5	1.82
B2	**8941**	2995.3	1.79
B3	**10287**	3015.4	1.79
C1	**8824**	2936.1	1.87
C2	**10127**	3101.1	1.75
C3	**10233**	3091.6	1.79
D1	**8355**	2897.2	1.81
D2	**8910**	2984.1	1.79
D3	**10098**	3011.7	1.82

Table 3-6: Quantification of amplified transcriptome on Qubit (Life Technologies, California, United States).

Sample ID	Participant ID	Amplified cDNA concentration after 100x dilution (ng/µl)
A1	**4450**	10.5
A2	**6618**	37.8
A3	**8908**	12.9
B1	**8903**	11.8
B2	**8941**	11.6
B3	**10287**	9.22
C1	**8824**	6.32
C2	**10127**	14.2
C3	**10233**	4.3
D1	**8355**	4.71
D2	**8910**	13.2
D3	**10098**	10.9

3.3.5. Library Preparations

The amplified cDNA samples were used as starting material for library preparations using Nextera XT DNA library preparation kit (Illumina Inc., California, United States). In order to produce good quality libraries of optimal and desired fragment sizes, (as recommended in the original Nextera XT protocol), the starting DNA material should be normalized to an input amount of 1 ng using 5 µl, meaning that samples would be diluted to a concentration of 0.2 ng/µl. However, a modified version of this protocol was adopted (NetoVIR protocol) in which the samples were normalized to a concentration of 1.2 ng/µl instead of 0.2 ng/µl.

3.3.5.1. Normalization of starting cDNA material

Table 3-7 below shows the dilution calculations used to normalize the samples to a desired concentration prior to library preparations.

Table 3-7: Dilution calculations of amplified transcriptome to 1.2 ng/µl.

Sample ID	Participant ID	cDNA concentration (Obtained from Qubit) in ng/µl	Desired final concentration of cDNA after dilution in ng/µl	Volume of undiluted cDNA sample in µl	Volume of cDNA sample after dilution in µl	Volume of buffer to be used as a diluent in µl
A1	**4450**	10.5	1.4	2.7	20	17.3
A2	**6618**	37.8	1.4	2.2	60	17.8
A3	**8908**	12.9	1.4	2.2	20	17.8
B1	**8903**	11.8	1.4	2.4	20	17.6
B2	**8941**	11.6	1.4	2.4	20	17.6
B3	**10287**	9.22	1.4	3.0	20	17.0
C1	**8824**	6.32	1.4	2.2	10	17.8
C2	**10127**	14.2	1.4	2.0	20	18.0
C3	**10233**	4.3	1.4	3.2	10	16.8
D1	**8355**	4.71	1.4	3.0	10	17.0
D2	**8910**	13.2	1.4	2.1	20	17.9
D3	**10098**	10.9	1.4	2.6	20	17.4

After normalizing the cDNA samples to a concentration of 1.2 ng/µl, Qubit assay was again performed to confirm the normalized concentrations. The concentrations obtained after normalization were recorded in **Table 3-8**.

Table 3-8: Quantification by Qubit to confirm the normalized samples.

Sample ID	Participant ID	Targeted cDNA concentration after dilution in ng/µl	Obtained concentration of cDNA after dilution in ng/µl
A1	**4450**	1.20	1.24
A2	**6618**	1.20	1.23
A3	**8908**	1.20	1.29
B1	**8903**	1.20	1.26
B2	**8941**	1.20	1.19
B3	**10287**	1.20	1.19
C1	**8824**	1.20	1.15
C2	**10127**	1.20	1.25
C3	**10233**	1.20	1.28
D1	**8355**	1.20	1.24
D2	**8910**	1.20	1.20
D3	**10098**	1.20	1.30

3.3.5.2. Tagmentation of normalized cDNA

Using the cDNA samples normalized to a concentration of approximately 1.2 ng/µl, libraries were prepared for next generation sequencing. The first step of the protocol was tagmentation, in which the starting cDNA material was enzymatically sheared while simultaneously tagging the produced shorter fragments with adaptors. The tagmentation step was then followed by indexing of the individual libraries.

3.3.5.3. Quantification of Fragmented and Indexed Libraries: HS dsDNA Assay by Qubit 3.0

The uniquely marked or indexed libraries were purified using magnetic beads and the concentration of the DNA libraries was determined on Qubit fluorometer. In order to proceed with downstream processes of the library preparations, it was very important to have good quality libraries. **Table 3-9** shows the obtained concentrations of barcoded libraries as determined on Qubit.

Table 3-9: Concentrations of DNA libraries measured on Qubit after library indexing and clean-up.

Sample ID	Participant ID	DNA library concentration after indexing in ng/µl
A1	**4450**	4.74
A2	**6618**	4.01
A3	**8908**	3.56
B1	**8903**	5.62
B2	**8941**	3.25
B3	**10287**	2.93
C1	**8824**	2.79
C2	**10127**	2.74
C3	**10233**	6.02
D1	**8355**	3.39
D2	**8910**	4.90
D3	**10098**	3.25

3.3.5.4. Quality Validation: Bioanalyzer

Prior to sequencing on an NGS platform, several quality control steps were required to ensure that the quality of prepared libraries were sufficiently good to avoid generating poor quality data or complete run failure. Thus, each of the individual libraries were validated on Agilent Bioanalyzer (Life Technologies, California, United States) to assess the quality of the libraries by determining the fragment size distribution after low molecular weight fragments and high molecular weight fragments were eliminated during the magnetic beads-based clean up. **Figure 3-5** shows a Bioanalyzer profile of the indexed and cleaned libraries, which is represented in the form of an electropherogram (peaks) and a gel image. Note that only 11 samples can be analysed at one time and results are generated in real-time. One sample was analysed separately (image not shown), although the fragment distribution was consistent with those shown in **Figure 3-5**.

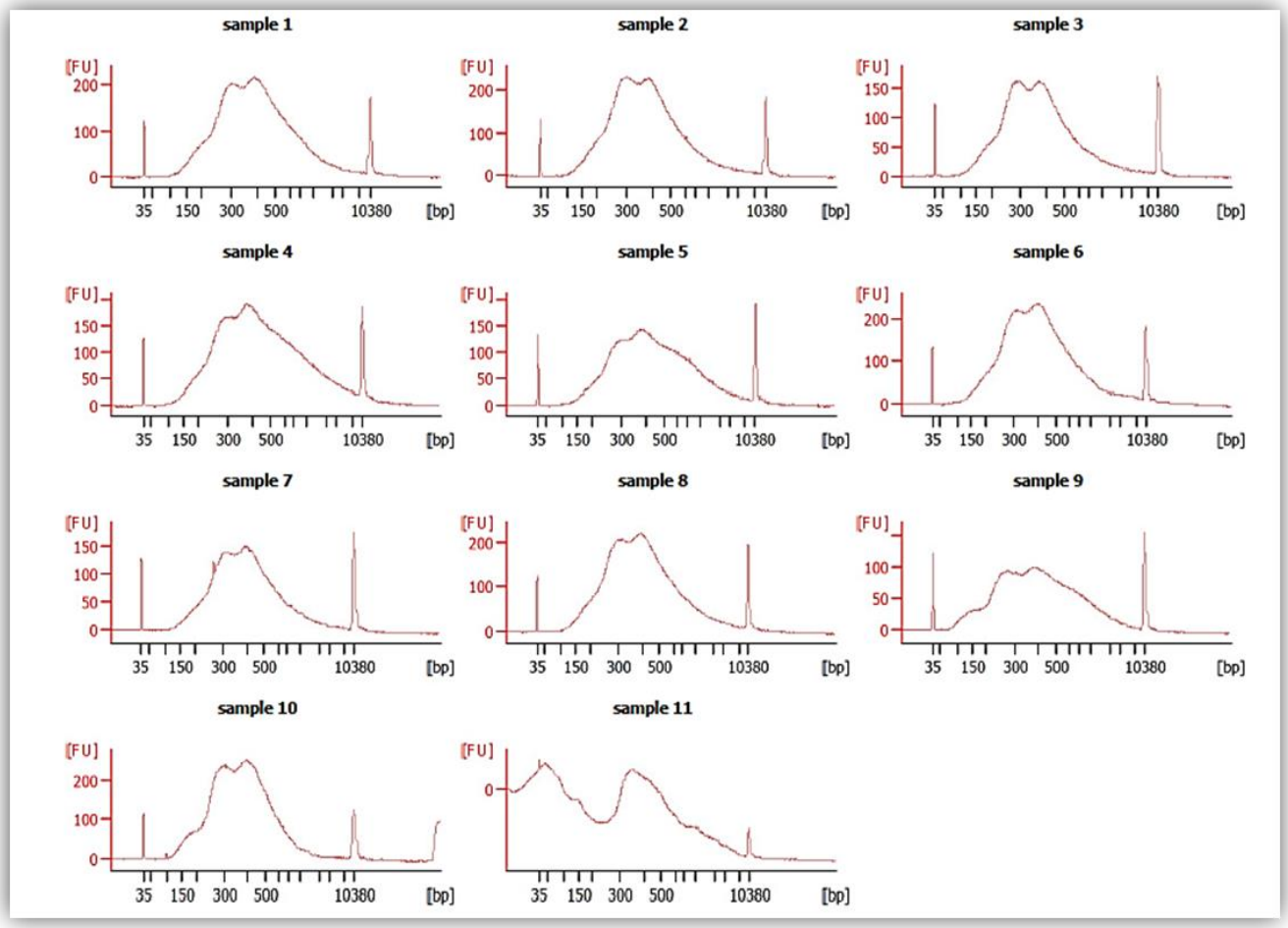

Figure 3-5: Bioanalyzer electropherogram showing library size distribution determined by high sensitivity dsDNA assay.

The electropherogram shows an average library size of around 450 bp for all samples validated. This was also shown on the gel image below with the library products enclosed in a yellow border. However, the library fragments range from around 150 bp to slightly over 600 bp. **Table 3-10** shows average library size for each sample.

Table 3-10: Average library size in base-pairs from Bioanalyzer validation.

Library	Sample ID	Average library size (bp)
A1	**4450**	416
A2	**6618**	394
A3	**8908**	415
B1	**8903**	442
B2	**8941**	414
B3	**10287**	429
C1	**8824**	436
C2	**10127**	371
C3	**10233**	423
D1	**8355**	390
D2	**8910**	412
D3	**10098**	429

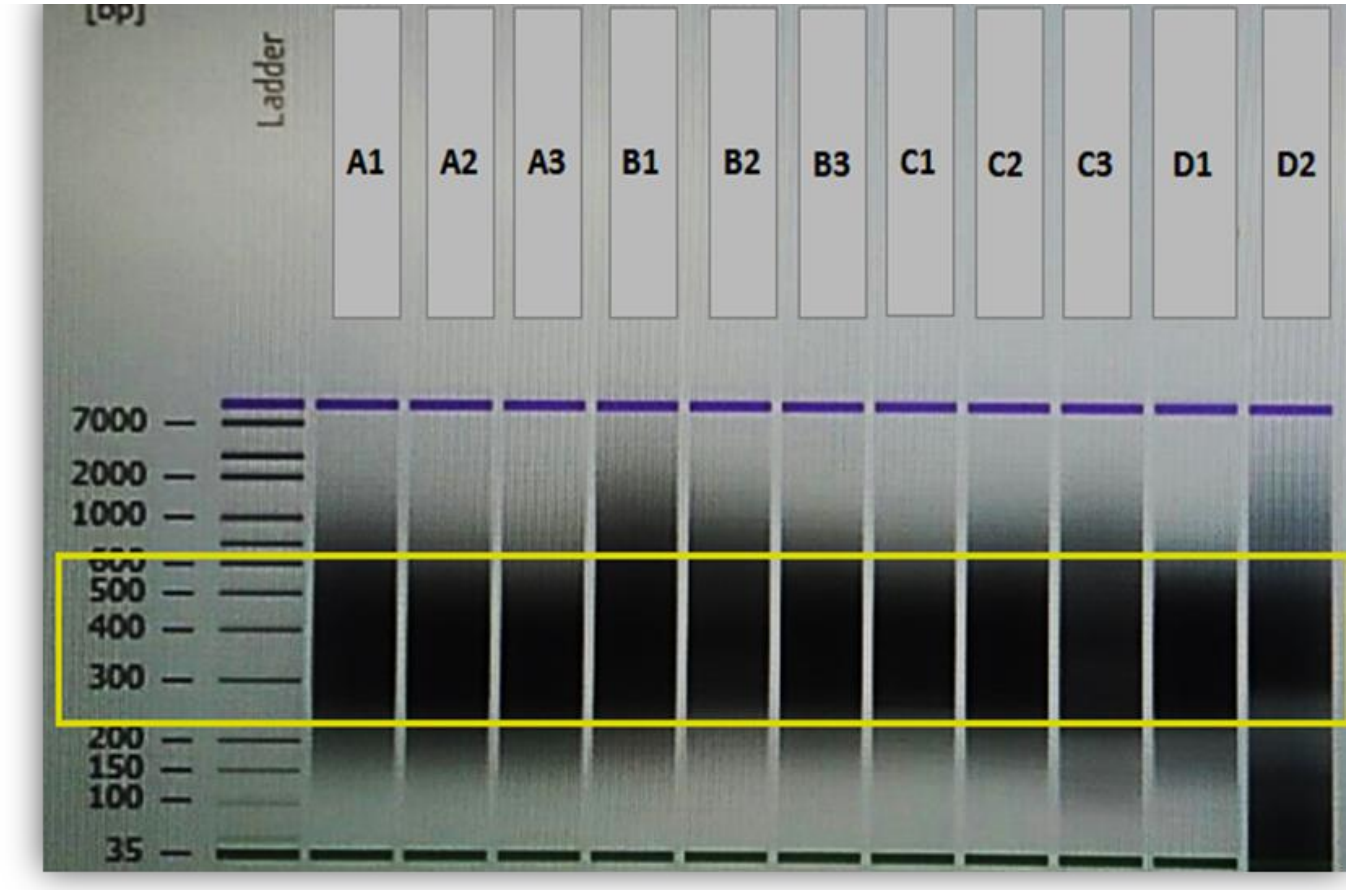

Figure 3-6: Bioanalyzer gel image showing the library size distribution.

3.3.5.5. Library Normalization to 4nM

Determination of DNA library concentration in nM

The formula illustrated below was used to determine the library concentrations in nanomolar. The obtained concentration readings were recorded in **Table 3-11**.

$$\frac{\boldsymbol{Concentration\ in\ ng/\mu l}}{(\boldsymbol{660g/mol\ x\ Average\ library\ size})}\ \boldsymbol{X\ 10^6 = Concentration\ in\ nM}$$

Table 3-11: Concentration of validated libraries in nanomolar (nM).

Sample ID	Participant ID	DNA library concentration after indexing in ng/µl	Average library library size in bp from Bioanalyzer	DNA library concentration after indexing in nM
A1	**4450**	4.74	416	17.3
A2	**6618**	4.01	394	15.4
A3	**8908**	3.56	415	13.0
B1	**8903**	5.62	442	19.3
B2	**8941**	3.25	414	11.9
B3	**10287**	2.93	429	10.3
C1	**8824**	2.79	436	9.7
C2	**10127**	2.74	371	11.2
C3	**10233**	6.02	423	21.6
D1	**8355**	3.39	390	13.2
D2	**8910**	4.90	412	18.0
D3	**10098**	3.25	429	11.5

Although all samples were normalized to nearly the same concentration prior to tagmentation, slight variations in concentrations were however expected after indexing and magnetic bead clean up. As a result, the libraries were normalized again to an equal concentration of 4 nM before multiplexing. This was done by diluting the samples in nuclease-free water using the formula illustrated below using at least 2 µl of each of the DNA libraries in every dilution. **Table 3-12** shows the concentration readings and volumes of DNA libraries and diluent used in the calculations.

$$\boldsymbol{C1.V1 = C2.V2}$$

Table 3-12: Normalization of libraries after indexing and clean up.

Sample ID	Participant ID	cDNA concentration (Obtained from Qubit) in nM	Desired final concentration of cDNA after dilution in nM	Volume of undiluted cDNA sample in µl	Final volume of cDNA sample after dilution in µl	Volume of nuclease-free water to be used as a diluent in µl
A1	**4450**	17.3	4	2.3	10	7.7
A2	**6618**	15.4	4	2.6	10	7.4
A3	**8908**	13.0	4	3.1	10	6.9
B1	**8903**	19.3	4	2.1	10	7.9
B2	**8941**	11.9	4	3.4	10	6.6
B3	**10287**	10.3	4	3.9	10	6.1
C1	**8824**	9.7	4	4.1	10	5.9
C2	**10127**	11.2	4	3.6	10	6.4
C3	**10233**	21.6	4	2.8	10	12.2
D1	**8355**	13.2	4	3.0	10	7.0
D2	**8910**	18.0	4	2.2	10	7.8
D3	**10098**	11.5	4	3.5	10	6.5

3.3.6. Cluster Generation and Illumina Sequencing

Sequencing on Illumina MiSeq platform (Illumina, Inc., California, United States) to produce 250 bp x 2 paired end reads was completed without errors. Preliminary analysis showed a good quality metagenomics data was generated, with a quality score (Q-score) of more than 30 for nearly 80 % of the bases. Q-score is the probability of an error in base calling, therefore Q score of 30 denotes an error of 1 in 1000 bases, translating to 99.9 % base calling accuracy. A cluster density of 1014 K/mm2 was achieved, with 95.1 % of clusters passing filter. In summary, the above sequencing parameters indicate a successful library preparation and NGS run with optimal clustering and good quality preliminary data.

3.4. Discussion

In theory, metagenomic analysis can be applied on any environmental and/or clinical sample including but not limited to marine water, blood, or faecal samples (Breitbart *et al*., 2003; Breitbart, 2005; Angly *et al*., 2006; Zhang *et al*., 2006; Breitbart *et al*., 2008; Li *et al*., 2009; Li *et al*., 2010). Bacterial and eukaryotic genomes are much larger compared to viral genomes which are relatively smaller. Consequently, bacterial and host genome interfere with the isolation and detection of virus nucleic material which is usually low in abundance (Mokili *et al*., 2013). Thus, for efficient recovery of viral genetic material, enrichment is always inevitable which basically aims to remove non-viral nucleic acids. To obtain better results, proper homogenization of faecal samples is always necessary as this step ensures even distribution of microbial particles in the sample (**Figure 3-2**). Although this was achieved by high speed vortexing which apparently does not provide a standardized approach for sample homogenization, visual observation of the faecal suspensions (prepared in 10 % PBS) showed thorough homogenization with no noticeable solid particles. Furthermore, it has been shown in literature that homogenization of stool in the presence of beads results in severe loss of viruses due to destruction of virus particles by beads, thereby releasing their nucleic material into the solution (Conceição-Neto *et al.*, 2015). Not only do they destroy virus particles, but homogenization with beads leads to massive increase in bacterial genome (Conceição-Neto *et al.*, 2015). In this study, homogenization was therefore performed in the absence of beads to avoid viral loss.

Furthermore, it is imperative that the centrifugation conditions, namely the time and speed, were cautiously chosen as they have a huge impact on the reduction in bacterial and viral particles. In this study, the resulting homogenate was centrifuged to precipitate the larger cell-sized particles and cellular debris while the viruses remain in solution. The centrifugation conditions (15 000 *x g* for 5 minutes) were ideal for enrichment of viral particles. The resulting homogenate was subsequently subjected to filtration procedures using a 0.45 µm filter. It is worth noting that as much as filtration is very efficient and extensively utilized to get rid of bacterial and eukaryotic cells, it will not completely remove all the unwanted bacterial particles since not all of them are larger than the filter pore size used. Moreover, filtration cannot get rid the unprotected nucleic material that might be originating from the bacteria, host, plants or other eukaryotic organisms. Therefore it is expected that quite a substantial amount of bacteria would still be present, albeit filtration aiding in reducing these microbes (Edwards *et al*., 2006; Thurber *et al*., 2009). Homogenization, centrifugation and filtration were essential applications in this metagenomics study and aided to increase the amount of virus particles, while minimizing bacterial and eukaryotic particles in the sample.

