Investigating the Human Intestinal Virome

Shen-Yuan Hsieh

B.Sc. and M.Sc.

A thesis submitted to the University of East Anglia for the degree of Doctor of Philosophy

Quadram Institute Biosciences,

Norwich Research Park, UK

December 2020

Abstract

The human intestinal virobiota consists of highly complex and diverse viruses and virus-associated genes (termed the "virome"), dominated by bacteriophages. Numerically, viruses have been considered the most abundant and diverse biological entities on Earth, estimated to be approximately 10^{31} in number. In the human gastrointestinal tract (GIT), virus-to-microbe ratio (VMR) may be close to 1:1, while it may reach 20:1 at mucosal surfaces and within the mucus layer, in total numbering 10^{10}-10^{15} virus-like particles (VLPs). Recent studies suggested that changes in the intestinal virome may lead to chronic GI-inflammation and intestinal microbial dysbiosis, thereby triggering diseases such as myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS). Thus, this thesis aimed to develop robust and reliable protocols for characterising the human faecal microbiome that can be applied to analysing the virome in patients with severe ME/CFS.

This thesis first aimed to develop a reliable and reproducible protocol for VLP isolation from human faeces and for VLP quantification using a digital image analysis (DIA)-based method. Protocols were then optimised for VLP DNA extraction to obtain DNA of sufficient quality and quantity for next generation sequencing (NGS). As part of these studies, I developed a bioinformatics pipeline for viral metagenomic analysis, and also, I determined the extent of PCR amplification bias in virome-enriched, uncultivated virus genomes (UViGs) by comparing the methods of linker amplified shotgun library (LASL) and non-amplified shotgun library (NASL) preparations. The optimised protocols and bioinformatics pipeline were then applied to the initial analysis of the faecal virome of severely affected ME/CFS patients and same household healthy control individuals (SHHC).

The optimised protocol is comprised of (1) homogenisation of faecal samples by vortexing without the use of bead-beating, followed by incubation on ice to facilitate the release of VLPs from solid materials; (2) partition of crude faecal matter, dietary debris and virions/VLPs by two-round of high-speed centrifugation; (3) sequential filtration using 0.8 µm and 0.45 µm filter; (4) PEG precipitation; (5) DNase and RNase treatment; (6) proteinase K digestion; (7) viral capsid lysis with SDS lysis buffer; (8) Phenol/Chloroform/Isoamyl alcohol extraction; (9) DNA purification using silica-based spin columns, and (10) DNA concentration using a vacuum-based condenser. Using three independent stool samples to evaluate reproducibility, VLP DNA yields were between 67.2 ng and 94.8 ng per gram of faeces. For VLP quantification, manual counting-based DIA method was more accurate and reliable than automated counting-based method.

Using an optimised bioinformatics pipeline to analyse UViGs from PCR and non-PCR virome-derived datasets, I found that misrepresentation of certain viruses may occur after

amplification in their relative abundance. In alpha diversity, the UViGs from non-PCR datasets generally have higher richness and diversity than those from PCR datasets, suggesting that PCR is likely to lower viral richness and diversity. Moreover, the major differences in beta diversity were more likely to be driven by a high level of intestinal virome individuality between donors, while amplification bias may have a minor effect on the beta diversity of viruses in the PCR datasets. In addition, in an initial analysis of comparing UViG similarity networks, I found that there is no significant difference between both datasets but further investigation is required.

In comparing faecal samples from ME/CFS and SHHC, VLPs with nucleic acid-containing capsids in SHHC samples were higher than those of severe ME/CFS patients, although the variation and diversity of VLP were seen in both sets of faecal samples. Transmission electron microscopy (TEM) analysis identified *Siphoviridae* as the most prominent virus in both ME/CFS patients and SHHC VLP samples. Moreover, giant Siphoviruses were occasionally detected, suggesting potential novel strains are present in these samples. The biological meaning of these findings is not clear and requires further investigation. In ongoing work, the optimised protocol and bioinformatic pipeline is being applied to investigating the composition of the intestinal virome in severe ME/CFS patients and SHHC.

List of Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I would like to show deep gratitude to my supervisor, Professor Simon R. Carding, for offering me this opportunity to enter PhD after a career in academia. I really appreciate all the help, support and reassurances he has given me with the practical work, and really appreciate the guidance he has given me during the provision of my thesis. I would also like to thank Simon for building me to be a researcher. Moreover, I would like to thank him for being more tolerant of myself during times when I had difficulty expressing and communicating in English, and found it difficult or made mistakes in research, and eventually he never gave up on myself.

I would like to thank Dr James Ebdon for kindly providing me with the phage stock, bacterial host strain and the protocol from the University of Brighton. I am so grateful to Professor Tom Wileman from UEA/Quadram, Professor Lesley Hoyles from Nottingham Trent University and Dr Mohammad Adnan Tariq from Quadram Institute for being the members of my supervisory panel to support and guide my research during my PhD. I would like to thank QIB Bioinformatics Team, particularly Dr Andrea Telatin and Dr Rebecca Ansorge, for supporting all computational and bioinformatic knowledge as well as techniques to help us develop the bioinformatic pipeline for virome analysis. I would also like to thank QIB Microscopy Team (Core Science Resources, CSR), particularly Dr Catherine Booth and Mrs Kathryn Gotts, for EFM training and supporting myself in TEM imaging. Moreover, I would like to thank many of my friends and colleagues from Carding group, including both of the former and the present, for all supporting me in practical work. Particularly, I am so grateful to the “Poo Crew” who have been my great lab partners, including Dr Mohammad Adnan Tariq, Fiona Newberry and Katharine Seton, for all helping and supporting myself in experimental work and bioinformatics stuff. I would also like to thank “Invest in ME Research” UK charity for all support and collaboration in ME/CFS research.

Finally, this thesis would not have been possible without the support from my beloved family. I would like to show eternal gratitude to my father Kevin and my mother Susan for the support they have given me throughout my PhD period both financially and emotionally. I would also like to thank my little brother John for all reassuring and supporting myself emotionally throughout tough times. This thesis will be dedicated to my beloved family. Thank you all.

List of Abbreviations

AAI: average amino acid identity

ADL: activities of daily living

AIDS: HIV-associated acquired immunodeficiency syndrome

APMV: *Acanthamoeba polyphaga mimivirus*

ART: anti-retroviral therapy

Bf: *Bacteroides fragilis*

bp: base pair

BPRM: Bacteroides Phage Recovery Medium

$CaCl_2 \cdot 2H_2O$: calcium chloride dihydrate

CCC: Canadian Consensus Criteria

CCD: charge-coupled device

CD: Crohn's disease

CDC: Centers for Disease Control and Prevention

CDI: *Clostridium difficile* infection

cDNA: complementary DNA

CFIDS: chronic fatigue and immune dysfunction

cfp: colony-forming particles

CMV: *Cytomegalovirus*

crAssphage: Cross-Assembly phage

CRISPR: clustered regularly interspaced short palindromic repeats

CRT: cyclic reversible termination

CsCl: caesium chloride or cesium chloride

CSR: Core Science Resources, QIB

CTAB: cetyltrimethylammonium bromide

d: day

DAPI: 4',6-diamidino-2-phenylindole

DEF/NFF: dead-end/normal flow filtration

DemoVir: Democratic taxonomic classification of viral contigs

DIA: digital image analysis

DM: diabetes mellitus

DMSO: dimethyl sulfoxide

DNA: deoxyribonucleic acid

dsDNA: double-stranded DNA

dsRNA: double-stranded RNA

DTRs: direct terminal repeats

EBV: *Epstein-Barr virus*

EDTA: ethylenediaminetetraacetic acid

EFM: epifluorescence microscopy

EGTA: ethylene glycol-bis(β-aminoethyl)-N,N,N',N'-tetra acetic acid

EtOH: ethanol

e.g.: exempli gratia

et al.: et alia

FF: faecal filtrate

FMH: Faculty of Medicine and Health Sciences

FMT: faecal microbiota transplantation

FOV: field of view

g: gram or gramme

Gb: giga bases

GC content: guanine-cytosine content

gDNA: genomic DNA

GIT: gastrointestinal tract (or GI-tract)

GRCh37: Genome Reference Consortium Human Build 37

GTC: guanidinium thiocyanate or guanidine thiocyanate

h or hr: hours

HADS: Hospital Anxiety Depression Scale

HBV: *Hepatitis B virus*

HCV: *Hepatitis C virus*

HERV-K: *Human endogenous retrovirus K*

HHV: *Human herpesvirus*

HIV: *Human immunodeficiency virus*

HMM: Hidden Markov Model

HQ: high quality

HRA: Health Research Authority

ICTV: International Committee on Taxonomy of Viruses

IBD: inflammatory bowel disease

IBS: irritable bowel syndrome

ICC: International Consensus Criteria

ICD: WHO's International Classification of Diseases

i.e.: id est

IMG/VR: Integrated Microbial Genome/Virus

IRAS: Integrated Research Approval System

ITRs: inverted terminal repeats

JSD: Jensen-Shannon divergence

kb or kbp: kilo base pair

kDa: Kilodalton

KEGG: Kyoto Encyclopedia of Genes and Genomes

KH_2PO_4: monopotassium phosphate

KGMMV: *Kyuri green mottle mosaic virus*

kV: kilovolts

L: litre or liter

LASL: linker amplification shotgun libraries

LSU-rRNA: large subunit rRNA

M: Molar

mbp: mega base pairs

Mb: mega bases

MDA: multiple displacement amplification

ME/CFS: myalgic encephalomyelitis/chronic fatigue syndrome

2-ME: 2-mercaptoethanol or β-Mercaptoethanol

MetaPhlAn: Metagenomic Phylogenetic Analysis

$MgCl_2 \cdot 6H_2O$: magnesium chloride hexahydrate

$MgSO_4 \cdot 7H_2O$: magnesium sulfate heptahydrate

ml: millilitre

mM: millimolar

MNV: *Murine norovirus*

MWCO: molecular weight cut-off

N/A: not applicable

NaCl: sodium chloride

Na_2CO_3: sodium carbonate

Na_2HPO_4: disodium phosphate

NaOH: sodium hydroxide

NASL: non-amplified shotgun library

NCBI: National Centre for Biotechnology Information

ng: nanogram

NGS: next-generation sequencing

NHS: United Kingdom National Health Service

nM: nanomolar

nm: nanometre

NRES: National Research Ethics Service

$OD_{620\,nm}$: Optical density 620 nanometres

ONT: Oxford Nanopore Technologies

Pa: *Pseudomonas aeruginosa*

PacBio: Pacific Biosciences

PBS: phosphate-buffered saline

PC: polycarbonate or protein cluster

P/C/I: phenol/chloroform/isoamyl alcohol

PCoA: principal coordinate analysis

PCR: polymerase chain reaction

PE: paired-end

PEG: polyethylene glycol

PEM: post-exertional malaise

PES: polyethersulfone

PF: PCR-free, same as non-PCR or non-amplified (or unamplified)

PFU or pfu: plaque-forming unit

pg: picogram

pH: potential of hydrogen or power of hydrogen

Phage: bacteriophage

PIFS: post-infectious fatigue syndrome

PK: proteinase K

PMMV: *Pepper mild mottle virus*

pOTUs: phage operational taxonomic units

PPT: precipitation

PT: patient

PTFE: polytetrafluoroethylene

PVDF: polyvinylidene fluoride or polyvinylidene difluoride

PVFS: post viral fatigue syndrome

Q30: Phred quality score 30

QC: quality control

QIB: Quadram Institute Biosciences

qPCR: quantitative PCR or real-time PCR

R^2: coefficient of determination

r: Pearson correlation coefficient

REC: Research Ethics Committee

RNA: ribonucleic acid

rRNA: ribosomal RNA

ROS: reactive oxygen species

RT-PCR: reverse transcription-polymerase chain reaction

RUTF: ready-to-use therapeutic food treatment

RVDB: Reference Viral Database

SAM: severe acute malnutrition

S.D.: standard deviation

SDS: sodium dodecyl sulfate

SE: single-end

S.E.M.: standard error of the mean

SHHC: same household healthy control

SIA: sequence independent amplification

SISPA: sequence-independent, single-primer amplification

SM buffer: sodium chloride-magnesium sulphate buffer

SMRT: single-molecule real-time

SNA: single-nucleotide addition

sp.: species

spp.: several species

ssDNA: single-stranded DNA

ssRNA: single-stranded RNA

SSU-rRNA: small subunit rRNA

SYBR: Synergy Brands, Inc.

Tb: tera bases

T1D: type I diabetes

T2D: type II diabetes

TE buffer: Tris-EDTA buffer

TEM: transmission electron microscopy

TFF: tangential flow filtration

TGS: third-generation sequencing

TMV: *Tobacco mosaic virus*

Tris-HCl: Tris-hydrochloride

U: enzyme units

UA: uranyl acetate

UARTO: Uganda AIDS Rural Treatment Outcomes

UC: ulcerative colitis

UEA: University of East Anglia

UK: United Kingdom

UN: United Nations

UNAIDS: United Nations Programme in HIV/AIDS

UNICEF: United Nations Children's Fund

UpViG(s): uncultivated provirus genome(s)

US or USA: United State of America

UViG(s): uncultivated virus genome(s)

VC: viral cluster

vConTACT: Viral CONTigs Automatic Clustering and Taxonomy

VLP(s): virus-like particle(s)

VMR: virus-to-microbe ratio

VOGDB: Virus Orthologous Groups Database

VP: viral particle

vs.: versus

v/v: volume to volume

WHO: World Health Organization

WGA: whole genome amplification

w/o: without

w/v: weight to volume

WTA: whole transcriptome amplification

XMRV: *Xenotropic murine leukemia virus-related virus*

ZR: Zymo Research

λ: lambda, representing wavelength

µg: microgram

µl: microlitre

µm: micrometre

µM: micromolar

#: number, count or quantity

n: sample size

p: a statistical measure of the probability

Φ: phi, representing a bacteriophage or phage

π: pi, a mathematical constant representing approximately 3.1416

1. Introduction

1.1. The Human Virome

1.1.1. Definition of Human Virome

Within the human body exists dynamic communities of commensal microbes which includes viruses that infect eukaryotic or prokaryotic cells (e.g. bacteria and archaea) (**Figure 1.1**). The virobiota defines the communities of commensal viruses and virus-like particles (VLPs) that exist in a particular environment such as the gastrointestinal tract (GIT) (White et al., 2012). VLP, which has been widely used to describe uncharacterised viral structure and morphology, empty viral capsid without genetic materials and live-attenuated viruses with the lack of core pathogenetic genomes, is generally considered to be “non-infectious” (Mohsen et al., 2018, Zeltins, 2013). Human virobiota and viral metagenomes are also collectively termed the “virome”, which widely includes: (1) eukaryotic viruses that cause latent, acute or chronic infection in humans, (2) prokaryotic viruses (bacteriophages) that infect bacteria, (3) archaeal viruses that infect archaea, and (4) virus-associated genes integrated in host genomes that alter host gene expression, encode essential viral proteins and generate endogenous pathogenic or non-pathogenic viruses such as prophages, endogenous retroviruses and/or endogenous viral elements (Virgin, 2014). A number of eukaryotic and prokaryotic viruses (or bacteriophages) as well as VLPs residing in the human oral cavity (Pride et al., 2012, Lazarevic et al., 2012), nasopharynx (Wang et al., 2016, Nokso-Koivisto et al., 2002), bloodstream (Stremlau et al., 2015, Popgeorgiev et al., 2013), respiratory tract (Lysholm et al., 2012, Willner et al., 2009), digestive tract (Shkoporov et al., 2019, Minot et al., 2011, Reyes et al., 2010), genitourinary tract (Santiago-Rodriguez et al., 2015, Wylie et al., 2012) and skin surface (Foulongne et al., 2012, Schowalter et al., 2010) found in asymptomatic and symptomatic individuals have been reported. Recent studies have revealed that the virome plays an important role in homeostatic regulation of the microbiota, modulating the immune system and contributing to human health and disease pathogenesis (Norman et al., 2015, MacDuff et al., 2015, Lim et al., 2015, Reyes et al., 2013).

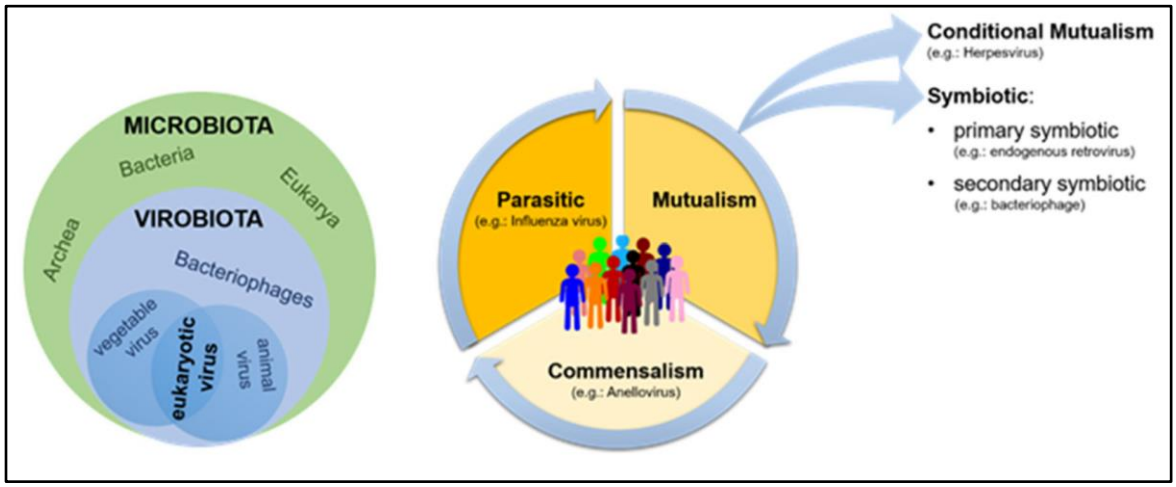


Figure 1.1. Human microbiota and virobiota. (Left) The green circle represents the community of the human microbiota (including bacteria, archaea and eukarya) and the human virobiota is a subset of the microbial community (including bacteriophages and eukaryotic viruses; blue-light circle). (Right) The circle shows dynamic interaction between microorganisms and humans: the microbial community can shift from parasitism to mutualism and commensalism in asymptomatic or healthy individuals, whereas healthy individuals may convert into a disease state if the microbial and/or viral communities move toward parasitism (Luganini and Gribaudo, 2020).

1.1.2. The Human Intestinal Virome

Different communities of eukaryotic single-stranded RNA viruses, single-stranded DNA viruses, double-stranded DNA viruses and some endogenous retroviruses have been identified in the intestines and faeces of healthy individuals (Minot et al., 2013, Reyes et al., 2010). Sequence analysis of eukaryotic virome in faecal samples from two healthy infant siblings has also identified the members of the families *Adenoviridae*, *Anelloviridae*, *Parvoviridae*, *Picornaviridae* and *Picobirnaviridae*, and the genera such as *Adenovirus, Aichi virus*, *Astrovirus* (or *Mamastrovirus*), *Bocavirus*, *Enterovirus*, *Parechovirus*, *Picobirnavirus* as well as *Rotavirus* (Kapusinszky et al., 2012). These eukaryotic viruses can be regarded as constituents of the human commensal virobiota as they can establish chronic, latent or persistent infection with or without symptoms in the human body (Lecuit and Eloit, 2013).

However, the existence and the role of eukaryotic viruses in the human GIT still remain unclear and need to be further characterised. A longitudinal study comparing the intestinal virome among four pairs of healthy infant twins has showed that the richness of the eukaryotic viral communities is relatively low in the earliest-in-life infant twins but increases thereafter, suggesting that the human eukaryotic virome is initially established due to environmental exposures (Lim et al., 2015). Also, the infant intestinal virome and

microbiome show more similarity between co-twins instead of unrelated infants. However, their findings were in disagreement with another earlier virome study of four pairs of adult female co-twins and their mothers, showing that the human virome is highly personalised without sharing more similarity with each other between these co-twins than unrelated subjects (Reyes et al., 2010). These different results can be interpreted by environmental exposures primarily driving virome composition as the infant twins generally live in the same household conditions (Lim et al., 2015). On the other hand, the species of anelloviruses were frequently found in infant specimens at six to twelve months of age and progressively decreased at around eighteen months of age that may result from the suppressed immune state of mothers as it coincides with the low titre of maternal immunoglobulin G (IgG) (Lim et al., 2015). However, sequence independent amplification (SIA) and multiple displacement amplification (MDA)-based methodologies used in this study may contribute to bias. Similarly, amplification bias may also exist using whole genome amplification (WGA) method in the study of Reyes *et al.* (2010), although they have considered and attempted to minimise potential bias in amplification process by conducting independent WGA reactions three times.

The first human virome was isolated from the faeces of a 33-year-old healthy male individual using a clone-based shotgun library, estimating that the human intestinal virobiota consists of approximately 1,200 viral genotypes with the most abundant virus accounting for around 4% of the total (Breitbart et al., 2003). The majority of viruses found in their study were *Siphoviridae* and prophages, in total accounting for around 81%. Also, sequences of *Myoviridae*, *Podoviridae* and *Microviridae* associated with the hosts *Listeria monocytogenes*, *Burkholderia thailandensis* and *Lactococcus lactis* were identified in this study (Breitbart et al., 2003). More details of the human intestinal virome and the related isolation methodology are described in **Chapter 3**.

1.2. Bacteriophage-Involved Human Intestinal Virome

1.2.1. Virus and Bacteriophage

Virus, which originally means “poison”, is a nano-sized, filterable and transmissible agent with simple genetic composition that can widely infect invertebrates, vertebrates, plants, bacteria, archaea, algae, fungi, yeast and protozoa (Fauquet, 1999), but can only multiply in living systems including eukaryotic and prokaryotic cells, originally identified by Dmitri Ivanovsky in 1892, in Russia, and then later confirmed and named by Martinus Beijerinck in 1898, in Netherlands (Beijerinck, 1898, Iwanowski, 1892). Typically, viruses have either DNA or RNA genome with simple essential viral components enclosed in a protein-based capsid. Each virion with a protective capsid only is called “naked” but some of them are “enveloped”, having a lipid-based envelope surrounding a naked virion that results from host cell membranes during viruses release from a cell (Orlova, 2012). Viruses generally are stable at pH 5 to 8 (Pirtle and Beran, 1991), while some (e.g. archaeal viruses or certain thermostable bacteriophages from hot springs) can be found in extreme environments such as hot springs or acidic conditions (Krupovic et al., 2018, Minakhin et al., 2008, Pirtle and Beran, 1991). Viruses are generally sensitive to detergents such as urea, sodium dodecyl sulfate (SDS), and chaotropic salts such as guanidinium thiocyanate (GTC) (Steward and Culley, 2010). Naked viruses (e.g. *Primate erythroparvovirus 1* or called *Parvovirus B19*) are generally resistant to chloroform (Conceicao-Neto et al., 2015), however, lipid-containing viruses including enveloped viruses (e.g. *Influenza virus*) and certain bacteriophages (e.g. *Pseudomonas virus PRD1*, *Tectiviridae*) are susceptible to chloroform and may lose infectivity (Olsen et al., 1974, Feldman and Wang, 1961, Morgan et al., 1956).

Viruses that infect bacterial hosts are called “bacteriophages” or “phages”, which means “eater of bacteria” (Orlova, 2012). Due to their lytic effect on their hosts, some virulent phages are therefore referred to as “lytic” (as opposed to “lysogenic”, which represents temperate phages), originally discovered by Frederick Twort in 1915 and then later confirmed and named by Felix d’Hérelle in 1917 (d’Hérelle, 1917, Twort, 1915). Viruses, particularly bacteriophages, have been considered the most abundant biological entities on the planet, estimated to be approximately 10^{30} to 10^{32} in number (Breitbart and Rohwer, 2005, Wommack and Colwell, 2000). The first survey of phage classification initially began in 1967 (Bradley, 1967), with the number of novel strains continuously accumulating over time. Although by original definition viruses and phages are filterable, some novel giant viruses (e.g. *Mimiviridae* family, around 0.8 μm in size) and jumbo phages have not been included (Yamada et al., 2010, Raoult et al., 2007). Thus, the definition of virus and phage needs to be modified to include giant viruses, proposed by a recent report (Raoult and Forterre, 2008).

To date, nearly 9,828 complete genomes of viruses have been published on the NCBI (National Centre for Biotechnology Information) viral RefSeq database, including ~224 human-related viruses, ~3,361 bacteriophages and ~80 archaeal viruses, which can be classified into ~7,571 taxa (NCBI, https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/, available in December 2020). Moreover, nearly 6,600 viral species have been reported by ICTV (International Committee on Taxonomy of Viruses, https://talk.ictvonline.org/taxonomy/, available in December 2020), and of these, over 6,300 prokaryotic viruses including bacteriophages and archaeal viruses have been morphologically characterised by electron microscopy (Ackermann and Prangishvili, 2012). **Figure 1.2** and **Table 1.1** show the classification of the common bacterial viruses and the representative examples for each type.