It was expected that treatment of the filtrate with a combination of nuclease enzymes would digest and therefore remove the unprotected nucleic acids that were free floating in the samples. This nuclease treatment also took

advantage of the hard viral capsid enclosing the viral genetic material, thereby protecting viruses from digestion. However, literature has shown that such treatment does not eliminate non-viral genetic material in the sample, but can result in the decrease of host background nucleic acids (Allander *et al*., 2001; Thurber *et al*., 2009). Nevertheless, in the RNA extraction step using the kit mentioned in the methodology section, there was a DNase treatment performed which was intended to digest DNA material and we therefore expected most of the non-viral along with the viral DNA to be destroyed and by so doing, we essentially further enriched for RNA.

Following RNA isolation, ribosomal RNA depletion was performed and this was intended to reduce the amount of host ribosomal RNA from the sample. From the quantification results by Biodrop, most of the samples, with the exception of one sample (4450), had RNA concentrations above 10 ng/µl. The large variations in the RNA concentration from sample to sample was likely to do with the initial RNA content of each sample than any possible bias that might have been introduced in the enrichments or extraction since all samples were processed in parallel. The absorbance ratio at 260 nm and 280 nm used to assess the quality and integrity of the isolated RNA was roughly within the acceptable range of 1.8 to 2.2 for 8 out of 12 samples and the rest were not far off, an indication of good quality RNA. Although the RNA isolation and the purification thereof were handled with caution to try and avoid any carry-over of phenols, there is always a possibility that negligible quantities of ethanol and/or isopropanol added during the isolation procedure were carried over leading to accurate measurements. Another possible cause for this could be the non-specificity of the assay in measuring nucleic acids. With the subsequent quantification of the RNA performed right after ribosomal RNA depletion, a decrease in RNA concentration was anticipated as most of the host rRNA was removed and such reduction was observed in 10 of the samples. On the contrary, sample B1 and B3 had increased levels of RNA concentration after rRNA deflection which we can speculate could be due to the inaccuracy of the Biodrop spectrophotometer. Another potential cause could be the efficient removal of digested rRNA and rRNA-specific probes by subsequent purification with magnetic beads. Nevertheless, the quality of the RNA was good as determined using the 260/280 absorbance ratio. Taking a closer look at the comparison between the RNA concentration before and after rRNA removal, we observed that for half the total number of samples, their RNA concentrations was reduced by more than 60 %. Three other samples showed a decrease in RNA concentration of 30 - 38 %, one sample had 10 % reduction whereas only two had slight increase in RNA concentration (6 % and 14% respectively). From these findings we could establish that most of the samples contained huge quantities of host ribosomal RNA which coincides with literature showing rRNA comprises between 80 % and 90 % of the total RNA in a given specimen. It is for this reason that removal of rRNA prior to sequencing is done, so that majority of the generated sequence data can consist of the more informative fractions of the transcriptome. The results obtained demonstrates the efficiency

of the rRNA depletion procedure, based on hybridization of rRNA-specific DNA probes to rRNA, in the enrichment of viral RNA, and the necessity of performing such a step in viral metagenomic studies.

In order to sequence viral RNA, reverse transcription needed to be performed to convert RNA to complementary DNA. Moreover, to obtain sufficient amount of input material random amplification of the synthesized viral cDNA was required. The WTA was thus performed and the extremely high concentration obtained with both Biodrop and Qubit quantification demonstrate the efficiency of WTA step. With the concentration of the amplified cDNA being out of the instrument's detection level, necessitating large dilution of the cDNA material, this is an indication that WTA method greatly increased the amount of genetic material present in the samples. Albeit a number of amplification methods been developed, not all of them are able to produce optimal desired results and this is largely due to the substantial bias introduced by some of these methods (Kim and Bae, 2011; Li *et al*., 2015). The consistent and uniform concentration readings and the good quality obtained across all amplified cDNA samples in this study shows that the QiaSeq FX RNA kit was ideal choice with little to no evidence of biased amplification.

In preparation of libraries for next generation sequencing using the Nextera XT DNA library kit, it is usually recommended that all (c)DNA samples to be sequenced be normalized to the same amount of starting material. In this study, the samples were normalized to approximately 1.2 ng/µl as opposed to 0.2 ng/µl, only half of the reagents volume were used. In the case of input DNA, 2.5 µl at a concentration of 1.2 ng/µl was used, translating to 3 ng input DNA amount and this helps to generate relatively longer fragments. Confirmation of the normalized samples has revealed only negligible variable with the minimum concentration of 1.19 ng/µl, maximum being 1.30 ng/µl and average was calculated to be 1.24 ng/µl which were optimal for the fragment sizes (>400 bp) we targeted. Tagmentation, which is the first step in the library preparation and was very crucial as it mostly determined the success of the library preparation. Although the Nextera XT library preparation kit is optimized for 1 ng dsDNA, however it is reported that only final libraries above 1.2 kb average size, including adapters, fail to cluster well on the Illumina flow cell (https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/Nextera-xt-troubleshooting-technical-note.pdf). Therefore, an ideal library size ranges from 200 bp to 1 kb. In this study the anticipated library size distribution was at least 350 bp to 600bp on bioanalyzer after clean-up of the indexed libraries with magnetic beads, a size range which would enable us to generate 2 x 250 bp or 2 x 300 bp paired end reads.

Quality assessment of the DNA after tagmentation/fragmentation and index PCR was done on Agilent 2100 bioanalyzer. This is an advanced technology that allowed the easy and reliable determination of a successful fragment size and distribution, and purity. The presence of additional peaks on the electropherogram trace can tell if there were any adaptors, shorter fragments or primers not efficiently removed during the clean-up step.

The average fragment size obtained in this study was approximately 450 bp for all the samples. The obtained library size was optimal for the desired read length of 250 bp paired end. The electropherogram shows that the quality of the fragmented and cleaned libraries was good, with a consistent size distribution across nearly all samples. Out of the twelve samples that were assessed for quality, only one showed an unusual library distribution, in which an additional peak (**Figure 3-5**) or band was observed (**Figure 3-6**), symbolizing inefficient removal of fragments below 200 bp. Nonetheless, since the average fragment size for this sample was within the range (412 bp), it was included for analysis with the remaining samples assuming that it would not compromise the quality of the endpoint data since it fell with the limits of detection.

A quality score (Q-score) of more than 30 for nearly 80 %, with a cluster density of 1014 K/mm2 and 95.1 % of clusters passing filter were indicative of good quality data generated from Illumina MiSeq instrument (Illumina, Inc., California, United States) with the sequencing run completed without any errors. It is important to produce sequence data of good quality especially for such complex studies in order to be able to adequately address the main objectives of such a study. Methodology on further data analysis using various bioinformatics tools and the findings thereof will be discussed in the next chapter.

3.5. Conclusion

Virome (viral metagenomics) analysis with next generation sequencing (NGS) has enabled metagenomics-based identification of viruses from a range of clinical and environmental samples (Breitbart *et al*., 2003; Zhang *et al*., 2006). Not only that, but viral metagenomics approaches have also been widely applied in the research of RNA viruses (Masson *et al*., 2014; Webster *et al*., 2015; Paez-Espino *et al*., 2016). Owing to its sequence independent nature, viral metagenomics has the potential to readily detect any virus, known or new, in any biological sample. Several studies have demonstrated the applicability of this approach in characterizing the virome from various biological samples including human and animal faeces (Breitbart *et al*., 2003; Finkbeiner *et al*., 2009; Phan *et al*., 2014; Smits *et al*., 2014; Bodewes *et al*., 2014; Moore *et al*., 2015). However, methods that are involved in metagenomics-based virome analysis are usually complex, hence a thoroughly optimized protocol was adapted to achieve overcome such challenges in this particular study.

Although viral metagenomics has become a promising technology in the discovery of viruses, sample type has a huge influence on the composition of sequencing reads. For instance, sequencing of nucleic acids obtained directly from biological samples often leads to high background of host and bacterial genomes, hampering efficient identification of viruses (Yang *et al*., 2011; Mokili *et al*., 2013). To overcome this and to ensure recovery of complete RNA virus genomes, removal of host background prior to sequencing is necessary. Common steps in

sample preparation for unbiased metagenomics sequencing are virus enrichment, isolation of nucleic material, cDNA synthesis, and random amplification. The most common methods used for virus enrichment include filtration and nuclease treatment. These methods take advantage of the small size of virus particles and the stability of the virus capsid protecting the virus genome (Delwart *et al*., 2007; Thurber *et al*., 2009; Hall *et al*., 2014; Kohl *et al*., 2015; Conceição-Neto *et al*., 2015). Moreover, loss of viruses and introduction of bias must be avoided at all cost during sample preparations.

Some of the challenges associated with virome analysis from clinical samples include the low quantity of viral RNA and its susceptibility to degradation, making preparation of libraries challenging. It was thus imperative, in this study, to ensure efficient construction of sequencing libraries for metagenomic characterization of RNA viruses from faecal samples. As mentioned before, a significant fraction of nucleic material in a biological sample is of non-viral origin, and the absence of a universally conserved genomic region together with the high genetic diversity of viruses make detailed analysis of the virome more complicated. In this viral metagenomics study, we enriched for virus particles in faecal samples. Filtration was aimed at reducing the amount of larger particles, mostly bacterial and other microorganisms. Microccocal and Benzonase nucleases were simultaneously added to the filtered sample to digest free-floating nucleic material which mostly originates from the host. The library preparation protocol was also modified by increasing the input cDNA amount and the number of PCR cycles with the aim of producing longer reads and generate data of significance for virome detection of RNA viruses and characterize the virome of RNA viruses in the four study participants. The length of the fragments from bioanalyzer showed that the modifications might have worked although there was no negative control with which to compare.

Lastly, although sample preparation methods have improved over the years, the hands-on time is still considerably long. Moreover, with the price of performing a sequence run continuously falling, library preparation costs for virus transcriptomics remains high.

The next chapter which is based on the virome analysis and characterization will confirm if the enrichment methods such as filtration, enzyme digestion, ribosomal RNA depletion and random amplification were helpful in enriching for virus-like particles or not.

3.6. References

Acar, E., Bulbul, O., Rayimoglu, G., Shahzad, M. S., Argac, D., Altuncul, H. and Filoglu, G. (2009). Optimization and validation studies of the MentypeR Argus X-8 kit for paternity cases. *Forensic Sci Int Genet Suppl Ser* **2**: 47-48.

Allander, T., Emerson, S. U., Engle, R. E., Purcell, R. H. and Bukh, J. (2001). A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. *Proc Natl Acad Sci U S A* **98**: 11609-11614.

Allander, T., Tammi, M. T., Eriksson, M., Bjerkner, A., Tiveljung-Lindell, A. and Andersson, B. (2005). Cloning of a human parvovirus by molecular screening of respiratory tract samples. *Proc Natl Acad Sci U S A* **102**: 12891-12896.

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., Chan, A. M., Haynes, M., Kelley, S. and Liu, H. *et al*. (2006). The marine viromes of four oceanic regions. *PLoS Biol* **4**: e368.

Barzon, L., Lavezzo, E., Militello, V., Toppo, S. and Palù, G. (2011). Applications of next- generation sequencing technologies to diagnostic virology. *Int J Mol Sci* **12**: 7861-7884.

Bodewes, R., Ruiz-Gonzalez, A., Schapendonk, C. M., van den Brand, J. M., Osterhaus, A. D. and Smits, S. L. (2014). Viral metagenomic analysis of feces of wild small carnivores. *Virol J* **11(1)**: 89.

Breitbart, M., Haynes, M., Kelley, S., Angly, F., Edwards, R. A., Felts, B., Mahaffy, J. M., Mueller, J., Nulton, J., Rayhawk, S. *et al*. (2008). Viral diversity and dynamics in an infant gut. *Res Microbiol* **159**: 367-373.

Breitbart, M. and Rohwer, F. (2005). Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. *Biotechniques* **39**: 729-736.

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J. M., Nulton, J., Salamon, P. and Rohwer, F. (2003). Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* **185(20)**: 6220-6223.

Cleaveland, S., Laurenson, M. K. and Taylor, L. H. (2001). Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. *Philos Trans R Soc Lond B: Biol Sci* **356**: 991-999.

Conceição-Neto, N., Zeller, M., Lefrère, H., De Bruyn, P., Beller, L., Deboutte, W., Yinda, C. K., Lavigne, R., Maes, P., van Ranst, M. *et al*. (2015). Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci Rep* **5**: 16532.

Delwart, E. L. (2007). Viral metagenomics. *Rev Med Virol* **17(2)**: 115-131.

Djikeng, A., Halpin, R., Kuzmickas, R., Depasse, J., Feldblyum, J., Sengamalay, N., Afonso, C., Zhang, X., Anderson, N.G., Ghedin, E. *et al.* (2008). Viral genome sequencing by random priming methods. *BMC Genomics* **9**: 5.

Donis-Keller, H. (1979). Site specific cleavage of RNA. *Nucleic Acids Res* **7(1)**: 179-192.

Edwards, R. A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D. M., Saar, M. O., Alexander, S., Alexander, E. C. and Rohwer, F. (2006). Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7**: 57.

Edwards, R. A., Rohwer, F. (2005). Viral Metagenomics. *Nat Rev Microbiol* **3(6)**: 504-10.

Finkbeiner, S. R., Holtz, L. R., Jiang, Y., Rajendran, P., Franz, C. J., Zhao, G., Kang, G. and Wang, D. (2009). Human stool contains a previously unrecognized diversity of novel astroviruses. *Virol J* **6(161)**: 1-5.

Frey, K. G., Herrera-Galeano, J. E., Redden, C. L., Luu, T. V., Servetas, S. L., Mateczun, A. J., Mokashi, V. P. and Bishop-Lilly, K. A. (2014). Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood. *BMC Genomics* **15**: 96.

Glasel, J. A. (1995). Validity of nucleic acid purities monitored by 260 nm/280 nm absorbance ratios. *BioTechniques* **18(1)**: 62-63.

Goodwin, E. C. and Rottman, F. M. (1992). The use of RNase H and poly(A) junction oligonucleotides in the analysis of in vitro polyadenylation reaction products. *Nucleic Acids Res* **20(4)**: 916.

Gubler, U. and Hoffman B. J. (1983). A simple and very efficient method for generating cDNA libraries. *Gene* **25(2-3)**: 263-269.

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

Hamza, I. A., Jurzik, L., Wilhelm, M. and Uberla, K. (2009). Detection and quantification of human bocavirus in riverwater. *J Gen Virol* **90(11)**: 2634-2637.

Huberman, J. A. (1995). Importance of measuring nucleic acid absorbance at 240 nm as well as at 260 and 280 nm. *BioTechniques* **18(4)**: 636.

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L. and Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature* **451**: 990-993.

Kim, K. H. and Bae, J. W. (2011). Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Applied and environmental microbiology* **77**: 7663-7668.

Kohl, C., Brinkmann, A., Dabrowski, P. W., Radonić, A., Nitsche, A. and Kurth, A. (2015). Protocol for metagenomic virus detection in clinical specimens. *Emerg Infect Dis* **21(1)**: 48-57.

Li, L., Kapoor, A., Slikas, B., Bamidele, O. S., Wang, C., Shaukat, S., Masroor, M. A., Wilson, M. L., Ndjango, J. B., Peeters, M. *et al*. (2009). Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. *J Virol* **84**: 1674-1682.

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**: 139-146.

Li, L., Victoria, J. G., Wang, C., Jones, M., Fellers, G. M., Kunz, T. H. and Delwart, E. (2010). Bat guano virome: predominance of dietary viruses from insects and plants plus novel mammalian viruses. *J Virol* **84**: 6955-6965.

Manchester, K. L. (1995). Value of A260/A280 ratios for measurement of purity of nucleic acids. *BioTechniques* **19(2)**: 208-210.

Manchester, K. L. (1996). Use of UV methods for the measurement of protein and nucleic acid concentrations. *BioTechniques* **20(6)**: 968-970.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9**: 387-402.

Masson, P., Hulo, C., de Castro, E., Foulger, R., Poux, S., Bridge, A., Lomax, J., Bougueleret, L., Xenarios, I. and Le Mercier, P. (2014). An integrated ontology resource to explore and study host-virus relationships. *PLoS One* **9(9)**: e108075.

McKnight, R. E., Gleason, A. B., Keyes, J. A. and Sahabi, S. (2006). Binding mode and affinity studies of DNA-binding agents using topoisomerase I DNA unwinding assay. *Bioorganic and Medicinal Chemistry Letters* **17(4)**: 1013-1017.

Minot, S. and Bryson, A. (2013). Rapid evolution of the human gut virome. *Proc Natl Acad Sci U S A* **110(30)**: 12450-12455.

Mommaerts. K., Sanchez, I., Betsou, F and Mathieson, W. (2015). Replacing β-mercaptoethanol in RNA extractions. *Anal Biochem* **479**: 51-53.

Mokili, J. L., Dutilh, B. E., Lim, Y. W., Schneider, B. S., Taylor, T., Haynes, M. R., Metzgar, D., Myers, C. A., Blair, P. J., Nosrat B. *et al*. (2013). Identification of a novel human papillomavirus by metagenomic analysis of samples from patients with febrile respiratory illness. *PLoS One* **8(3)**: e58404.

Moore, N. E., Wang, J., Hewitt, J., Croucher, D., Williamson, D. A., Paine, S., Yen, S., Greening, G. E. and Hall, R. J. (2015). Metagenomic analysis of viruses in feces from unsolved outbreaks of gastroenteritis in humans. *J Clin Microbiol* **53(1)**: 15-21.

Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N. N., and Kyrpides, N. C. (2016). Uncovering earth's virome. *Nature* **536(7617)**: 425-430.

Phan, T. G., Nordgren, J., Ouermi, D., Simpore, J., Nitiema, L. W., Deng, X. and Delwart, E. L. (2014). New astrovirus in human feces from Burkina Faso. *J Clin Virol* **60(2)**: 161-164.

Radford, A. D., Chapman, D., Dixon, L., Chantrey, J., Darby, A. C. and Hall, N. (2012). Application of next-generation sequencing technologies in virology. *J Gen Virol* **93**: 1853-1868.

Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F. and Gordon, J. I. (2012). Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat Rev Microbiol* **10**: 607-617.

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

Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27(6)**: 863-864.

Schweitzer, C. and Scaiano, J. C. (2003). Selective binding and local photophysics of the fluorescent cyanine dye PicoGreen in double-stranded and single-stranded DNA. *Physical Chemistry Chemical Physics* **5**: 4911-4917.

Smits, S. L., Schapendonk, C. M. E., van Beek, J., Vennema, H., Schürch, A. C., Schipper, D., Bodewes, R., Haagmans, B. L., Osterhaus, A. D. M. E. and Koopmans, M. P. (2014). New viruses in idiopathic human diarrhea cases, the Netherlands. *Emerg Infect Dis* **20(7)**: 1218-1222.

Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. and Rohwer, F. (2009). Laboratory procedures to generate viral metagenomes. *Nat Protoc* **4(4)**: 470-483.

Webster, C. L., Waldron, F. M., Robertson, S., Crowson, D., Ferrari, G., Quintana, J. F., Brouqui, J. M., Bayne, E. H., Longdon, B., Buck, A. H. *et al*.(2015). The discovery, distribution, and evolution of viruses associated with drosophila melanogaster. *PLoS Biol* **13**: e1002210.

Woolhouse, M. E. J. and Gowtage-Sequeria, S. (2005). Host range and emerging and reemerging pathogens. *Emerg Infect Dis* **11**: 1842-1847.

Yang, J., Yang, F., Ren, L., Xiong, Z., Wu, Z., Dong, J., Sun, L., Zhang, T., Hu, Y., Du, J. *et al*. (2011). Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. *J Clin Microbiol* **49(10)**: 3463-3469.

Zhang, T., Breitbart, M., Lee, W. H., Run, J. Q., Wei, C. L., Soh, S. W., Hibberd, M. L., Liu, E. T., Rohwer, F. and Ruan, Y. **(2006).** RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol* **4**: e3.

Chapter 4: Human gut virome analysis by deep metagenomics sequencing

4.1. Introduction

Viruses, the most abundant biological entities in planet earth, have been classified and determined in a variety of host species and environments, such as the gastrointestinal tract (GIT), in freshwater lakes and the marine ecosystems (Wommack and Colwell, 2000; Zhang *et al*., 2006; Breitbart *et al*., 2008). However, due to the high diversity of their genetic material and morphology, it is difficult to cultivate novel viruses or deeply explore their populations. The metagenomic approach, which is an advanced method developed over a decade ago, has been very instrumental in providing an in-depth look at the genomic diversity of viruses in different environments, including the human GIT (Breitbart *et al*., 2002; Minot *et al*., 2011).

Although the microbiota plays a critical role in the human GIT and the human health in general, the process through which this microbial community develops during infancy is not well understood. Particularly, little is known about how the amount and type of viruses present in the human gut, otherwise known as the virome, changes throughout this period and about the role this collection of viruses may play in the assembly of the gut virome. The patterns of changes and dynamics of viral community occurring in the infant's gut can be analyzed in a birth cohort of infants during the first year of life.

In order to comprehensively describe the human viral populations, suitable and efficient molecular tools are required. In the past years, the main techniques used in classical virology entailed viral isolation from cells followed by observation of cytopathic effects on cell lines or the intracerebral inoculation of suckling mice. The detection of viral antigens was then performed using serological techniques such as hemagglutination or seroneutralization (Specter and Lanz, 1992). Later on, several advancements were made in the field of molecular biology, with polymerase chain reaction (PCR) being the preferred methods used in detecting viral agents from diverse environmental and clinical samples (Ratcliff *et al*., 2007). However, a major setback was identifying novel viruses that could not be cultivated. Nevertheless, the advent of high-throughput sequencing or next-generation sequencing approaches has enabled the identification and characterization of viral populations in a given sample without prior knowledge about a specific virus, including faecal samples (Zhang *et al*., 2006). The human gastrointestinal tract (GIT) remains the most studied part of the human body as far as virome studies is concerned. This is because the GIT can be easily sampled and the material is sufficient, allowing for the analysis of the viral composition and dynamics in the gut (Popgeorgiev *et al*., 2013).

There are several approaches that can be used to analyze high throughput sequence data to characterize the human gut virome composition and detect new viruses. A number of steps are involved in the analysis of this type of data generated using the metagenomics approach. Firstly, the adapter sequences which were added during the library preparation stage must be removed and low quality reads trimmed. BLAST can be utilized to filter out reads

that are of human origin (Altschul *et al*., 1990). Shorter sequence reads can be analyzed individually, or assembled into larger contiguous sequences (contigs) that represent parts of a genome (Luo *et al*., 2012; Namiki *et al*., 2012; Peng *et al*., 2012). The longer contigs provide a longer sequence for similarity searches using BLAST and allows for more sensitive tracking of viruses. Methods for constructing contigs are continuously being optimized although multiple challenges remain (Nagarajan and Pop, 2013). For example, sequence heterogeneity and relative abundance of genomes can affect the outcome. Downstream, BLAST (Altschul *et al*., 1990) or other methods such as Bowtie (Langmead and Salzberg, 2012) can all be used to detect sequence homology of reads and contigs to reference sequences in the viral database and thus quantify abundance and composition. Open reading frames (ORFs) can also be called on contigs to predict and identify viral genes of interest (Wang *et al*., 2010). The NCBI Genome database includes the reference whole genome sequences of 7 321 viruses as of 2017. In addition, viral protein sequences are available in Refseq (O'Leary *et al*., 2016), UniProt (Hulo *et al*., 2011). Custom databases of viral proteins are also available for samples from the ocean (Roux *et al*., 2016), humans and various environments (Paez *et al*., 2016). However, alignment to these databases is often challenging when sequence identity is less than 30 %. Viruses often accumulate nucleotide substitutions at high rates (Duffy *et al*., 2008). RNA viruses replicate using error prone RNA-dependent RNA polymerases (Lauring *et al*., 2013) and retroviruses error prone reverse transcriptases (Svarovskaia *et al*., 2003). In addition, ssDNA viruses also show high rates of substitution (Duffy *et al*., 2008).

Within the past few years, several analysis pipelines have been developed which combine multiple programs for pre-processing, assembly and annotation (Lorenzi *et al*., 2011; Wommack *et al*., 2012; Ho and Tzanetakis, 2014; Roux *et al*., 2014; Rampelli *et al*., 2016; Zhao *et al*., 2017). However, all these methods fail to provide a solution for viral dark matter (unknown sequences), which is not surprising considering the large number of viruses on earth, most of which are not present in databases (Aggarwala *et al*., 2017). Nevertheless, this is not a major issue for researchers who are focusing on medically important viral pathogens that infect humans, most of which have been well studied.

The present chapter describes the viral metagenomic analysis pipeline used in this study and presents the characterization of infant's gut viral communities. This chapter focuses on the taxonomic classification of viral sequences in order to establish the composition of the virome present in each sample and to assess the viral dynamics over time.

4.2. Materials and Methods

4.2.1. Quality Control and Trimming/Filtering

Modern high throughput sequencing instruments can generate a massive amount of genomic data in a single run. However, downstream sequence analysis is compromised by low-quality sequences, which in most cases results in bad assembly and ultimately incorrect conclusions. To overcome these issues, efficient tools for quality control and pre-processing of all genomic data must be used. Quality control for genomic datasets includes determining sequence length, GC content, quality score, sequence duplication and contamination among other things.

Therefore, before analyzing any genomic sequence data to draw biological conclusions it is always imperative to perform quality control checks to ensure that the quality of the raw data is good and there are no issues or biases in the data. For our quality assessment of the data generated, a program called FASTQC version 0.11.7 (Andrews, 2010; http://www.bioinformatics.babraham.ac.uk/projects/FASTQc) was used. This program provides a simple way to do some quality control checks on raw sequence data from NGS run. It generates a QC report on the quality of the data, and this helps in identifying problems which originate either in the sequencer or in the starting library material. The mode chosen for our QC was a stand-alone interactive application which is suitable for immediate analysis of small numbers of FASTQ files. FASTQC mainly imports data from FASTQ files and provide a quick overview in a form of summary graphs and tables to quickly assess your data. It also exports results to an HTML based permanent report.

Furthermore, preprocessing which was performed following quality control involved trimming of the sequence ends and filtering of unwanted sequences. Based on the FASTQC report, poor quality reads were trimmed by filtering at phred quality score of 20. Pinseq-lite version 0.20.4 (Schmieder and Edwards, 2011) was used to perform the trimming and filtering of the genomic data to remove sequence duplicates, short or long sequences, sequences with N's, low-quality sequences and adapters. This program generated summary statistics of the sequences in graphical and tabular format. The input format of the raw sequence data was in FASTQ. The summary statistics included read length, GC content, sequence complexity and quality score distributions, number of read duplicates, occurrence of Ns and poly-A/T tails, assembly quality measures and tag sequences. Each of the two QC and pre-processing steps were performed by running scripts on Linux command line to generate detailed quality reports.

4.2.2. *De novo* Assembly

Although metagenomics has become a technology of choice for analysis of microbial communities, the assembly of metagenomic data remains challenging, thus making it difficult for researchers to draw biological conclusions. The challenge emerges due to the huge volume of metagenomics data produced from sequencing projects. Therefore, *de novo* assembly of this metagenomics data has become an effective solution not only in reducing the amount of data, but also in improving the quality of data for downstream analysis, such as annotation. Although several programs exist for the assembly metagenomics data, it is always imperative to choose an assembler that is suitable for a specific goal. In contrast to reference based assembly which requires mapping of a query sequences against a reference genome, *de novo* assembly takes short quality-filtered reads and assemble them into longer, contiguous sequences (contigs) without any reference. Here, *de novo* assembly was performed using MetaSPAdes mode from SPAdes Assembler version 3.11.1 (Bankevich *et al*., 2012; Nurk *et al*., 2017) by executing a script on Linux command line with paired-end reads in FASTQ format as input to generate assembled contigs in FASTA format.

4.2.3. Taxonomic Classification

The alignment of sequencing reads against a protein reference database is a major drawback in metagenomics projects. Despite promising improvements offered by recent tools over the gold standard BLASTX, very often performance speed and sensitivity become a stumbling block. In our study, the assembled FASTA files were subjected to BLASTX search for contigs annotation and taxonomic classification using DIAMOND version 0.9.22 (Buchfink *et al*., 2015). DIAMOND is an open-source algorithm based on alignment of translated protein against the NCBI protein reference database and it is reported to be 20 000 times faster than BLASTX on short reads with the same level of sensitivity.

4.2.4. Statistical Analysis

Annotated contigs that were identified as viruses from the DIAMOND BLASTX results were retained and further analyzed to determine the proportions of different viruses detected and classified them according to their virus families. For our statistical analysis the annotated virus contigs were classified according to their target hosts, for instance, these were grouped into human viruses, non-human mammalian viruses, plant viruses, insect viruses, bacterial viruses, fungal viruses and others. Furthermore, contigs of the different viruses in each group were statistically analyzed using a python script in Jupyter Notebook, an open-source web application used for data cleaning and manipulation, numerical simulation, statistical modeling and data visualization, to determine the proportions of the virus populations identified from all 12 faecal specimens. To determine the prevalence at family

level, classification of virus contigs into respective virus families was done manually. Focusing on the human viruses, contigs of these viruses were further classified into enteric and non-enteric viruses. Moreover, the distribution of enteric or gut viruses was determined in each study participant across the three collection points at family and genus/species level to establish which were the most and least predominant of gut viruses. Each study participant's gut virome was analyzed individually to study the changes in virome composition and how it changed over time and also how it compared to other participants.

4.3. Results

4.3.1. Abundance of viral contigs and non-viral contigs

This section is based on the results obtained after analyzing the metagenomic sequence data from twelve human faecal. As mentioned in the methodology section of this chapter, several bioinformatics tools were utilized to explore the gut virome of the four infants under study. To characterize this viral component of the human gut microbiome, generated paired-end reads were assembled into contigs and annotated by BLASTX search against the NCBI protein database to enable taxonomic classification. Based on the BLASTX results obtained, a total of 92 185 contig hits were obtained from the 12 human faecal samples. Our analysis revealed that 11.6 % of the hits mapped to viruses, which translated to a total of 10 648 viral contigs as summarized in **Table 4-1**.

Table 4-1: Summary of the distribution of assembled contigs obtained from the twelve faecal samples.

Sample #	Sample ID	Collection time	Total assembled contigs	Total viral contigs	Total non-viral contigs	Percentage viral contigs
A1	**4450**	6 weeks	5549	68	5481	**1.2**
A2	**6618**	10 weeks	6255	28	6227	**0.4**
A3	**8908**	36 weeks	9805	489	9316	**5.0**
	Total		***21 609***	***585***	***21 024***	***2.7%***
B1	**8903**	10 weeks	12445	2504	9941	**20.1**
B2	**8941**	14 weeks	15184	7064	8120	**46.5**
B3	**10287**	24 weeks	5636	138	5498	**2.4**
	Total		***33 265***	***9706***	***23559***	***29.2%***
C1	**8824**	6 weeks	4368	87	4281	**2.0**
C2	**10127**	14 weeks	8637	60	8577	**0.7**
C3	**10233**	20 weeks	3638	67	3571	**1.8**
	Total		***16 643***	***214***	***16 429***	***1.3%***
D1	**8355**	6 weeks	7405	36	7369	**0.5**
D2	**8910**	14 weeks	7357	37	7320	**0.5**
D3	**10098**	20 weeks	5906	70	5836	**1.2**
	Total		***20 668***	***143***	***20 525***	***0.7%***
Total contigs overall			***92185***	***10648***	***81537***	***11.6%***

To break down the obtained contigs by focusing at each study participant and their individual samples, participant B out of the four participants had the highest number of total contigs (33 265), of which at 29 % of these belonged to viruses across the three collection time points. The viral contigs for this infant increased from 2 504 in the sample collected at 10 weeks to 7 064 in the sample collected at 14 weeks of age, and a drastic decrease in the number of viral contigs was seen during the last collection (24 weeks) where 138 viral sequences were present. Participant A had the second highest number of contigs (21 609) of which 2.7 % or 585 mapped to viruses. As indicated in **Table 4-1**, participant A and B had varying collections time frames with the baseline for participant A being four weeks earlier than that of participant B, and the last faecal samples had a difference of 11 weeks between the two participants.

Furthermore, participants C and D, who had similar collection time frames, showed the least number of contigs at 16 643 and 20 668 total contigs, respectively. Although participant D had nearly four thousand more contigs than participant C, however the latter had a higher percentage of contigs that mapped to viruses at 1.3 %, whereas only 0.7 % of the contigs from participant C were assigned to viruses. There was no distinct pattern exhibited in terms of the change in viral contigs over time in the four study participant. To explain this further, in participant A the percentage of viral contigs dropped by more than half in the second collection (10 weeks). It then increased again by almost 10-fold in the last collection (36 weeks). A different trend was observed in participant B, whereby the percentage of viral contigs increased in the second collection (14 weeks) and was at its lowest in the final collection (24 weeks). With regard to participant C, the percentage of viral contigs was highest at baseline of 6 weeks being 2 %, it then dropped to 0.7 % by the second collection and a rise in viral contigs percentage was observed again by the last collection (1.8 %). On the other hand, participant D had a nearly the same amount of viral contigs between the first and second faecal collection. However, although the total number of contigs decreased in the last collection, a rise in the percentage of viral contigs, from 0.5 % to 1.2 % was observed.

4.3.2. Taxonomic classification of viral contigs

Table 4-2 gives a summary of the viral populations obtained from each of the faecal samples collected from the four infants. The data in **Table 4-2** includes the detected virus genera and/ or species classified into their respective virus families. In addition, the identified viruses were also categorized based by the nature of their nucleic material, which consisted of double-stranded (ds)RNA viruses, single-stranded (ss)RNA viruses, DNA viruses and other RNA viruses. The different viruses present in each faecal sample was highlighted in **Table 4-2** using a unique colour and the number of contigs for that specific virus was also indicated. As mentioned above and indicated in **Table 4-1**, there was a total of 92 185 assembled contigs, of which 10 648 mapped to viruses. These viral contigs were further analyzed to determine the composition of the human enteric virome.

Table 4-2: Taxonomic distribution of detected viruses from assembled contigs per collected faecal sample

Virus family					*Picornaviridae*									*Reoviridae*	*Astroviridae*		*Retroviridae*		*Caliciviridae*	*Flaviviridae*	*Poxviridae*	*Virgaviridae*	*Rhabdoviridae*	*Siphoviridae*	*Amalgaviridae*	*Partitiviridae*	*Totiviridae*	*Peribunyaviridae*	*Unclassified Picornavirales*		*Unclassified*	*Unclassified*
Sample number	**Sample ID**	**Sample collection time (weeks)**	**Total number of contigs**	**Contigs assigned to viruses**	Coxsackievirus A, B	Enterovirus A, B, C, D	Echovirus E	Poliovirus 1, 2, 3	Parechovirus 1, 3 8, 17, 19	Rhinovirus A, B, C	Swine vesicular disease virus	Bat picornavirus	Marmot sapelovirus	Rotavirus A	Astrovirus	Mamastrovirus	Human endogenous retrovirus	Equine infectious anaemia virus	Norovirus GI, GII.4	Hepatitis GB virus B	Vaccinia virus	Pepper mild mottle virus	Wheat rosette stunt virus	Geobacillus virus	Zygosaccharomyces bailli Z virus	Botryosphoeria dothidea virus	Scheffersomyces segobiensis virus	Shuangao insect virus	Apodemus agrarius picornavirus	Niniventer confucianism	Hubei partiti-like virus	Giant panda associated partiti-like virus
Genome type					(+)ssRNA									dsRNA	(+)ssRNA		RNA	(+)ssRNA	(+)ssRNA	(+)ssRNA	dsDNA	(+)ssRNA	(-)ssRNA	dsDNA	dsRNA	dsRNA	dsRNA	(-)ssRNA	(+)ssRNA	(+)ssRNA	RNA	RNA
A1	***4450***	**6**	5549	68	2	0	0	0	0	1	0	0	0	58	0	0	0	1	0	2	0	0	0	0	0	0	5	0	0	0	0	0
A2	***6618***	**10**	6255	28	1	0	0	0	0	0	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A3	***8908***	**36**	9805	489	84	34	309	0	0	0	2	0	0	59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B1	***8903***	**10**	12445	2504	1386	146	5	914	0	0	0	0	0	51	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B2	***8941***	**14**	15184	7064	3953	549	22	2333	0	88	3	1	1	90	9	1	0	1	0	5	0	0	0	0	0	0	2	1	1	1	0	1
B3	***10287***	**24**	5636	138	2	1	0	4	94	0	0	0	0	20	0	0	0	0	4	6	1	1	0	0	0	0	3	0	0	0	1	0
C1	***8824***	**6**	4368	87	5	4	20	1	0	0	0	0	0	52	0	0	0	0	0	3	0	0	0	0	0	1	0	0	0	0	0	0
C2	***10127***	**14**	8637	60	2	0	0	1	0	0	0	0	0	57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C3	***10233***	**20**	3638	67	0	1	1	0	0	1	0	0	0	43	0	0	0	17	0	2	0	0	1	0	1	0	0	0	0	0	0	0
D1	***8355***	**6**	7405	36	0	0	0	0	0	0	0	0	0	32	0	0	1	0	0	2	0	0	0	1	0	0	0	0	0	0	0	0
D2	***8910***	**14**	7357	37	0	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D3	***10098***	**20**	5906	70	0	0	0	0	0	0	0	0	0	67	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Total contigs/virus				**10648**	**5435**	**735**	**357**	**3253**	**94**	**90**	**5**	**1**	**1**	**593**	**11**	**1**	**3**	**19**	**4**	**21**	**1**	**1**	**1**	**1**	**1**	**1**	**10**	**1**	**1**	**1**	**1**	**1**

Based on the results shown in **Table 4-2**, all the detected viruses could be classified into fourteen (14) different virus families based on contig annotation by BLASTX search. *Picornaviridae,* which is a family of positive sense ssRNA viruses, was the most predominant. A total of nine different viruses that belonged to the *Picornaviridae* family were identified from various faecal samples (**Table 4-2**). At least one virus from *Picornaviridae* virus family was present in nine out of the twelve samples. Noteworthy, the nine viruses from *Picornaviridae* were detected in only three of the four infants, namely participant A, participant B and participant C. Furthermore, of the 10 648 contigs that mapped to viruses, 9 974 (93.7 %) were assigned to viruses that fall under the *Picornaviridae* family. Of these 9 974 contigs, 5 435 (54.5 %) were assigned to coxsackievirus type A and B combined, which were detected in eight samples, making species of coxsackieviruses the most predominant overall. Within this virus family, the second most abundant viral contigs belonged to polioviruses, with a total of 3 253 out of the 9 974 (32.6 %) contigs from *Picornaviridae* family. The poliovirus sequences recovered comprised poliovirus types 1, 2 and 3, detected from five samples. The other viruses classified under the *Picornaviridae* family that had more than hundred contigs were enteroviruses (735 contigs) and echoviruses (357 contigs) which were detected in six and five samples, respectively. These consisted of four enterovirus species, namely, enterovirus A, B, C and D (**Table 4-2**).

Following *Picornaviridae* as the most predominant virus family, the second was *Reoviridae* virus family with 593 contigs detected. In contrast to *Picornaviridae* family under which nine different viruses were identified, there was only one identified under the *Reoviridae* family, namely rotavirus A. The remaining twelve virus families consisted of only less than 25 contigs per virus family. Of these twelve families, only four virus families consisted of ten or more viral contigs. These virus families included, in decreasing order of contigs prevalence, *Retroviridae, Flaviviridae*, *Astroviridae* and *Totiviridae*. Furthermore, majority of the remaining virus families identified in this study comprised of only one viral contig, making them the least abundant virus families (**Table 4-2**).

4.3.3. Virus distribution based on the genome type

Different viruses harbour different types of nucleic material. In this virome study, there were 28 different species of viruses detected across the 12 faecal samples samples collected from four infants. Analysis was done to determine the number of viruses with a specific type of genome and these were classified accordingly, as presented in **Table 4-3**. These included positive-sense single-stranded RNA ((+)ssRNA) viruses, negative-sense single-stranded RNA ((-)ssRNA) viruses, double-stranded RNA (dsRNA) viruses, DNA viruses and other RNA viruses which could not be assigned families. Data presented in **Table 4-3** revealed that more than one different type of virus, based on the nature of the virus genomic material, were detected in each of the twelve stool samples.

Table 4-3: Different types of viruses detected from twelve faecal samples categorized based on the genome type.