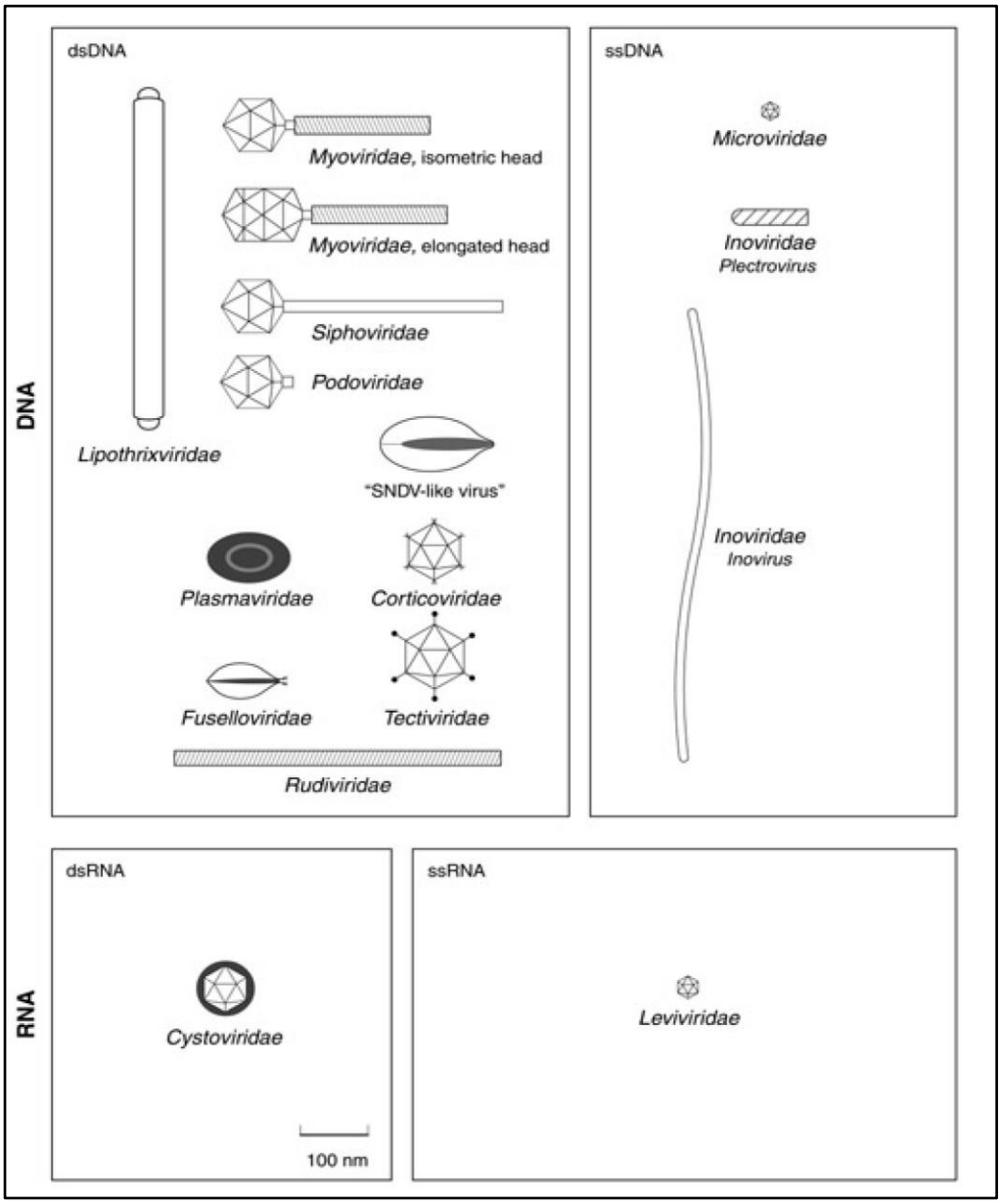


Figure 1.2. Families and genera of bacterial viruses. This diagram illustrates morphological types of bacterial viruses (phages) that can infect the bacterial hosts, including DNA and RNA phages. Image has been reproduced from (Fauquet, 1999), with permission from Elsevier (Copyright © 1999).

Table 1.1. Examples of bacterial viruses (Aksyuk and Rossmann, 2011)

DNA genome						RNA genome	
dsDNA				ssDNA		dsRNA lipid containing *Cystoviridae*	ssRNA *Leviviridae*
Caudovirales (dsDNA tailed phages)			dsDNA lipid containing *Tectiviridae*	Filamentous *Inoviridae*	Icosahedral *Microviridae*		
Podoviridae	*Myoviridae*	*Siphoviridae*	**PRD1**	**M13, fd, f1**	**phiX174**	**phi6, phi8, phi12**	**MS2, Qβ, fr, GA, R17, f2, phiCb5**
P22, T7, phi29, N4	**T4, Mu, G, P2/P4**	**Lambda, SPP1, HK97, T5, p2**					

1.2.2. Intestinal DNA Virome

According to ICTV classification guidelines, the classification of viruses is based on several criteria such as host range, morphology, genome type and viral structures (Fauquet, 1999). Of these classification criteria, viral morphological observation by electron microscopy and the nature of the genetic materials analysis by sequencing have been the most universal approaches to characterising viruses.

To date, most DNA phages found in the human GIT can be taxonomically classified into two main categories based on both of the composition of genetic materials and the morphological properties: (1) double-stranded DNA (dsDNA) phages, which are divided into the families of *Myoviridae* (non-enveloped, contractile tails composed of a sheath and a central tube), *Siphoviridae* (non-enveloped, long, non-contractile tails) and *Podoviridae* (non-enveloped, short, non-contractile tails), under the order *Caudovirales* (which have been grouped into the clades of the *Duplodnaviria* > *Heunggongvirae* > *Uroviricota* > *Caudoviricetes* > *Caudovirales*, updated as from 2019, ICTV), and (2) single-stranded DNA (ssDNA) phages, particularly the *Microviridae* family, under the order *Petitvirales* (which has been grouped into the clades of the *Monodnaviria* > *Sangervirae* > *Phixviricota* > *Malgrandaviricetes* > *Petitvirales*, updated as from 2019, ICTV) (Fauquet, 1999, Ackermann, 1998).

1.2.2.1. Double-Stranded DNA Viruses

The vast majority of known bacteriophages on Earth are tailed and constitute the order *Caudovirales*, accounting for around 96% (Ackermann, 2007). Of these tailed phages observed by electron microscopy, 61% are of the *Siphoviridae* family, 25% the *Myoviridae* family and 14% belong to the *Podoviridae* family (Ackermann, 2007). All members of the order *Caudovirales* are non-enveloped, dsDNA viruses with larger genomes, of between 15 and 500 kb, including giant phages (Hatfull and Hendrix, 2011, Hendrix, 2009, Casjens, 2005). The VLP morphology and the approximate size of members of the order *Caudovirales* observed by transmission electron microscopy (TEM) are described in detail in **Chapter 5**. The order *Caudovirales* has a very broad host range that includes the vast majority of the bacterial phyla identified in the human GIT such as *Firmicutes*, *Bacteroidetes*, *Proteobacteria* and *Actinobacteria* (Shkoporov and Hill, 2019). The most typical member of the *Caudovirales* in the human GIT is the *Siphoviridae* family which has been detected using either *de novo* or from reference database-based approaches (Shkoporov et al., 2018b, Minot et al., 2013), or by infecting common intestinal bacterial hosts such as *Bacteroides* that have been isolated and propagated in culture (Ogilvie et al., 2012). Recently, some atypical, novel viruses and phages have also been identified in the human intestinal virome

studies, such as giant viruses/jumbo phages and crAssphages, which likely play a vital role in balancing the relationships between virus, microbe and host in the human GIT. These atypical giant viruses, jumbo phages and crAss-like phages are described in detail in the following sections.

(1) Giant Viruses and Jumbo Phages

In the past decades, “giant viruses” (**Figure 1.3**) and “jumbo phages” (**Figure 1.4**) have been reported in the human intestinal virome. For example, a novel viral order called *Megavirales*, including the families of *Mimiviridae* and *Marseilleviridae* that can infect human intestinal parasites such as amebae from aquatic environments, has been proposed (Colson et al., 2013a,b). All members of the order *Megavirales* are dsDNA viruses with giant dimensions (~0.8 μm) and larger genomes (>300 kb) unable to pass through a filter smaller than 0.45-μm pore size, making them frequently excluded in filtration-based virome studies. Recently, the megavirome (e.g. *Mimivirus*) isolated from human faeces has also been associated with human intestinal diseases such as diarrhoea (Colson et al., 2013b), but further investigation for their existence and roles in the human GIT is required. Moreover, a group of tailed bacteriophages with large virion sizes and dsDNA genomes larger than 200 kb is referred to as “jumbo phages” (Yuan and Gao, 2017, Hendrix, 2009). The jumbo phages are isolated from diverse environmental samples such as soil, sewage and seawater (Drulis-Kawa et al., 2014, Yamada et al., 2010, Sullivan et al., 2010, Serwer et al., 2007). Also, more recently, some “megaphages” (>540 kb in genome size) infecting the Gram-negative intestinal bacteria, *Prevotella*, were identified in the human intestinal microbiome datasets from the faeces of ten Bangladeshi male adults and twenty-seven Tanzanian individuals as well as in two African baboon social groups and Danish pigs, suggesting that these giant phages are likely to be widespread in the GIT of diverse species but may have been overlooked (Devoto et al., 2019).

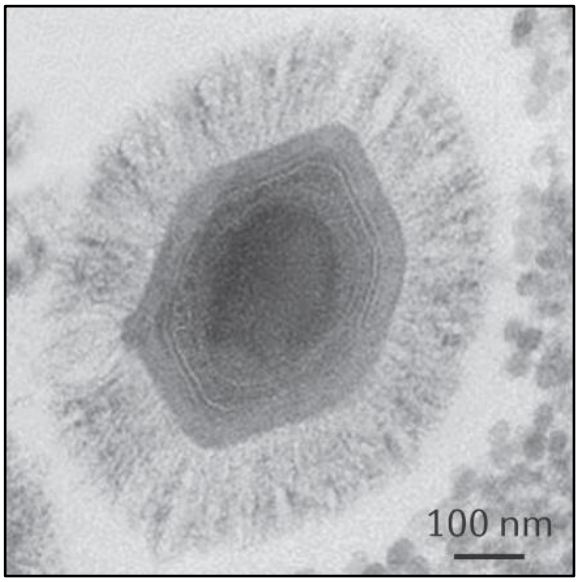


Figure 1.3. Transmission electron micrograph of a giant virus. This image is an example of a *Mimivirus* particle (belonging to *Mimiviridae* family) called *Acanthamoeba polyphaga mimivirus* (APMV) infecting the host *Acanthamoeba* sp. Cells. Scale bar: 100 nm. Image has been reproduced from (Colson et al., 2017), with permission from Springer Nature (Copyright © 2017).

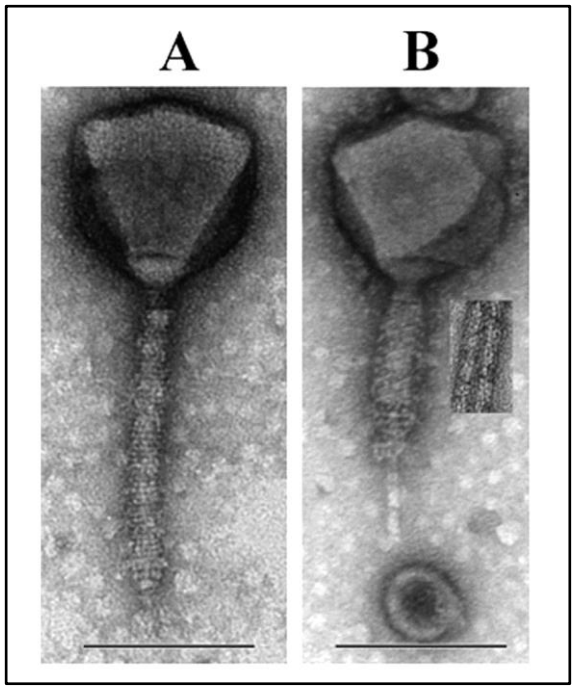


Figure 1.4. Transmission electron micrographs of a jumbo phage. An example of a giant *Pseudomonas aeruginosa* phage PaBG (belonging to *Myoviridae* family) isolated from the samples of lake water and found to infect the host *P. aeruginosa*. A PaBG virion with the extended (A) and with the contracted (B) tail sheath. Scale bar: 100 nm. Image has been reproduced from (Kurochkina et al., 2018), with permission from Elsevier (Copyright © 2017).

(2) crAss-Like Phages

In 2014, a study led by Dutilh *et al.* validated a group of highly abundant sequences (up to 90% of viral reads predominantly in the human intestines) present in the majority (>50%) of human population with unknown identities in previously published human faecal metagenomes from Western cohorts, which were termed "crAssphage" (Cross-Assembly) (Dutilh et al., 2014). The whole ~97 kb of DNA genome cross-assembled from twelve independent viromes has no similarity to any known viruses and phages, which has been predicted to infect the phylum *Bacteroidetes* based on co-abundance and clustered regularly interspaced short palindromic repeats (CRISPR) (Cinek et al., 2018, Dutilh et al., 2014). Based on their findings, Manrique and colleagues identified the crAssphages and several crAss-like phage genomes in the majority of the healthy human population (Manrique et al., 2016). To date, crAssphage family has been found in diverse environments such as mammalian faeces and aquatic systems (Stachler et al., 2017). To further characterise the crAssphages, Yutin and colleagues used sequence-based taxonomic classification methods to predict their taxonomy and found that crAssphages would be assigned to a novel family level with its *Podo*-like phenotypic morphology (Yutin et al., 2018). More recently, Shkoporov and colleagues have successfully isolated the first human intestinal crAssphage (ΦCrAss001), having a ~102 kb of circular dsDNA genome and *Podo*-like morphology from enriched human faecal filtrates, and propagated in a pure culture of *Bacteroides intestinalis* (Shkoporov et al., 2018a) (**Figure 1.5**). They demonstrated that ΦCrAss001 can stably co-replicate with its host in equilibrium *in vitro* for several weeks without identifying any distinct lysogeny genes. The finding of crAssphages indicates that most human intestinal viromes are likely to be diverse and specific to each individual (Manrique et al., 2016, Reyes et al., 2010), with some viruses or phages (e.g. crAss-like phages and *Microviridae*) tending to predominate and be conserved over time in the whole human population, particularly in healthy individuals (Shkoporov et al., 2019); in agreement with previous findings of high virome stability in a single healthy adult as well as in a cohort study of co-twins (Minot et al., 2013, Reyes et al., 2010).

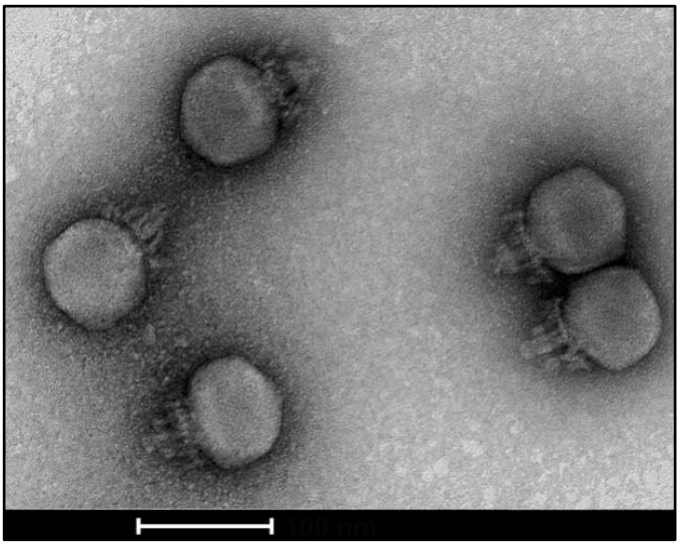


Figure 1.5. Morphology of crAssphage. The first isolated and propagated human intestinal crAssphage (ΦCrAss001) infecting the host *B. intestinalis* has *Podo*-like morphology (Shkoporov et al., 2018a). Scale bar: 100 nm.

1.2.2.2. Single-Stranded DNA Viruses

Members of *Microviridae* family are tailless, non-enveloped, round, cubic or filamentous ssDNA viruses with icosahedral symmetric structures and smaller genomes, approximately from 3 to 7 kb (Roux et al., 2012) (**Figure 1.6**). An earlier report has specifically investigated the ssDNA viruses in the human faecal virome and has showed that sequences of *Microviridae* accounted for 27-49% of those identified in fresh faecal samples from five healthy Koreans. However, the abundance of ssDNA viruses in the human GIT may have been overestimated in this study due to methodological bias generated by the use of phi29 polymerase that preferentially amplifies short circular ssDNA (Kim et al., 2011). More recently, Shkoporov and colleagues also identified several virulent *Microviridae* phages predominantly present in the human GIT (Shkoporov et al., 2019). However, bias may be still present in this study due to utilising multiple displacement amplification (MDA) with phi29 polymerase, although they claimed that MDA has an advantage of converting single-stranded complementary DNA (cDNA) into a double-stranded form in reverse transcription step (Shkoporov et al., 2018b). Some up-to-date protocols for analysing ssDNA viruses in environmental samples have been improved to reduce biases and can therefore be applied to human virome study to avoid overestimation of the ssDNA viruses in the human GIT (Roux et al., 2016, Zhong et al., 2015, Gansauge and Meyer, 2013).

More recently, a study identified sequences of certain ssDNA filamentous phages from 56,868 microbial genomes and 6,412 shotgun metagenomic datasets, particularly the *Inoviridae* family (inoviruses) (Roux et al., 2019). Inoviruses are able to establish a chronic infection without killing their bacterial hosts or interfering with their replication and cell

division, which is known to infect many Gram-negative intestinal bacterial pathogens such as *Vibrio cholerae*, *Escherichia coli*, *Salmonella*, *Pseudomonas aeruginosa*, *Clostridium* and other Enterobacteria (Fauquet et al., 2005), implying that certain ssDNA phages or *Inoviridae* may be considered a noticeable part of the human intestinal virome (Sausset et al., 2020).

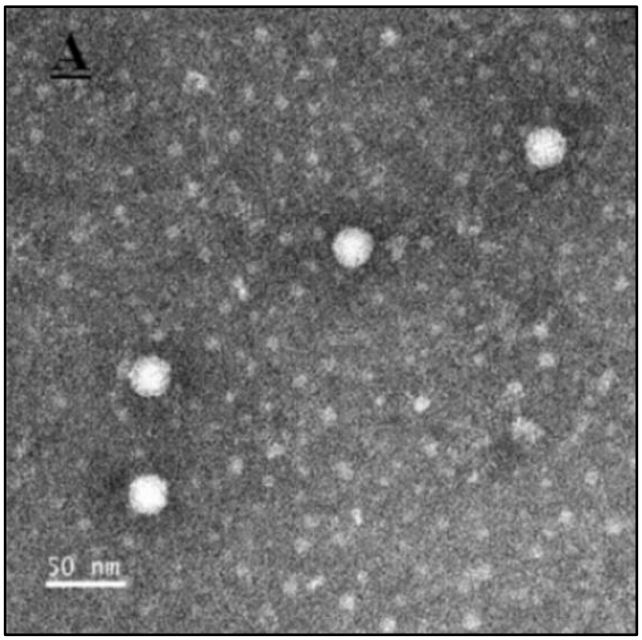


Figure 1.6. Morphology of Microviridae. An example of a novel ssDNA virus vB_Cib_ssDNA_P1 (belonging to *Microviridae* family) isolated from marine samples and found to infect the host *Citromicrobium bathyomarinum* RCC1878, and other members of the genus (Zheng et al., 2018). Scale bar: 50 nm.

1.2.3. Intestinal RNA Virome

In addition to the intestinal DNA virome, RNA viruses have been found in the human GIT, revealed by recent virome analysis, but much less common than dsDNA virome (i.e. the *Caudovirales*) (Lim et al., 2015, Zhang et al., 2006b, Shkoporov et al., 2018b). A plant-pathogenic RNA virus found in the human GIT called *Pepper mild mottle virus* (PMMV) that can infect all species of the genus *Capsicum* is likely a result of daily dietary ingestion (Zhang et al., 2006b). This finding indicated that the majority of enteric RNA viruses may come from exogenous sources such as food, including fruits and vegetables (Zhang et al., 2006b, Reyes et al., 2012). Another study also identified several plant-pathogenic RNA viruses such as PMMV, *Tobacco mosaic virus* (TMV) and *Kyuri green mottle mosaic virus* (KGMMV) in the faeces of two Japanese patients using random reverse transcription-polymerase chain reaction (RT-PCR) analysis (Nakamura et al., 2009). Although they also identified several typical eukaryotic RNA viruses in human faecal samples such as *Norovirus*, *Coronavirus* as well as endogenous retroviruses (e.g. *Human endogenous retrovirus K*, HERV-K), bias may exist in this study due to the use of random RT-PCR amplification with whole transcriptome amplification (WTA) kit for cDNA synthesis (Nakamura et al., 2009). An earlier study reported that PMMV from the faeces of healthy individuals may be associated with specific immune responses and clinical symptoms, suggesting that there is a direct or indirect pathogenic role of plant viruses in humans (Colson et al., 2010), but more compelling evidence is required to identify the relationships between plant viruses and human diseases.

Human eukaryotic RNA viruses include some common intestinal pathogens such as the species of reoviruses, rotaviruses, enteroviruses, noroviruses, bocaviruses and some possible retroviruses (e.g. *Human immunodeficiency virus*, HIV) that can cause asymptomatic, latent, acute or chronic infection and have also been reported in recent intestinal/faecal virome studies (Lim et al., 2015, Reyes et al., 2010, Victoria et al., 2009), implying that eukaryotic RNA viruses would be likely to lie dormant in the human GIT until reactivation (Duerkop and Hooper, 2013). However, RNA phages are rarely seen in human intestinal viral communities, likely due to low viral loads and the limitation of detection/isolation methodology.

1.2.4. The Life Cycle of Phages

Based on different infection lifestyles, bacteriophages can be functionally divided into lytic (virulent), lysogenic (temperate) or chronic types (Sausset et al., 2020) (**Figure 1.7**). With regard to lytic phages, most belong to tailed *Caudovirales* infecting bacterial host cells by hijacking and exploiting the host replication mechanisms to synthesise and assemble new virions. Once newly synthesised phage DNA are packaged and newly formed virions are assembled, virions are released by lysing the host cells. Lysogenic phages can stay in a dormant or latent infective state without killing the hosts through either integrating their nucleic acids into bacterial chromosomes or forming independently episomal DNA or plasmids, replicating along with host cells and forming a stable “prophage”. However, in response to environmental stressors such as temperature, oxidative stress and bacterial DNA damage resulting from antibiotics treatment (e.g. quinolones), prophages can therefore be induced and may restore a lytic or chronic cycle (Matos et al., 2013, Selva et al., 2009, Cowlishaw and Ginoza, 1970). In addition, certain filamentous bacteriophages, particularly the *Inoviridae* family, are likely able to establish a chronic infection of intestinal bacterial pathogens (e.g. *V. cholerae*, *P. aeruginosa* or others). They can continue the production of new virions alongside the reproduction of the host cells remaining infected without killing or disrupting their replication and cell division, either in a viral-productive state or in a latent prophage state (Sausset et al., 2020, Fauquet et al., 2005).

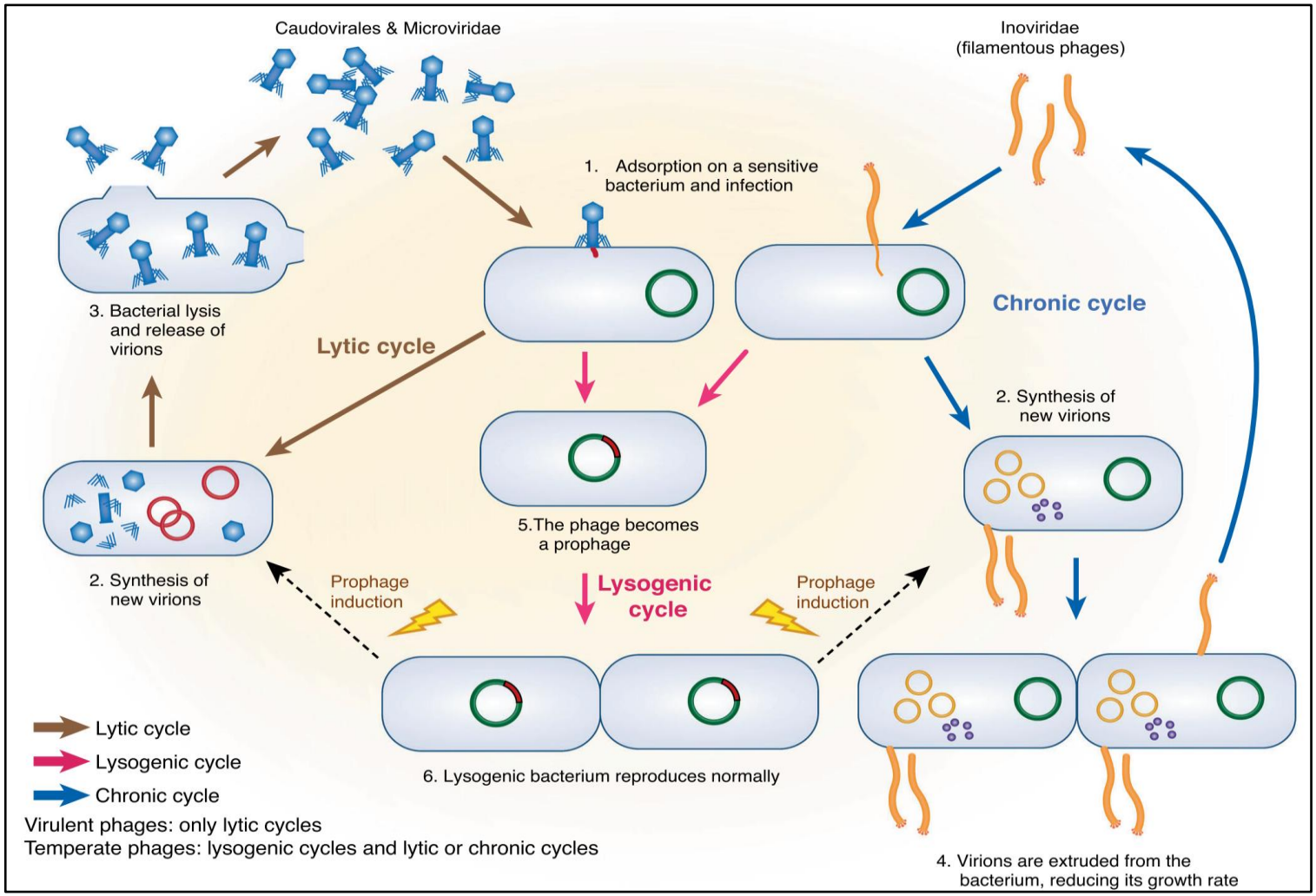

Caudovirales & Microviridae
Inoviridae
(filamentous phages)
1. Adsorption on a sensitive bacterium and infection
3. Bacterial lysis and release of virions
Lytic cycle
Chronic cycle
2. Synthesis of new virions
5.The phage becomes a prophage
2. Synthesis of new virions
Prophage induction
Lysogenic cycle
Prophage induction
6. Lysogenic bacterium reproduces normally
4. Virions are extruded from the bacterium, reducing its growth rate
Lytic cycle
Lysogenic cycle
Chronic cycle
Virulent phages: only lytic cycles
Temperate phages: lysogenic cycles and lytic or chronic cycles

Figure 1.7. Potential models of life cycles of virulent and temperate phages in the human intestines. This diagram illustrates the production of new progeny virions reproduced through either lytic (virulent; brown arrows), lysogenic (temperate; pink arrows) or chronic (blue arrows) cycles. All life cycles begin with phage recognition and adsorption to a targeted bacterial host cell and phage DNA translocation into the host (1), followed by phage DNA synthesis, packaging and assembly of new virions (2). In the lytic cycle, new virions (e.g. *Caudovirales* and *Microviridae*) are released by lysing the host (3), while in the chronic cycle, new virions of filamentous ssDNA phages (e.g. *Inoviridae*) are released from the bacteria using a specific secretion mechanism without killing the host (4). Some phages are able to enter the lysogenic cycle, whereby they stay in a dormant or latent state in the infected host and form the "prophages" (5) by either integrating into the bacterial genomes or in an episomal state, replicated with the host chromosome during cell reproduction and division (6). For lysogenic or temperate phages, once the host cells encounter environmental stress, the prophages can be induced and may resume a lytic or a chronic cycle (Sausset et al., 2020).