Sample number	Sample ID	Collection time (Weeks)	Positive-sense ssRNA viruses	Negative-sense ssRNA viruses	Double-stranded RNA Viruses	DNA Viruses	Unclassified RNA Viruses	Different viruses detected overall
A1	**4450**	**6**	3	-	2	-	-	***5***
A2	**6618**	**10**	1	-	1	-	-	***2***
A3	**8908**	**36**	5	-	1	-	-	***6***
B1	**8903**	**10**	5	-	1	-	-	***6***
B2	**8941**	**14**	13	1	2	-	1	***17***
B3	**10287**	**24**	8	-	2	1	1	***12***
C1	**8824**	**6**	5	-	2	-	-	***7***
C2	**10127**	**14**	2	-	1	-	-	***3***
C3	**10233**	**20**	5	1	2	-	-	***8***
D1	**8355**	**6**	1	-	1	1	1	***4***
D2	**8910**	**14**	-	-	1	-	-	***1***
D3	**10098**	**20**	1	-	1	-	1	***3***

According to the results tabulated in **Table 4-3**, sample B2 (14 weeks old) of participant B had the highest number of viruses detected in which 17 different viruses were present in this particular sample. This was followed by sample B3 of the same participant, with 12 different viruses (**Table 4-3**), whereas all the remaining samples consisted of no more than ten viruses per sample, with the least number of viruses identified in a given sample being one, detected in sample D2 (14 weeks) of participant D (**Table 4-3**).

Moreover, the results recorded in **Table 4-3** above showed that the most detected viruses were (+)ssRNA viruses. All except one faecal sample had at least one virus with of this genome type, with the highest being 13 viruses detected in sample B2 (14 weeks) of participant B (**Table 4-3**). This was then followed by sample B3 (24 weeks) of

the same study participant with eight viruses of (+)ssRNA genome being detected. Three samples which had the lowest (+)ssRNA viruses were A2 (10 weeks) for participant A, D1 (6 weeks) and D3 (20 weeks) both for participant D. All these had only one virus of this type of nucleic material.

Following the (+)ssRNA viruses, dsRNA viruses were the second most abundant. These were present in all 12 samples, however there was two or less viruses with a dsRNA genome across all the samples (**Table 4-3**). Seven out of the 12 samples had only one type of dsRNA virus (rotavirus) whereas the remaining five samples had two different viruses of this genome. Furthermore, participant D had only one double-stranded RNA virus for each of the three collection time points. In participant A on the other hand, two dsRNA viruses in the first sample, then only one virus was present in each of the next two collections. In participant B and C, dsRNA viruses were detected twice in two samples, respectively. Moreover, (-)ssRNA viruses could also be detected in two samples (B2 and C3), and only one virus was present in each. Similarly, DNA viruses were also present in two samples (B3 and D1), with one virus identified in each collected sample. Lastly, low detection of unclassified viruses was observed in participant B and D as shown in (**Table 4-3**).

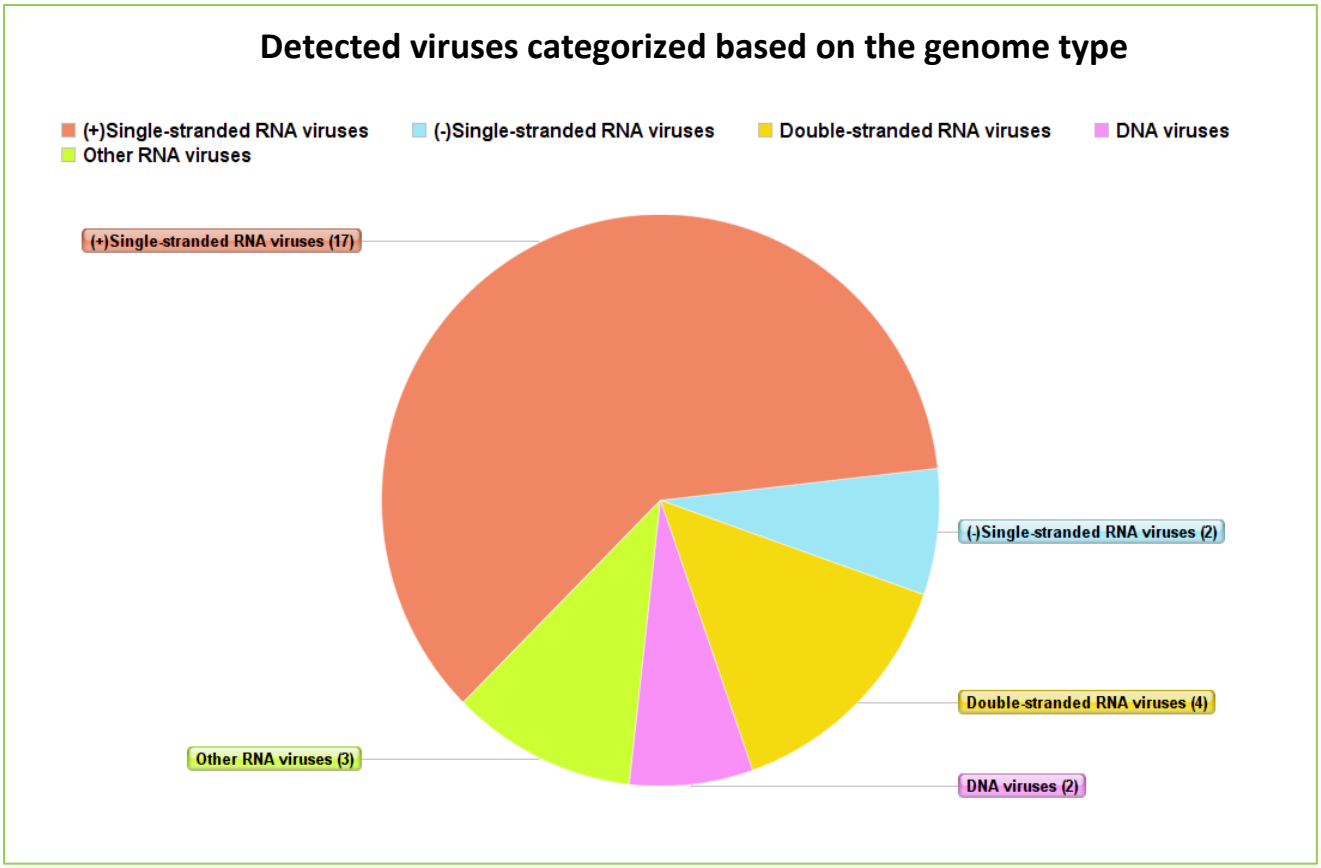

Figure 4-1: Pie chart showing the proportion of detected viruses from twelve samples classified based on the genome type.

Figure 4-1, which is in correlation with data captured in **Table 4-2**, shows the proportion of viruses with specific nucleic acid material out of the 28 different virus species identified in this metagenome study. The most prevalent type of viruses were (+)ssRNA viruses, whereby 17 of the 28 (60.71 %) viruses encapsidated this type of genome (**Figure 4-1**). This were followed by dsRNA viruses which were four. There were three different viruses which were unclassified. The least prevalent were (-)ssRNA viruses and DNA viruses, of which there was only two viruses, each.

4.3.4. Virus detection rate

Part of this enteric virome characterization study was to analyze the viral population of each sample and determine the detection rate of the identified viruses. Therefore, an attempt was made to establish in how many sample(s) out of the 12 samples of each of the 28 viruses were detected. A bar graph (**Figure 4-2**) shows the rate of detection of all the viruses. It was noted that rotaviruses had a 100 % detection rate, thus rotaviruses were present in all 12 of the faecal samples analyzed. This was followed by coxsackieviruses being detected in eight out 12 samples, translating to 66.7 % detection rate. Hepatitis GB virus B was among the viruses with had over 50 %

detection rate as these were found in seven out of 12 samples. Furthermore, enteroviruses were detected in 50 % of the samples, whereas polioviruses and echoviruses had 41.7 % detection rate, meaning they could only be identified in five samples.

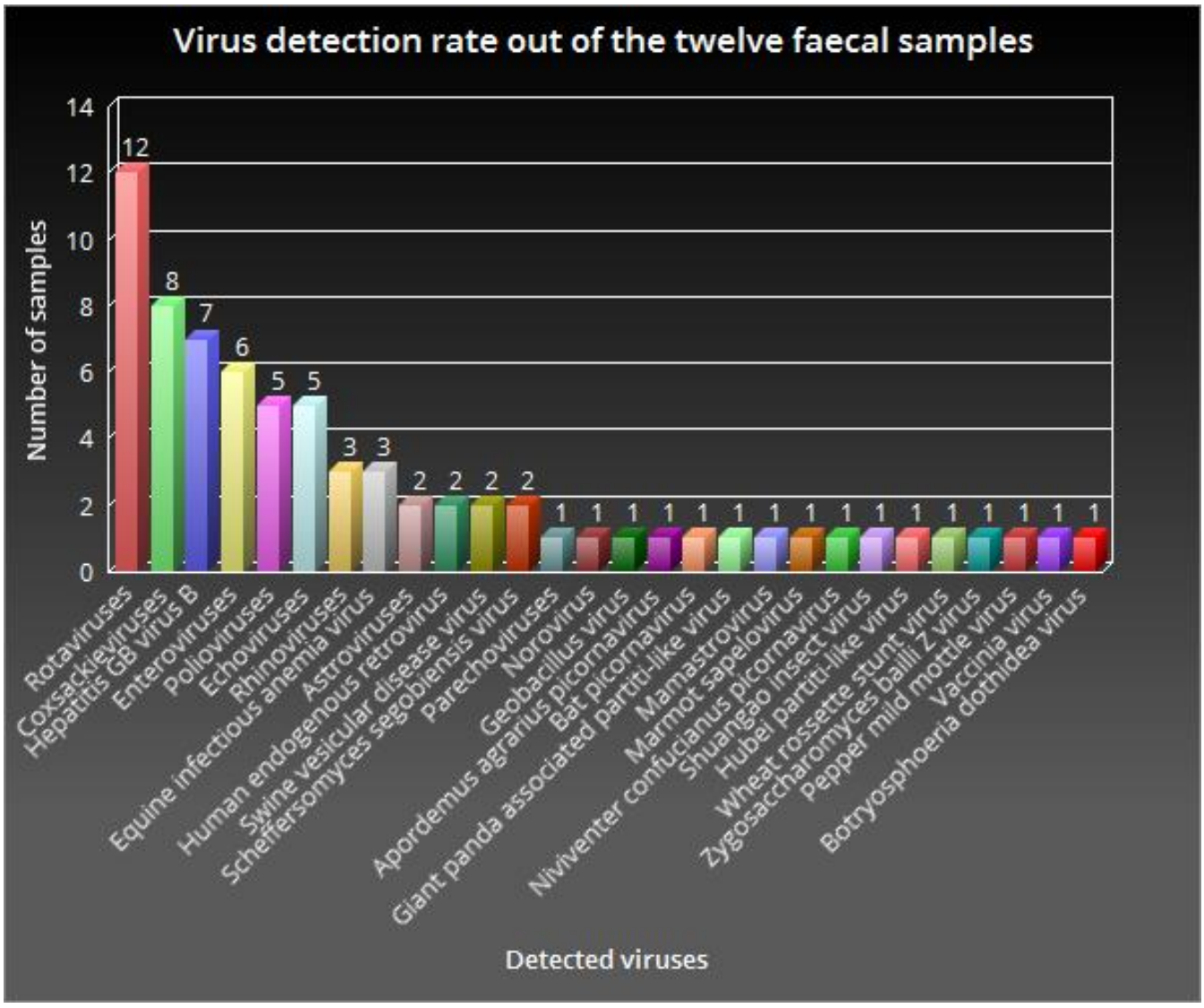

Figure 4-2: Bar graph showing the detection rate of each virus within the twelve faecal samples.

The remaining 22 different viruses were only detected in three samples or less, of which 16 of the 22 were only detected once (8.33 %). Among these, an important human enteric pathogen include noroviruses and astroviruses, detected in one and two samples, respectively.

4.3.5. Virus abundance by host-specificity

Viruses are host specific and although a number of viruses can infect multiple hosts, others can only be restricted to one natural host. This host-specificity is one of the several methods used in viral studies to classify viruses. As a result, this classification approach was also used in this viral metagenomics study to determine the prevalence of each virus and categorize them according to their specific hosts. **Figure 4-3** is a pie chart that shows the prevalence of viruses infecting different hosts. Mammalian viruses were the most abundant in this virome study, out of the 28 different viral genera/species, 20 of them infect mammals including humans. Bacterial and fungal viruses including few which were unclassified came the second most abundant. Following these, there were two plant viruses which included pepper-mild-mottle virus and wheat rosette stunt viruses each was detected in only one sample. Lastly, there was only insect virus detected in one of the infant, and this was shuangao insect virus.

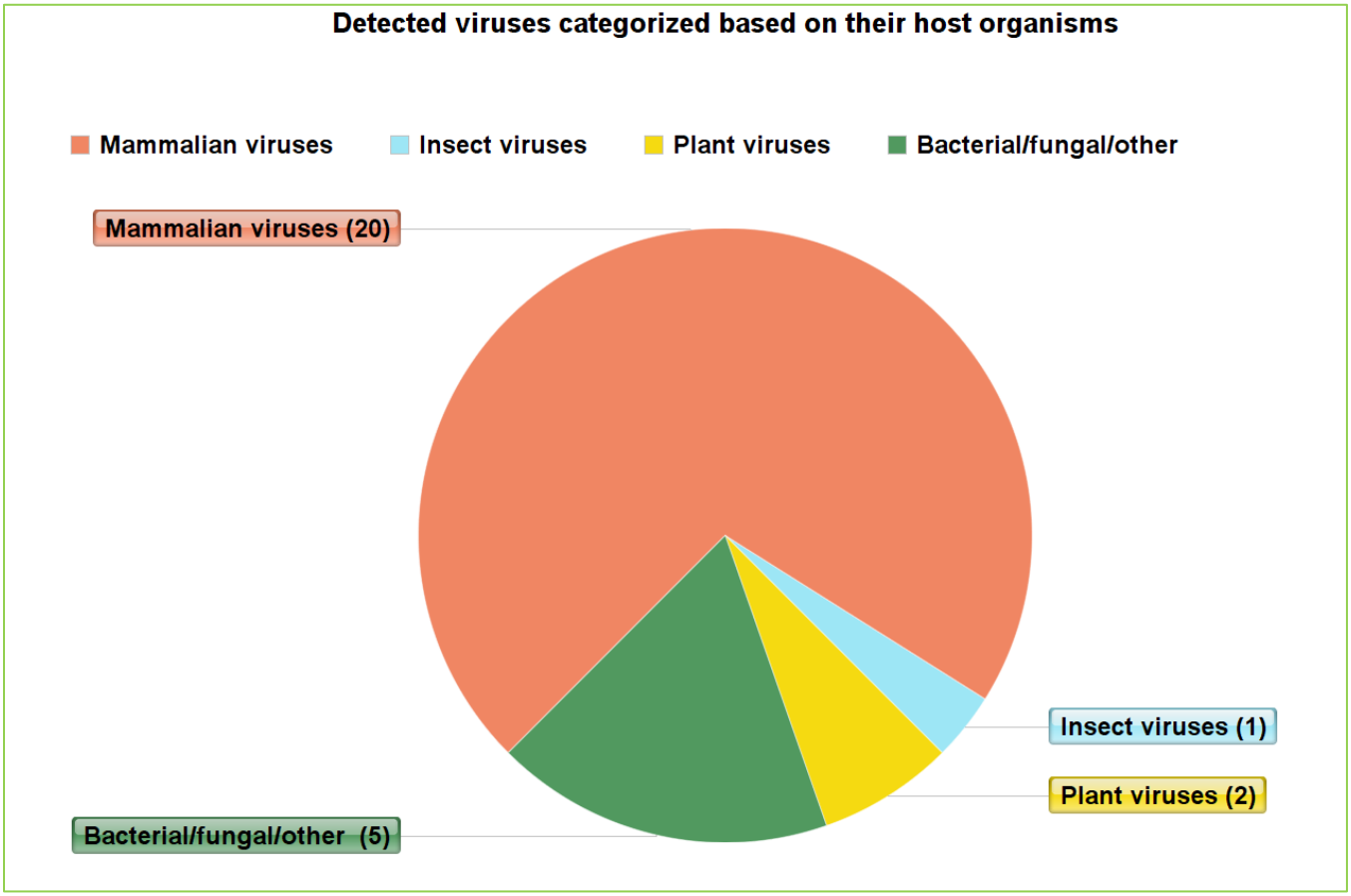

Figure 4-3: Pie chart of the different types viruses from 12 faecal samples categorized based on their natural hosts.

4.3.6. Viral taxonomic distributions

Although analysis of the complete (DNA and RNA) virome is imperative, the current research study was fundamentally aimed at determining and characterizing the infant's gut RNA virome. Therefore, this study was focusing more on the mammalian gut RNA viruses, which play a very significant role on human health and disease. **Figure 4-4** is a pie chart showing the distribution of contigs from RNA viruses that belong to different virus families. As shown on the Figure, a total of seven different families of mammalian RNA viruses were present in the human faeces.

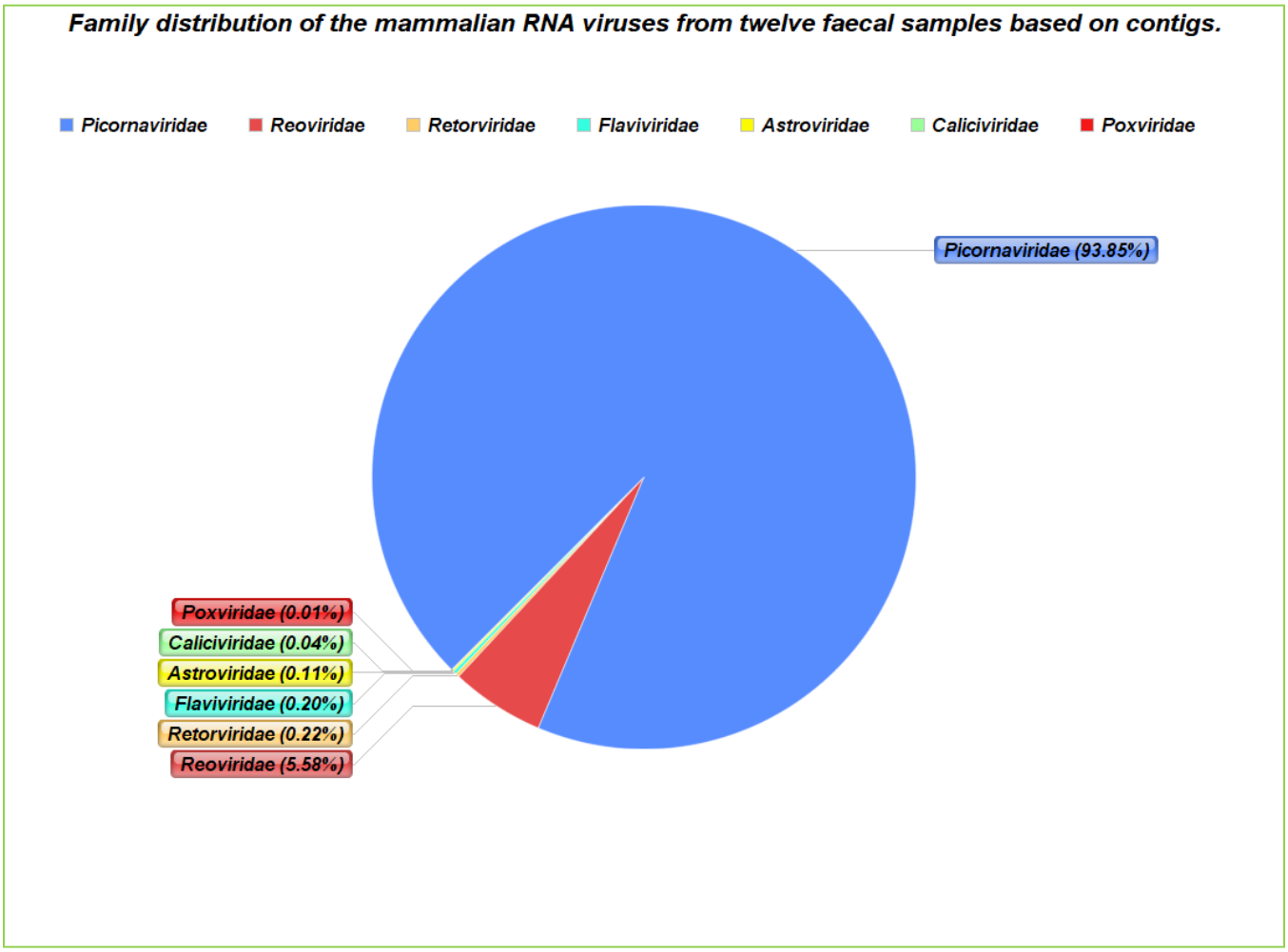

Figure 4-4: Pie chart showing percentages of contigs of the RNA viruses from all twelve faecal samples classified into viral families.