1.3. Intestinal Virome to Human Health

1.3.1. Transkingdom Interactions Between Virus, Microbe and Host

Intestinal virome including phages (or phageome) can regulate or be regulated by other microbiomes such as bacteria (or bacteriome) via direct or indirect 'transkingdom' interactions. "Transkingdom" interactions defines members of two or more microbial communities from different kingdoms dynamically interacting with the viromes within the human body, particularly in GIT, thereby involved in balancing human health and disease (Pfeiffer and Virgin, 2016, Virgin, 2014). For instance, intestinal antiviral mechanisms of *Drosophila* can be regulated via the signalling of Gram-negative bacteria to restrict intestinal viral infections (Sansone et al., 2015). Conversely, intestinal viruses may also play a role in protecting the host GIT from dysbiosis and pathogens (Kernbauer et al., 2014). In some cases, commensal bacteria are likely to facilitate viral replication and infection such as norovirus in the host GIT (Baldridge et al., 2015, Jones et al., 2014). Moreover, intestinal phageome may play a role in maintaining the homeostasis between bacteria and host by altering the balance of intestinal bacteriome via predator-prey interactions and in disease pathogenesis such as human inflammatory bowel disease (Norman et al., 2015). Thus, in some cases, viruses directly interact with microorganisms within the intestinal microbial communities. In other cases, intestinal microbiome may indirectly affect intestinal viral infection by regulating host immune system. Moreover, individual genetic variations between hosts may also involve viral transkingdom interactions, thereby contributing to clinically symptomatic or asymptomatic phenotypes (Pfeiffer and Virgin, 2016, Virgin, 2014).

1.3.2. Phage-Mediated Intestinal Dysbiosis

Alterations (gain or loss) in the composition of intestinal microbial communities is often referred to as "dysbiosis" (Petersen and Round, 2014). Recently, several theoretical models that potentially mediate phage-associated intestinal dysbiosis have been proposed (De Paepe et al., 2014): First, "kill the winner" model (**Figure 1.8.A**) in which phages infect and kill the most dominant commensal bacteria in the GIT to maintain a balanced ecosystem and prevent bacterial over growth, which is triggered when bacterial populations reach a replication threshold, assumes that phages can opportunistically meet and infect their bacterial hosts as long as the richness and/or abundance of specific bacterial population is relatively high, based on the model of predator-prey dynamics (also known as the Lotka-Volterra equations) (Lotka, 1910). Several studies that support this model have been reported, particularly occurring in the early life of infants or children and involved in disease pathogenesis (Norman et al., 2015, Lim et al., 2015, Reyes et al., 2013, Breitbart et al., 2008).

Second, “biological weapon” or “kill the relative” model (**Figure 1.8.B**) assumes that intestinal commensal bacteria exploit their own species-specific phages or prophages as a weapon to destroy the related competitors which are susceptible to phages, thereby reducing the abundance and diversity of intestinal bacteria and leading to intestinal dysbiosis. An earlier evidence which may support this model revealed that *in vivo* competition assays between lysogen and susceptible *Enterococcus faecalis* strains in the GIT of monoxenic mice showed a transient, ~1.5-fold enrichment of the lysogen higher than the susceptible strains after 24-hour co-colonisation, implying that the wild-type *E. faecalis* strain can utilise a composite phage as a weapon to infect and lyse the related susceptible *E. faecalis* and mutant strains during co-colonisation (Duerkop et al., 2012).

Third, “community shuffling” model (**Figure 1.8.C**) in response to environmental stresses such as oxidative stress, antibiotic treatment or inflammation hypothesises that environmental stressors induce prophages to kill related commensal bacteria by activating the lytic cycles, thereby reducing the abundance and diversity of intestinal bacterial community and contributing to dysbiosis. For example, antibiotic treatment such as quinolone or beta-lactams related to prophage induction has been found in several bacterial species, including *E. coli* (Zhang et al., 2000), *Clostridium difficile* (Meessen-Pinard et al., 2012), *E. faecalis* (Matos et al., 2013) and *Staphylococcus aureus* (Goerke et al., 2006).

Fourth, “emergence of new bacterial strain” or “invade the relative” model (**Figure 1.8.D**) has been proposed that temperate phages may invade bacterial population to produce lysogeny or prophages instead of causing cell lysis, and novel bacterial strains with the capability of antibiotic resistance or pathogenic toxin production may therefore emerge by obtaining additional genetic materials from other related strains. Recently, several metagenomic studies have highlighted some antibiotic resistance genes involving the intestinal virome (Modi et al., 2013, Minot et al., 2011), implying that some phages may horizontally transfer bacterial DNA with these antibiotic resistance genes between hosts via phage-mediated transduction, thereby generating new strains with the capability of antibiotic resistance (De Paepe et al., 2014).

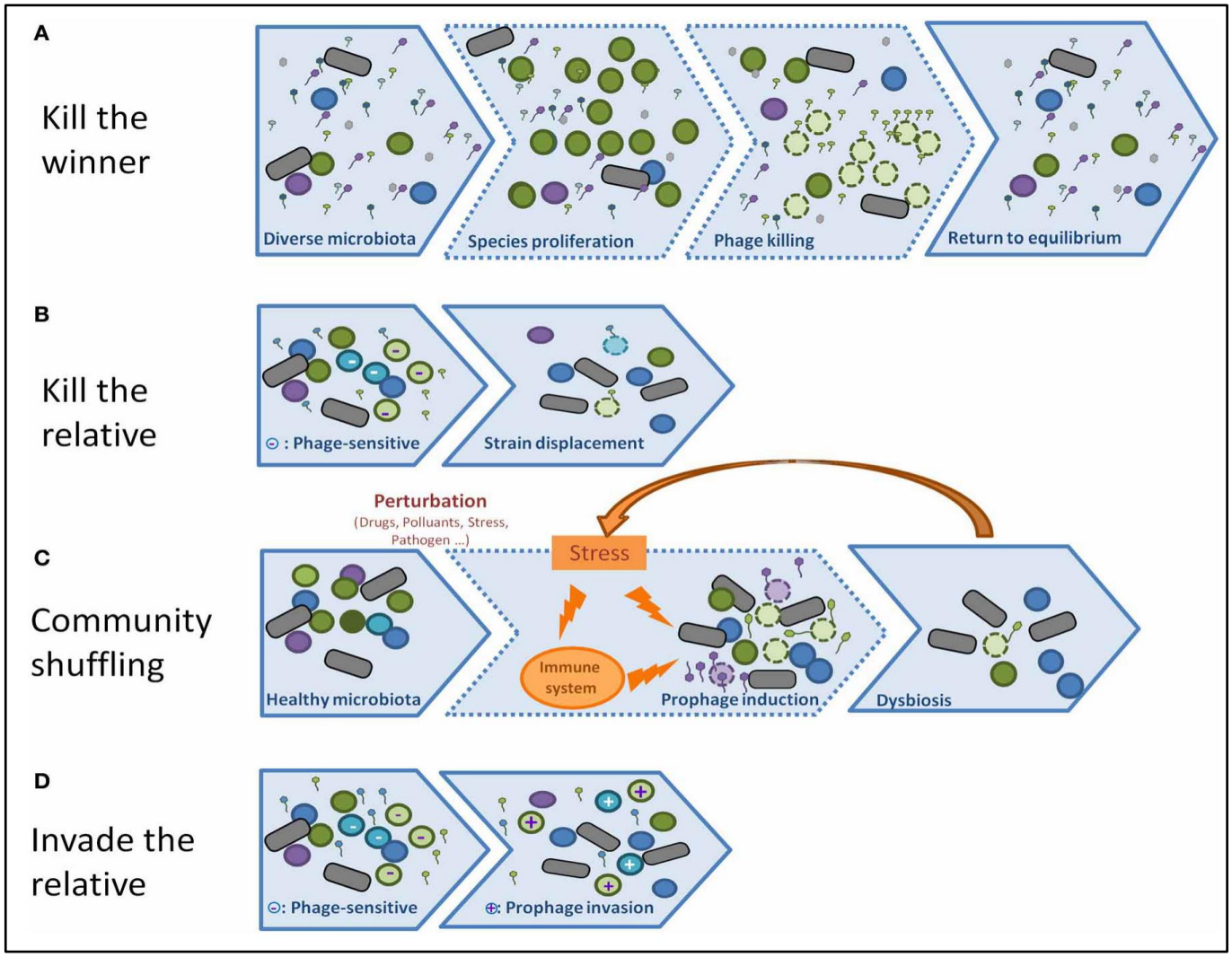

A
Kill the winner
Diverse microbiota
Species proliferation
Phage killing
Return to equilibrium
B
Kill the relative
⊖ : Phage-sensitive
Strain displacement
C
Community shuffling
Perturbation
(Drugs, Polluants, Stress, Pathogen ...)
Stress
Healthy microbiota
Immune system
Prophage induction
Dysbiosis
D
Invade the relative
⊖: Phage-sensitive
⊕: Prophage invasion

Figure 1.8. Hypothetical models of intestinal phage-bacteria dynamic interactions. (A) The schematic diagram represents the “kill the winner” model that phages tend to hunt the most dominant commensal bacteria that may overgrow and reach the replication threshold, thus balancing their numbers and returning to the norm. (B) “Biological weapon” or “kill the relative” model illustrates that commensal bacteria are likely to drive strain-specific phages or prophages as a weapon to displace or kill other phage-sensitive bacterial competitors that also colonise in the same area such as intestine. (C) “Community shuffling” model illustrates that intestinal dysbiosis may be triggered due to prophage induction by introducing environmental stressors such as antibiotics, pollutants, oxidative stress or inflammation. (D) “Invade the relative” or “emergence of novel bacterial strain” model hypothesises that novel bacterial strains may emerge by obtaining additional genetic materials (e.g. antibiotic resistance genes or pathogenic toxin genes) from other related strains via phage or prophage-mediated transduction and invasion without cell lysis (De Paepe et al., 2014).

1.3.3. Human Intestinal Virome-Associated Diseases

Recent studies have proposed that reductions in richness and diversity of the human intestinal microbiome and alterations in the intestinal virome may be associated with disease pathogenesis (Zuo et al., 2017, Zhao et al., 2017, Norman et al., 2015). The main disease focus has been *Clostridium difficile* infection (CDI) (Zuo et al., 2017, Meessen-Pinard et al., 2012), inflammatory bowel disease (IBD) (Clooney et al., 2019, Norman et al., 2015, Lepage et al., 2008), HIV-associated acquired immunodeficiency syndrome (AIDS) (Monaco et al., 2016, Li et al., 2012), type I diabetes (T1D) (Zhao et al., 2017, Kramna et al., 2015, Foxman and Iwasaki, 2011), type II diabetes (T2D) (Ma et al., 2018), severe acute malnutrition (SAM) (Reyes et al., 2015) and myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) (Giloteaux et al., 2016b, Fremont et al., 2009, Chia and Chia, 2008). The details of intestinal virome-associated diseases are described in the following sections from **1.3.3.1** to **1.3.3.6**, respectively.

1.3.3.1. *Clostridium difficile* Infection (CDI)

Clostridium difficile or *Clostridioides difficile* infection (CDI) is one of the vital nosocomial pathogens resulting from antibiotic resistance that poses a threat to global public health and hospitalised patients by leading to severe, life-threatening diarrhoea (Centers for Disease Control and Prevention, 2019). According to an epidemiological estimation, there were around 223,900 CDI cases confirmed in 2017 in the US, of which at least 12,800 patients died from CDI (Centers for Disease Control and Prevention, 2019). In the UK, the number of CDI-confirmed cases were 12,275 in 2018/19, of which around 4,200 inpatient cases with CDI were confirmed in 2018/19, but CDI still remains a serious worldwide healthcare challenge (Public Health England, 2019).

A recent study showed that intestinal virome dysbiosis occurs in patients with CDI, characterised by dramatically increasing in the abundance of the *Caudovirales* and decreasing in the richness and diversity of the *Caudovirales*, compared with healthy household individuals (Zuo et al., 2017). However, the treatment of faecal microbiota transplantation (FMT) may have a protective outcome associated with alterations in the virome composition of patients with recurrent CDI (Zuo et al., 2017). When healthy donors had higher richness of *Caudovirales* than that of recipients, FMT treatment had positive outcome; in contrast, over 50% of the recipients who displayed higher richness of intestinal virome than that of donors suffered from recurrent CDI after FMT treatment (Zuo et al., 2017). It has been hypothesised that higher richness of the *Caudovirales* in donors than that in CDI-recipients may play a key role in FMT efficiency, while the underlying roles and mechanisms of the intestinal *Caudovirales* in CDI still remain unknown (Zuo et al., 2017).

Moreover, intestinal virome including prophage induction are likely to involve CDI progression (Meessen-Pinard et al., 2012). Meessen-Pinard and colleagues have isolated and characterised four CDI-associated prophage strains (i.e. ΦMMP01, ΦMMP02, ΦMMP03 and ΦMMP04) from faecal samples of the patients with CDI (Meessen-Pinard et al., 2012). These prophages had *Myo*-like morphology and were able to lysogenise in *C. difficile* without identifying significant virulent factors (Meessen-Pinard et al., 2012). Their findings also suggested that antibiotic treatment such as fluoroquinolones may increase *in vivo* prophage induction and may facilitate phage-mediated gene transfer during CDI (Meessen-Pinard et al., 2012).

1.3.3.2. Inflammatory Bowel Disease (IBD)

Inflammatory bowel disease (IBD) is a chronic intestinal disorder including Crohn's disease (CD) and ulcerative colitis (UC), likely caused by dysbiosis (Tamboli et al., 2004). Although the aetiology of IBD is still unclear, many studies have hypothesised that changes in intestinal microbiome and virome are likely to be associated with disease progression, potentially involving reductions in diversity and/or abundance of certain intestinal bacteria such as *Firmicutes* and *Bacteroides* with an overall increase in the diversity of intestinal virome, particularly the order *Caudovirales* (Norman et al., 2015). An animal-based study provided compelling evidence to support that intestinal virome can play a role in chronic GI-inflammation (Cadwell et al., 2010). In this study, transgenic mice with a mutation in the autophagy gene *Atg16L1*, a human risk allele for a predisposition to CD, grew normally and were asymptomatic. However, CD was triggered by an intestinal *Murine norovirus* (MNV) infection, demonstrating that viral infection interacting with host genome is likely to be associated with disease development, particularly in genetically susceptible individuals (Cadwell et al., 2010).

One patient-based study indicated that viruses may have higher survival rates in non-ulcerated tissues (Lepage et al., 2008). In this study, VLPs (mainly phages) were significantly observed in the colonic biopsies of nineteen CD patients more than those found in fourteen healthy individuals (2.9×10^9 vs. 1.2×10^8 VLPs/biopsy); however, for the CD patients, fewer VLPs were identified in ulcerated mucosa than nonulcerated mucosal area (2.1×10^9 vs. 4.1×10^9 VLPs/biopsy) (Lepage et al., 2008). Moreover, Norman and colleagues found that the composition of the intestinal virome was abnormal in CD and UC patients, compared to household controls, implying that changes in intestinal virome may contribute to GI-inflammation and bacterial dysbiosis, thereby likely triggering disease (Norman et al., 2015). There was an increased tendency in the virome-associated richness (predominantly *Caudovirales* and *Microviridae*) of IBD patients compared with those of

controls. Their findings were confirmed using two independent and geographically distinct (Boston and Chicago, USA) patient cohorts with matched controls (Norman et al., 2015). Moreover, their findings that reductions in richness and diversity of intestinal microbiota accompanied by increased richness of the faecal virome can be interpreted by predator-prey dynamic interactions (Norman et al., 2015), but the potential roles and mechanisms of these intestinal viromes associated with IBD are still uncertain.

More recently, a published dataset of healthy and IBD intestinal viromes from Norman *et al.* (2015) was reanalysed using whole virome analysis-based approach, compared with an in-house longitudinal cohort study composed of forty UC subjects (Clooney et al., 2019). By analysing both datasets, Clooney and colleagues found that there were no significant alterations in the whole-virome richness and alpha diversity between UC and healthy subjects, while significant compositional changes in the intestinal viromes between IBD patients and healthy subjects were observed by beta diversity analysis. Moreover, they proposed that a stable core of intestinal virome predominates in healthy individuals (e.g. crAss-like phages or *Microviridae*) which tends to be replaced with virulent temperate phages (e.g. some *Siphoviridae* and *Myoviridae*) in IBD patients. Their findings implied that a stable and conserved core of intestinal virome may play a key role in maintaining intestinal homeostasis predominantly in healthy individuals, but disease-associated lysogenic phages induced by environmental stresses such as reactive oxygen species (ROS) become lytic and virulent and may be involved the pathogenesis of IBD (Clooney et al., 2019).

1.3.3.3. Acquired Immune Deficiency Syndrome (AIDS)

Depletion of $CD4^+$ T cells caused by HIV infection is a defining feature of AIDS which has been used to classify stages of HIV infection (i.e. stage 0, 1, 2, 3 or unknown) (Centers for Disease Control and Prevention, 2014). According to an epidemiological estimation of Joint United Nations Programme in HIV/AIDS (UNAIDS), around 1.7 million newly confirmed cases of HIV infection were identified in 2019, with a total of around 38 million people suffering from HIV, of which 25.4 million cases are being treated. However, 690,000 people still died from AIDS worldwide in 2019 (Joint United Nations Programme on HIV/AIDS, 2020).

A low level of $CD4^+$ T cells leads to HIV/AIDS patients being susceptible to viral or microbial infection, potentially involving GIT dysfunction, dysbiosis and translocation of microbial products across impaired intestinal mucosal barrier, thereby likely triggering chronic GI-inflammation (Brenchley, 2013, Brenchley et al., 2006). Hence, it has been proposed that HIV/AIDS can be associated with chronic GI-inflammation, translocation of microbes and their products and changes in intestinal bacteriome and virome (Dinh et al., 2015, Dillon et

al., 2014, Handley et al., 2012). Human pathogenic intestinal viruses can directly cause enteropathy such as gastroenteritis, enteritis or colitis. However, if or how intestinal virome including phages engages in intestinal microbial dysbiosis and leads to HIV-related immunodeficiency is unclear (Monaco et al., 2016).

A patient-based plasma virome study has shed light on the compositions of the plasma bacteriome and the DNA virome associated with HIV/AIDS using next-generation sequencing (NGS) technologies (Li et al., 2012). In this study, ten HIV/AIDS patients and ten healthy controls were examined. Li and colleagues found that the plasma bacteriome in HIV/AIDS patients shared a similar composition with the intestinal bacteriome, implying that intestinal microbes and/or their products are likely to be translocated from GIT to blood, and the plasma DNA virome in HIV/AIDS patients as well as healthy adults were similar to common eukaryotic viruses. By drawing a comparison between HIV/AIDS patients and healthy controls, the compositions of the plasma DNA virome were different: *Anelloviridae* belonging to eukaryotic ssDNA virus rather than phages were predominantly identified in the healthy adults; in contrast, the phageome including *Pseudomonas* and *Enterobacteria* phages as well as *Human endogenous retrovirus K* (HERV-K) in the plasma of HIV/AIDS patients, which are likely to be involved in HIV/AIDS progression and/or opportunistic infection (Li et al., 2012).

Recently, another patient-based cohort study indicated that changes in intestinal virome and bacteriome are more likely to be associated with HIV-caused immunodeficiency, triggering AIDS-associated enteropathy and disease development (Monaco et al., 2016). Matched faecal and plasma samples were collected from 122 individuals (Uganda AIDS Rural Treatment Outcomes, UARTO) composed of 82 HIV-positive Ugandan patients including 40 subjects on anti-retroviral therapy (ART) treatment for at least five years, 42 HIV-nontreated subjects, and additional 40 HIV-negative individuals as healthy controls (Siedner et al., 2016). Monaco and colleagues re-examined this cohort investigating the HIV/AIDS-associated intestinal virome and bacteriome using NGS technologies. The vast majority of the VLP-enriched, intestinal virome identified in both HIV-positive and HIV-negative groups were phages, dominated by members of the order *Caudovirales* (dsDNA phages) and the *Microviridae* family (ssDNA phages), along with several eukaryotic viruses such as *Adenoviridae* and *Anelloviridae* family (Monaco et al., 2016).

Furthermore, they revealed that many *Adenoviridae* sequences were detected in HIV-positive cases with <200 $CD4^+$ T cells, compared with both HIV-positive subjects with $CD4^+$ T cells >200 and HIV-negative subjects, implying that an increased abundance of *Adenoviridae* is associated with HIV-caused immunodeficiency (Monaco et al., 2016). However, there were no significant differences in *Adenoviridae* between ART-treated and

nontreated subjects, suggesting that this increase is associated with human immunity instead of therapy (Monaco et al., 2016). Also, the abundance of enteric *Anelloviridae* was significantly different between $CD4^+$ T cell counts and treatment status, and largely increased in HIV-untreated cases with $CD4^+$ T cells <200, suggesting that an expansion of *Anelloviridae* is also likely to involve HIV-caused immunodeficiency (Monaco et al., 2016). Moreover, they showed significant reductions in phylogenetic diversity and richness of the intestinal bacteriome in HIV-infected cases with $CD4^+$ T cells <200, indicating that alterations in intestinal bacteriome, particularly the depletion in *Ruminococcus* (e.g. *R. callidus* and *R. bromii*), is also associated with HIV-caused immunodeficiency (Monaco et al., 2016).

1.3.3.4. Diabetes

Diabetes mellitus (DM) or diabetes refers to as a series of metabolic disorder defined by the presence of hyperglycaemia (i.e. elevated levels of blood glucose), and is a chronic, progressive disease that poses a serious threat to global public health (World Health Organization, 2019). The updated classification of diabetes includes type 1 diabetes (T1D), type 2 diabetes (T2D), hybrid forms of diabetes (e.g. ketosis prone T2D), other specific types (e.g. infections), unclassified types and gestational diabetes, in accordance with the newest guidance (World Health Organization, 2019). According to an epidemiological estimation of WHO (last updated in 2016), more than 400 million people aged over 18 years were suffering from diabetes in 2014 worldwide and totally around 1.6 million people directly die of diabetes in 2016, with the age-standardised prevalence in adults having substantially increased from 4.7% in 1980 to 8.5% in 2014, particularly rising faster in low- and middle-income countries than high-income countries (World Health Organization, 2016). To date, both the number of clinical cases and the prevalence of diabetes have been steadily and continuously growing (World Health Organization, 2016).

Type 1 diabetes (also known as insulin-dependent), one of the metabolic autoimmune diseases involving the human immune system self-attacking and damaging the insulin-secreting β-cells of the islet of Langerhans of the pancreas, is referred to as insulin deficiency in the human body, thereby leading to reliance on daily administration of exogenous insulin for survival, the majority usually occurring in children and adolescents (Zhao et al., 2017, World Health Organization, 2016). Recently, alterations in intestinal microbiota which involve intestinal dysbiosis and increased intestinal permeability have been correlated with T1D pathogenesis (Mejia-Leon et al., 2014, Murri et al., 2013). An increase in certain *Bacteroides* species (e.g. *B. ovatus*) and a reduction in lactate-producing (e.g. *Bifidobacterium* spp.) as well as butyrate-producing species (e.g. *Eubacterium*,

Roseburia or *Faecalibacterium* spp.) have been observed in T1D-related cases (de Goffau et al., 2013, Brown et al., 2011).

Alterations in intestinal eukaryotic viruses and disease-associated phages have also been identified in T1D pathogenesis. Previous animal-based studies have demonstrated that viral infection can damage pancreatic cells (Oldstone et al., 1991). Moreover, many observations have also implicated intestinal eukaryotic viruses such as *Coxsackie B virus* in T1D progression (Coleman et al., 1973). In addition, epidemiological evidence indicated that the presence of enterovirus infection may involve the progress of T1D in genetically predisposed children (Stene et al., 2010). A recent case-control study suggested that no significant alterations in intestinal virome were observed in early development of T1D autoimmunity (Kramna et al., 2015). Kramna and colleagues analysed the faecal virome of nineteen children with early development of islet autoimmunity in comparison with nineteen matched healthy controls using NGS to characterise intestinal virome and real-time PCR to verify the human eukaryotic viruses. The majority of the viruses detected in both T1D subjects and matched healthy controls were phages but no significant differences in the compositions of faecal viromes in both participant groups were seen, suggesting no associations with early development of T1D in children (Kramna et al., 2015).

To further assess whether or how intestinal virome involves the progress of T1D autoimmunity, one VLP-enriched virome study collected faecal samples from eleven children with the presence of serum autoantibodies representing the development of T1D recruited from a previous cohort and additional eleven healthy controls matched with gender, HLA (human leukocyte antigen) genotype, age, delivery route and country for NGS analysis (Zhao et al., 2017). Zhao and colleagues showed that reductions in the diversity of intestinal virome were seen in T1D cases rather than in healthy controls (Zhao et al., 2017). They also noticed a significant increase in the abundance of *Circoviridae*-related eukaryotic viral contigs in controls more than that in T1D cases (Zhao et al., 2017). Moreover, their findings indicated that alterations in the diversity and richness of intestinal viromes (e.g. *Microviridae*, *Myoviridae* and *Podoviridae*) of both cases and controls were different over time: higher Shannon diversity and richness were seen in healthy controls, compared with T1D cases (Zhao et al., 2017). Furthermore, several disease-discriminatory phage contigs that can be correlated to specific bacterial species such as *Bacteroides* and *Bifidobacterium* were identified, implying an association with disease progression (Zhao et al., 2017).

Globally, the vast majority (90-95%) of the diabetes cases are type 2 diabetes (also known as insulin-resistant) that involves multiple risk factors such as overweight and obesity, commonly occurring in adults, resulting from β-cell dysfunction, thereby leading to ineffective utilisation of insulin in the body (World Health Organization, 2019). Recently,

intestinal microbial dysbiosis has been associated with T2D pathogenesis (Qin et al., 2012), but the roles of intestinal virome in the development of T2D are still unknown. To further characterise intestinal virome or phageome in T2D and evaluate if the intestinal phageome is able to be correlated to T2D pathogenesis, Ma and colleagues re-analysed the metagenomic sequences collected from a previously published Chinese cohort from Qin *et al*. using whole-community metagenomic sequencing-based approaches (Ma et al., 2018). In this study, a significant enrichment of intestinal phageome was seen in 71 T2D subjects compared with 74 healthy adults, with the identification of seven specific phage operational taxonomic units (pOTUs), defined by a group of phages sharing with homologous phage taxon-specific genes and the same genus of bacterial hosts, followed by a further validation from independent VLP-based metagenomic sequencing datasets generated by three previous studies (Norman et al., 2015, Minot et al., 2013, Minot et al., 2011). The majority of these pOTUs were of the order *Caudovirales*, with the most abundant phageome being *Siphoviridae*, followed by *Myoviridae* and *Podoviridae* families (Ma et al., 2018). Moreover, they revealed that dramatic changes in intestinal virome in the T2D cases were seen and validated, with a significant increase in the relative abundance of members of the order *Caudovirales* in T2D groups compared with healthy controls, of which there were seven pOTUs specifically observed in the T2D samples rather than in the healthy controls, including four *Siphoviridae*, two *Podoviridae* and one unclassified family, which can potentially be assigned to several bacterial taxa including *Enterobacteria*, *Escherichia*, *Lactobacillus*, *Pseudomonas* and *Staphylococcus* (Ma et al., 2018). Thus, an altered intestinal virome and bacteriome have been associated with the progression of T2D, while the underlying roles and mechanisms of the intestinal virome, particularly T2D-associated phages, are not yet been clear (Ma et al., 2018).

1.3.3.5. Malnutrition

Malnutrition, one of the global health issues predominantly contributing to child death, particularly in children aged under five years living in low-income and middle-income countries, is a diet-dependent disorder referred to as nutritional deficiency, including wasting, stunting and obesity (UNICEF, 2020, Bhutta et al., 2017, Black et al., 2013). According to United Nations Children's Fund (UNICEF), WHO and World Bank Group in 2020, around 144 million children (<5 years) worldwide were stunted in 2019, 47 million children were wasted, of which 14.3 million were severe, and 38.3 million were overweight (United Nations Children's Fund et al., 2020). Based on children's body size and the presence of oedema, malnutrition can be divided into different severity, with the most severe form referred to as severe acute malnutrition (SAM) that can be highlighted in some key syndromes, including progressive wasting (marasmus) or oedematous malnutrition

(kwashiorkor), which can largely increase the risk of illness, long-term developmental delays and mortality in infants and children (Bhutta et al., 2017, Reyes et al., 2015).

Recently, the pathogenesis of SAM has been associated with the maturation of intestinal microbiota (Subramanian et al., 2014, Smith et al., 2013), and changes in intestinal microbiota may lead to intestinal mucosal barrier impairment, thereby likely triggering intestinal dysfunction and chronic GI-inflammation in SAM children (Bhutta et al., 2017, Ahmed et al., 2014). The intestinal microbiota involved with metabolism and vitamin biosynthesis in Malawian and Bangladeshi children with SAM was more immature than that in healthy children, potentially involving developmental abnormality as well as disease progression in these children (Subramanian et al., 2014, Smith et al., 2013).

To characterise the development of intestinal microbiome and virome and to investigate whether or how intestinal virome interacts with microbiome in malnourished infant and children in comparison with the healthy child individuals, VLP-enriched intestinal viromes and microbiomes were investigated (Reyes et al., 2015). Faecal samples were collected from a subset of a large twin cohort (317 twin pairs totally), including eight monozygotic and twelve dizygotic Malawian twins, and the sequences of bacteria and VLPs were analysed. Reyes and colleagues identified certain phages (e.g. mostly lysogenic phages or prophages) and viruses (e.g. *Anelloviridae* and *Circoviridae*) that are likely to help distinguish the SAM-discordant co-twins (i.e. one developed kwashiorkor or marasmus and the other sibling remained healthy) from eight other concordant healthy pairs (Reyes et al., 2015). This study also revealed that the developmental program of assembly and maturation of intestinal virome seemed delayed and immature in both members of the SAM-discordant pairs, which cannot be restored by ready-to-use therapeutic food treatment (RUTF) (Reyes et al., 2015).

1.3.3.6. Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS)

(1) History and Epidemiology of ME/CFS

The psychopathologic term “neurasthenia” was coined and independently published by both the Americans E. H. Van Deusen and George Beard in 1869 to illustrate an elusive status of illness with chronic fatigue based on clinical observations (Van Deusen, 1869, George, 1869). This hereafter became a popularised diagnosis and spread to Europe between 1910 and 1980. Previous observations have found that many conditions of neurasthenia bear several striking similarities to chronic fatigue syndrome (CFS) as well as myalgic encephalomyelitis (ME), post viral fatigue syndrome (PVFS), chronic fatigue and immune

dysfunction (CFIDS), post-infectious fatigue syndrome (PIFS), fibrositis and myalgia (Bansal, 2016).

Myalgic Encephalomyelitis and chronic fatigue syndrome (ME/CFS) were derived from two early epidemic outbreaks: one was in 1955, in London, UK (Compston, 1978) and another similar outbreak occurred later in 1984, in Nevada, US (Holmes et al., 1988). ME/CFS is a chronic, disabling and debilitating, heterogeneous disorder featuring a variety of symptoms that lead to the disturbance of central nervous system and immune system, myalgia, post-exertional malaise (PEM), sleep disturbance, gastrointestinal dysfunction, neurological and cognitive impairment, and unexplained fatigue, with the estimated worldwide prevalence of 0.4-1% (Jason et al., 1999). However, the local prevalence widely varies from country to country between 0.2% and 2.6%, depending on the diagnostic criteria used (Nacul et al., 2011, van't Leven et al., 2010, Steele et al., 1998, Wessely et al., 1997). According to the US Centers for Disease Control and Prevention (CDC) report, an estimation of ME/CFS cases in the US is in excess of 800,000 and up to 2.5 million, with the majority (80%) of affected individuals being female (Clayton, 2015, Jason et al., 1999). Only 5-6% or less of ME/CFS cases are able to return to healthy status or to mitigate the symptoms; by contrast, the prognosis in children is dramatically better than adults, with 80% of ME/CFS children rehabilitating to normal health or being much improved with mild symptoms (Cairns and Hotopf, 2005, Bell et al., 2001, Reid et al., 2000). To date, no effective treatments are available to relieve or cure ME/CFS, and no distinctive ME/CFS biomarkers exist for physical examination and clinical diagnosis. Hence, there is a considerable socioeconomic impact associated with ME/CFS, with an earlier report estimating the total economic cost of ME/CFS to be between $17 and $24 billion annually (Jason et al., 2008). In the UK, ME/CFS also imposes a financial burden on patients and their families, as well as on the National Health Service (NHS) (Bansal, 2016). Although there is a lack of epidemiological evidence for the UK, recent surveys have estimated that over 250,000 of the UK citizens are likely to suffer from ME/CFS, with a minimum prevalence of ~0.2% (Pemberton and Cox, 2014, Nacul et al., 2011).

(2) Clinical Criteria of ME/CFS

The clinical definition of ME/CFS is still controversial with a series of distinct and strict criteria being proposed. Holmes and colleagues first proposed the initial definition and criteria in *Epstein-Barr virus* (EBV)-related CFS (Holmes et al., 1988). Fukuda *et al.* subsequently established more distinct criteria to define CFS (Fukuda et al., 1994) without considering the core symptoms of ME/CFS such as PEM and neurocognitive symptoms. Additionally, the Fukuda criteria did not exclude some conditions overlapping with psychiatric symptoms such as depression or anxiety (Carruthers et al., 2011). As a result,

an accurate clinical diagnosis of ME/CFS is very difficult. The Canadian Consensus Criteria (CCC) then attempted to distinguish ME patients from those with psychiatric symptoms including depression and anxiety (Carruthers et al., 2011). Recently, WHO classified ME/CFS as a neurological disorder according to WHO's International Classification of Diseases (ICD G93.3), and the International Consensus Criteria (ICC) was subsequently developed (Carruthers et al., 2011). To date, the diagnosis of ME/CFS is in accordance with the clinical criteria of Fukuda *et al.* (1994), Oxford (Sharpe et al., 1991) and ICC (Carruthers et al., 2011). Currently, the symptom severity of ME/CFS is divided into four levels, from mild (around 50% reduced activities of daily living), moderate (mostly reduced activities of daily living and become housebound), to severe (few activities of daily living and becoming bedbound; refer to **Section 2.1.2.2**) and very severe (lack activities of daily living, becoming bedbound and unable to care for themselves), based on ICC with PEM condition (Carruthers et al., 2011). Of note, a key clinical indicator that distinguishes between depression, fatigue and ME/CFS is not only neuropsychological symptoms such as PEM, but also fatigue lasting for six months or more in adults or at least three months in children. However, a recent clinical standard has been modified for ME/CFS cases in the UK, with the updated criteria being over four months in adults and three months in children (Bansal, 2016).

(3) Potential Aetiology of ME/CFS

The aetiology of ME/CFS is still unclear, with several potential hypotheses including eukaryotic/prokaryotic virus infection (Rasa et al., 2018), intestinal microbial dysbiosis (Armstrong et al., 2017, Giloteaux et al., 2016a, Shukla et al., 2015), metabolic disorder (Germain et al., 2017, Armstrong et al., 2017, Sheedy et al., 2009), mitochondrial dysfunction (Myhill et al., 2009, Behan et al., 1991), and autoimmunity (Morris et al., 2014, Lorusso et al., 2009, Klimas et al., 1990). Of these, the involvement of viral infection may play a potential role in the progress of ME/CFS as clinical and laboratory observations indicated that many ME/CFS cases begin with a flu-like illness, implying that viral infection may trigger ME/CFS (Underhill, 2015).

The first was reported for a polio-like outbreak in 1934, in Los Angeles, US (Meals et al., 1935a, Meals et al., 1935b). This outbreak was considered to be atypical due to inconsistent symptoms observed in patients, with the lack of flaccid paralysis used to define poliomyelitis (Mateen and Black, 2013). Patients were diagnosed with having symptoms of acute upper respiratory infection accompanied by initial toxaemia, diarrhoea, aching muscles with paresis, severe headache, sore throat, malaise, significant chronic fatigue and hyperesthesia (Meals et al., 1935a, Meals et al., 1935b). Another epidemic outbreak for atypical simulating poliomyelitis cases called "Iceland disease" subsequently occurred in

Akureyri, Iceland, between 1948 and 1949, with overlapping similarities to cases in 1934 (Sigurdsson et al., 1950). Over the period between 1934 and 1990, it has been estimated that 61 other similar epidemic outbreaks were reported (Underhill, 2015). Of these, the most severe outbreak was in 1955 in London, UK, during which 255 medical staff of the Royal Free Hospital were admitted to hospital suffering from the disease (Medical staff of the Royal Free Hospital, 1957, The Lancet, 1956). This unexplained, atypical polio-like syndrome was formally renamed “myalgic encephalomyelitis” (Compston, 1978), followed by a later extension to chronic fatigue syndrome (Holmes et al., 1988). A recent outbreak associated with CFS was in 2004 in Bergen, Norway, with at least 5% of patients previously suffering from a waterborne epidemic of *Giardia lamblia* gastroenteritis (Naess et al., 2012).

Several human eukaryotic viruses were identified as potential aetiological agents of ME/CFS, including *Human herpesvirus (HHV)-6*, HHV-7, EBV, *Cytomegalovirus* (CMV), *Parvovirus B19* and enteroviruses, based on detection by real-time PCR (Fremont et al., 2009). Moreover, herpesviruses, polyomaviruses, anelloviruses, adenoviruses, papillomaviruses, *Hepatitis B virus* (HBV), *Hepatitis C virus* (HCV), and HIV may involve chronic and systemic GI-infections (Virgin, 2014). An earlier report attempted to connect ME/CFS with *Xenotropic murine leukemia virus-related virus* (XMRV) found in prostate cancer samples (Lombardi et al., 2009), which has been refuted and attributed to laboratory contamination (Alter et al., 2012, Paprotka et al., 2011).

(4) Current Study: Intestinal Virome and ME/CFS

The majority of ME/CFS patients have persistent or intermittent symptoms of gastrointestinal dysfunction such as irritable bowel syndrome (IBS) (Aaron et al., 2000), likely involving chronic enteroviral infection (Giloteaux et al., 2016b, Fremont et al., 2009, Chia and Chia, 2008). Thus, it has been proposed that changes in the human intestinal viruses and virome may contribute to the pathogenesis of ME/CFS by alterations in intestinal microbiota and leading to dysbiosis.

As of 2020, eight original research reports that involve the investigation of virus and virome in ME/CFS have been published (Schreiner et al., 2020, Williams et al., 2019, Rodrigues et al., 2019, Giloteaux et al., 2016b, Loebel et al., 2014, Fremont et al., 2009, Chia and Chia, 2008, Lane et al., 2003). Of these, three studies mainly focused on the human intestinal/faecal virome in ME/CFS, with one study using real-time quantitative PCR to determine viral loads (Fremont et al., 2009), one other using RT-PCR and immunostaining to detect viruses (Chia and Chia, 2008), and another relying on NGS analysis (MiSeq) (Giloteaux et al., 2016b). Giloteaux and colleagues (2016b) revealed an increased richness in bacteriophages of the order *Caudovirales*, of which the majority were *Siphoviridae* and

Myoviridae families in ME/CFS patients. Fremont *et al.* (2009) used real-time PCR analysis to determine viral loads of eukaryotic viruses in the human GIT, with 48 CFS patients and 35 healthy individuals. Their findings indicated that *Parvovirus B19* DNA was detected in 40% of the CFS cases, while only in 15% of healthy controls, suggesting that *Parvovirus B19* may be associated with the pathogenesis of ME/CFS. Chia and colleagues (2008) used RT-PCR and immunoperoxidase staining for GIT biopsies to directly examine intestinal viruses in 165 CFS patients and 34 healthy subjects. They found that not only *Enterovirus* VP1 (viral capsid protein 1) was detected in 82% of patients, with that in 20% of healthy controls, but also *Enterovirus* RNA was detected with RT-PCR. While these amplification-based methods may introduce bias in the study, the hypothesis has been proposed that alterations in intestinal viruses and virome may directly or indirectly influence human immunity and may cause dysbiosis, thereby leading to ME/CFS.

To date, the influence of intestinal viruses and virome on human health shows great promise for ME/CFS research due to recent advances in the development of NGS technologies and bioinformatic tools, in addition to culture- or clone-based approaches. As a result, it is now possible to investigate the potential link between the human intestinal virome and ME/CFS through a large cohort study using sequence-based approaches.

1.4. Studying Human Virome: Up-to-Date DNA Sequencing Technologies

Over the past one to two decades, advances in the development of shotgun metagenomics and high-throughput, pyrosequencing technologies (so-called "Next-Generation Sequencing", NGS) in combination with improved bioinformatics tools and computational approaches, as well as expanded viral reference databases to identify and predict novel viruses and phages have significantly advanced the study of viral metagenomes (i.e. virome or phageome) (Virgin, 2014, Reyes et al., 2012). To date, three types of NGS systems have been commonly applied for whole-genome microbial metagenomic study, including 454 pyrosequencing system, Illumina sequencing system and Applied Biosystems SOLiD™ System (Goodwin et al., 2016, Oulas et al., 2015, Mardis, 2008). More recently, "Third-Generation Sequencing" (TGS) technologies have been developed and introduced for long-length sequence read analysis, such as nanopore technology-based minION™ system first announced by Oxford Nanopore Technologies (ONT) in 2014 (Jain et al., 2016, Mikheyev and Tin, 2014) or single-molecule real-time (SMRT) technology-based system (e.g. PacBio RS series) first commercialised by Pacific Biosciences (PacBio) in 2011 (Schadt et al., 2010, Eid et al., 2009) (**Table 1.2**).

The accuracy, performance and costs of these sequencing platforms vary, of which the error rate of the Illumina platforms is generally lower than other platforms (particularly TGS platforms) having most bases scoring at Q30 (Phred quality score) or above, representing a probability of 1 in 1,000 that a base is incorrectly called by the sequencer with a corresponding base call accuracy of 99.9% (Illumina, 2011, Ewing et al., 1998, Ewing and Green, 1998). Although NGS platforms are able to generate considerable sequence output, error rates are still higher and have much shorter read lengths, compared with the Sanger method-based, first-generation sequencers (Goodwin et al., 2016). However, these short reads enable the assembly of full-length metagenomes by *de novo* approaches – making it possible for virome and microbiome analysis.

Table 1.2. Comparisons of current sequencing platforms

Platform	Generation	Read length	# of reads per run	Output (# of bases) per run	Accuracy	Time per run	Sequencing principle	Brand/Suppliers
96-capillary ABI 3730xl	1st	400-900 bp	96	Various	99.9%	0.5-3 h	Sanger's chain termination	Thermo Fisher/Life Technologies
Ion Torrent GeneStudio S5 series	2nd	200-600 bp (SE)	2-130 million	0.3-50 Gb	99%	2.5-4 h	Sequencing-by-synthesis: SNA	Thermo Fisher/Life Technologies
454 GS FLX+ series (pyrosequencer)	2nd	Up to 1,000 bp (SE) or 2 x 700 bp (PE)	~1 million	700 Mb	99%	23 h	Sequencing-by-synthesis: SNA	Roche
MiSeq v3	2nd	Maximum 2 x 300 bp (PE)	1-25 million	0.5-15 Gb	99.9%	4-56 h	Sequencing-by-synthesis: CRT	Illumina
HiSeq 2500 (High Output)	2nd	Maximum 2 x 125 bp (PE)	Up to 6-8 billion (PE)	540-1,000 Gb	99.9%	6-11 d	Sequencing-by-synthesis: CRT	Illumina
HiSeq 3000/4000	2nd	Maximum 2 x 150 bp (PE)	Up to 5-10 billion (PE)	650-1,500 Gb	99.9%	<24 h to 3.5 d	Sequencing-by-synthesis: CRT	Illumina

HiSeq X series	2nd	2 x 150 bp (PE)	5.3-6 billion	1.6-1.8 Tb	99.9%	<3 d	Sequencing-by-synthesis: CRT	Illumina
NextSeq 550 (High Output)	2nd	Maximum 2 x 150 bp (PE)	Up to 800 million (PE)	100-120 Gb	>99%	29 h	Sequencing-by-synthesis: CRT	Illumina
SOLiD 5500xl Wildfire	2nd	2 x 50 bp (PE)	~2.8 billion	240 Gb	≥99.9%	10 d	Sequencing-by-ligation	Thermo Fisher/Life Technologies
MinION MK1 series	3rd	Various, longest >2 mbp	Various, >100,000	Up to 30 Gb per flow cell	~88%	Up to 48 h	SMRT	Oxford Nanopore
PacBio RS II: P6-C4	3rd	Various	Various, ~50,000	0.5-1Gb per SMRT cell	86-87%	Up to 4 h	SMRT	Pacific BioSciences
PacBio Sequel series	3rd	Various	Various, ~500,000	Up to 20 Gb per SMRT cell	86-87%	Up to 4 h	SMRT	Pacific BioSciences

All information was obtained from manufacturer's data and literature reviews (Goodwin et al., 2016, Ghurye et al., 2016, Rhoads and Au, 2015), with permission from Springer Nature (Copyright © 2016). SNA: single-nucleotide addition; CRT: cyclic reversible termination; SMRT: single-molecule real-time approach; SE: single-end; PE: paired-end; bp: base pair; mbp: mega base pairs; Gb: giga bases; Mb: mega bases; Tb: tera bases; h: hours; d: day

1.5. Study Aims

1. To develop reliable and reproducible protocols for VLP isolation from human faeces and for faecal VLP quantification using a digital image analysis (DIA)-based method

2. To optimise protocols for VLP DNA extraction to obtain DNA of sufficient quality and quantity for NGS

3. To determine the extent of PCR-amplification bias in sequence-based, virome-enriched metagenomes

4. To develop a bioinformatic pipeline for virome-enriched metagenomic analysis

5. To apply the optimised protocols to the analysis of the faecal virome of severely affected ME/CFS patients and same household healthy control individuals (SHHC)

2. General Materials and Methods

2.1. Sample Collection

2.1.1. Protocol Optimisation

The original study proposal was reviewed and subsequently approved by the University of East Anglia (UEA) Faculty of Medicine and Health Sciences (FMH) Research Ethics Committee (REC) in 2014 (reference FMH20142015-28) and by Health Research Authority (HRA) NRES Committee London Hampstead in 2017 (reference 17/LO/1102; IRAS ID: 218545) (**Appendix 1**). To develop and optimise protocols, faecal samples were obtained from three “healthy” adult males aged between 31 and 39 years following informed consent.

2.1.2. ME/CFS Study

2.1.2.1. Participant Recruitment

Two cohorts of ME/CFS-related participants were recruited at different time points and were reviewed and approved by different Research Ethics Committees. In total seventeen subjects composed of nine ME/CFS patients with severe symptoms and eight healthy controls living in the same household were analysed.

For a batch of faecal samples collected from four ME/CFS-related house-matched pairs and one unpaired ME/CFS patient with severe symptoms in 2017, this study was also reviewed and approved by the University of East Anglia Faculty of Medicine and Health Sciences Research Ethics Committee in 2014 (reference FMH20142015-28). For depositing collected faecal samples in the Norwich Biorepository, this was approved by the Cambridge East Committee of the National Research Ethics Service (NRES). For another batch of faecal samples collected from five ME/CFS-related, house-matched pairs during 2018 and 2019, this study was reviewed and approval by Health Research Authority (HRA) NRES Committee London Hampstead in 2017 (reference 17/LO/1102; IRAS project ID: 218545). Informed written consent was obtained from all participants registered at the Chronic Fatigue Service of the Epsom and St Helier NHS Foundation Trust University Hospital (Carshalton, UK).

2.1.2.2. Diagnostic Criteria and Severity for Selecting ME/CFS Patients

Patients were evaluated based on the criteria described previously (Bradley et al., 2013). Inclusion/exclusion criteria included significant clinical depression and anxiety, clinical history and the Hospital Anxiety Depression Scale (HADS). Patients receiving probiotics or antibiotics for six weeks prior to participating in the research were excluded, except for one participant (sample 18QI103; ME patient with severe symptoms). Recently, Bansal (2016) published a robust ME/CFS diagnostic scoring system to help clinicians diagnose and distinguish other unrelated conditions in those suffering from fatigue. Using this approach a diagnosis of ME/CFS was based on the presence of disabling fatigue for longer than four months and post exertional malaise (PEM) after either physical, mental or emotional over activity. The presence of inflammation, immune activation, organ dysfunction, endocrine dysfunction, gluten sensitivity and autoimmunity were excluded and considered as unrelated causes of fatigue. The modified Bansal's diagnostic scoring system is shown in **Appendix 2**. The disease severity was categorised into three levels according to the following criteria:

- **Mild:** mobile, self-caring, light domestic duties, may be working but to detriment of social, family and leisure activities.

- **Moderate:** Reduced mobility, not working, reduced activities of daily living (ADL), sleeping in daytime, peaks and troughs of activity.

- **Severe:** Few or no ADL, severe cognitive difficulties, wheelchair dependent for mobility, rarely leave their houses, or bed bound and can require someone else to wash, toilet and feed them. Often significant worsening of symptoms with any mental or physical exertion and extreme cases are unable to tolerate any noise and are light sensitive.

2.1.2.3. Same Household Healthy Controls

The same household healthy controls (SHHC) engaged in the study were asked to satisfy the inclusion/exclusion criteria below:

Inclusion criteria:

- Men or women aged between 18 and 70 years

- No current or ongoing medical conditions

- Able to provide informed consent

Exclusion criteria:

- Long-term medical conditions, in particular, affecting the stomach and bowel
- Previously diagnosed with autoimmune diseases, for example, systemic lupus erythematous or rheumatoid arthritis
- Suffer from significant anxiety or depression
- In receipt of immunomodulatory drugs, statins, beta blocker or steroids
- Consumed probiotic capsules or antibiotics for six weeks prior to joining the study

2.1.2.4. Home Visit, Sample Collection and Storage

Severe ME/CFS patients who satisfied the study criteria were identified by clinicians at the Epsom and St Helier CFS Service and occupational therapists at the ME/CFS Service at East Coast Community Healthcare (Lowestoft, UK). The health professionals within these NHS organisations notified patients registered and posted them the details and information. Patients and household individuals willing to participate in the study were asked to contact the research team to arrange a home visit to be taken through the consenting process.

At the first home visit, eligibility was confirmed and volunteers gave consent. An interview was then taken by our research team. Faecal samples were collected in the Fecotainer® (Excretas Medical BV, Enschede, Netherlands) that were provided to participants within 24 hours of the home visit. Subsequent home visits were scheduled to collect samples and ensure volunteers have not received antibiotics or antivirals within the past six weeks, and have completed a 48-hour food diary control. All faecal samples were kept at 4°C for a maximum of 24 hours after collection prior to aliquoting, VLP isolation, viral nucleic acids extraction. All faecal aliquots were then stored at -70°C prior to analysis (for no longer than 24 months). The details of samples are summarised in **Table 2.1**. In this collaborative work, I mainly dedicated myself to the process of samples collected, including sample aliquoting and storage, VLP and VLP DNA isolation, and also got involved in home visits, participant interviews and sample collection with other team members for several times.

Table 2.1. Summary of severe ME/CFS patients and same household healthy controls

Sample	Sample barcode	Health status	Matched pair	Bristol Stool Scale	Gender	Age at time of collection
1	18QI101	severe, PT	1	2	F	59
2	18QI102	SHHC	1	7	M	62
3	18QI103	severe, PT (+antibiotics)	2	6	F	38
4	18QI104	SHHC	2	5	M	38
5	18QI105	severe, PT	3	1	F	22
6	18QI106	SHHC	3	6	F	57
7	18QI109	severe, PT	4	5	F	26
8	18QI110	SHHC	4	2	M	23
9	17TB106	severe, PT	5	no record	F	37

10	17TB107	SHHC	5	F	64
11	17TB108	severe, PT	6	F	23
12	17TB109	SHHC	6	F	60
13	17TB132	severe, PT	7	F	18
14	17TB133	SHHC	7	F	55
15	KM	severe, PT	8	F	44
16	JM	SHHC	8	F	70
17	17TB130	severe, PT	unmatched	F	37

PT: ME/CFS patient; SHHC: same household healthy control

2.2. Materials

2.2.1. Sterilisation

All buffers used were sterilised by 0.22 µm (PES) filtration, followed by autoclaving at 121°C for 15 minutes. All media and glassware were autoclaved using the same conditions.