The pie chart distinctly shows that contigs that belong to viruses within the family of *Picornaviridae* were the most abundant. *Reoviridae* which is a family of dsRNA viruses was the second most abundant and although this virus family consists of only 5.58 % of the total viral contigs, rotavirus which belongs to this family was the only virus

detected in all 12 faecal samples. The remaining five other RNA virus families comprised of less than 1 % of the viral contigs. These families were, in decreasing order of prevalence, *Retroviridae* (0.22 %)*, Flaviviridae* (0.20 %)*, Astroviridae* (0.11 %), *Caliciviridae* (0.04 %), *Poxviridae* (0.01 %).

Further analysis was performed to determine the distribution of viral contigs at genus and/or species level, focusing specifically on mammalian RNA viruses. **Figure 4-5** is a pie chart showing the distribution of these detected virus genera and/or species across the 12 human stools. **Figure 4-5** shows data of the 10 virus genera/species that were detected and others were not specified because they had at most two contigs, mostly in a single human faecal sample. According to **Figure 4-5**, it is apparent that nearly 50 % of the viral contigs were derived from coxsackieviruses, a species of positive-sense single-stranded RNA viruses. This was followed by polioviruses, with nearly 30 % of the viral contigs mapping to this species of positive-sense single-stranded RNA. Furthermore, enteroviruses and rotaviruses showed nearly 1 % difference between the one another, both consisting of 6.93 % and 5.59 % of the viral contigs respectively. Echoviruses were the fifth most abundant viruses with 3.37 %, whilst the remaining five families consisted of less than 1 % of the viral contigs. Interestingly, one of the most important human enteric pathogens, norovirus, was detected although there were only four contigs that mapped to this virus detected in one sample.

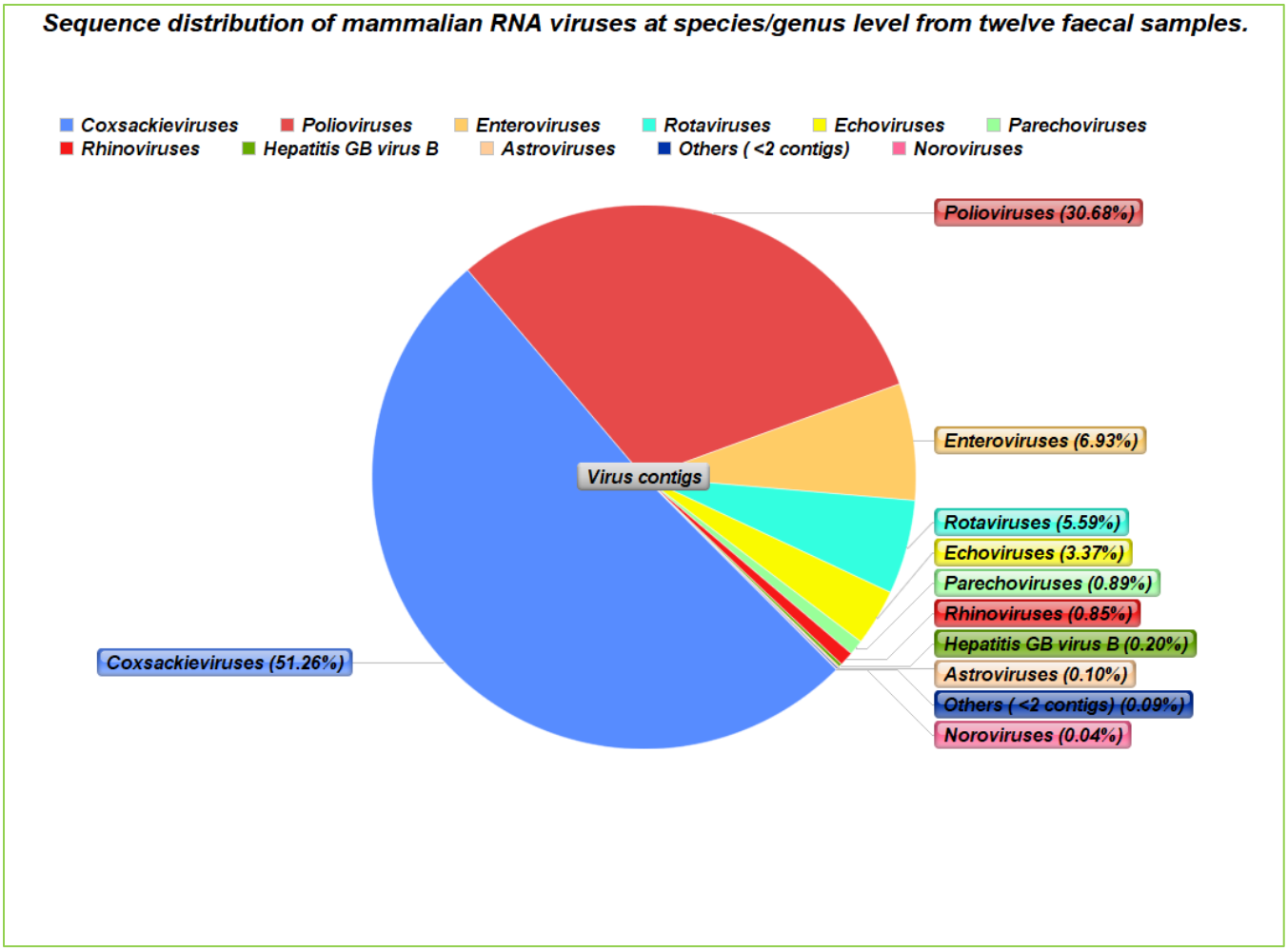

Figure 4-5: Sequence (contigs) distribution of the detected mammalian RNA viruses at genus/species level from twelve faecal samples.

4.3.7. Gut virome composition and dynamics over time

Understanding the role of gut viral populations in healthy and disease states of humans requires profound knowledge about their composition and dynamics. In order to understand this and establish the virome patterns within each individual, it is crucial to monitor and compare the changes in viral populations between collection time points. Therefore, in this study analysis was done to compare the virome composition between the three collection time points for each of the four participant. **Figure 4-6**, **Figure 4-7**, **Figure 4-8** and **Figure 4-9** are the bar graphs that demonstrate how the viral populations changed over time for particpant A, B, C and D focusing specifically on the RNA virome. The detected viruses are shown on the x-axis and the number of assembled contigs

for each virus on the y-axis. Some viruses which were detected in certain individuals were omitted in the graphs below because of their very low abundance whereas others were not mammalian viruses.

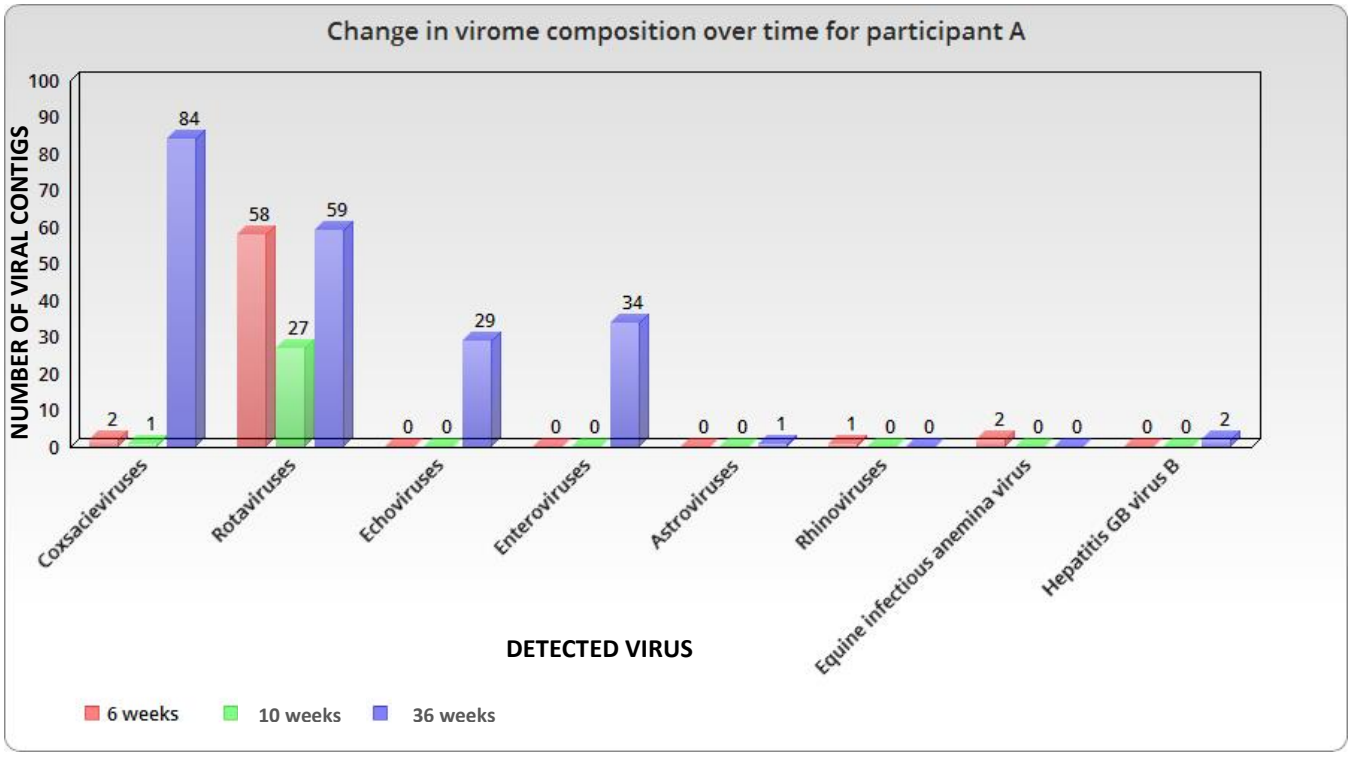

Figure 4-6: Bar graph showing the changes in the faecal virome composition throughout the three collection time points for participant A.

Participant A was a female infant. Changes in virome composition for this participant throughout the three time points from a baseline of 6 weeks, then second collection at 10 weeks and the last collection at 36 weeks are show in (**Figure 4-6**). The changes in contig abundance of eight mammalian viruses that were detected in each of the three samples in participant A are indicated (**Figure 4-6**). The ninth virus that was not included in this part of the analysis was Scheffersomyces segobiensis virus L which is a fungi-infecting dsRNA virus under the family *Totivirdae*. Nevertheless, looking at **Figure 4-6**, most of the viral contigs seemed to have been low at baseline (6 weeks) for majority of the viruses. Five contigs detected in this sample belonged to three of the mammalian viruses, namely coxsackievirus (2), equine infectious anemia virus (2) and from rhinoviruses (1). Furthermore, 58 contigs were attributed exclusively to rotaviruses. There were no contig(s) for the remaining four viruses as indicated by the red bars (**Figure 4-6**). During the second collection done at 10 weeks of age, represented by the green bars, very minimal viral contigs could be detected. Only rotaviruses were present in abundance although there was a decline, with nearly half the number of contigs from 58 contigs at 6 weeks to 27 contigs at 10 weeks old.

Apart from rotavirus, there was no significant difference in coxsackievirus detection as there was only one contig in the second collection. These were the only two viruses that could be detected in the stool sample of participant A. Surprisingly, at 36 weeks of age (blue bars) a rapid increase in the number of viral contigs was observed for 50% of the viruses in this participant. Some of these viruses could not be detected in the first two faecal collections. The most abundant viral contigs belonged to coxsackievirus which increased from just a single contig in the second collection to 84 contigs in the final collection. Moreover, in the final collection rotavirus maintained nearly the same number of contigs (59 contigs) as in the beginning (58 contigs) despite a drop by almost half observed in the second collection (27 contigs). Interestingly, echoviruses and enteroviruses were not present in the first two faecal samples. However, a significant number of contigs that mapped to these viruses were detected in the last collection. There were 29 assembled contigs for echoviruses and 34 contigs for enteroviruses. The other two viruses that were only detected in the last collection were reported in small numbers, included astrovirus (1 contig) and hepatitis GB virus B (2 contigs). Lastly, contigs for rhinoviruses and equine infectious anemia viruses were still not detected at 36 weeks.

On the other hand, participant B (**Figure 4-7**) showed a slightly different pattern in terms of the dynamics of viral populations across the three collection time points. Faecal samples for this participant were collected at 10, 14 and 24 weeks as indicated below, and these time points varied considerably with those of individual A. In contrast to participant A, the changes in the gut virome composition for participant B involved ten different viruses, which was two more than the first infant (**Figure 4-7**).

To some extent, a particular pattern could be established from **Figure 4-7** in terms of how the viral populations changed between the specified time points. Significantly, 90 % of the viruses showed an increase in the number of contigs between the first and second collection, followed by a drastic drop in the final collection (**Figure 4-7**).

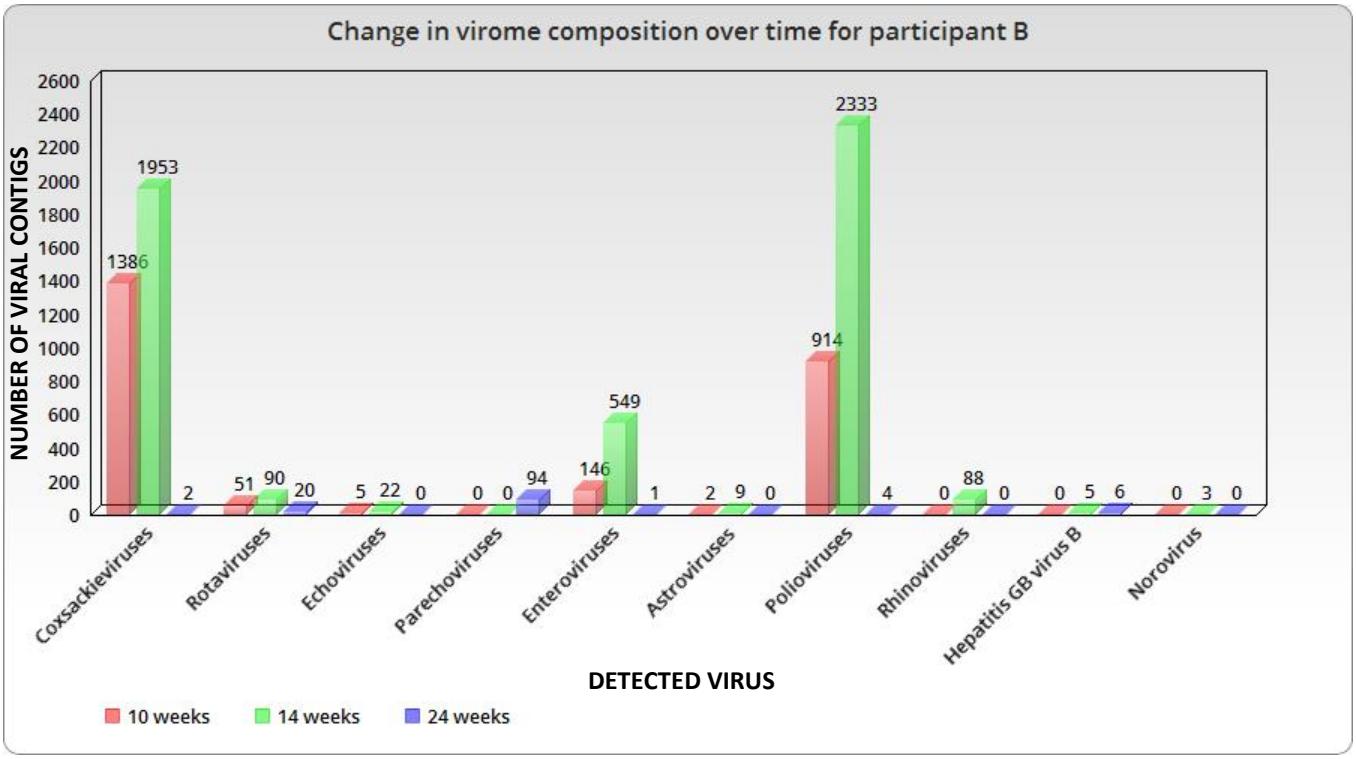

Figure 4-7: Bar graph showing the changes in the faecal virome composition throughout the three collection time points for participant B.

At 10 weeks (red bars) 1386 contigs that belonged to coxsackievirus were detected making this RNA virus the most predominant at baseline for study participant B. This was followed by polioviruses with 914 contigs, then enteroviruses and rotaviruses with 146 and 51 viral contigs, respectively. Contigs that belonged to echoviruses and astroviruses were only present in very low numbers, each having 5 and 2 contigs respectively. At baseline, there were no contigs detected for the four other viruses that was present in this participant, namely parechovirus, rhinovirus, hepatitis GB virus B and norovirus (**Figure 4-7**).

At 14 weeks of age, a sharp rise in the number of contigs was observed in the infant's stool. The contigs for all except one virus (parechovirus) were detected in much more quantities than in the first collection. This includes even viruses that were not present at baseline. The highest number of viral contigs detected at 14 weeks were observed in polioviruses, increasing from 914 in the first collection to a total of 2 333 contigs in the second collection. This was followed by coxsackieviruses which increased from 1 386 to 1 953 contigs. Enteroviruses and rotaviruses followed with 549 and 90 contigs, respectively. The other two viruses (rhinoviruses and echoviruses) had at least 20 contigs, whereas the remaining three, apart from parechovirus, had less than 10 contigs each. Parechovirus was an exception in this case because there was not a single contig detected for this virus in the first two collected faecal samples (**Figure 4-7**).

Interestingly, a rapid decrease in the number of viral contigs was seen in the last collection time point for all the viruses that showed a rise in contig numbers in the second collection. The decrease in viral contigs was so dramatic that some of the viruses were completely cleared at 24 weeks, whilst some were very few. At 24 weeks of age, there were only two contigs for coxsackievirus dropping from 1953 at 14 weeks. Polioviruses also showed a significant decrease from 2 333 contigs at 14 weeks to only four contigs detected in the final collection. Some of the viruses that exhibited a large decrease in the amount of contigs included enteroviruses and rhinoviruses both being completely cleared from the faecal samples in the last collection time point. Although there was a decrease in rotavirus contigs, a significant number of contigs (22) were still present in the sample during the third collection. Unexpectedly, parechovirus contigs which were not detected in the first two collected samples, suddenly emerged in the third time point with a total of 94 contigs having being detected. Moreover, astrovirus and norovirus contig were not present in the third time collection time point, whilst hepatitis GB virus B contigs showed no significant change in the number of contigs between the last two collection times although the viral contigs were lowly detected (**Figure 4-7**).