2.2.2. TBT Buffer

Table 2.2. Stock solutions for TBT buffer

Ingredients	Formula
1 M Tris-HCl	24.228 g of Tris base (Sigma-Aldrich, UK) dissolved in 200 ml of distilled water and added 5 M HCl to adjust pH to 8.0
1 M NaCl	11.688 g of NaCl (Sigma-Aldrich, UK) dissolved in 200 ml of distilled water
0.5 M $MgCl_2 \cdot 6H_2O$	10.165 g of $MgCl_2 \cdot 6H_2O$ (Sigma-Aldrich, UK) dissolved in 100 ml of distilled water

Table 2.3. TBT buffer

Ingredients	Amounts	Final concentration
1 M Tris-HCl (pH 8.0)	100 ml	100 mM
1 M NaCl	100 ml	100 mM
0.5 M $MgCl_2 \cdot 6H_2O$	20 ml	10 mM
Distilled water	780 ml	
Total volume	1 L	

2.2.3. Phage Buffer

Table 2.4. Phage buffer (as described in Ogilvie et al., 2012)

Ingredients	Amounts	Final concentration
Na_2HPO_4 anhydrous (Sigma-Aldrich, UK)	1.384 g	19.5 mM
KH_2PO_4 anhydrous (Sigma-Aldrich, UK)	1.497 g	22 mM
NaCl	2.498 g	85.5 mM
$MgSO_4 \cdot 7H_2O$ (Sigma-Aldrich, UK)	0.123 g	1 mM
$CaCl_2 \cdot 2H_2O$ (Sigma-Aldrich, UK)	0.007 g	0.1 mM
Distilled water	Up to 500 ml	

2.2.4. SDS Lysis Buffer

Table 2.5. Stock solutions for SDS lysis buffer

Ingredients	Formula
1 M EDTA	14.61 g of EDTA anhydrous (Sigma-Aldrich, UK) dissolved in 50 ml of distilled water by adding a few NaOH tablets to alter pH, so EDTA dissolved into solution and pH reached ~8.0
20% (w/v) SDS	5 g of SDS (Sigma-Aldrich, UK) dissolved in 25 ml of distilled water

Table 2.6. SDS lysis buffer

Ingredients	Amounts	Final concentration
1 M Tris (pH 8.0)	10 ml	400 mM
1 M EDTA (pH 8.0)	2.5 ml	100 mM
20% (w/v) SDS	12.5 ml	10% (w/v)
Total volume	25 ml	

→ SDS lysis buffer was heated at 55°C for 10 minutes to dissolve all constituents prior to sterilisation.

2.2.5. GTC Buffer and Phage Disruption Buffer

Table 2.7. Stock solutions for GTC buffer

Ingredients	Formula
0.5 M Sodium citrate	7.35 g of Sodium citrate dihydrate (Sigma-Aldrich, UK) dissolved in 50 ml of distilled water and adjusted pH to 7.0
10% (w/v) Sarcoryl	1 g of N-lauroylsarcosine sodium salt (Sigma-Aldrich, UK) dissolved in 10 ml of distilled water

Table 2.8. GTC buffer

Ingredients	Amounts	Final concentration
6 M Guanidine thiocyanate (Sigma-Aldrich, UK)	22.5 ml	~4 M
0.5 M Sodium citrate (pH 7.0)	1.76ml	26.11 mM
10% (w/v) Sarcoryl	2.64 ml	0.78% (w/v)
Distilled water	6.8 ml	
Total volume	33.7 ml	

Table 2.9. Phage disruption buffer

Ingredients	Amounts
GTC buffer	1 ml
2-mercaptoethanol (Sigma-Aldrich, UK)	7.2 µl

2.2.6. Bacteroides Phage Recovery Medium (BPRM)

Table 2.10. Stock solutions for BPRM

Ingredients	Formula
0.45 M $CaCl_2$	5 g of $CaCl_2$ (Sigma-Aldrich, UK) dissolved in 100 ml of distilled water
1% (w/v) Hemin solution	0.1 g of Hemin (Sigma-Aldrich, UK) dissolved in 0.5 ml of 1 M NaOH and added 99.5 ml of distilled water
10.6% (w/v) Na_2CO_3	10.6 g of Na_2CO_3 (Sigma-Aldrich, UK) dissolved in 100 ml of distilled water

Table 2.11. BPRM agar plate (bottom layer; pH 7.0)

Ingredients	Formula
BPRM powder	29.4 g of BPRM powder (Formedium Ltd, UK) dissolved in 1 L of distilled water
Agar (1.5%, w/v)	15 g of agar (Oxoid, UK) dissolved in 1 L of distilled water
0.45 M $CaCl_2$ solution	1 ml
Distilled water	Up to 1 L

→ Once autoclaved, added 10 ml of 1% (w/v) hemin solution and 25 ml of 10.6% (w/v) Na_2CO_3 solution per litre of media

Table 2.12. BPRM semi-solid overlays (top layer; pH 7.0)

Ingredients	Formula
BPRM powder	5.88 g of BPRM powder dissolved in 200 ml of distilled water
Agar (0.35%, w/v)	0.7 g of agar dissolved in 200 ml of distilled water
0.45 M $CaCl_2$ solution	1 ml
Distilled water	Up to 200 ml

→ Once autoclaved, added 2 ml of 1% (w/v) hemin solution and 2.5 ml of 10.6% (w/v) Na_2CO_3 solution per litre of media

Table 2.13. BPRM broth (pH 7.0; based on ISO 10705-4:2001)

Ingredients	Formula
BPRM powder	29.4 g of BPRM powder dissolved in 1 L of distilled water
0.45 M $CaCl_2$ solution	1 ml
Distilled water	Up to 1 L

→ Once autoclaved, added 10 ml of 1% (w/v) hemin solution and 25 ml of 10.6% (w/v) Na_2CO_3 solution per litre of media

2.2.7. Preparation of Phage Stock

The bacteriophage ΦB124-14 (Ogilvie et al., 2012) was obtained using the host strain *Bacteroides fragilis* GB-124, which was first isolated from untreated municipal sewage from a treatment plant located in East Sussex (SE England, UK) (Ebdon et al., 2007). ΦB124-14 was provided by Dr James Ebdon (School of Environment and Technology, University of Brighton). Briefly, single ΦB124-14 plaque was picked from a BPRM agar plate and suspended in 200 µl of phage buffer, and then incubated at 4°C for 16-24 hours. The phage suspension was purified three times using plaque assay to generate fresh plaques. Once purified, 5-8 ml of phage buffer was added into the final agar plates left at 20°C for 1 hour, followed by harvesting liquid and top semi-solid agar layer into a 50-mL centrifuge tube (Corning, UK), mixing briefly and incubating at 20°C (room temperature) for 30 minutes. The mixtures were centrifuged at 3,000 x g for 20 minutes at 20°C to remove bacterial debris and agar, followed by passing the supernatants through 0.22 µm PES syringe filters and stored at 4°C (Purnell et al., 2015). The titre of phage suspension was determined using plaque assays every 3 months (approximately 1 x 10^9 to 1 x 10^{10} pfu/ml for use).

2.2.8. Bacterial Strain

Bacteroides fragilis (Bf) GB-124 was used as a host for the phage, ΦB124-14, and was provided by Dr James Ebdon. In brief, frozen stock was placed in sterile BPRM broth and incubated under anaerobic condition (5% CO_2, 5% H_2 and 90% nitrogen at ~25 psi pressure) at 37°C for 16-24 hours. 1 ml of overnight inoculum was subcultured into fresh BPRM broth under anaerobic condition at 37°C and OD_{620nm} was checked every 30 minutes. Once the absorbance reached 0.3-0.33 corresponding to a cell density of approximately 2 x 10^8 cfp (colony-forming particles) per ml, it was used for phage spiking-and-recovery assays.

2.2.9. Reagents and Chemicals

Table 2.14. Reagents used

Reagent	Suppliers
SYBR™ Gold stock solution (10,000× concentrated in DMSO)	Thermo Fisher Scientific, UK
DNase I (1 U/µl)	Promega, UK
RNase A (10 mg/ml)	Thermo Fisher Scientific, UK
Proteinase K (20 mg/ml)	Life technologies, UK

Lysozyme (10 mg/ml): 10 mg dissolved in 1 ml of 10 mM Tris/1 mM EDTA solution (pH 8.0)	Sigma-Aldrich, UK
Invitrogen™ Qubit BR reagents: for sample concentration from 100 pg/µl to 1,000 ng/µl	Thermo Fisher Scientific, UK
Invitrogen™ Qubit 1X dsDNA HS assay kit: for sample concentration from 10 pg/µl to 100 ng/µl	Thermo Fisher Scientific, UK
TURBO™ DNase (2 U/µl)	Thermo Fisher Scientific, UK
Ambion™ RNase I (100 U/µl)	Thermo Fisher Scientific, UK
10X TURBO™ DNase buffer	Thermo Fisher Scientific, UK
Ambion™ nuclease-free water	Thermo Fisher Scientific, UK
KAPA pure beads	Roche, South Africa
Invitrogen™ Quant-iT dsDNA assay kit	Thermo Fisher Scientific, UK
Fluoromount-G® antifade mounting reagent	SouthernBiotech, US

Table 2.15. Chemicals used

Chemical	Suppliers
PEG 8,000 (polyethylene glycol)	Sigma-Aldrich, UK
Chloroform	Sigma-Aldrich, UK
UltraPure™ phenol/chloroform/ isoamyl alcohol (25:24:1, v/v)	Thermo Fisher Scientific, UK
20 mM EGTA solution	Promega, UK
3 M Sodium acetate solution (pH 5.2)	Merck, UK
Absolute ethanol (EtOH)	VWR, UK
0.5% (w/v) Uranyl acetate (UA)	BDH/VWR, UK
2.5% (v/v) Glutaraldehyde	Agar scientific, UK

2.2.10. Commercial Kits

Table 2.16. Nucleic acids purification/isolation kits

Kit	Suppliers
Zymo Research genomic DNA Clean & Concentrator™-25	Cambridge Bioscience, UK
PowerViral® environmental RNA/DNA isolation kit	MO BIO, UK
ZR viral DNA/RNA kit	Zymo Research, UK
Phage DNA Isolation kit	Norgen, UK

Table 2.17. Commercial kits for sequencing library preparation

Kit	Suppliers
NEBNext® Ultra™ II DNA library prep kit	New England Biolabs Ltd., US

2.3. General Methods

2.3.1. Faecal VLP Isolation: Route B (see Figure 3.1)

To develop and optimise protocols for faecal VLP and VLP DNA isolation, ~5 g of faeces was initially used, as previously described (Thurber et al., 2009). After optimising stool sample size (see **Section 2.3.6**), each 3-4 g faecal aliquot was homogenised in 10× volumes (w/v) of sterile TBT buffer in a pre-weighed 50-mL centrifuge tube and then kept on ice for 1 hour. Faecal homogenates were then centrifuged at 11,200 x g for 30 minutes at 10°C and supernatants transferred to sterile centrifuge tubes, followed by a centrifugation again under the same conditions. Supernatants were first passed through 0.8 µm PES (polyethersulfone) syringe filters (Sterlitech, US), followed by 0.45 µm PES syringe filters (Starlab, UK).

2.3.2. PEG Enrichment

To enrich VLPs from faecal filtrates (FF), NaCl (final concentration 6%, w/v) was added to faecal filtrates and mixed gently until dissolved completely, followed by addition of PEG 8,000 (final concentration 10%, w/v). The samples were left at 4°C for at least 16 hours. PEG-precipitated VLPs were then harvested by centrifugation at 4,500 x g for 1 hour at 4°C. The phage-containing pellets were resuspended in 500 µl of TBT buffer and then used immediately or stored at 4°C for use within 24 hours.

2.3.3. Efficiency of VLP Isolation

2.3.3.1. Spiking-and-Recovery Assay (a) – by Plaque Assay

Five grams of faecal samples collected from healthy donors was spiked with 1 ml of the phage ΦB124-14 of known titre after homogenisation. Phage titres were determined in the following samples: 100 µl of phage stock, the raw faecal homogenate, the spiked supernatant collected after centrifugation, the faecal filtrate collected from 0.8 µm filtration, and PEG-VLP suspensions (**Figure 2.1**). Phage recovery was also determined for the dual filtration (0.8 µm combined 0.45 µm filter) protocol using plaque assays (**Figure 2.2**).

Phage plaque forming assay was carried out using 100 µl of 10-fold dilution series of spiked samples from 10^{-1} to 10^{-9} mixed with 200 µl of the host GB-124 in 3 ml of melted overlays and then poured into agar plates. Once solidified, these agar plates were incubated at 37°C for 16-24 hours under anaerobic condition, followed by plaques counted and plaque forming units (PFU) per ml determined.

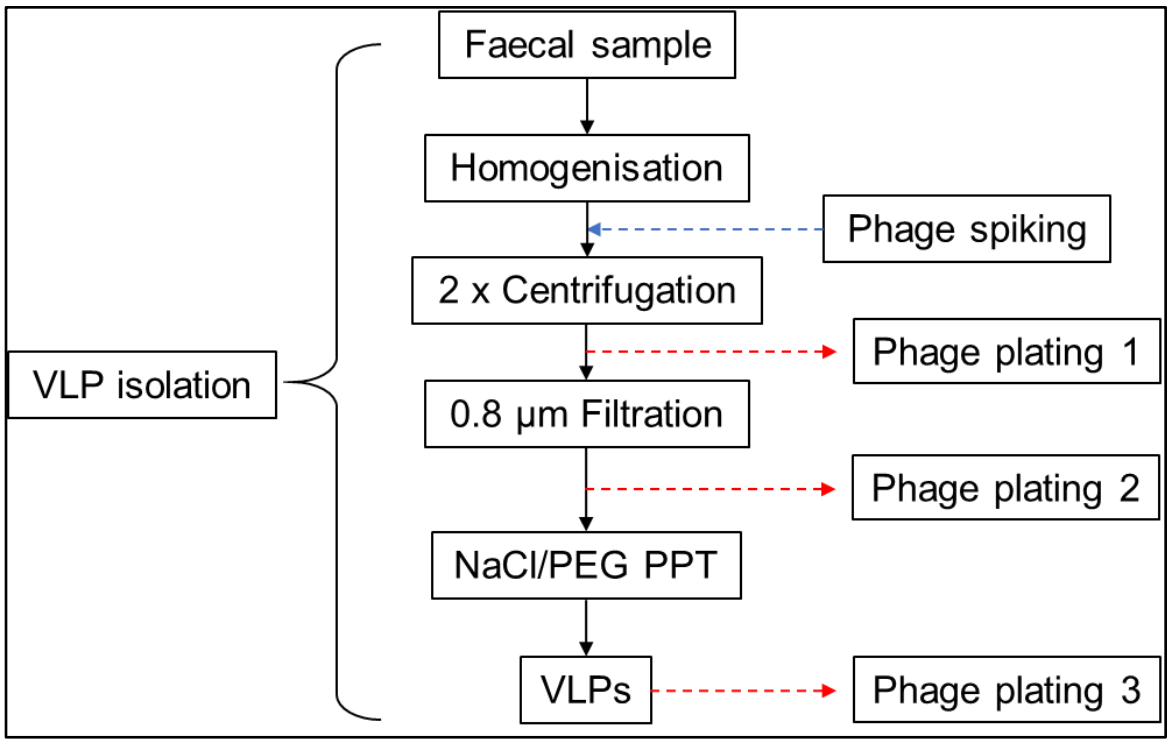


Figure 2.1. Workflow of phage spiking procedure for determining the efficiency of 0.8-µm filtration. 1 ml of the phage stock (~1.2 x 10^{10} pfu/ml) was added to faecal homogenate with aliquots collected after centrifugation, 0.8 µm filtration and PEG precipitation (PPT) for plaque assay (Route C, see **Figure 3.1**).

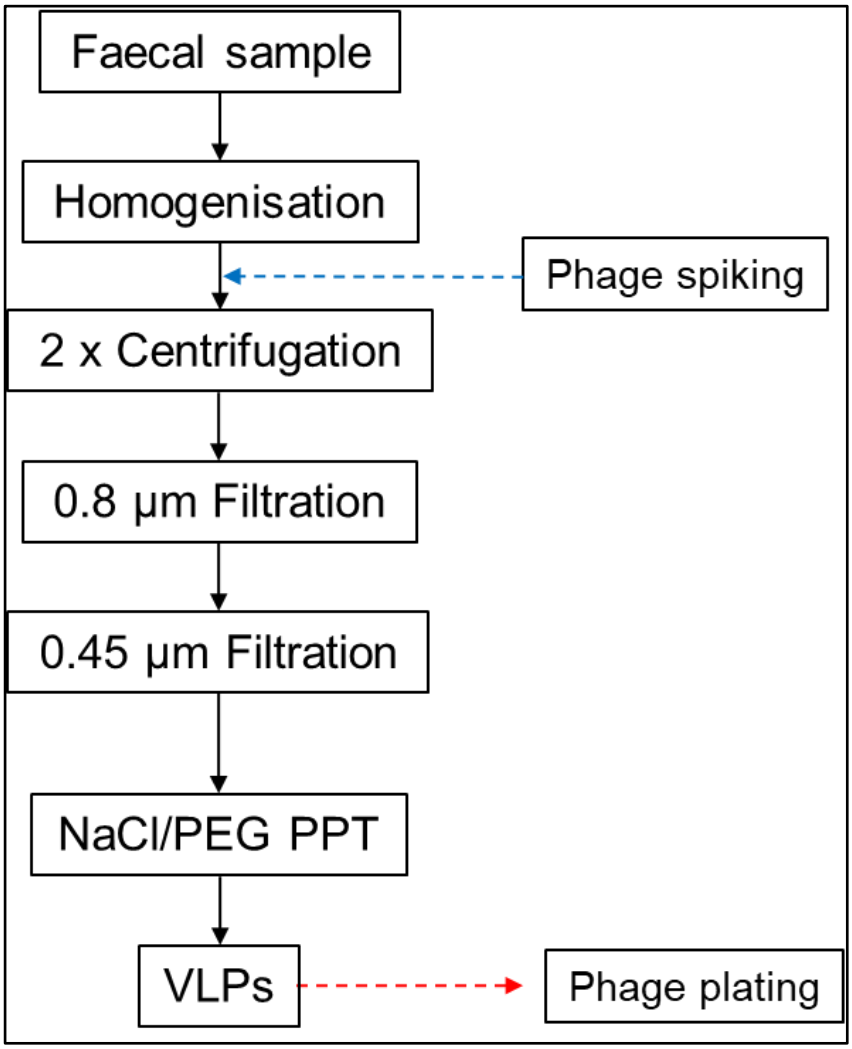


Figure 2.2. Workflow of phage spiking procedure for determining the efficiency of dual filtration (0.8- and 0.45-µm). 1 ml of the phage stock (~8.7 x 10^9 pfu/ml) was added to 3 independent faecal homogenates, followed by centrifugation, dual filtration and PEG precipitation (PPT) with PEG-VLP aliquots collected for plaque assay (Route B, see **Figure 3.1**).

2.3.3.2. Spiking-and-Recovery Assay (b) – by Epifluorescence Microscopy (EFM)

Twenty grams of faeces from a healthy donor was spiked with the reference phage ΦB124-14 of known titre after homogenisation (**Figure 2.3**). First, 900 µl of viral stock solution was incubated with 100 µl of SYBR Gold secondary stock (0.25%, v/v) in the dark at 4°C for 24 hours. Faecal homogenates were then spiked with the SYBR Gold-labelled phage and VLPs were isolated. After PEG enrichment, the phage-containing pellets were resuspended in 2 ml of TBT buffer. 20 µl of PEG-VLP suspensions was diluted to 1:50 (v/v) in nuclease-free water and then fixed onto an Anodisc membrane by vacuum filtration, and were then visualised and enumerated by EFM. The average number of VLPs per field was multiplied by sample dilution factor and microscope conversion factor (area of 13-mm Anodisc filter / area of FOV), and divided by the sample volume (Budinoff et al., 2011).

$$\frac{\text{Average number of VLPs/field} \times \text{dilution factor} \times \text{microscope conversion factor}}{\text{Volume of sample (ml)}}$$

1. The diameter of field of view (FOV) of 100X oil immersion objective (at 1,000X magnification) is 0.178 mm, so the area = $(0.178/2)^2 \times \pi = 0.025$ mm^2

2. The diameter of the Anodisc filter is 13 mm, making the area = $(13/2)^2 \times \pi = 132.73$ mm^2. Therefore, the microscope conversion factor = area of 13-mm Anodisc filter / area of FOV = 5309.3

3. Converting VLP/ml to VLP/g faeces: VLP/ml x 25 (20 µl VLPs out of 500 µl PEG-suspensions) and divided by 4 (g of faeces)

Details of SYBR Gold staining, filtration system, EFM observation, the methods of image post-analysis and manual counts with ImageJ were described in **Section 2.3.8**.

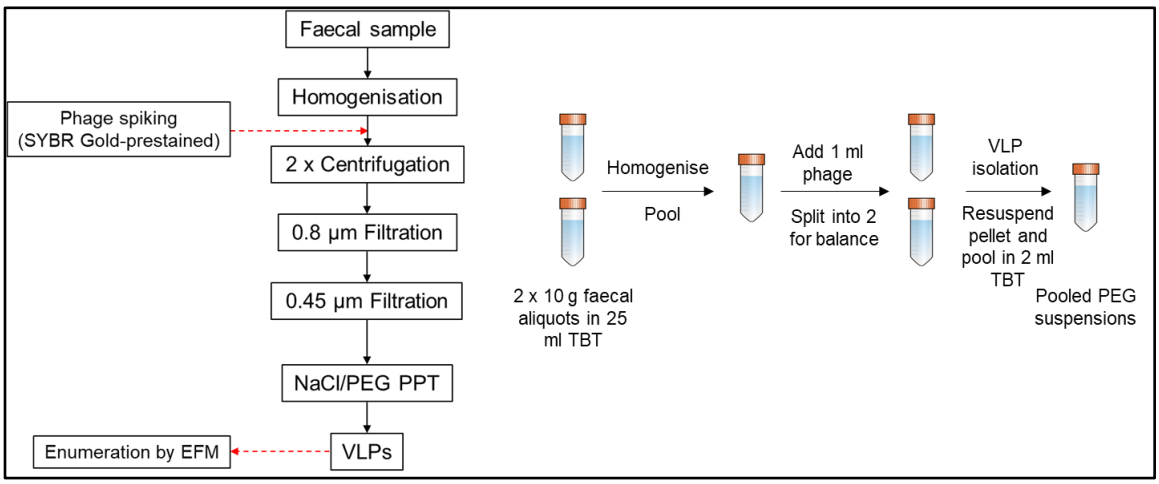


Figure 2.3. Workflow of phage spiking procedure for EFM analysis. The SYBR Gold-prestained phage suspension was added to faecal homogenates, followed by two-round centrifugation, dual filtration, and collecting the PEG-VLP suspensions after PEG precipitation for observation by EFM.

2.3.4. Procedures to Remove Contaminants

(1) Chloroform and Ultrafiltration Treatment: Route E and G (see Figure 3.1)

VLPs were isolated from 4-5 g of faeces and the virus-containing PEG-suspended pellets were resuspended in 1 ml of TBT buffer for chloroform treatment and in 2 ml of TBT for ultrafiltration. VLP-suspensions were centrifuged at 1,000 x g for 5 minutes at 20°C, followed by chloroform treatment or centrifugal ultrafiltration using Vivaspin®-20 polyethersulfone 300K (Sartorius, UK) and 1,000K molecular weight cut-off (MWCO) filters (Sartorius, UK). For evaluation of chloroform treatment, VLP-suspension was treated with an equal volume of chloroform and then centrifuged at 12,000 x g for 5 minutes at 20°C and then repeated. For ultrafiltration, 300K and 1,000K MWCO filters were selected, based on the size of the most phages. VLP-suspensions were added into the top of filter columns and then centrifuged at 6,000 x g for 1-1.5 hours at 10°C until the volumes of VLP-suspensions were reduced to 400-500 µl. The upper reservoirs were removed and the bottom of upper reservoir was sealed with parafilm. 100-200 µl of TBT buffer was added to wash the inner surface of the filters, followed by vortexing at 1,500 rpm for 20-30 seconds. Concentrated faecal VLP-suspensions (500-600 µl) were then recovered and centrifuged at 1,000 x g for 1 minute at 20°C to remove debris.

VLP samples were then treated with 1 U of DNase I and 10 µg/ml of RNase A at 37°C for 1 hour, followed by addition of 1 µl of 20 mM EGTA to stop the reaction and heated at 70°C for 10 minutes. 0.5% (w/v) of SDS and 80 µg of Proteinase K (Ambion/Thermo Fisher Scientific, UK; adapted from Shkoporov et al., 2018b) were then added and incubated at 56°C for 20 minutes to inactivate nucleases and disrupt viral capsids. VLP samples were subsequently treated with an equal volume of GTC/2-ME solution and incubated at 20°C for 10 minutes prior to phenol/chloroform extraction and DNA precipitation. The workflow is described below (**Figure 2.4**).

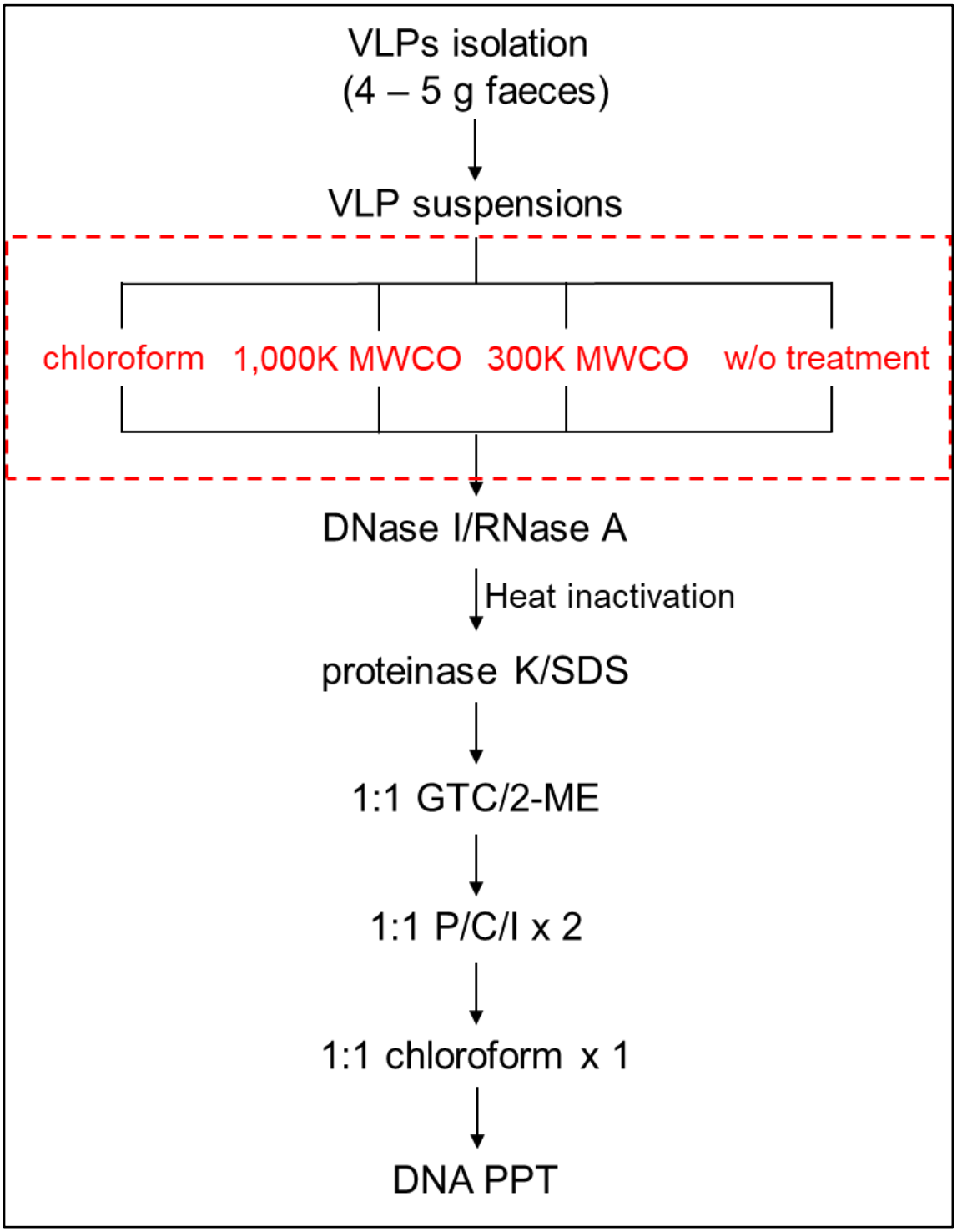


Figure 2.4. Overview of experimental design for evaluation of chloroform treatment and ultrafiltration. VLPs were isolated from 4-5 g faecal aliquots and were enriched by PEG. PEG-VLP suspensions were then treated with chloroform and were concentrated by 300K and 1,000K MWCO filters, followed by the addition of nucleases, SDS/proteinase K, GTC/2-ME prior to phenol/chloroform extraction and ethanol precipitation (PPT). An aliquot of VLP-suspension prior to treatment acted as process control.

(2) GTC/2-ME Treatment

To investigate if guanidinium thiocyanate and 2-mercaptoethanol (GTC/2-ME) treatment is detrimental to viral genomic DNA integrity, bacterial genomic DNA isolated from *Bacteroides thetaiotaomicron* (VPI-5482; DNA concentration ~608 ng/µl) was used. Briefly, 100 µl of bacterial genomic DNA (~6 µg) and 150 µl of nuclease-free water were mixed with an equal volume (250 µl) of GTC/2-ME mixture and incubated at 20°C for 10 minutes. Afterwards, DNA aliquots were purified using Zymo Research genomic DNA Clean & Concentrator^{TM}-25 columns and then eluted in 65 µl of nuclease-free water. Spectrophotometer and gel electrophoresis were used to determine DNA quality and quantity.

2.3.5. Optimising Faecal VLP DNA Extraction

To optimise the viral DNA extraction protocol and obtain the highest quality DNA from VLP samples using 0.8 µm filtration-based isolation protocol, six methods (see **Figure 3.1**) were compared side by side: phenol/chloroform extraction (Sambrook and Russell, 2001), GTC-based phenol/chloroform extraction (Murphy et al., 2013), SDS-based phenol/chloroform extraction (Miller et al., 1999) and three commercial DNA/RNA extraction kits. DNA quantity and quality were determined using Nanodrop (Thermo Scientific), Qubit (Invitrogen) and gel electrophoresis.

(1) P/C/I Extraction (Sambrook and Russell, 2001): Route D

An equal volume of phenol/chloroform/isoamyl alcohol (P/C/I, 25:24:1, v/v/v) was mixed with a lysozyme/nuclease-treated VLP sample in 2 ml of Quantabio phase lock gel tube (VWR, UK) and centrifugated at 15,000 x g for 5 minutes at 20°C, which was repeated once to ensure all proteins were degraded. Chloroform extraction was then performed prior to precipitating viral DNA using 1/10 volume of 3 M sodium acetate (pH 5.2) and 2.5 volumes of ice-cold absolute ethanol, leaving on dry ice for 30 minutes, followed by centrifugation at 15,800 x g for 30 minutes at 20°C. The pellet was washed twice using 1 ml of 70% (v/v) ethanol in combination with centrifugation at 15,800 x g for 5 minutes at 20°C. After 15 minutes air-drying, viral DNA was resuspended in an appropriate volume (50-100 µl) of TE buffer (pH 8.0), and then stored at 4°C for later use or at -70°C for long-term storage.

(2) GTC-Based P/C/I Extraction (Murphy et al., 2013): Route E and G

An equal volume of phage disruption buffer, prepared by adding 7.2 µl of 2-mercaptoethanol to 1 ml of GTC stock solution, was mixed with chloroform- or 300/1,000K MWCO-treated VLPs after nucleases and SDS/proteinase K treatment, then an equal volume of P/C/I was added in a 2 ml of phase lock gel tube and centrifugated at 15,000 x g for 5 minutes at 20°C, followed by chloroform extraction and DNA precipitation. DNA was eluted in 50-100 µl of TE buffer (pH 8.0).

(3) SDS-Based P/C/I Extraction (Miller et al., 1999): Route H

Two 5 g of faecal aliquots were used to isolate VLPs. After PEG enrichment, PEG pellets were sequentially resuspended and pooled in 500 µl of TBT buffer. The pooled and concentrated VLPs were then treated with 4 U of DNase I and 2 µg of RNase A in 1X nuclease buffer at 37°C for 1 hour, followed by adding 1 µl of 20 mM EGTA to stop reaction and heat inactivation at 70°C for 10 minutes. 300 µl of 10% (w/v) SDS lysis buffer was

added to the VLP-suspension and then incubated at 20°C for 10 minutes, followed by addition of 300 µl of P/C/I and mixed gently [(SDS+P/C/I):VLP = 1:1], and then centrifuged at 8,000 x g for 5 minutes at 20°C. The resulting upper aqueous phase was transferred to a fresh 2-mL centrifugal tube and repeated again by treating an equal volume of P/C/I. The resulting aqueous phase was then subjected to final round of purification using ZR genomic DNA Clean & Concentrator™-25 following manufacturer's instruction with a final elution volume of 60 µl (**Figure 2.5**).

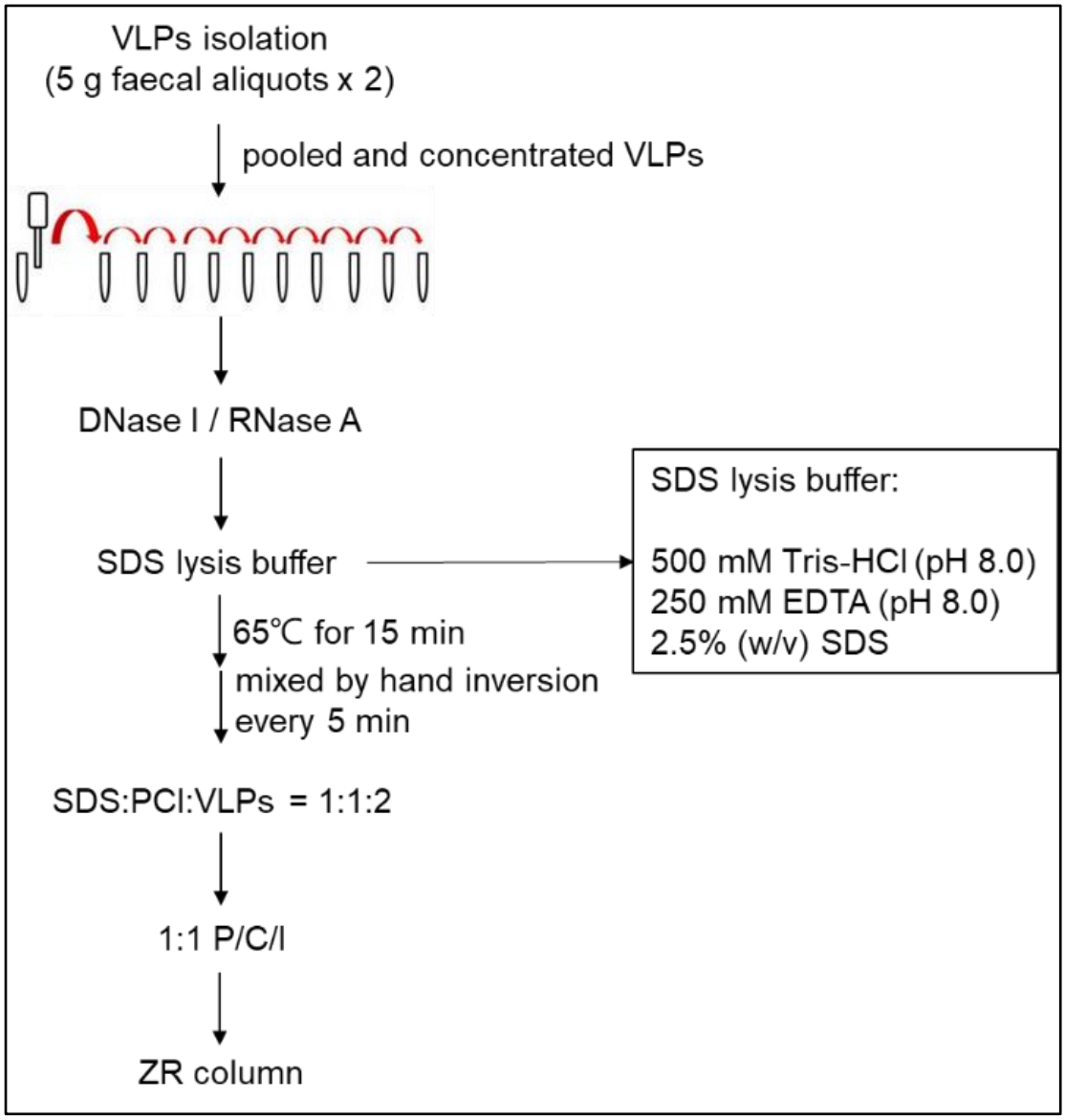


Figure 2.5. Overview of experimental design for SDS-based viral DNA extraction. VLPs were isolated from ~5 g faecal aliquots. After PEG enrichment, PEG pellets were sequentially resuspended and pooled in ~500 µl of TBT buffer. Concentrated VLP-suspension (~500 µl) was then treated with DNase I and RNase A, followed by disrupting viral capsids and decontaminating proteins/enzymes using SDS lysis buffer with P/C/I. The resulting aqueous phase was then subjected to final round of purification using ZR genomic DNA Clean & Concentrator™-25 kit.

(4) PowerViral® Environmental RNA/DNA Isolation Kit (MO BIO): Route I

VLPs isolated from 5 g of faeces and the resulting VLP-pellet was resuspended in 500 µl of TBT buffer after PEG enrichment. Approximately 600 µl of nuclease-treated VLP samples were mixed with 300 µl of 10% (w/v) SDS lysis buffer and 300 µl of saturated phenol solution (pH 8.0), and then incubated at 20°C for 10minutes. The mixture was centrifuged at 8,000 x g for 5 minutes at 20°C, and then the resulting aqueous phase (top) was subjected to MO BIO environmental RNA/DNA isolation kit according to manufacturer's instruction with a final elution volume of 50 µl (**Figure 2.6**).

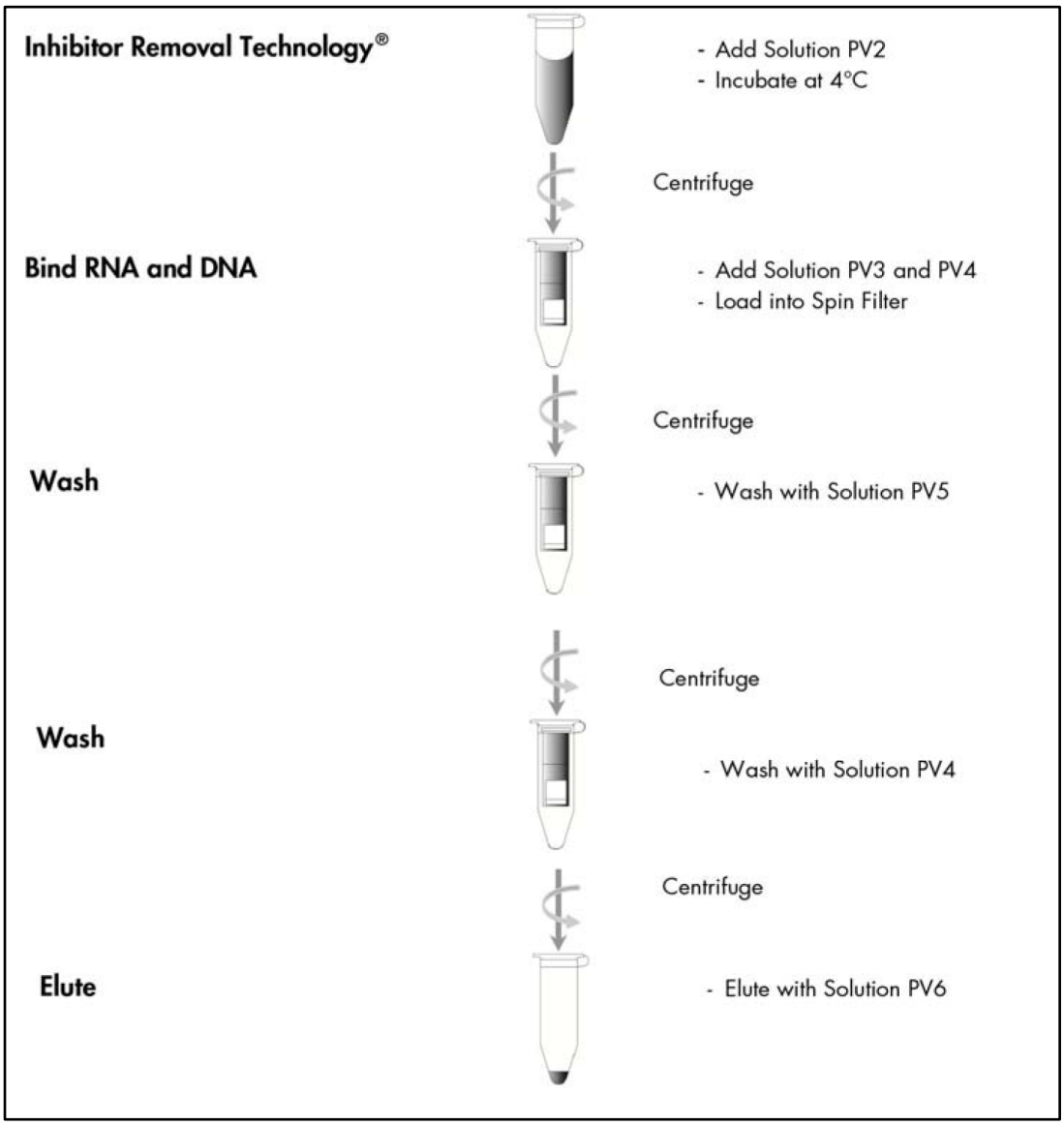


Figure 2.6. Workflow for viral DNA isolation using MO-BIO PowerViral environmental RNA/DNA isolation kit. Image was taken from manufacturer's protocol (https://www.qiagen.com/us/resources/resourcedetail?id=404e2342-cf6c-4709-b127-fb22f088f296&lang=en, available in December 2020).

(5) ZR Viral DNA/RNA Kit (ZYMO RESEARCH): Route J

VLPs isolated from 5 g of faeces and the resulting VLP-pellet was resuspended in 500 µl of TBT buffer after PEG enrichment. Approximately 600 µl of nuclease-treated VLP samples were then subjected to ZR viral DNA/RNA kit according to manufacturer's instruction with a final elution volume of 50 µl.

(6) Phage DNA Isolation Kit (Norgen): Route K

VLPs isolated from 5 g of faeces and the resulting VLP-pellet was resuspended in 500 µl of TBT buffer after PEG enrichment. Approximately 600 µl of nuclease-treated VLP samples were then subjected to Norgen phage DNA isolation kit according to manufacturer's instruction with a final elution volume of 75 µl (**Figure 2.7**).

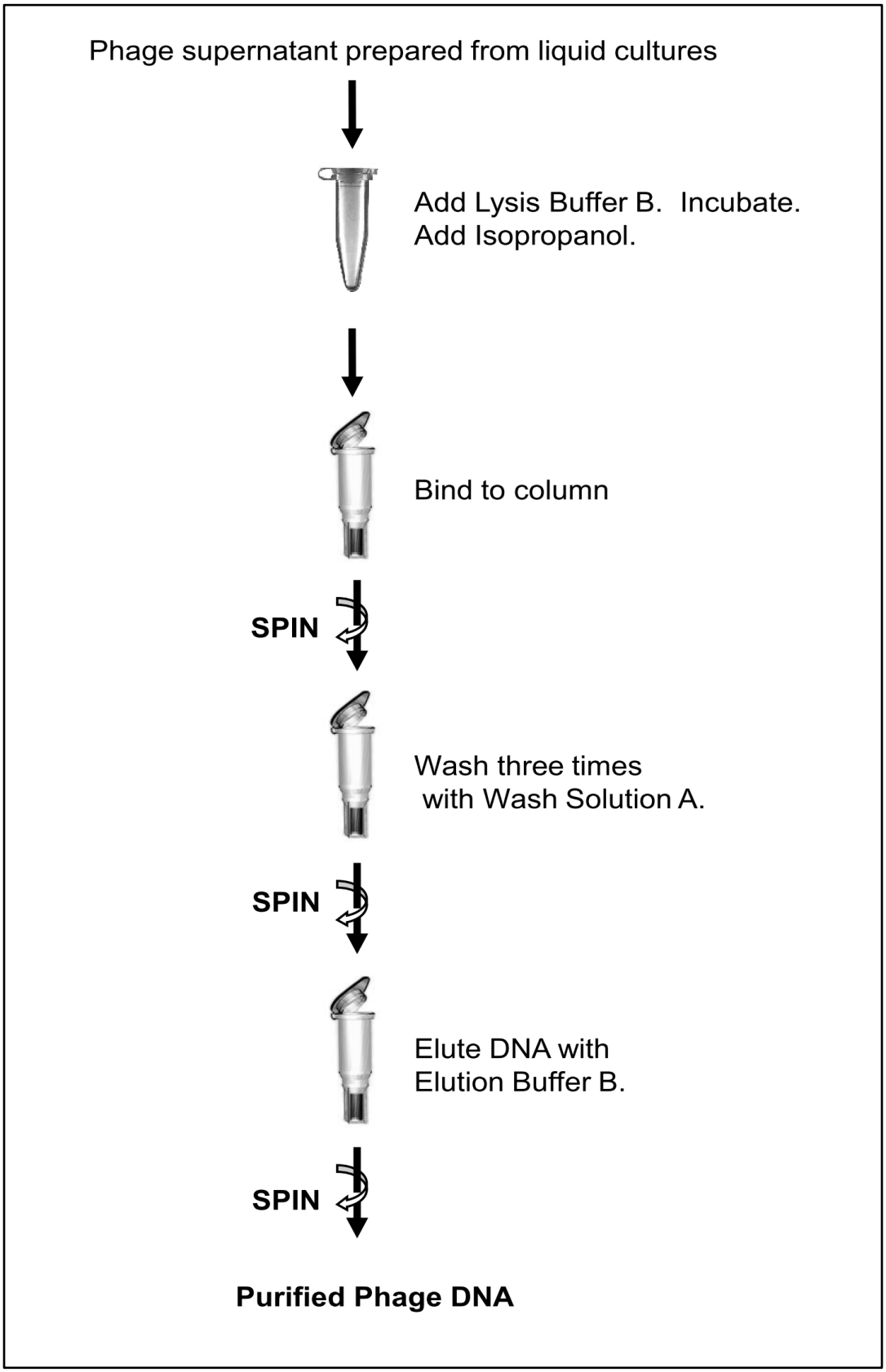


Figure 2.7. Workflow for viral DNA isolation using Norgen phage DNA isolation kit. Image was taken from manufacturer's protocol (https://norgenbiotek.com/sites/default/files/resources/Phage-DNA-Kit-Insert-PI46800-5-M14.pdf, available in December 2020).

2.3.6. Optimising Faecal Sample Size

To determine optimal faecal sample size, 0.5, 1.5, 3.0, 5.0, 10.0 and 15.0 g of faeces were used to isolate VLPs and extract DNA, visualised by gel electrophoresis.

2.3.7. Quantity and Quality of Faecal VLP DNA

To determine the quantity and quality of extracted faecal VLP DNA, Nanodrop and Qubit were used. The absorbance-based spectrophotometer Nanodrop measures the quantity and quality of extracted viral DNA. An absorbance ratio 260/280 between around 1.8 and 2.0 indicates that extracted viral DNA is “pure” without protein contamination in combination with an absorbance ratio 260/230 above 2.0 representing high purity DNA without salt, nucleotide, carbohydrate, and/or phenol contamination. The fluorescence-based Qubit™ system determines amounts of viral DNA using BR reagents for sample concentration from 100 pg/µl to 1,000 ng/µl, or 1X dsDNA HS Assay Kit for sample concentration from 10 pg/µl to 100 ng/µl. For the preparation of non-PCR sequencing libraries, the sequencing service (Novogene Ltd., Hong Kong) required a minimum of 1.5 µg DNA and a minimal concentration of 30 ng/µl, and sample volume of 50 µl; for the preparation of PCR-based libraries, at least 0.8 µg of DNA with a minimal concentration of 30 ng/µl and volume of 50 µl were required.

2.3.8. Enumerating Faecal VLPs in ME/CFS Patients and Same Household Healthy Controls

2.3.8.1. SYBR Gold Staining

To estimate the number of faecal VLPs from PEG-enriched VLP suspensions, a combination of the methods described previously (Hoyles et al., 2014, Budinoff et al., 2011, Thurber et al., 2009, Chen et al., 2001) was used for SYBR Gold staining and EFM imaging. Briefly, the primary SYBR™ Gold stock solution (concentration 10,000X) was diluted to a secondary stock solution (concentration 25X or 0.25%, v/v) with sterile Ambion™ nuclease-free water. SYBR Gold solution was defrosted at 20°C in the dark for around 15 minutes prior to use. 20 µl of PEG-enriched VLP suspensions were diluted in 900 µl of nuclease-free water. 900 µl of diluted VLP suspensions were stained with 100 µl of SYBR Gold secondary stock solution (final concentration 0.025%, v/v) and mixed gently, then incubated in the dark for 15 minutes at 20°C.

2.3.8.2. Vacuum-Based Filtration and EFM Analysis

An Omnipore 0.45-µm, 13-mm PTFE backing filter (Millipore/Sigma-Aldrich, UK) was placed onto the top of Swinnex filter holder (Millipore/Sigma-Aldrich, UK) and silicone gasket (Millipore/Sigma-Aldrich, United Kingdom) and was rinsed with nuclease-free water, followed by gently placing a 0.02-µm white Whatman™ Anodisc 13-mm filter membrane (Sigma-Aldrich, UK) on the top of the backing filter. The Anodisc filter was rinsed by nuclease-free water under a low-vacuum pressure (2-4 psi or ~20 kPa) (Budinoff et al., 2011, Hoyles et al., 2014). The diluted, stained VLP sample was then placed on the filter. The Millivac-Maxi vacuum pump (Millipore/Sigma-Aldrich Ltd., UK) was continued for additional 30 seconds once all liquids had passed through the filter to ensure all VLPs were fixed onto the filter. With the vacuum still on, 1 ml of nuclease-free water was added and passed through the membrane, followed by addition of 30 seconds to remove excessive dye. The filter was gently removed from the Swinnex outlet using forceps and transferred to a Whatman® filter paper (Sigma-Aldrich, United Kingdom) and then left to dry on the filter paper for 1 minute. The vacuum-based filtration device is showed below (**Figure 2.8**).

A drop (~20 µl) of Fluoromount-G® antifade mounting reagent (SouthernBiotech, US) was spotted on a microscope slide and the dried, stained Anodisc filter was then placed on the mountant droplet. A drop of antifade mounting reagent was added to the filter surface prior to covering with a coverslip. The slide was then left at 20°C in the dark for 16-20 hours to allow the mountant to set. All slides were observed and imaged under a Zeiss Axio Imager

M2 widefield epifluorescence microscope with the Alexa Fluor 488 channel selected and the 100X high-resolution oil objective lens.

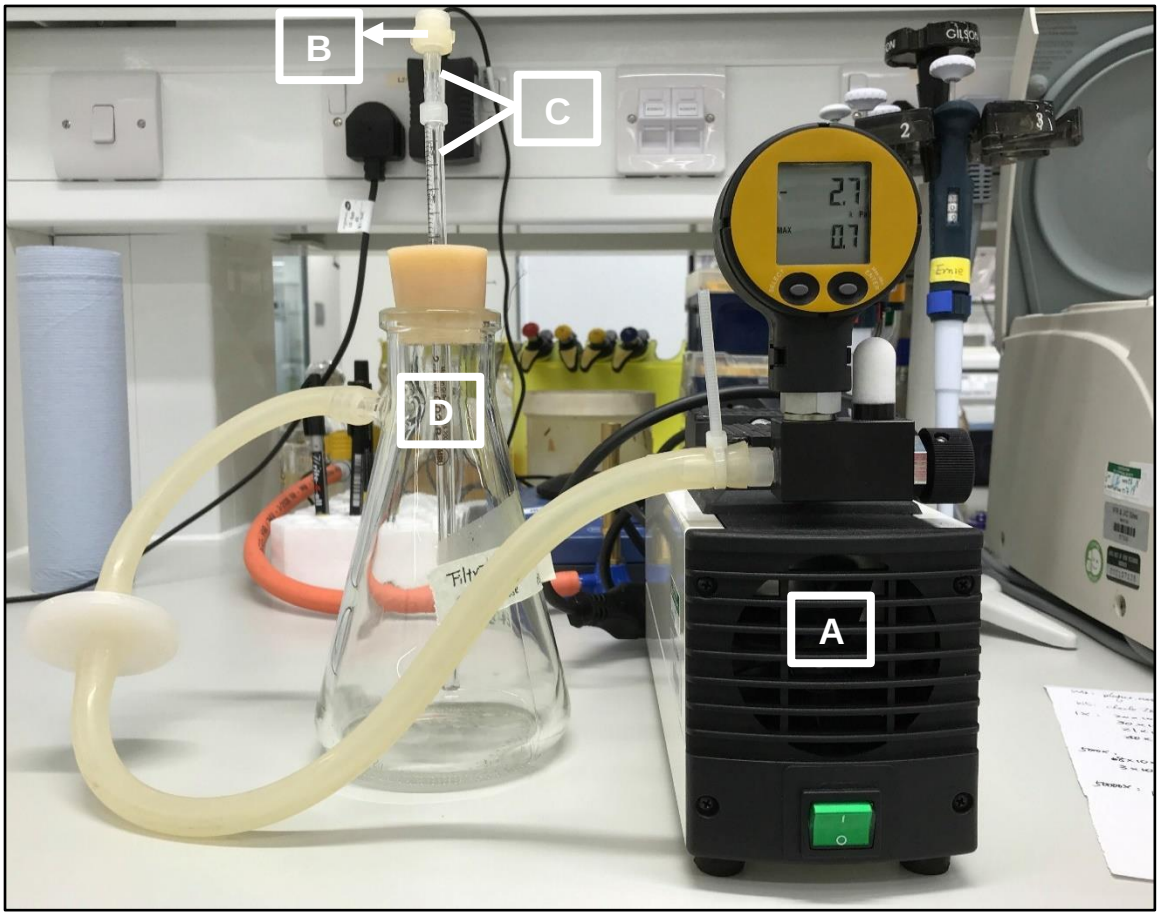


Figure 2.8. The vacuum-based filtration device used to fix viral particles onto the Anodisc 13-mm filter membrane. (A) A vacuum pump with low vacuum presure (2-4 psi or ~20 kPa). (B) An Anodisc 13-mm diameter filter membrane and a 0.45-µm, 13-mm diameter PTFE backing filter are put onto the Swinnex filter holder with a silicone gasket. (C and D) A filtration device including an adapter connecting a filter holder and a glass-graduated column and a filtering flask.