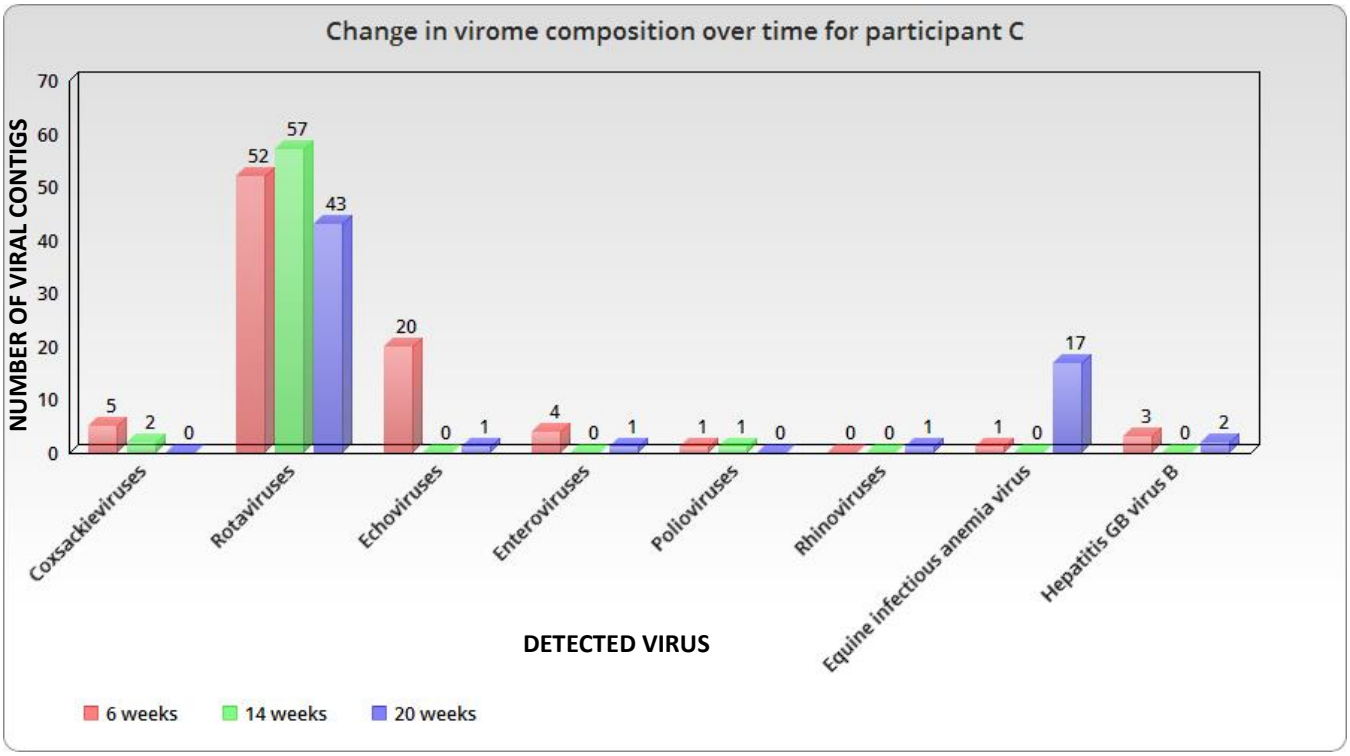

Figure 4-8: Bar graph showing the changes in the faecal virome composition throughout the three collection time points for participant C.

Figure 4-8 shows the changes in viral populations over three collection time points for participant C. The baseline for this individual was 6 weeks, second fecal sample was collected at 14 weeks and lastly 20 weeks of age. In this

study participant, eight different viruses were used in monitoring the gut virome dynamics over time. It is clear from the bar graph in **Figure 4-8** that majority of the viruses had minimal number of contigs, with only rotavirus being present in abundance (**Figure 4-8**). At 6 weeks of age, the two most predominant viral contigs belonged to rotaviruses and echoviruses. There were 52 contigs that were assigned to rotaviruses and 20 contigs assigned to echoviruses, while the other viruses had contigs ranging from 1 to 5 contigs at baseline. These viruses were in decreasing order of contig hits detected as follows: coxsackievirus (5), enterovirus (4), hepatitis GB virus B (3), poliovirus (1) and equine infectious anemia virus (1) (**Figure 4-8**).

Furthermore, during the second collection time point (14 weeks), most of the viruses could not be detected. There was only three viruses detected in participant C of which rotaviruses were the most prevalent with a slight increase in the number of contigs from 52 to 57 between the baseline and second collection. Apart from rotavirus, there were only two coxsackievirus contigs and one contig that belonged to poliovirus detected in participant C at 14 weeks (**Figure 4-8**). Although echovirus was the second most abundant virus at baseline with 20 contigs coming after rotavirus, there were no contigs that belonged to echoviruses in the second collection time point. In addition, enteroviruses, equine infectious anemia virus and hepatitis GB virus B were also not present in the second collection (**Figure 4-8**).

In the third collection time point there was no significant difference in the number of viral contigs, except for the appearance of equine infection anemia virus increasing to 17 contigs. On the other hand, although there was a fair decrease in the number of rotavirus contigs, they were still present in abundance with 43 contigs being detected in the last collection.

Figure 4-9 demonstrates the viral population changes over time for participant D. As it appears on this figure, participant D had the least number of viruses compared to the other three study participants. It is also important to note that the collection time points here are similar to those of participant C (6, 14 and 20 weeks). **Figure 4-9** indicates that rotavirus was the most predominant virus in terms of the detected number of contigs, similar to participant C. **Figure 4-9** reveals that the number of rotavirus contigs are increasing progressively over time starting with 32 rotavirus contigs at a baseline of six weeks. Furthermore, the rotavirus contigs increased to 37 at 14 weeks and in the third collection time point a total of 67 rotavirus contigs were detected.

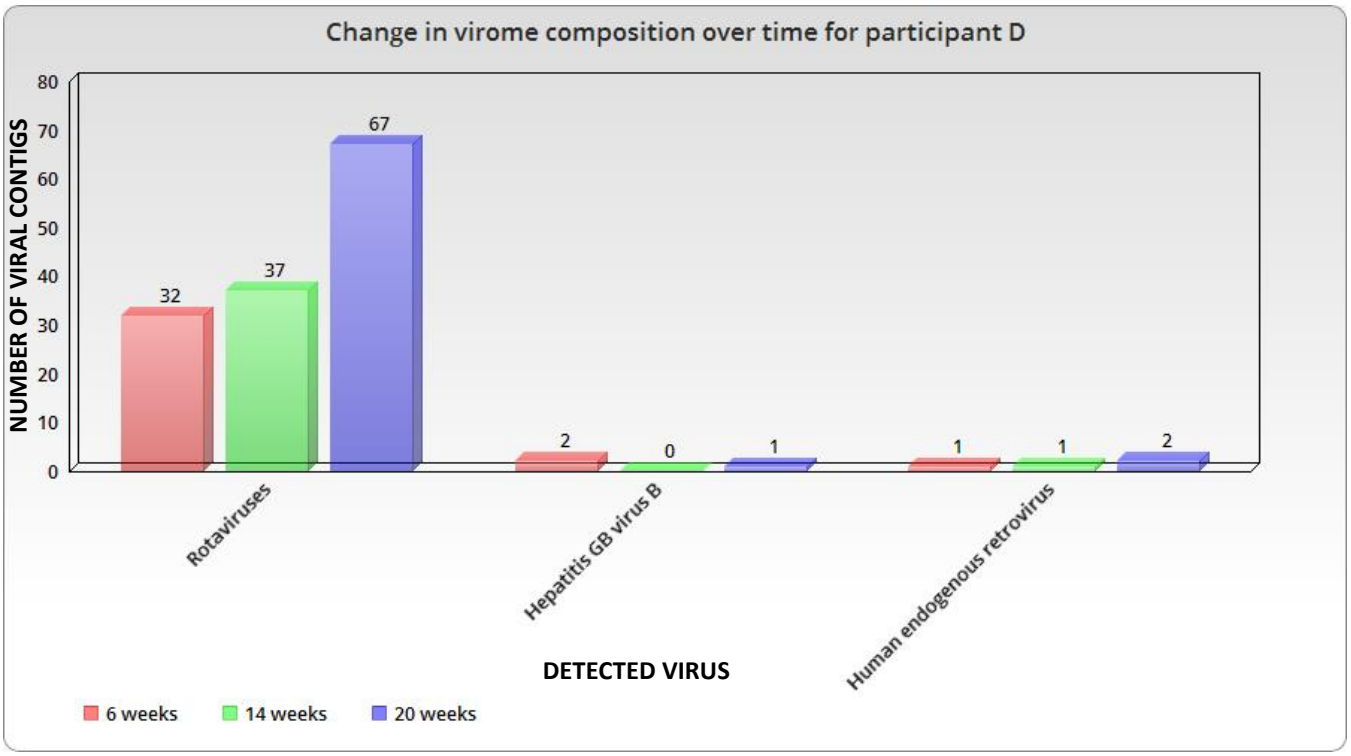

Figure 4-9: Bar graph showing the changes in the faecal virome composition throughout the three collection time points for participant.

Furthermore, minimal number of contigs that belonged to hepatitis GB virus B as well as human endogenous retrovirus were detected. These remained constant throughout the three collection time points. As illustrated on **Figure 4-9**, there were two hepatitis GB virus B at 6 weeks, which disappeared at 14 weeks and only one contig was detected in the third collection. Likewise, there was only one contig for human endogenous retrovirus that was detected in the baseline and again in the second collection time points. This increased to two in the last faecal sample.

4.4. Discussion

The study of the virome by several metagenome scientists has gained momentum, albeit partially understood. This study undertook to characterize the taxonomic compositions of the infant's faecal virome, particularly targeting the RNA virome. The *de novo* assembly and BLASTX annotation using DIAMOND revealed a wide diversity of eukaryotic viruses in the stool of the infants, majority of which were RNA viruses. The most commonly detected viral family was *Picornaviridae*, predominantly coxsackieviruses with over 5 000 contigs assigned to coxsackieviruses by BLASTX. Although coxsackieviruses were detected in three of the four infants, a vast majority of coxsackievirus contigs (98.3 %) were detected in participant B only, whereas less than 2 % of the coxsackievirus contigs were distributed between participant A and C.

Within the family *of Picornaviridae*, contigs assigned to polioviruses were the second most predominant. However, only one of the four infants (participant B) presented large quantities of poliovirus contigs in their faecal samples, whereas a negligible quantity of poliovirus contigs were observed in the first two samples of participant C (Table 2). A secondary BLASTN search of randomly selected contigs that were identified as polioviruses suggested that the poliovirus sequences detected in the few stool samples of the study infants were likely oral poliovirus vaccine strains (OPV) and not wild-type polioviruses. The South African Department of Health report has shown that there has been no transmission of wild-type poliovirus since 1989 (WHO, 2017). In other countries in Africa, such as the Democratic Republic of Congo (DRC), the last confirmed case of wild-type poliovirus paralysis was in 2011 (Gumede *et al*., 2013). On the other hand, cases of vaccine-derived polioviruses in the DRC have been reported for the past 14 years.

In this study, the increase in poliovirus sequences in participant B between 10 and 14 weeks was attributed to replication of the oral poliovirus vaccine strains in the infant's intestines (**Table 4-2**). The drastic decline in poliovirus contigs at 24 weeks is in correlation with previous reports that after exposure to the oral poliovirus vaccine, immuno-competent individuals excrete poliovirus vaccine strains for a short period of time, often for less than two months (Kew *et al*., 1998, Hovi *et al*., 2004). On the contrary, several immuno-deficient OPV recipients been reported to excrete polioviruses for several years (Kew *et al*., 1998, Bellmunt *et al*., 1999, Shulman *et al*., 2000).

Importantly, several studies have also reported that the oral poliovirus vaccines can undergo genetic mutations resulting in reversion to neuro-virulence, thereby leading to the emergence of vaccine-derived polioviruses with transmissibility properties similar to that of wild-type poliovirus strains (Buttinelli *et al*., 2003, Hovi *et al*., 2004). Such events have been reported to occur after extended replication in the gut of immune-deficient persons

following vaccination or during transmission between individuals in populations of low immunity (Gumede *et al*., 2013; Diop *et al*., 2015; Jorba *et al*., 2016, Jorba *et al*., 2017).

Moreover, human enteroviruses which are commonly transmitted via the faecal-oral route and occasionally by respiratory droplets (Rotbart, 2000; Julbert and Lipton, 2013) were also among the three most prevalent viruses under the *Picornaviridae* family of single-stranded RNA viruses. Interestingly, as shown in **Table 4-2** enteroviruses were consistently detected alongside echoviruses which belong to the same virus family. In five out of six samples where enteroviruses were detected, echoviruses were also present (**Table 4-2**). Although the infants under study were asymptomatic and no clinical history was available, enteroviruses are known to cause a range of illnesses including among others respiratory infection, hand-foot-and-mouth disease and aseptic meningitis (Pallansch *et al*., 2013; Pons-Salort *et al*., 2015). A recent study has reported meningitis outbreaks linked to human enteroviruses in South Africa (Smuts *et al.*, 2018). Another earlier study had also reported on enterovirus-associated meningitis in young children in other parts of South Africa (Wolfaardt *et al*., 2014). Nevertheless, numerous studies have established that many of the enteric viruses are persistently shed in healthy humans (Kapusinszky *et al*., 2012). Human enterovirus (Witsø *et al*., 2006) and parechovirus (Kolehmainen *et al*., 2012), for instance, are shed by children aged 0-5 years with no evidence of association with disease. A one-year NGS longitudinal faecal virome study of two healthy infant siblings demonstrated continuous shedding of diverse gut viruses (Kapusinszky *et al*., 2012).

A recent faecal virome study which was aimed at characterizing viruses from paediatric patients with severe hand-foot-and-mouth disease reported that nearly 60 % of the identified viruses were positive-sense single-stranded RNA viruses belonging to the *Picornaviridae* family, such as enteroviruses, coxsackieviruses, echoviruses and human rhinoviruses (Wang *et al*., 2018). In mild cases of hand-foot-and-mouth disease, viral populations were more complex, with the detection of nine families of RNA viruses, namely, *Picornaviridae*, *Reoviridae, Astroviridae, Caliciviridae*, *Virgaviridae, Tymoviridae, Orthomyxoviridae*, *Tombusviridae* and *Alphaflexiviridae* (Wang *et al*., 2018). Although it is not clear if the infants in the current study had previously been diagnosed with hand-foot-and-mouth disease, several of the viral pathogens belonging to *Picornaviridae* detected in their faecal samples have been reported as the causative agents of hand-foot-and-mouth disease in young children, including human enteroviruses and coxsackieviruses (Hamaguchi *et al*., 2008; Zhu *et al*., 2008).

In contrast to coxsackievirus and poliovirus sequences that were recovered in eight and five samples, respectively, a 100% detection rate of the genus rotavirus was attained. That is to say, rotavirus contigs were detected in all three samples collected from each of the four infants. Group A rotaviruses are the leading cause of acute gastroenteritis in young children across the world, with up to 128 500 deaths occurring every year among children

under 5 years old, most frequently in developing countries of Sub-Saharan Africa and Asia (Troeger *et al*., 2018). Although no prevalence studies were done at genotype level for rotaviruses, secondary BLAST search analysis of the randomly selected rotavirus contigs performed against the NCBI database revealed the presence of G1P[8] strains among others, which is the same genotype combination as that included in the Rotarix (GlaxoSmithKline, Rixensart, Belgium) vaccine. Rotarix is a two-dose schedule live-attenuated human rotavirus oral vaccine administered to infants from 6 to 24 weeks (European Medicines Agency, 2016) and at 6 and 14 weeks in South Africa (Seheri *et al*., 2012). It is possible that there could be vaccine-derived strains present in these samples since the study participants were all vaccinated and the sample collection times were very close to the vaccination time, In addition, a vaccine-derived rotavirus can replicate in the gut of a vaccinated individual and be shed in their stool (Sakon *et al*., 2017). However, it is difficult to make a conclusive statement on whether or not the strains detected in this study were vaccine-derived because only a fraction of the contigs was used in BLAST search and also no reference mapping was done to confirm this. It has been shown in a burden of disease surveillance study conducted in the Gauteng and the North West provinces of South Africa less than a decade ago that 90 % of children under 24 months old visiting the outpatient department suffered from rotavirus infection, which occurs as early as 2 months of age (Seheri *et al*., 2010). In South Africa, most incidences of rotavirus infection have been recorded to occur among children between 3 and 17 months of age (Steele *et al*., 1986; Seheri *et al*., 2010). This varied slightly in comparison to developed countries with the age ranging from 6 to 23 months old (Giaquinto and Van Damme, 2010). These findings, to some extent, support the speculation that the infants under study might have also suffered asymptomatic infection with rotavirus at some point or recovering from rotavirus disease albeit not having diarrhoea at the time of stool collection.

Nevertheless, unlike the *Picornaviridae* family under which there were several different viruses detected, rotavirus was the only genus detected under the *Reoviridae* with 593 rotavirus contigs in the 12 faecal samples. Of note, these contigs were all identified as group A rotaviruses by BLAST annotation. In addition, BLAST identification indicated that some rotavirus sequence could be of non-human mammalian origin such as bovine, porcine and rhesus, suggestive of possible interspecies transmission. However, considering the very young age of the participants, the speculation is that these infants might have potentially acquired the viruses elsewhere, eventually ending up in their gut.

Other viral families that constituted the infant's gut virome in this study included *Astroviridae* and *Caliciviridae*, and the viruses classified under these two families were astrovirus and norovirus, respectively. Even though the viral contigs for these two genera (norovirus and astrovirus) were detected in low prevalence, they are important and well-characterized viral agents that have been implicated in cases of childhood diarrhoea. Although this gut

viral metagenomic study involved healthy infants, norovirus is known as the most common agent of acute gastroenteritis in humans of all ages worldwide (van Beek *et al*., 2013). It is highly contagious and can be transmitted in several ways such as by contact with infected persons, contaminated environments, or by consumption of contaminated foods (Parashar *et al.,* 1998). However, norovirus infections can also result in subclinical symptoms, thereby going undiagnosed (Ayukekbong *et al*., 2015). In this study, norovirus contigs were only detected in one stool sample. Several studies have shown that asymptomatic norovirus infections are common (García *et al.,* 2006), and a study conducted in Cameroon reported high asymptomatic norovirus prevalence of 27 % in adults and around 30 % in children (Ayukekbong *et al*., 2014a). In the current study, the norovirus contigs that were detected in participant B (24 weeks) was likely asymptomatic infection, most probably acquired from the parent or the environment.

Expectedly, few contigs mapping to bacteriophages and plant viruses were identified. The detection of bacteriophages and plant viruses in the human gut virome has been described before (Minot *et al*., 2011; Cotton *et al*., 2014). Only one sample each from two infants (participant B and C) contained one plant virus, namely pepper mild little virus and wheat rosette stunt virus, respectively (**Table 4-2**). Coincidentally, the two viruses were detected at the same age (14 weeks old) in the respective infants. Pepper mild mottle virus, which is a well-characterized plant pathogen that infects all species in the genus *Capsicum*, has previously been reported to be present in abundance in the faeces of healthy humans (Zhang *et al*., 2006). These plant viruses have never been associated with any human disease. The presence of plant viruses in stool samples of humans is often a reflection of acquisition from diet. However, in the current study the participants were only two months old, and with their dietary information missing, we speculated that these viruses were introduced through weaning or breastfeeding.

Although bacteriophages have been documented as the major constituent of the intestinal virome, accounting for nearly 90 % of its composition (Reyes *et al*., 2012), only one contig of geobacillus virus which belongs to the *Siphoviridae* family of dsDNA bacterial viruses in the order *Caudovirales* was reported in this study. This was unsurprising since this study was restricted to RNA viruses. As described in the methodology section of Chapter 3, the protocol used to enrich for RNA virus particles involved filtration and digestion of non-encapsidated nucleic material using nuclease enzymes. Consequently, bacterial cells on which bacteriophages depend for their replication were filtered, and the prophages that are integrated into the genomes of bacteria were eliminated along with their prokaryotic host cells leading to low phage detection. Apart from successful enrichment of virome, the low prevalence of phages in this virome studies could also be due to the absence of reference sequences as many virome studies have reported that the bacteriophage sequences are mostly unknown due to the unavailable

of homologs in the public database (Breitbart *et al*., 2003; Zhang *et al*., 2006;; Minot *et al*., 2012; Minot *et al*., 2013; Reyes *et al*., 2013).

To compare the changes and dynamics of the viral populations over time, it was observed that of the nine different viruses that were detected in participant A, only coxsackieviruses and rotaviruses were present throughout the entire collection period (**Table 4-2**, **Figure 4-6**). One would actually expect the viral sequences to decrease with time due to a maturing immune system. However this was not the case, in fact, the viral sequences were mostly detected in the last collection time point, at the age of 36 weeks (equivalent to nine months). Conversely, in the second collection, there was minimal diversity in terms of viral populations, with only two viruses detected. In the last collection, enterovirus contigs were the most abundant in this sample. Interestingly and in correlation with our findings with regards to sample A3, a related study that investigated the gut virome composition and dynamics of healthy infants reported that eukaryotic viral population richness of the infant's gut was low early in life and increased thereafter. Based on this observation, the assumption was that the eukaryotic virome is established through environmental exposure (Lim *et al.*, 2015). Furthermore, in that particular study, enteroviruses and parechoviruses were among the most commonly detected viruses and since similar viral sequences were also detected in our study, these findings could suggest that these enteric viruses form part of the viral flora in the gut of healthy infants.