2.3.8.3. Post-Analysis of Images and Estimating VLP Counts

For each slide, a minimum of 20 images were captured by digital camera connected to microscope and saved as ".czi files" in the "Acquisition Mode", and then particle sizes were determined using ImageJ. Pure Bf phage (ΦB124-14) was used to decide the size of the observed particles in ImageJ (NIH Image, US). To assess the accuracy of the method of VLP enumeration using ImageJ, manual counting and running an automated programme in ImageJ were performed. For automated counting, the "particle size" (μm^2) was set between 0 and 0.20 and the "circularity" was set between 0 and 1 (i.e. 1 represents a perfect circle and 0 represents a straight line), based on the images of the pure phage (no bacterial cells were seen) in "Analyze Particles". The automated script was established in "Macro mode" as follows:

```
run("Subtract Background...", "rolling=50");
setAutoThreshold("Triangle dark");
//run("Threshold...");
run("Convert to Mask");
run("Analyze Particles...", "size=0.00-0.20 show=[Count Masks] summarize");
run("RGB Color");
```

For manual counting using ImageJ, each of the conditions/thresholds including "Subtract Background", "Threshold", and "Analyze Particles" was manually set and optimised, and VLPs were tracked and counted by human eyes, compared with the original raw images. To investigate faecal VLP counts recovered from patient and SHHC samples (see **Chapter 5**), the formula was used to estimate VLP counts described in **Section 2.3.3.2**.

2.3.9. Characterising Faecal VLPs by Transmission Electron Microscopy (TEM)

To determine phage and VLPs morphology, the training for TEM operation was received from Elaine Barclay (Cell and Developmental Biology, John Innes Centre, Norwich). 200 µl of faecal filtrates (FFs) were prepared for negative staining with uranyl acetate (UA) and TEM imaging in collaboration with QIB Microscopy Team (Core Science Resources). In brief, 5 µl of diluted FFs were applied to the carbon-film on copper 400 mesh grids (EM Resolutions, UK) for 1 minute and excess FFs removed by wicking the edge of the grid with Whatman filter paper, followed by 2 minutes incubation with 0.5% (w/v) UA solution. Excess UA was removed by wicking with filter paper and grids were allowed to dry thoroughly. Each of the grids was vapour fixed by adding 1 ml of 2.5% (v/v) glutaraldehyde to the dish containing the dried grids for a minimum of 2 hours. Imaging was then performed using a Talos F200C TEM microscope at 200 kV with a “Gatan One View” digital camera. The raw digital micrograph files (.DM4) were converted to .tiff file format using Digital micrograph software.

2.4. Bioinformatics: Viral Metagenomic Sequencing Analysis

2.4.1. Library Preparations and Shotgun Metagenomic Sequencing

Three DNA samples isolated from independent faecal samples were subjected to both shotgun metagenomic PCR-based sequencing library (LASL) and amplification-free library (NASL) preparations, respectively, with an average size of 350 bp generated by NEBNext® Ultra™ II DNA Library Prep Kit, following the standard manufacturer's instruction performed by the sequencing service, Novogene Ltd., Hong Kong. In brief, input viral genomic DNA samples were randomly fragmented, followed by end-repairing, 5'-phosphorylation and A-tailing at 3'-end, and ligating adapters. Adapter-ligated DNA was then size-selected, followed by i5/i7 index primer sets being used to barcode, enrich and amplify input DNA templates with PCR for PCR-based libraries and with omitting PCR steps for unamplified libraries. PCR products were then cleaned-up, followed by determining their quality and quantity using Agilent 2100 Bioanalyzer (Agilent Technologies, US) and real-time PCR.

In total, six virome-related libraries were then sequenced using 2 x 150 bp paired-end chemistry (PE150) on the Illumina HiSeq X Ten platform at Novogene Ltd., Hong Kong. The output of raw data had a Q30 score of >90% for the libraries constructed. Paired-end sequencing reads were provided as "FASTQ" format. Prior to sequencing, we designed this comparative study and made strategies for library preparations and sequencing in discussion with Novogene Ltd. Details of library preparations and sequencing were provided by the sequencing service.

2.4.2. Quality Controls for Raw Sequence Reads

All raw sequencing reads were pre-processed to trim and filter the reads with adapters, low quality (Q-value ≤38) and N nucleotides by Novogene Ltd. using readfq and fqcheck tools. Human DNA identified by Kraken 2 (v2.0.8) (Wood et al., 2019, Wood and Salzberg, 2014) against the Genome Reference Consortium Human Build 37 database (GRCh37/hg19) was removed using the confidence at 0.5 to reduce the false positives, processed by Dr Andrea Telatin. Sequence reads were further cleaned using fastp (v0.21.0) with a quality cut-off of 20 (Chen et al., 2018), prior to genome assembly. In parallel, both Kraken 2 and MetaPhlAn 3 (Beghini et al., 2020) were also used to taxonomically classify cleaned reads using default parameters (see **Supplementary Material S1:** https://www.dropbox.com/sh/5qyessiu7ezfvsw/AACuDdNaRfb8MbjYK369SXpsa?dl=0, available in December 2020). Moreover, VLP enrichment was evaluated using ViromeQC (v1.0) in default mode (Zolfo et al., 2019).

2.4.3. Genome Assembly by *De novo* Approach

MEGAHIT assembler (v1.2.9) (Li et al., 2015) was used to assemble cleaned reads into longer contigs with default parameters and was performed by Dr Andrea Telatin. QUAST (v5.0.2) was used to assess the quality of assembled genomic contigs using default parameters (see **Supplementary Material S2:** https://www.dropbox.com/sh/u7bozhrz79q0rd5/AABEdYGuz2lDfwZy1FbhjfLoa?dl=0**,** available in December 2020) (Gurevich et al., 2013, Mikheenko et al., 2018).

2.4.4. Identification of Viruses

VirSorter (v1.0.3) (Roux et al., 2015) and VirFinder (v1.1) (Ren et al., 2017) were used to identify potential viral contigs from the whole MEGAHIT-assembly pool, performed by Dr Andrea Telatin. In this study, those putative viral contigs sorted and classified in VirSorter categories 1 to 6 (including all completed viruses and prophages), and those piped through VirFinder under the appropriate sorting criteria (i.e. score ≥0.7 and p <0.05), were considered viral and were used for further investigation.

2.4.5. Assessing the Quality of Viral Genomes

CheckV (v0.7.0) was introduced to identify non-viral regions integrated in proviral genomes, estimate the completeness of viral genomes, predict closed genomes and assess the quality of viral genomic contigs, using end-to-end mode with default parameters (Nayfach et al., 2020). The output tables were produced in ".tsv" files (see **Supplementary Material S3**: https://www.dropbox.com/sh/9p5vfi8zm5yka7c/AACkZEPceAJhCuIuo95AAXaya?dl=0, available in December 2020)..

2.4.6. Read Mapping Against Reference Genomes

First, cleaned reads were individually mapped to total contigs of each corresponding library dataset as the reference genomes. In parallel, a pooled, non-redundant viral contig file across all datasets were generated using CD-HIT-EST (v4.8.1) with certain parameters used (i.e. -p 1 -g 1 -aS 0.9 -c 0.95 -M 0 -T 0) (Fu et al., 2012, Li and Godzik, 2006) and the viral reads detected by VirSorter and VirFinder were then mapped back to the pooled, non-redundant viral contigs as the viral reference genomes using BWA (v0.7.17) with default bwa-mem mode for paired-end manner (Li and Durbin, 2009), followed by using SAMtools (v1.10) (Li et al., 2009) to sort and index alignments and to convert the output to ".bam"

format. Both SAMtools and seqkit (v0.12) (Shen et al., 2016) were also used to calculate the numbers of mapped and unmapped reads.

2.4.7. Cluster Analysis and Taxonomic Annotation for the Viromes

To compare PCR and non-PCR library datasets, taxonomic classification to viruses and prophages identified by both VirSorter and VirFinder was conducted using vConTACT 2.0 and was performed by Dr Mohammad Adnan Tariq (Carding group, QIB), for hierarchically clustering and analysing the similarity of the UViG sequences (Jang et al., 2019). Briefly, prior to performing vConTACT 2.0, protein coding genes were initially predicted using Prodigal (v2.6.3) (Hyatt et al., 2010) and the vcontact_gene2genome script was used to produce the input files for vConTACT 2.0. ProkaryoticViralRefSeq94-merged database was used in the DIAMOND mode for aligning amino acid homology (Buchfink et al., 2015) and the ClusterONE mode for viral cluster (VC) construction (Nepusz et al., 2012) to classify and cluster viral genomes, based on amino acid homology. The “nodes” defined as the UViGs and the “edges” defined as the strength of the relationships between the genomes within a network (Bolduc et al., 2017) was then generated. In addition, the output of CheckV was incorporated into the clustering matrices and was used to add relative genome sizes for the nodes. Finally, Cytoscape (v3.8.2) was used to visualise the sequence clustering networks (Shannon et al., 2003).

In parallel, DemoVir (https://github.com/feargalr/Demovir, available in December 2020) was utilised to compare their amino acid homology against the non-redundant viral references from the TrEMBL database and assign taxonomy at the order and family levels using default parameters. For those viral contigs that could not be identified and annotated by DemoVir, vConTACT 2.0 was then used to annotate the remaining unknown sequences at the family level. If neither both were able to assign taxonomy to the viral genomes, the viral contigs were labelled as “unassigned”.

2.4.8. Analysis of Relative Abundance, Alpha and Beta Diversity

Relative abundance, alpha and beta diversity were generated mainly using PhyloSeq (v1.30.0) (McMurdie and Holmes, 2013) and Tidyverse (v1.3.0) (Wickham et al., 2019) with other essential packages such as ggplot 2 (v3.3.2) (Wickham, 2016). All plots were visualised in R and performed by Dr Rebecca Ansorge. The relative abundance analysis for virome-derived datasets was visualised in bubble plots showing the sorted top 25 viral contigs. For alpha diversity, all measures of each intra-subject of the virome library datasets were calculated with “observed richness”, “Chao1”, “Shannon” and “Simpson” indices. With regard to beta diversity, distances among inter-subjects of virome library datasets were

calculated using the matrices of Bray-Curtis dissimilarities and Jensen-Shannon divergence (JSD), and the ordination plots were drawn based on principal coordinate analysis (PCoA).

3. Optimising Human Faecal VLP Isolation and DNA Extraction

3.1. Introduction

3.1.1. The Human Intestinal Virome

The human intestinal virome is a highly complex, diverse and stable community of viruses, dominated by bacteriophages (Shkoporov et al., 2019). Numerically, viruses have been considered the most abundant and diverse biological entities on Earth, estimated to be approximately 10^{30} to 10^{32} in number (Breitbart and Rohwer, 2005, Wommack and Colwell, 2000). In many environments such as oceans, viruses outnumber bacteria on average by 10:1 (Wommack and Colwell, 2000). In the human gastrointestinal tract (GIT), virus-to-microbe ratio (VMR) may be close to 1:1 (Reyes et al., 2010), although it may reach at least 20:1 at the mucosal surfaces and within the mucus layer (Barr et al., 2013), in total numbering 10^{10}-10^{15} virus-like particles (VLPs) in human GIT (Hoyles et al., 2014, Dalmasso et al., 2014, Lepage et al., 2008).

The first human faecal virome metagenomic study in 2003, which estimated that the human intestinal viral community consists of approximately 1,200 viral genotypes with the most abundant virus accounting for around 4% of the total, was generated from the stools of a 33-year-old healthy male donor (Breitbart et al., 2003). Dysbiosis in the human intestinal virome has been reported in diseases, such as inflammatory bowel disease (IBD; including Crohn's disease and ulcerative colitis) (Norman et al., 2015, Lepage et al., 2008, Clooney et al., 2019), type I diabetes (T1D) (Zhao et al., 2017, Kramna et al., 2015, Foxman and Iwasaki, 2011), type II diabetes (T2D) (Ma et al., 2018), human immunodeficiency virus (HIV)-associated acquired immunodeficiency syndrome (AIDS) (Monaco et al., 2016, Li et al., 2012), malnutrition (Reyes et al., 2015), and myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) (Giloteaux et al., 2016b, Fremont et al., 2009, Chia and Chia, 2008). The more details of these virome-associated diseases are described in **Chapter 1**. However, the human intestinal virome is still overlooked in most microbiome studies due to limited tools and methods available for viral detection and cultivation. It therefore remains largely uncharacterised and is often referred to as "viral dark matter" (Reyes et al., 2012, Pedulla et al., 2003), with 40-90% sequences being unidentifiable (Reyes et al., 2015, Minot et al., 2013). Traditionally, virologists relied upon direct observation and enumeration using transmission electron (TEM) and epifluorescence microscopy (EFM), as well as culture-based approaches to screen, isolate and characterise lytic VLPs that kill specific bacterial hosts (Shkoporov et al., 2018b, Hoyles et al., 2014, Lepage et al., 2008, Castro-Mejia et al., 2015, Tartera et al., 1992). However, many intestinal viruses, particularly

temperate/lysogenic phages, are difficult to culture due to their non-lytic nature and/or low infective titre.

Recent advances in high-throughput pyrosequencing (next generation sequencing, NGS) technologies (Margulies et al., 2005, Ronaghi et al., 1996), viral bioinformatics tools and pipelines, as well as expanded reference viral genome databases, have enabled viral metagenomic analysis to be improved and to shed light on “viral dark matter” (Shkoporov and Hill, 2019, Reyes et al., 2012). However, some obstacles still need to be overcome. A major hurdle in using culture-independent, VLP-enriched metagenomics is the use of different protocols for virus isolation resulting in conflicting results, primarily due to contamination and biases (Shkoporov et al., 2018b). Thus, I aimed to develop a simple and reproducible protocol to isolate and enrich sufficient numbers of VLPs from human stools minimising contamination and the loss of VLPs to obtain sufficient high quality DNA for NGS.

3.1.2. Method Development of VLP Isolation

Published methods for faecal VLP isolation, enrichment and purification are diverse, but share basic concepts and procedures. Typically, faeces are homogenised, centrifuged and filtrated prior to VLP enrichment and contaminant removal steps followed by DNA extraction. In this study, I have developed an optimised protocol for faecal VLP isolation and downstream VLP DNA extraction based the core features of six published methods (**Table 3.1**).

Faeces are homogenised in buffers such as sodium chloride-magnesium sulphate (SM) buffer (Sambrook and Russell, 2001, Weigle et al., 1959), phosphate buffer or phosphate-buffered saline (PBS) (Clowes and Hayes, 1968, Breitbart et al., 2003) and TBT buffer (Biswal et al., 1967, Hoyles et al., 2014) with or without the use of bead-beating, followed by several rounds of low- or high-speed centrifugation to remove large debris and faecal matter. This is usually followed by syringe filtration using 0.22- and/or 0.45-µm (Shkoporov et al., 2018b, Hoyles et al., 2014, Reyes et al., 2010) or 0.8-µm filter (Conceicao-Neto et al., 2015) with low protein-binding membranes (e.g. polyethersulfone, cellulose acetate or polyvinylidene fluoride etc.) to remove human and microbial cells and residual faecal materials, as well as to concentrate VLPs from faecal filtrates (FFs).

Faecal VLPs are further enriched from FFs using polyethylene glycol (PEG) precipitation, dead-end/normal flow filtration (DEF/NFF), tangential flow filtration (TFF) or centrifugal ultrafiltration (Thurber et al., 2009, Colombet et al., 2007, Suttle et al., 1991). Chloroform is commonly used to extract lipids from cells and tissues (McKibbin and Taylor, 1949) and has

been applied to remove lipid contaminants and bacterial cells from VLP suspensions (Shkoporov et al., 2018b, Reyes et al., 2010), however, lipid-containing viruses may be sensitive to chloroform and may lose infectivity (Feldman and Wang, 1961). Alternatively, lysozyme is also used to remove bacterial cells (Zuo et al., 2017, Norman et al., 2015), followed by nuclease treatment (DNase and/or RNase) to degrade free non-viral capsid-associated nucleic acids (Shkoporov et al., 2018b, Reyes et al., 2010, Thurber et al., 2009). Moreover, some protocols rely on CsCl density gradient ultracentrifugation to remove residual contaminants such as nucleases. This technique leads to relatively pure VLPs but is laborious, has poor reproducibility and the loss of VLPs of atypical densities, thereby causing bias (Kleiner et al., 2015, Castro-Mejia et al., 2015).

Although recent studies have focused on protocol optimisation for faecal VLP isolation and enhancing recovery of VLP DNA yields for high-throughput studies (Shkoporov et al., 2018b, Kleiner et al., 2015, Conceicao-Neto et al., 2015, Castro-Mejia et al., 2015, Hoyles et al., 2014, Thurber et al., 2009), the extent of contamination, bias and recovery efficiency of most approaches for faecal VLP isolation and faecal VLP DNA extraction/purification have not been fully evaluated.

Table 3.1. Published methods for recovering VLPs and their nucleic acids from environmental samples

Sample source(s)	Sample sizes	Homogenisation	Centrifugation	Filtration	Enrichment	Purification/ cell lysis	Nucleic acid extraction	References
Soils & sediments	100 mg soils & 50 mg sediments	Bead beating	N/A	N/A	N/A	Optimal: SDS-phenol or SDS-chloroform	Various	Miller et al. (1999)
Multiple sources	10 L seawater	Homogeniser (PowerGen 125)	2,500 x g for 5-10 min	0.45 & 0.22 µm	TFF-PEG 8,000/NaCl	CsCl-DNase-Tris/ EDTA/formamide-SDS/PK-CTAB/NaCl-chloroform	P/C/I	Thurber et al. (2009)
Faeces & caecal effluent	25 g faeces & 10-30 ml caecal effluent	Homogeniser (Stomacher)	11,180 x g at 10°C for 30 min (x 2)	0.45 µm	PEG 8,000/ NaCl	Thurber et al. (2009)	P/C/I	Hoyles et al. (2014)
Faeces	5 g	Homogeniser (Stomacher)	5,000 x g at 4°C for 45 min	N/A	1) TFF (100 kDa MWCO) 2) PEG 6,000	Thurber et al. (2009)	Formamide-EtOH -ZR DNA clean & concentrator	Castro-Mejía et al. (2015)

Mock virome community	200 µl	Homogeniser (Minilys) with or w/o bead-beating	100 x g or 17,000 x g for 3 min or 30 min	Various filters (PES, PC, PVDF) & pore sizes (0.8-, 0.45-, 0.22 µm)	N/A	Benzonase, microccocal nuclease	Qiagen QIAamp viral RNA mini kit	Conceição-Neto et al. (2015)
Faeces	0.5 g	Vortexing	5,000 rpm at 4°C for 10 min	0.45 µm x 2	PEG 8,000/ NaCl	Chloroform-DNase/ RNase-SDS/PK-GTC/2-ME	P/C/I-Qiagen DNeasy blood & tissue kit	Shkoporov et al. (2018)

PEG: polyethylene glycol; TFF: tangential flow filtration; SDS: sodium dodecyl sulfate; PK: proteinase K; CTAB: cetyltrimethylammonium bromide; GTC/2-ME: guanidinium thiocyanate/2-mercaptoethanol; P/C/I: phenol/chloroform/isoamyl alcohol (25:24:1, v/v/v); N/A: not applicable; w/o: without

3.2. Aim

In this chapter, I set out to establish a simple and reproducible protocol for VLP isolation and VLP DNA extraction from human faeces to obtain DNA of sufficient quality and quantity for virome-enriched shotgun metagenomic sequencing and then to apply it to investigating the composition of the intestinal virome in the ME/CFS patients and same household healthy subjects.

3.3. Study Design

Faecal samples were obtained from three “healthy” adult volunteers (three males aged between 31 and 39 years). The following flow diagram below (**Figure 3.1**) provides an overview of protocol optimisation and of the individual steps that have been evaluated.

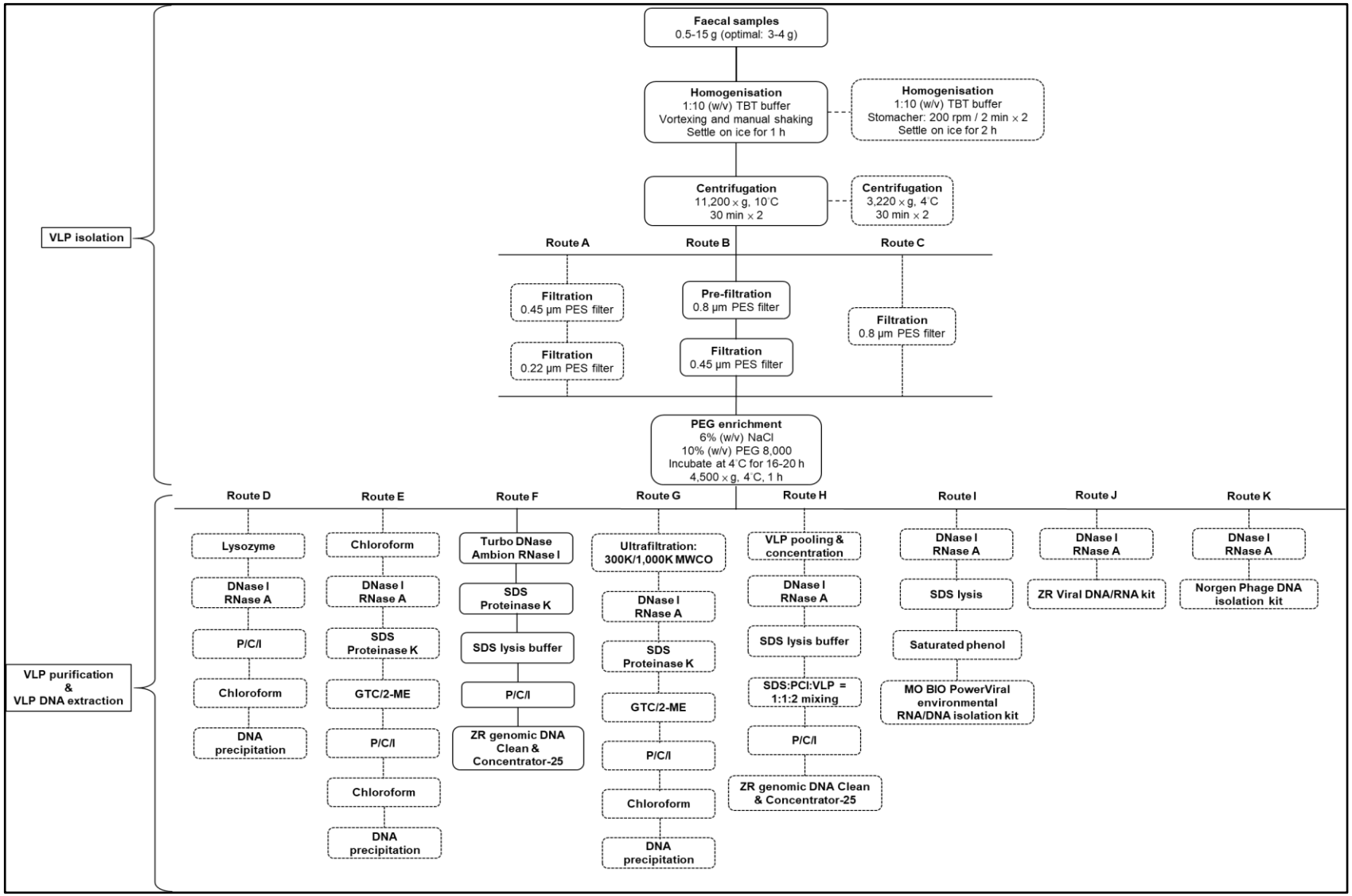

VLP isolation
Faecal samples
0.5-15 g (optimal: 3-4 g)
Homogenisation
1:10 (w/v) TBT buffer
Vortexing and manual shaking
Settle on ice for 1 h
Homogenisation
1:10 (w/v) TBT buffer
Stomacher: 200 rpm / 2 min × 2
Settle on ice for 2 h
Centrifugation
11,200 × g, 10°C
30 min × 2
Centrifugation
3,220 × g, 4°C
30 min × 2
Route A
Route B
Route C
Filtration
0.45 µm PES filter
Filtration
0.22 µm PES filter
Pre-filtration
0.8 µm PES filter
Filtration
0.45 µm PES filter
Filtration
0.8 µm PES filter
PEG enrichment
6% (w/v) NaCl
10% (w/v) PEG 8,000
Incubate at 4°C for 16-20 h
4,500 × g, 4°C, 1 h
VLP purification & VLP DNA extraction
Route D
Lysozyme
DNase I RNase A
P/C/I
Chloroform
DNA precipitation
Route E
Chloroform
DNase I RNase A
SDS Proteinase K
GTC/2-ME
P/C/I
Chloroform
DNA precipitation
Route F
Turbo DNase Ambion RNase I
SDS Proteinase K
SDS lysis buffer
P/C/I
ZR genomic DNA Clean & Concentrator-25
Route G
Ultrafiltration: 300K/1,000K MWCO
DNase I RNase A
SDS Proteinase K
GTC/2-ME
P/C/I
Chloroform
DNA precipitation
Route H
VLP pooling & concentration
DNase I RNase A
SDS lysis buffer
SDS:PCI:VLP = 1:1:2 mixing
P/C/I
ZR genomic DNA Clean & Concentrator-25
Route I
DNase I RNase A
SDS lysis
Saturated phenol
MO BIO PowerViral environmental RNA/DNA isolation kit
Route J
DNase I RNase A
ZR Viral DNA/RNA kit
Route K
DNase I RNase A
Norgen Phage DNA isolation kit

Figure 3.1. Overview of optimisation of faecal VLP isolation and DNA extraction. Protocol optimisation focuses on two main stages, VLP isolation (Route A-C), and VLP purification-capsid disruption and DNA extraction (Route D-K). The optimal procedures are indicated by the solid lined boxes with the dashed lined boxes indicating non-optimal procedures as described in the results section.