Although human rhinoviruses, which are single-stranded RNA viruses within the *Picornaviridae* family, are considered as respiratory pathogens since it has been reported that their replication in the intestinal mucosa is hampered by their sensitivity to low pH, our study and other studies have reported their detection in faecal samples of young children (Salminen *et al*., 2004; Harvala *et al*., 2012; Honkanen *et al*., 2013). The detection of human rhinoviruses in this gut virome study implies that stool samples can offer a useful addition to respiratory samples for researchers who are investigating rhinoviruses in infants and young children. Although human rhinovirus sequences were not present in high frequencies in two of the participants, the fact that they could be detected in abundance in the stool of participant C suggests that they can either replicate in the intestines or are ingested by the infants and make their way into the GIT. The possibility of sample contamination with human rhinoviruses during faecal collection cannot be ignored. Previous studies have shown that the intestinal pH of infants is not as low as that of adults (Wills and Paterson, 1926; Maffei and Nobrega, 1975), which could be the reason for the frequent detection of rhinoviruses in the faecal samples of infants. In addition to that, the human rhinovirus genome has a protein capsid protecting viral RNA from degradation, thereby allowing the virus to reach the stomach without any loss of genome integrity (Wills and Paterson, 1926; Maffei and Nobrega, 1975). That means even if the virus may lose its infectivity, the viral genome enclosed within the capsid would remain

unaffected.

Apart from the above-mentioned human viruses discussed so far, there was also a detection of a few sequences that mapped to non-human mammalian viruses in the stool samples of participant A. These included equine infectious anaemia virus, a horse-infecting viral pathogen from the family *Retroviridae* commonly transmitted by blood-sucking insects such as horse fly and deer fly. The other virus detected was swine vesicular disease virus, a pig enterovirus from the family *Picornaviridae*. The latter has previously been reported to be a porcine variant of human coxsackievirus B5 (Lin and Kitching, 2000), a member of the *Picornaviridae* which was detected in majority of the samples in this study. Additionally, sequences that belonged to scheffersomyces segobiensis virus L were present in stool samples of participant A and B. This is a double-stranded RNA virus member of the genus Totivirus to which fungi serve as natural hosts.

The largest number of different species of viruses were identified in Participant B, with 23 different viruses detected from the three stool samples. Participant B had the most diverse population of viruses, most of which were eukaryotic RNA viruses, particularly single-stranded RNA viruses. To some extent, this infant exhibited a characteristic pattern over the three collection time points wherein there was a significant increase in the viral contigs between 10 and 14 weeks for certain viruses including coxsackieviruses, rotaviruses, echoviruses, enteroviruses, astroviruses, polioviruses, rhinoviruses and noroviruses. A sudden decline in the viral contigs was observed at 24 weeks with some of the viruses going undetected (**Table 4-2**, **Figure 4-7**). Interestingly, these eight viruses, with the exception of rotaviruses, were positive-sense single-stranded RNA viruses, and five of them were members of the genus *Enterovirus* in the family of *Picornaviridae*, implying that the infant's enteric RNA virome composition is quite diverse and it is dominated by ssRNA viruses that belongs to the *Picornaviridae* family. Although participant B did not exhibit any symptoms of diarrhoea, the presence of norovirus sequences in the last collected sample might suggest asymptomatic infection or the viral load of norovirus was too low to induce diarrhoea. Norovirus, member of the *Caliciviridae* family, was detected in only one of the 12 faecal samples collected from four infants. Asymptomatic cases of norovirus infections have been reported before. In fact, a recent study has estimated the prevalence of asymptomatic infection based on the analysis of 71 studies to be 7 % (Qi *et al*., 2018). In that particular study, higher prevalence was observed more in children than in adults, linking asymptomatic infection to immunity, since children are more susceptible to viral diseases due to their immature immunity. The detection of norovirus in the faecal sample on participant B could be shedding from a previously symptomatic infection. The history on previous infections was not available, therefore it was not possible to conclude on that. Furthermore, food contamination could be another source of norovirus transmission, bearing in mind common mode of transmission is via faecal-oral-route, poor hygiene by food handlers can contribute to

norovirus infections (Hall *et al*., 2014). More interestingly, an exception to the gut RNA virome composition of participant B was the recovery of a single contig that belonged to a DNA virus, vaccinia virus classified under *Poxviridae* family was detected in the last collection time point.

In the same participant, several non-human eukaryotic viruses were mostly detected in the second sample collection but not in the other two stool samples. These viruses were predominantly (+)ssRNA virus two of which included unclassified members of the order *Picornavirales*, namely apodemus agrarius picornavirus and niniventer confucianus picornavirus. The former is known to infect striped field mice, whilst the latter infects a species of rodents in *Muridae* family. Other non-human eukaryotic virus sequences recovered in participant B included bat picornavirus, marmot sapelovirus, genus *Mamastrovirus*, giant panda-associated partiti-like virus, swine vesicular disease virus, and shuangao insect virus. Although the quantity of contigs for each of these non-human eukaryotic viruses in participant B did not exceed five, the diversity of viruses especially from field animals suggests possible interspecies transmission due to indirect environmental exposure. Considering that Oukasie is a small poor Township in the rural North West, one can speculate that the home of some participants are in close proximity with a dumping site, fields or bushy areas which are often associated with rodents and other field animals. Other contributing factors may include poor hygiene and unavailability of proper sanitation or leaking sewage systems that can grossly contaminate the environment.

Parechoviruses, another (+)ssRNA viruses within *Picornaviridae* was the only virus found exclusively in participant B. It has previously been detection in asymptomatic subjects. More specifically, a case-control study was recently done in Ghana to determine the prevalence of human parechoviruses as well as its association with diarrhoea in children (Graul *et al*., 2017). The study reported high detection rate and diversity of human parechovirus in asymptomatic controls. Although in our study, the detection rate of parechoviruses was low (1/12 samples). The fact that the virus was recovered from an asymptomatic sample, an observation similar to that in the Ghanaian study, support the view that non-human parechoviruses is not associated with acute diarrhoea.

As opposed to participant B, relatively less diverse viral populations were observed in participant C, and the amount of contigs of detected viruses were also very low throughout the three collected samples. The same enteric viruses that were described in the previous two participants were also present in participant C, most of which belonged to *Picornaviridae* family. However, rotavirus was the only virus that could be detected throughout the three collection time points. Intriguingly, three viruses that could be detected exclusively in participant C were all non-human eukaryotic viruses. One was a plant-infecting virus (wheat rosette stunt virus), the other two were double-stranded RNA fungal-infecting viruses in the family *Amalgaviridae* and *Partitivirade,* respectively, both were detected in the last collection time point (20 weeks).

Participant D had the lowest diversity of all, with rotavirus being the only virus detected in all three-stool samples. Human endogenous retrovirus is commonly found integrated into the host genome and sequences of this virus were discovered in the first and the last collected sample of participant D. Specially, participant B was one of the two that had a DNA virus and the only infant in which a bacteriophage sequence was recovered (geobacillus virus).

This study has provided a description of the enteric virome of infants from Oukasie Township. To our knowledge, this study represents the first cohort metagenomics characterization of the RNA enteric virome of infants from South Africa. The current study has demonstrated that even in asymptomatic cases, the human gastrointestinal tract is still colonized by diverse populations of eukaryotic viruses. However, due to the viral loads being probably below detection the viruses identified in this study might represent minimum quantities of the infant's virome and it is possible that some viruses may have gone undetected. Furthermore, taking into account the enormous size and diversity of viral populations that have been discovered thus far, very often a significant proportion of the sequences obtained from metagenomic studies of the human virome cannot be annotated and classified into taxonomic groups. Moreover, the fact that only a few thousand viral reference sequences exist in genome databases contributes to this, meaning that any potential new virus obtained from the human or environmental sample will usually lack resemblance to genomes in the databases.

4.5. Conclusions

The human gut virome remains largely unexplored. Several drawbacks like the financial load of metagenomics studies, limited number of study participants, unavailability of clinical information and inadequate analysis tools, makes it difficult to draw general conclusions from viral metagenomics studies. Nevertheless, a lot of progress has been made with some intriguing findings that include discovering the changes and dynamics of viral populations throughout the lifetime of a human being, to how age and geographical differences influence virome compositions of an individual. However, most virome research studies have focused more on characterization of DNA viruses, particularly the diversity of bacteriophages. Consequently, characterization of the RNA virome still lags behind. Moreover, the sequence databases currently available are known to contain sequence information of known viruses, especially those that are deemed clinically and economically relevant. As a result, it is challenging to assign viruses recovered from metagenomic studies into viral families.

In this study, we investigated the composition of the enteric virome of four healthy infants under the age of one year in a follow up study. Known enteric viral pathogens were detected in all study participants, majority of which were single-stranded RNA viruses as anticipated. The only two DNA viruses included a bacteriophage and vaccinia virus. The family of *Picornaviridae* was the most predominant, with coxsackieviruses and polioviruses standing out as two viruses whose contigs were significantly higher than the other detected viruses. Further analysis is needed to confirm that the poliovirus sequences detected in two of the study subjects were oral poliovirus vaccine strains. In terms of detection rate, *Reoviridae* family of double-stranded RNA viruses were dominating with group A rotavirus contigs present in all 12 faecal samples. As the leading cause of severe diarrhoea in children under five years old, the presence of rotavirus A in healthy subjects could be asymptomatic infection, vaccine-derived rotavirus strains, or viruses that did not clear from a recovering symptomatic infection. The same can be the situation for norovirus, except that no vaccine is available for norovirus. Several other human enteric viruses, mainly members of the genus *Enterovirus*, which are known to cause a variety of human diseases, were detected but no association with disease could be established. These viruses, which might have arisen as a result of asymptomatic infection, included parechoviruses, enterovirus species and echoviruses. *Astroviridae*, another family of RNA viruses, was the least detected of known enteric pathogens commonly implicated in cases of gastroenteritis. Under this family, only two samples from the same subject was reported to contain astrovirus sequences in the first two collection time points, implying that the virus was clearing from the gut of the infant concerned. Asymptomatic individuals may play an important role in the transmission of enteric pathogens. It is therefore imperative to pay more attention on understanding asymptomatic infections and the establishment thereof. Being able to interpret asymptomatic infections and transmission of enteric virus can be very useful in

the development of transmission prevention strategies. As mentioned before, many of the human enteric viruses characterized in this study are commonly transmitted by the faecal-oral route, therefore hygiene status is also an important factor contributing to the transmission of viral agents. Conversely, the detection of plant viruses in the infants may not necessarily represent an infection but could also be due to ingestion of contaminated food products. This further underscores the fact that composition of the human intestinal virome can vary between humans as a result of diet.

There is still a wealth of viral populations to be discovered. Although viral metagenomics methods can provide insights into the composition and structure of viral communities colonizing the gastrointestinal tract of humans, factors such as diet, nutrition status, geographical location, health including immune status, socioeconomic group must be taken into consideration as these can influence composition of the human gut virome. More metagenomics research is needed in order to fully characterise the human gut virome. Individuals in areas with poor sanitation and crowded living conditions, and the immuno-compromised are expected to harbour different viral communities and probably higher viral loads compared to those who are immuno-competent, living in better living conditions with good health care services.

The current gut RNA virome study has shown that even asymptomatic infants seem to harbour a diverse community of viruses, which should be explored further not only to expand the knowledge we already have but also to discover new viruses and add more viral sequences to the current databases. The detection of viruses that are naturally infecting non-human mammals warrant further investigation to understand how they are transmitted.

4.6. References

Acar, E., Bulbul, O., Rayimoglu, G., Shahzad, M. S., Argac, D., Altuncul, H. and Filoglu, G. (2009). Optimization and validation studies of the MentypeR Argus X-8 kit for paternity cases. *Forensic Sci Int Genet Suppl Ser* **2**: 47-48.

Aggarwala, V., Liang, G. and Bushman F. D. (2017). Viral communities of the human gut: metagenomic analysis of composition and dynamics. *Mobile DNA* **8**:12.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215:** 403-10.

Andrews, S. (2010). FASTQC: a quality control tool for high throughput sequence data. [Accessed 23-04-2017]. http://www.bioinformatics.babraham.ac.uk/projects/FASTQc.

Ayukekbong, J. A., Mesumbe, H. N., Oyero, O. G., Lindh, M. and Bergström, T. (2015). Role of noroviruses as aetiological agents of diarrhoea in developing countries. *J Gen Virol* **96(8)**: 1983-1999.

Ayukekbong, J. A., Andersson, M. E., Vansarla, G., Tah, F., Nkuo-Akenji, T., Lindh, M. and Bergström, T. (2014). Monitoring of seasonality of norovirus and other enteric viruses in Cameroon by real-time PCR: an exploratory study. *Epidemiol Infect* **42(7)**: 393-1402.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S. and Prjibelski, A. D. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**: 455-477.

Bellmunt, A., May, G., Zell, R., Pring-Akerblom, P. and Verhagen, W. (1999). Heim Evolution of poliovirus Type I during 5.5 years of prolonged enteral replication in an immunodeficient patient. *Virology* **265:** 178-184.

Breitbart, M., Haynes, M., Kelley, S., Angly, F., Edwards, R. A., Felts, B., Mahaffy, J. M., Mueller, J., Nulton, J., Rayhawk, S. *et al*. (2008). Viral diversity and dynamics in an infant gut. *Res Microbiol* **159**: 367-373.

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J.M., Nulton, J., Salamon, P. and Rohwer, F. (2003). Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* **185**: 6220-6223.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., Azam, F. and Rohwer, F. (2002). Genomic analysis of uncul- tured marine viral communities. *Proc Natl Acad Sci U S A* **99**: 14250-14255.

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

Buttinelli, G., Donati, V., Fiore, S., Marturano, J., Plebani, A., Balestri, P., Soresina, A. R., Vivarelli, R., Delpeyroux, F. and Martin, J. (2003). Fiore Nucleotide variation in Sabin type 2 poliovirus from an immunodeficient patient with poliomyelitis. *J Gen Virol* **84**: 1215-1221.

Conceição-Neto, N., Zeller, M., Lefrère, H., De Bruyn, P., Beller, L., Deboutte, W., Yinda, C. K., Lavigne, R., Maes, P., van Ranst, M. *et al*. (2015). Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci Rep* **5**: 16532.

Cotton, M., Oude Munnink, B., Canuti, M., Deijs, M., Watson, S. J., Kellam, P. and van der Hoek, L. (2014). Full genome virus detection in fecal samples using sensitive nucleic acid preparation, deep sequencing, and a novel iterative sequence classification algorithm. *PLoS ONE* **9(4)**: e93269.

Diop, O. M., Burns, C. C., Sutter, R. W., Wassilak, S. G. and Kew, O. M. (2015). Update on Vaccine-Derived Polioviruses - Worldwide, January 2014-March 2015. *MMWR Morb Mortal Wkly Rep* **64(23)**: 640-646.

Duffy, S., Shackelton, L. A. and Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* **9**: 267-276.

Equine Infectious Anemia. (2016). Introduction**.** *The Merck Veterinary Manual*. [Accessed: 14-09-2018]. Available from: https://www.merckvetmanual.com.

European Medicines Agency. (2016). Summary of product characteristics. *Rotarix*. [Accessed: 28-10-2019]. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Product_ Information/human/000639/WC500054789.pdf.

García, C., DuPont, H. L., Long, K. Z., Santos, J.I. and Ko, G. (2006). Asymptomatic norovirus infection in Mexican children. *J Clin Microbiol* **44(8)**: 2997-3000.

Giaquinto, C. and van Damme, P. (2010). Age distribution of paediatric rotavirus gastroenteritis cases in Europe: The REVEAL study. *Scand J Infect Dis* **42(2)**: 142-147.

Glasel, J. A. (1995). Validity of nucleic acid purities monitored by 260 nm/280 nm absorbance ratios. *BioTechniques* **18(1)**: 62-63.

Graul, S., Böttcher, S., Eibach, D., Krumkamp, R., Käsmaier, J., Adu-Sarkodie, Y., May, J., Tannich, E. and Panning, M. (2017). High diversity of human parechovirus including novel types in stool samples from Ghanaian children. *J Clin Virol* **96**: 116-119.

Gumede, N., Lentsoane, O., Burns, C. C., Pallansch, M., de Gourville, E., Riziki Yogolelo, Muyembe-Tamfum, J. J., Puren, A., Schoub, B. D. and Venter, M. (2013). Emergence of Vaccine-derived Polioviruses, Democratic Republic of Congo, 2004-2011. *Emerg Infect Dis* **19(10)**: 1583-1589.

Hall, A. J., Wikswo, M. E., Pringle, K., Gould, L. H. and Parashar, U. D. (2014). Vital signs: Foodborne norovirus outbreaks - United States, 2009–2012. *MMWR Morb Mortal Wkly Rep* **63(22)**: 491-495.

Hamaguchi, T., Fujisawa, H., Sakai, K., Okino, S., Kurosaki, N., Nishimura, Y., Shimizu, H. and Yamada, M. (2008). Acute encephalitis caused by intrafamilial transmission of enterovirus 71 in adult. *Emerg Infect Dis* **5**: 828-830.

Hamza, I. A., Jurzik, L., Wilhelm, M. and Uberla, K. (2009). Detection and quantification of human bocavirus in riverwater. *J Gen Virol* **90(11)**: 2634-2637.

Harvala, H., McIntyre, C. L., McLeish, N. J., Kondracka, J., Palmer, J., Molyneaux, P., Gunson, R., Bennett, S., Templeton, K. and Simmonds, P. (2012). High detection frequency and viral loads of human rhinovirus species A to C in fecal samples; diagnostic and clinical implications. *J Med Virol* **84**: 536-542.

Ho, T. and Tzanetakis, I. E. (2014). Development of a virus detection and discovery pipeline using next generation sequencing. *Virology* **471-473**: 54-60.

Honkanen, H., Oikarinen, S., Peltonen, P., Simell, O., Ilonen, J., Veijola, R., Knip, M. and Hyöty, H. (2013). Human rhinoviruses including group C are common in stool samples of young Finnish children. *J Clin Virol* **56**: 250-254.

Hovi, T., Lindholm, N., Savolainen, C. and Stenvik, M. (2004). Burns Evolution of wild-type 1 poliovirus in two healthy siblings excreting the virus over a period of 6 months. *J Gen Virol* **85**: 369-377.

Huberman, J. A. (1995). Importance of measuring nucleic acid absorbance at 240 nm as well as at 260 and 280 nm. *BioTechniques* **18(4)**: 636.

Hulo, C., de Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I. and Le Mercier, P. (2011). ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res* **39**: D576-D582.

Jorba, J., Diop, O. M., Iber, J., Henderson, E., Sutter, R. W., Wassilak, S. G. F. and Burns, C. C. (2017). Update on vaccine-derived polioviruses-worldwide, January 2016-June 2017. *MMWR Morb Mortal Wkly Rep* **66**: 1185-1191.

Jorba, J., Diop, O. M., Iber, J., Sutter, R. W., Wassilak, S. G. and Burns, C. C. (2016). Update on vaccine-derived polioviruses-worldwide, January 2015-May 2016. *MMWR Morb Mortal Wkly Rep* **65**: 763-769.

Julbert, B. and Lipton, H. L. (2013). Enterovirus/Picornavirus infections. Chapter 18 In: *Handb Clin Neurol* 123 (3rd series). Edited by Tselis, A.C. and Booss, J.: Neurovirology.

Kapusinszky, B., Minor, P. and Delwart, E. (2012). Nearly constant shedding of diverse enteric viruses by two healthy infants. *J Clin Microbiol* **50(11)**: 3427-3434.

Kew, O. M., Sutter, R. W., Nottay, B. K., MacDonough, M. J., Prevots, D. R. Quick, L. and Pallansch, M. A. (1998). Prolonged replication of a type 1 vaccine-derived poliovirus in an immunodeficient patient. *J Clin Microbiol* **36**: 2893-2899.

Kolehmainen, P., Oikarinen, S., Koskiniemi, M., Simell, O., Ilonen, J., Knip, M., Hy€oty, H. and Tauriainen, S. 2012. Human parechoviruses are frequently detected in stool of healthy Finnish children. *J Clin Virol* **54**: 156-161.

Langmead. B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357-359.

Lau, S. K., Yip, C. C., Lung, D. C., Lee, P., Que, T. L., Lau, Y. L. Chan, K. H., Woo, P. C. and Yuen, K. Y. (2012). Detection of human rhinovirus C in fecal samples of children with gastroenteritis. *J Clin Virol* **53**: 290-296.

Lauring, A. S., Frydman, J. and Andino, R. (2013). The role of mutational robustness in RNA virus evolution. *Nature Rev Microbiol* **11**: 327-336.

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. *Nature Med* **21(10)**: 1228-1234.

Lin, F. and Kitching, R. P. (2000). Swine vesicular disease: an overview. *Vet J* **160(3)**: 192-201.

Lorenzi, H. A., Hoover, J., Inman, J., Safford, T., Murphy, S., Kagan, L. and Williamson, S. J. (2011). The Viral Metagenome Annotation Pipeline (VMGAP): An automated tool for the functional annotation of viral metagenomic shotgun sequencing data. *Stand Genomic Sci* **4**: 418-429.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y. *et al*. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience* **1**: 18.

Maffei, H. V. and Nobrega, F. J. (1975). Gastric pH and microflora of normal and diarrhoeic infants**.** *Gut* **16**: 719-726.

Manchester, K. L. (1995). Value of A260/A280 ratios for measurement of purity of nucleic acids. *BioTechniques* **19(2)**: 208-210.

Manchester, K. L. (1996). Use of UV methods for the measurement of protein and nucleic acid concentrations. *BioTechniques* **20(6)**: 968-970.

McKnight, R. E., Gleason, A. B., Keyes, J. A., and Sahabi, S. (2006). Binding mode and affinity studies of DNA-binding agents using topoisomerase I DNA unwinding assay. *Bioorganic Med. Chem. Lett* **17(4)**: 1013-1017.

Minot, S., Bryson, A., Chehoud, C., Wu, G. D., Lewis, J. D. and Bushman, F. D. (2013). Rapid evolution of the human gut virome. *Proc Natl Acad Sci U S A* **110**: 12450-12455.

Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D. and Bushman, F. D. (2012). Hypervariable loci in the human gut virome. *Proc Natl Acad Sci U S A* **109**: 3962-3966.

Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S. A., Wu, G. D., Lewis, J. D. and Bushman, F. D. (2011). The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res* **21**: 1616-1625**.**

Mommaerts, K., Sanchez, I., Betsou, F. and Mathieson, W. (2015). Replacing β-mercaptoethanol in RNA extractions. *Anal Biochem* **479**: 51-53.

Nagarajan, N. and Pop, M. (2013). Sequence assembly demystified. *Nat Rev Genet* **14**: 157-167.

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

Nurk, S., Meleshko, D., Korobeynikov, A. and Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27(5)**: 824-834.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Hadad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B. and Ako-Adjei, D. *et al.* (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**: D733-D745.

Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N. N. and Kyrpides, N. C. (2016). Uncovering Earth's virome. *Nature* **536**: 425-30.

Pallansch, M. A., Oberste, M. S. and Whitton, J. L. (2013). Enteroviruses: polioviruses, coxsackieviruses, echoviruses, and newer enteroviruses. In *Fields Virology*. pp. 491-530. Edited by Knipe, D. M. and Howley P. M. Lippincott Williams and Wilkins, Philadelphia.

Parashar, U. D., Bresee, J. S., Gentsch, J. R. and Glass, R. I. (1998). Rotavirus. *Emerg Infect Dis* **4(4)**: 561-570.

Patel, M. M., Widdowson, M. A., Glass, R. I., Akazawa, K., Vinje, J. and Parashar U. D. (2008). Systematic literature review of role of noroviruses in sporadic gastroenteritis. *Emerg Infect Dis* **14**: 1224-1231.

Peng, Y., Leung, H. C. M., Yiu, S. M. and Chin, F. Y. L. (2012). IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**: 1420-1428.

Pons-Salort, M., Parker, E. P. and Grassly, N. C. (2015). The epidemiology of non-polio enteroviruses: recent advances and outstanding questions. *Curr Opin Infect Dis* **28(5)**: 479-487.

Popgeorgiev, N., Temmam, S., Raoult, D. and Desnues, C. (2013). Describing the silent human virome with an emphasis on giant viruses. *Intervirology* **56(6)**: 395-412.

Qi, R., Huang, Y., Liu, J, Sun, Y., Sun, X., Han, H., Qin, X., Zhao, M., Wang, L., Li, W. *et al*. (2018). Global Prevalence of Asymptomatic Norovirus Infection: A Meta-analysis. *E Clin Med* **2-3**: 50-58.

Rampelli, S., Soverini, M., Turroni, S., Quercia, S., Biagi, E., Brigidi, P. and Candela, M. (2016). ViromeScan: a new tool for metagenomic viral community profiling. *BMC Genomics* **17**: 165.

Ratcliff, R. M., Chang, G., Kok, T. and Sloots, T. P. (2007). Molecular diagnosis of medical viruses. *Curr Issues Mol Biol* **9**: 87-102.

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. and Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**: 69.

Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F. and Gordon, J. I. (2012). Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat Rev Microbiol* **10**: 607-617.

Reyes, A., Wu, M., McNulty, N. P., Rohwer, F. L. and Gordon, J. I. (2013). Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut. *Proc Natl Acad Sci U S A* **110(50)**: 20236-20241.

Rotbart, H. A. (2000). Viral meningitis**.** *Semin Neurol* **20**: 277-292.

Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A., Poulos, B. T., Solonenko, N., Lara, E. and Poulain, J. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**: 689-693.

Roux, S., Tournayre, J., Mahul, A., Debroas, D. and Enault, F. (2014). Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **15**: 76.

Sakon, N., Miyamoto, R. and Komano, J. (2017). An infant with acute gastroenteritis caused by a secondary infection with a Rotarix-derived strain. *Eur J Pediatr* **176(9)**: 1275-1278.

Salminen, K. K., Vuorinen, T., Oikarinen, S. Helminen, M., Simell, S., Knip, M., Ilonen, J., Simell, O. and Hyöty, H. (2004). Isolation of enterovirus strains from children with preclinical Type 1 diabetes. *Diabet Med* **21**: 156-164.

Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27(6)**: 863-864.

Schweitzer, C. and Scaiano, J. C. (2003). Selective binding and local photophysics of the fluorescent cyanine dye PicoGreen in double-stranded and single-stranded DNA. *Phys Chem Chem Phys* **5**: 4911-4917.

Seheri, L. M., Page, N., Dewar, J. B., Geyer, A., Nemarude, A. L., Bos, P., Esona, M. and Steele, A. D. (2010). Characterization and molecular epidemiology of rotavirus strains recovered in Northern Pretoria, South Africa during 2003-2006. *J Infect Dis* **202**: S139-S147.

Seheri, M., Page, N. A., Mothahadini, P., Mawela, B., Mphahlele, J. M. and Steele, D. A. (2012). Rotavirus vaccination within the South African Expanded Programme on Immunisation. *Vaccine* **30(3)**: C14-C20

Shulman, L. M., Manor, J., Handsher, R., Delpeyroux, F., MacDonough, M. J., Halmut, T., Silberstein, I., Alfandari, J., Quay, J., Fisher, T. *et al*. (2000). Mendelson Molecular and antigenic characterisation of a highly evolved derivative of the type 2 oral polio vaccine strain isolated from sewage in Israel. *J Clin Microbiol* **38**: 3729-3734.

Smuts, H., Cronje, S., Thomas, J., Brink, D., Korsman, S. and Hardie, D. (2018). Molecular characterization of an outbreak of enterovirus-associated meningitis in Mossel Bay, South Africa, December 2015 - January 2016. *BMC Infect Dis* **18**: 709.

Specter, S. and Lanza, G. J. (1992). In *Clinical Virology Manual*. 2nd Ed. pp. Edited by Specter, S. and Lanz, G. J. New York, Elsevier.

Steele, A. D., Alexander, J. J. and Hay, I. T. (1986). Rotavirus-associated gastroenteritis in black infants in South Africa*. J Clin Microbiol* **23**: 992-994.

Svarovskaia, E. S., Cheslock, S. R., Zhang, W. H., Hu, W. S. and and Pathak, V. K. (2003). Retroviral mutation rates and reverse transcriptase fidelity. *Front Biosci J Virtual Libr* **8**: 117-134.

Troeger, C., Khalil, I. A., Rao, P. C., Cao, S., Blacker, B. F., Ahmed, T., Armah, G., Bines, J. C., Brewer, T. G., Colombara, D. V. *et al*. (2018). Rotavirus vaccination and the global burden of rotavirus diarrhea among children younger than 5 years. *JAMA Pediatr* **172(10):** 958-965.

van Beek, J., Ambert-Balay, K., Botteldoorn, N., Eden, J. S., Fonager, J., Hewitt, J., Iritani, N., Kroneman, A., Vennema, H., Vinjé, J. *et al*. (2013). Indications for worldwide increased norovirus activity associated with emergence of a new variant of genotype II.4, late 2012. *Euro Surveill* **18**: 8-9.

Wang, C., Zhou, S., Xue, W., Shen, L., Huang, W., Zhang, Y., Li, X., Wang, J., Zhang, H. and Ma, X. (2018). Comprehensive virome analysis reveals the complexity and diversity of the viral spectrum in pediatric patients diagnosed with severe and mild hand- foot-and-mouth disease. *Virology* **518:** 116-125.

Wang, S., Sundaram, J. P. and Spiro, D. (2010). VIGOR, an annotation program for small viral genomes. *BMC Bioinformatics* **11**: 451.

Wills, L. and Paterson, D. (1926). A study of gastric acidity in infants. *Arch Dis Child* **1**: 232-244.

Witsø, E., Palacios, G., Cinek, O., Stene, L. C., Grinde, B., Janowitz, D., Lipkin, W. I. and Rønningen, K. S. (2006). High prevalence of human enterovirus A infections in natural circulation of human enteroviruses. *J Clin Microbiol* **44**: 4095-4100.

Wolfaardt, M., Buchner, A., Myburgh, M., Avenant, T., du Plessis, N. M. and Taylor, M. (2014). Molecular characterization of enteroviruses and clinical findings from a cluster of paediatric viral meningitis cases in Tshwane, South Africa 2010 – 2011. *J Clin Viral* **61**: 400-405.

Wommack, K. E. and Colwell, R. R. (2000). Virioplankton: viruses in aquatic ecosystems. *Microbiol Mol Biol Rev* **64**: 69-114.

Wommack, K. E., Bhavsar, J., Polson, S. W., Chen, J., Dumas, M., Srinivasiah, S., Furman, M., Jamindar, S. and Nasko, D. J. (2012). VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Stand Genomic Sci* **6**: 427-439.

World Health Organization. (2017). South Africa: expanded programme on immunisation and vaccine preventable disease surveillance. [Accessed: 12 May 2018]. Available from: http://www.afro.who.int/countries/south-africa.

Zhang, T., Breitbart, M., Lee, W. H., Run, J. Q., Wei, C. L., Soh, S. W., Hibberd, M. L., Liu, E. T., Rohwer, F. and Ruan, Y. (2006). RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol* **4:** e3.

Zhao, G., Wu, G., Lim, E. S., Droit, L., Krishnamurthy, S., Barouch, D. H., Virgin, H. W. and Wang, D. (2017). VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology* **503**: 21-30.

Zhu, L. Y., Ding, Z. T. and Wan, J. F. (2008). Epidemiology investigation of severe cases of HFMD infected with EV71 in FUYANG city. *J Anhui Med* **29**: 595-596.

Chapter 5: General discussion and conclusions

5.1. General discussion and concluding remarks

Viral metagenomics by high-throughput sequencing (HTS) is an exceptional approach that enables characterization of the virome in diverse host organisms. The human virome is a massive component of our metagenome which is continuously evolving at a faster rate than all other microbes (Virgin *et al*., 2014). The role and impact of the virome on human health and disease are only beginning to be understood. In this study, we demonstrated the practicability of viral metagenomic methods utilizing the next generation sequencing platform to provide the viral profiles of healthy infants. Using Illumina MiSeq platform, we obtained good quality metagenomic data which has enabled us to identify viral pathogens present in the faecal samples of infants.

A recently developed protocol was successfully applied to enrich for virus particles in human faeces. Application of NetoVIR protocol (Conceicao-Neto *et al*., 2015) was important for the enrichment of viruses by reducing rRNA and other microbial genomes in the samples. The key steps included the digestion of unprotected nucleic acids by nuclease digestion and filtration to reduce the larger eukaryotic and bacterial cell-sized particles. In addition, the host ribosomal RNA depletion was a convenient and user-friendly procedure that has enabled a significant reduction of the human host genomes in the faecal samples. Whole-transcriptome-amplification aided to increase the quantity of starting RNA material and in turn the recovery of RNA viral sequences. However, this procedure might have also potentially increased the bacterial sequences in the samples by amplifying the 16S ribosomal RNA. The modifications to the Nextera XT DNA Library preparation protocol were aimed at increasing the library fragment sizes. In this study, an optimal average library size of 450 bp was obtained and sequenced on Illumna MiSeq platform. An in-house developed bioinformatics pipeline, comprising of a combination of programs and web-based analytical tools, was successfully applied to analyze the generated metagenomic data. The programs involved were less computationally demanding and could be executed on a standard computer.

Although the analysis revealed that the data was dominated by bacterial sequences, a significant fraction of viral sequences were recovered. Metagenomic analysis revealed that known viral contigs constituted 11 % of the pre-processed and assembled data, as determined by BLASTX annotation against the NCBI non-redundant protein database. However, it is possible that the unknown fraction of sequences is likely to include viruses that could not be identified due to the unavailability of reference sequences in the database. Using Diamond as an annotation tool, viruses from six different families, which are known to infect mammals including humans, were detected from twelve faecal samples. These virus families were *Picornaviridae*, *Reoviridae*, *Astroviridae*, *Caliciviridae*, *Flaviviridae* and *Poxviridae*. In addition, within the *Retroviridae* family, sequences of retroviruses were recovered from two of the samples. Few unclassified viruses in the order of *Picornavirales* were also identified. Although *Reoviridae* family was detected in 100 % of the stools, the most abundant family in terms of the quantity of contigs,

was *Picornaviridae*, predominantly coxsackievirus sequences. The study was limited to RNA viruses, hence bacteriophages could not be detected in abundance. There was only a single phage contig (from the family *Siphoviridae*) which was detected in one of the samples.

Further analysis and characterization of the poliovirus strains and norovirus strains that were detected in the infants' needs to be done. The detection of polioviruses and other enteric viruses highlights the importance of surveillance studies to determine the strains circulating in developing countries including South Africa. The detection of animal viruses should also be further investigated for their transmission dynamics and zoonotic potential. Considering the presence of plant-derived viral sequences detected in some individuals one can hypothesize that diet could regulate composition of the gut virome, as previously reported (Zhang *et al*., 2006).

Nevertheless, the data obtained from this study can serve a baseline reference for future studies aimed at characterizing the gut virome of humans. The results from this study have shown the high prevalence of RNA viruses in gastrointestinal tract of infants, which have often been ignored by metagenomic analysis of the virome (Virgin *et al*., 2014; Rascovan *et al*., 2016). This indicates that research should focus more on the characterizing RNA viral flora as many of the RNA viruses are known to cause serious diseases in humans and animals, with several being known aetiologic agents of acute gastroenteritis, such as rotavirus and norovirus (Biscaro *et al*., 2018).

Despite several issues and challenges encountered in metagenomic analysis that needs to be addressed, Next generation sequencing-based metagenomic analysis provides a promising tool for future study of viromes and their dynamics. The applicability of the metagenomic techniques in a wide range of research studies makes it a valuable tool that can be used to answer difficult questions, such as how the viruses interact with host organisms and the transmission of pathogens between humans and other mammalian animals. In general terms, the metagenomic based approaches serve as important tools in the diagnosis of mild to severe diseases in human and other organisms, since such techniques do not target a specific pathogen. In one of the infants both coxsackievirus and norovirus were detected in the same sample, thus the detection of viral pathogens, not routinely tested for, in suspected samples can be clinically relevant. In addition, metagenomic-based methods can be useful in the identification of the disease-causing agents during disease outbreaks, the only setback for this is the high cost of sequencing projects, the amount of time, and the labor intensive nature of sample preparations and sequencing.

Furthermore, the current metagenomic study has demonstrated that despite the absence of symptoms, RNA viruses are still prevalent in the gastrointestinal tract of infants, giving the implication that even in health the gut mucosa is still subjected to frequent viral infections which play an important role in shaping the human intestinal

virome. Indeed, the implementation of NGS methodologies in virology has allowed researchers the opportunity to explore the previously inaccessible aspects of viral dynamics, and such approaches can certainly provide the scientific community with new insight on the transmission patterns of various pathogens inhabiting humans and animals. The success of viral metagenomics in the identification of novel viruses from various specimens opens doors to new application areas that seeks to prevent viral transmitted diseases and improve human health.

5.2. Limitations and Future perspectives

Viral metagenomics is providing crucial data for understanding the viral component of the human virome. The advent of high-throughput sequencing has made it easier for researchers to explore and characterize viruses using metagenomics. A major bottleneck for virus identification is the capacity of the viral metagenomics databases. As the identification of the viruses is done by sequence comparison with homologous reference genomes in the database, an increase of annotated complete viral genomes is desirable. If only short genome fragments are present in the database and the generated NGS read maps to another region of the viral genome, no identification of the virus is possible. Furthermore, the detection of novel viruses is not easy with any analytical tool. Although the genomic material of those viruses will be sequenced, it is only the viruses that shares certain degree of sequence resemblance to viruses in the used databases can be identified by performing a BLAST search. In fact, in order to be able to identify new viruses using the applied metagenomic techniques, novel bioinformatics tools must be developed.

For improvement of virome studies in future, efforts to address the following aspects will be helpful. **(1)** The enrichment and isolation of viral genomic material should be developed to enable bias-free virome representation prior to next generation sequencing. **(2)** Ensuring that the bioinformatics tools to be used are standardized. **(3)** In the proposal development phase of a virome characterization and analysis project, the importance of several aspects such as metadata collection, sequencing depth, replicates and controls must be taken into account. Paying attention and ensuring that these aspects are efficiently addressed and included in the study design would make comparison of virome data easier and conclusions would be made with more confidence. **(4)** More importantly, since infants are at a greater risk of viral infections, an initiative aimed at comprehensively analyzing and following up the dynamics of viral communities in the gut of infants and young children overtime, can help researchers discover ways to manipulate their virome, ultimately leading to improved management of human health and disease. Such initiative can contribute to the efforts in minimizing childhood mortalities and hospitalizations, especially in resource-poor countries of Africa. Indeed, making attempts to unravel and understand the complex community of viruses colonizing the gastrointestinal tract of humans can expand our knowledge and shed light on the health and disease of humans. In summary, there is a great need for thorough screening of infants and young

children in order to establish the prevalence and diversity of viruses colonizing their gut system. These efforts could allow for efficient measures and prevention of viral-induced illnesses including acute diarrhoea in children.

5.3. References

Biscaro, V., Piccinelli, G., Gargiulo, F., Ianiro, G., Caruso, A., Caccuri, F. and De Francesco, M. A. (2018). Detection and molecular characterization of enteric viruses in children with acute gastroenteritis in Northern Italy. *Infect Genet Evol* **60** 35-41.

Conceição-Neto, N., Zeller, M., Lefrère, H., De Bruyn, P., Beller, L., Deboutte, W., Yinda, C. K., Lavigne, R., Maes, P., van Ranst, M. *et al*. (2015). Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci Rep* **5**: 16532.

Rascovan, N., Duraisamy, R. and Desnues, C. (2016). Metagenomics and the Human Virome in Asymptomatic Individuals. *Annu Rev Microbiol* **70**: 125-41.

Virgin, H. W. **(2014).** The virome in mammalian physiology and disease. *Cell* **157(1)**: 142-150.

Zhang, T., Breitbart, M., Lee, W. H., Run, J. Q., Wei, C. L., Soh, S. W., Hibberd, M. L., Liu, E. T., Rohwer, F. and Ruan, Y. (2006). RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol* **4**: e3.