3.4. Results

3.4.1. Evaluating Efficiency of VLP Isolation

3.4.1.1. Evaluating Homogenisation and Centrifugation

With the aim of producing homogenous suspension of faeces (~5 g) and to disperse VLPs, the “Stomacher” was initially used. However, TEM analysis revealed many detached viral capsids and tails indicative of damage to and disruption of VLPs (data not shown). The more moderate procedure of vortexing and manual shaking was then adopted to homogenise faecal samples. I also considered the use of low-speed centrifugation (~3,200 x g) to avoid disrupting VLPs and minimising the loss. However, low-speed centrifugation was unable to effectively remove large faecal particulates resulting in the clogging of filters in subsequent filtration step. I therefore adopted high-speed centrifugation (11,200 x g) in this protocol.

3.4.1.2. Initial Evaluation of VLP Recovery by 0.45 µm and 0.22 µm Filtration

With the aim of evaluating VLP recovery, a published protocol for VLP isolation using serial 0.45 µm and 0.22 µm filtration described by Hoyles et al. (2014) was initially used (Route A, **Figure 3.1**). TEM images taken by Kathryn Gotts (QIB microscopy scientist) after serial 0.45 µm and 0.22 µm filtration (**Figure 3.2**) showed that very few intact VLPs in faecal filtrates (FFs) isolated from three frozen faecal samples provided by different healthy donors were seen, including two *Podoviridae*-like VLPs with isometric heads of approximately 90 nm and 100 nm in diameter and tail lengths of approximately 40 nm and 20 nm, respectively (**Figure 3.2.A and F**), and four *Siphoviridae*-like VLPs with isometric heads of between 60 nm and 75 nm in diameter and tail lengths of between 110 nm and 250 nm (**Figure 3.2.B-E**). This indicated that VLP loss has occurred after serial 0.45 µm and 0.22 µm filtration. Therefore, different filtration methods were further evaluated.

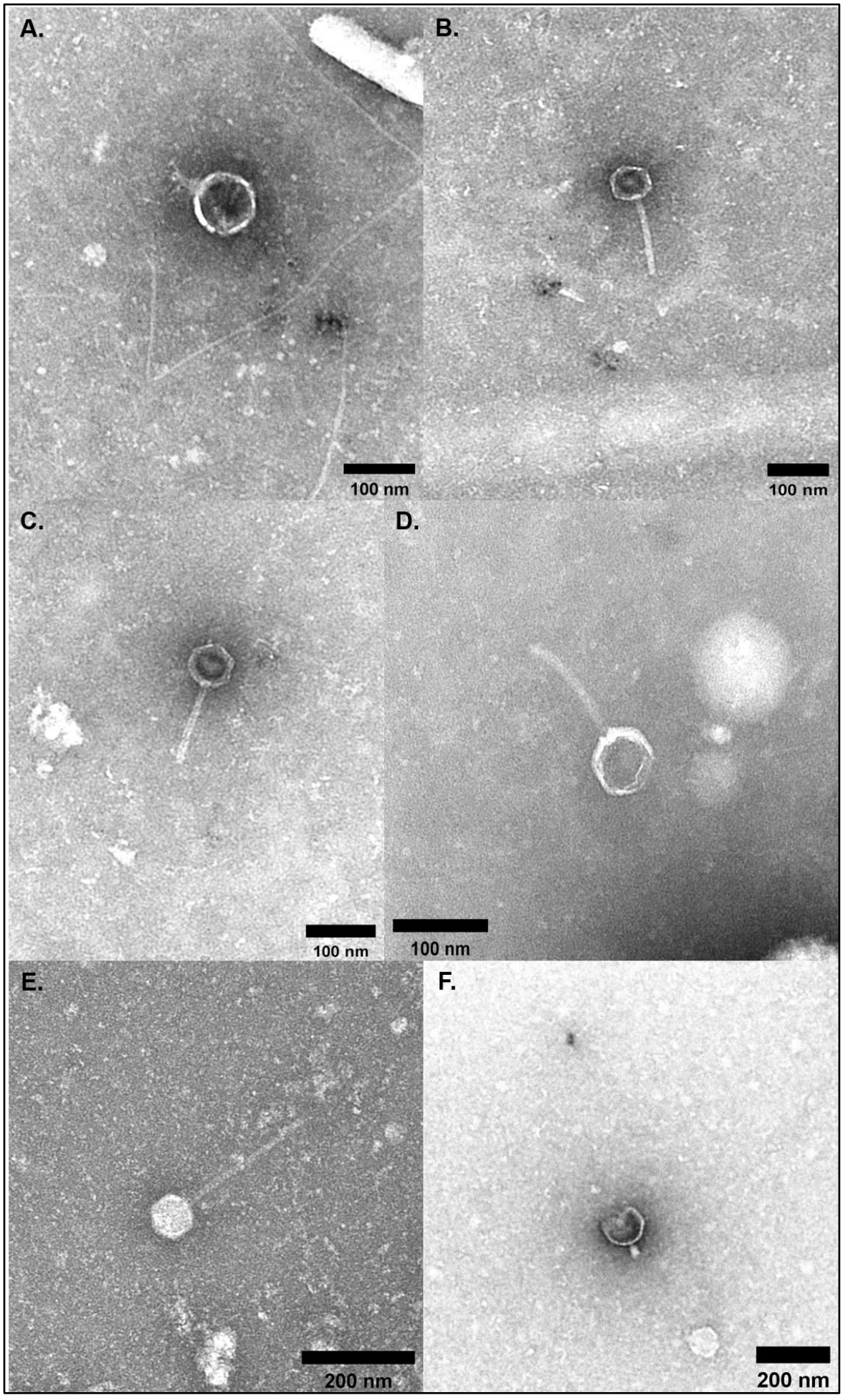

A.
100 nm
B.
100 nm
C.
100 nm
D.
100 nm
E.
200 nm
F.
200 nm

Figure 3.2. Transmission electron micrographs of VLPs in faecal filtrates isolated from three independent healthy donors after serial 0.45 µm and 0.22 µm filtration (Route A). (A) A *Podoviridae*-like phage (~130 nm in size). (B) A *Siphoviridae*-like phage (~200 nm in size). (C) A *Siphoviridae*-like phage (~170 nm in size). (D) A *Siphoviridae*-like phage (~200 nm in size). (E) A *Siphoviridae*-like phage (~325 nm in size). (F) A *Podoviridae*-like phage (~120 nm in size). Scale bar: (A-D) 100 nm; (E and F) 200 nm.

3.4.1.3. Improving VLP Recovery by 0.8 µm Filtration

With the aim of recovering sufficient VLPs from human faeces for VLP DNA isolation, the method of filtration was modified by introducing 0.8 µm filter (Route C, **Figure 3.1**). TEM analysis of FFs showed an increase in number and variety of VLPs (**Figure 3.3**), compared to initial TEM analysis of FFs collected from sequential fltration (0.45 µm and 0.22 µm). TEM images after 0.8 µm filtration (**Figure 3.3**) showed that the majority of faecal VLPs were bacteriophages, including *Siphoviridae* with isometric heads of approximately 140 nm and 57 nm in diameter and tail lengths of between 250 nm and 1,700 nm (**Figure 3.3.A and F**). Several different morphotypes of *Myoviridae*-like VLPs were also observed, with icosahedral heads, ranging from approximately 140 nm to 200 nm in diameter, with various tail lengths of between 170 and 250 nm (**Figure 3.3.B-D and G-I**). Some *Myoviridae*-like virions displayed radial whisker-like structures attached to their capsids (**Figure 3.3.D, H and I**). In addition, other spike (**Figure 3.3.B, D and G-I**) and fibre (**Figure 3.3.C**) structures were evident on the tails. Interestingly, a filamentous bacteriophage was identified which appeared to be a member of the *Inoviridae* family (positive ssDNA viruses) that infects *Enterobacteria*, with spherical structures observed on the end of the virion, which facilitate phage-mediated adsorption onto their hosts, of approximately 1,200 nm in length (**Figure 3.3.E**). However, using a large pore-size filter also resulted in contaminating solid materials including food debris, precipitants and bacteria (**Figure 3.3.J**). TEM analysis for **Figure 3.3** was performed by myself, Kathryn Gotts and Dr Catherine Booth (QIB microscopy scientists).

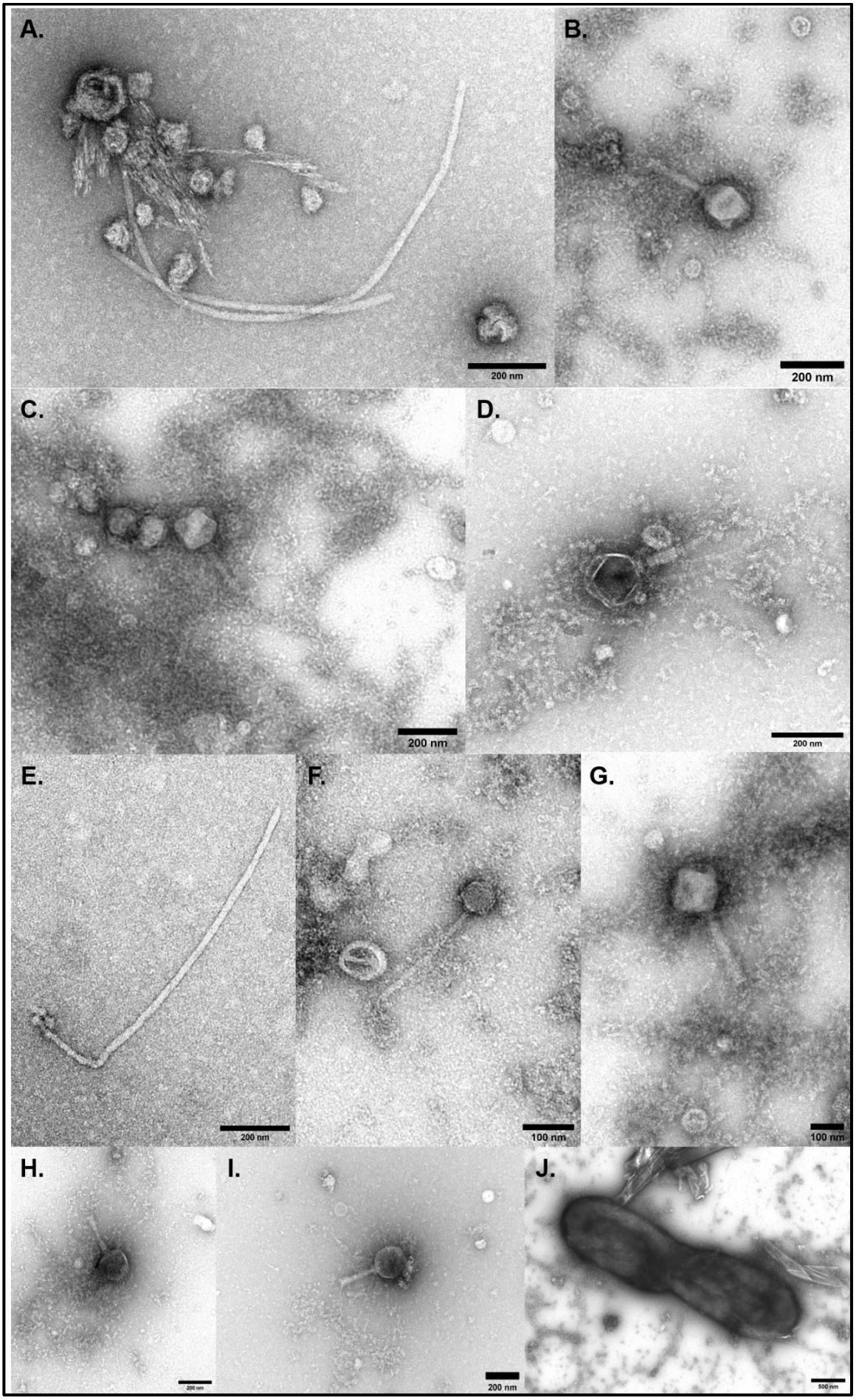

A.
200 nm
B.
200 nm
C.
200 nm
D.
200 nm
E.
200 nm
F.
100 nm
G.
100 nm
H.
200 nm
I.
200 nm
J.
500 nm

Figure 3.3. Transmission electron micrographs of VLPs in 0.8 µm FFs (Route C). (A) A giant *Siphoviridae*-like VLP (~1,800 nm in size) surrounded by several virions. (B) A *Myoviridae*-like VLP with the spikes on the end of the tail. (C) A *Myoviridae*-like VLP with fibres. (D) A *Myoviridae*-like VLP with radial whisker-like structures attached to capsid and spikes on the end of the tail. (E) An *Inoviridae*-like filamentous VLP (>1,200 nm in size) with spherical structures on the end of the virion. (F) A *Siphoviridae*-like VLP. (G) A *Myoviridae*-like VLP with spikes on the end of the tail. (H and I) Two *Myoviridae*-like VLPs with radial whiskers attached to capsids have different types of tails. (J) A dividing bacterial cell. Scale bar: (A-I) 100-200 nm; (J) 500 nm.

To determine efficiency of VLP isolation after 0.8 µm filtration, a faecal homogenate from a healthy donor was spiked with the reference *B. fragilis* (Bf) phage, ΦB124-14, with aliquots (100 µl) taken after centrifugation, 0.8 µm filtration and PEG precipitation for plaque assays. Throughout the entire isolation procedure, loss of the spiked phage was minimal with the phage-containing supernatant collected after centrifugation containing approximately 83% of original spiked Bf phages with ~11% of the spiked phage recovered in the pellet (**Table 3.2**). Approximately 85% and 92% of the spiked phages were recovered after 0.8 µm filtration and PEG precipitation, respectively (**Table 3.2**). This spiking-and-recovery protocol was then repeated three times using faecal samples from three independent healthy donors (**Table 3.3**). **Table 3.3** showed that approximately 69% to 86% of the spiked phages were recovered across these three samples, with the average titres ranging from $9.7 \times 10^9 \pm 4.9 \times 10^9$ to $1.2 \times 10^{10} \pm 2.7 \times 10^9$ (pfu/ml; mean ± S.D.). Faecal homogenates without phage spiking (triplicates for each of three faecal samples) were also used as negative control to monitor any aboriginal hosts in the human gut which were also infected by ΦB124-14, but no plaques were detected.

Table 3.2. Recovery of spiked phage after key steps in VLP isolation after 0.8 µm filtration (Route C)

Treatment	Phage titre (pfu/ml; n = 1)	Recovery (%)*
Original phage stock for spiking	1.2×10^{10}	100.0
PFU after centrifugation – supernatant	1.0×10^{10}	83.3
PFU after centrifugation – pellet	1.3×10^9	10.8
PFU after 0.8 µm filtration	1.0×10^{10}	85.0
PFU after PEG precipitation – pellet	1.1×10^{10}	91.7
PFU after PEG precipitation – supernatant	1.3×10^8	1.1

***** As determined by dividing PFU of original phage stock into PFU of each treatment then multiplying by 100%

Table 3.3. Efficiency of VLP isolation after 0.8 µm filtration determined by plaque assays (Route C)

	Titre (pfu/ml)			Mean (n = 3)	S.D. (n = 3)	Recovery (%)*
Original stock for spiking	1.6×10^{10}	1.5×10^{10}	1.0×10^{10}	1.4×10^{10}	3.2×10^{9}	100.0
Replicate 1: PEG suspensions	1.5×10^{10}	1.1×10^{10}	1.0×10^{10}	1.2×10^{10}	2.7×10^{9}	85.7
Replicate 2: PEG suspensions	1.3×10^{10}	1.2×10^{10}	4.0×10^{9}	9.7×10^{9}	4.9×10^{9}	69.3
Replicate 3: PEG suspensions	1.1×10^{10}	9.6×10^{9}	1.0×10^{10}	1.0×10^{10}	7.2×10^{8}	71.4

***** As determined by dividing average PFU of original phage stock into average PFU after PEG precipitation then multiplying by 100%

3.4.1.4. Further Improvement to Reduce Solid Materials and Bacterial Contamination by Dual Filtration

To further reduce solid materials and bacterial contamination in FFs, VLP isolation was modified by introducing serial 0.8 µm and 0.45 µm filtration steps (or dual filtration; Route B, **Figure 3.1**). **Table 3.4** showed that subsequent PEG-VLP suspensions contained between 30% and 40% of the original spiked Bf phages using dual filtration. Compared to the results of using 0.8 µm filters only, the overall recovery rate was therefore reduced by ~50%. This further reduction in VLP recovery may be a consequence of VLP retention within 0.45 µm filters.

Table 3.4. Efficiency of VLP isolation after serial 0.8 µm and 0.45 µm filtration determined by plaque assays (Route B)

	Titre (pfu/ml)			Mean (n = 3)	S.D. (n = 3)	Recovery (%)*
Original stock for spiking	8.0×10^9	8.5×10^9	9.6×10^9	8.7×10^9	8.2×10^8	100.0
w/o spiking	2.3×10^9	3.3×10^9	2.8×10^9	2.8×10^9	5.0×10^8	32.2
Replicate 1: PEG suspensions	3.3×10^9	3.1×10^9	2.6×10^9	3.0×10^9	3.0×10^8	34.5
Replicate 2: PEG suspensions	3.0×10^9	2.8×10^9	2.7×10^9	2.8×10^9	7.5×10^8	32.6
Replicate 3: PEG suspensions	3.5×10^9	3.0×10^9	4.0×10^9	3.5×10^9	1.0×10^9	40.2

***** As determined by dividing average PFU of original phage stock into average PFU after PEG precipitation then multiplying by 100%

To corroborate the accuracy of VLP recovery determined using plaque assay, fluorescence-stained spiked phage and epifluorescence microscopy (EFM) were used to visually enumerate VLPs in PEG-suspensions. ΦB124-14 was stained with SYBR Gold prior to spiking faecal homogenates and then enumerated after dual filtration and PEG precipitation by EFM. The number of VLPs detected by EFM was equivalent to a recovery of 95.5% (**Table 3.5**).

Table 3.5. Efficiency of VLP isolation after serial 0.8 µm and 0.45 µm filtration using reference phages and EFM (Route B)

	Sample spiked	**Process control (w/o spiking)**	**Phage stock**
Total no. of particles	13,846	13,363	14,501
Mean (n = 20)	692.3	668.2	725.1
Viral counts (VP/ml)	1.84×10^8	1.77×10^8	1.92×10^8
Recovery (%)*****	95.5	92.2	100.0

***** As determined by dividing PFU of original phage stock into PFU after PEG precipitation then multiplying by 100%

3.4.1.5. Characterisation of Faecal VLPs by TEM

Intact VLPs were observed in all three FFs after serial 0.8 µm and 0.45 µm filtration using TEM (**Figure 3.4**). The majority of faecal VLPs were bacteriophages, including *Siphoviridae*-like phages with isometric heads from 50 nm to 200 nm in diameter and various tail length sizes of between 180 nm and 600 nm (**Figure 3.4.A-C**). Different morphotypes of *Myoviridae* with icosahedral heads between 80 nm and 100 nm in diameter and diverse tail lengths, ranging from 100 to 200 nm, were also observed (**Figure 3.4.D and E**). Some *Myoviridae*-like VLPs also displayed spikes on the end of their tails attaching to membranous-like materials (**Figure 3.4.D and E**). In addition, many separate viral capsids and tails were found in these samples, consistent with 0.45 µm filtration negatively affecting structural stability of viral particles. However, as the images showed, this step was beneficial in removing solid materials as well as bacterial cells. TEM analysis for **Figure 3.4** was performed by myself and Dr Catherine Booth.

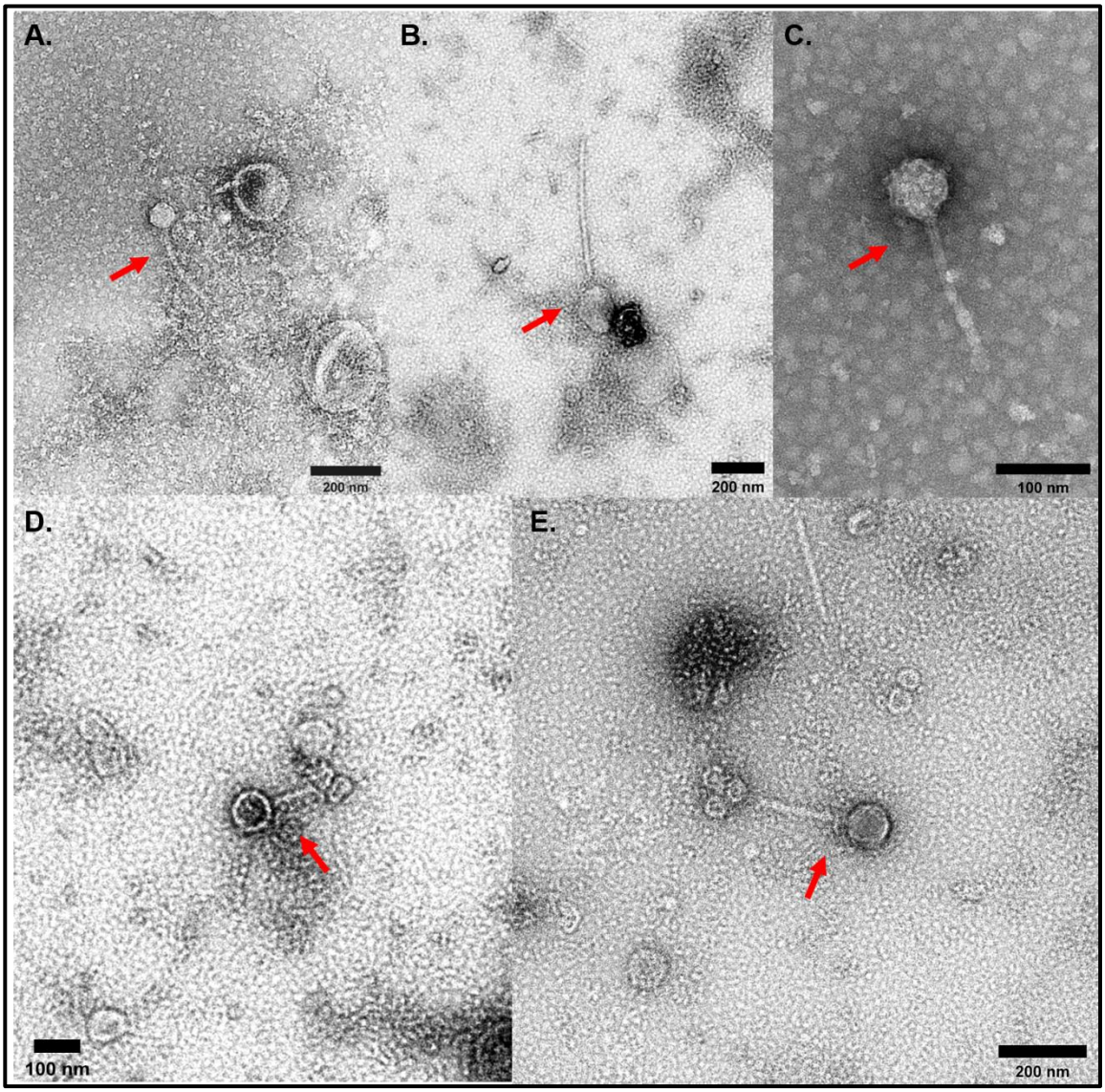


Figure 3.4. Transmission electron micrographs of VLPs in FFs after dual filtration in three healthy donors (Route B). Red arrows indicate intact virions and virus-like structures with 100-200 nm of the scale bars. (A) A *Siphoviridae*-like phage (~480 nm in size) found in donor 2; (B) A *Siphoviridae*-like phage (~880 nm) found in donor 3; (C) A *Siphoviridae*-like phage (~300 nm) found in donor 1; (D) A *Myoviridae*-like phage attaching to mambrane vesicle-like materials (~200 nm) in donor 3; (E) A *Myoviridae*-like phage with spikes on the end of the tail (~300 nm) found in donor 3.

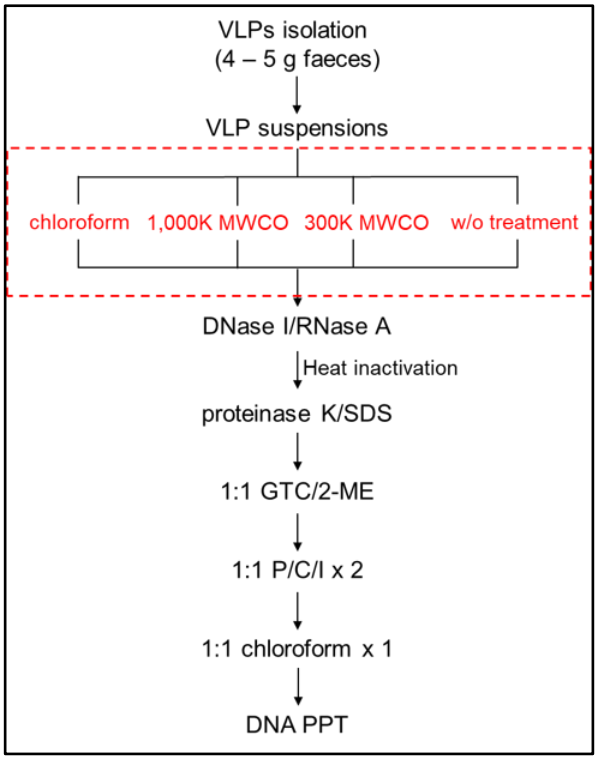
3.4.2. Optimising Procedures to Remove Contaminants

After isolating faecal VLPs using 0.8 μm filtration-based protocol, to minimise and remove potential contaminants and to increase efficiency of downstream DNA extraction, published protocols for the removal of free non-viral capsid-contained nucleic acid as well as protein and lipid contaminants were compared (refer to Route D-K identified in **Figure 3.1**).

(1) Chloroform and Ultrafiltration Treatment: Route E and Route G

PEG-VLP suspensions isolated from 4-5 g of faeces from a healthy donor were treated with chloroform and filtered through 300K and 1,000K MWCO filters (**Figure 3.5.A**) with extracted DNA samples visualised on 1% agarose gel (**Figure 3.5.B**). **Table 3.6** showed that the viral DNA concentrations and total yields measured by Qubit ranged from 52.4 ng to 404 ng. The results of chloroform treatment suggested that chloroform may disrupt certain types of viruses resulting in lower yields of viral genomic DNA, particularly lipid-enveloped viruses and chloroform-sensitive phages. Both 300K MWCO and 1,000K MWCO filters appeared to be more effective at enriching VLPs than chloroform treatment as the total yields of viral DNA were higher. However, total DNA yield after ultrafiltration were low and insufficient for NGS.

A.



B.

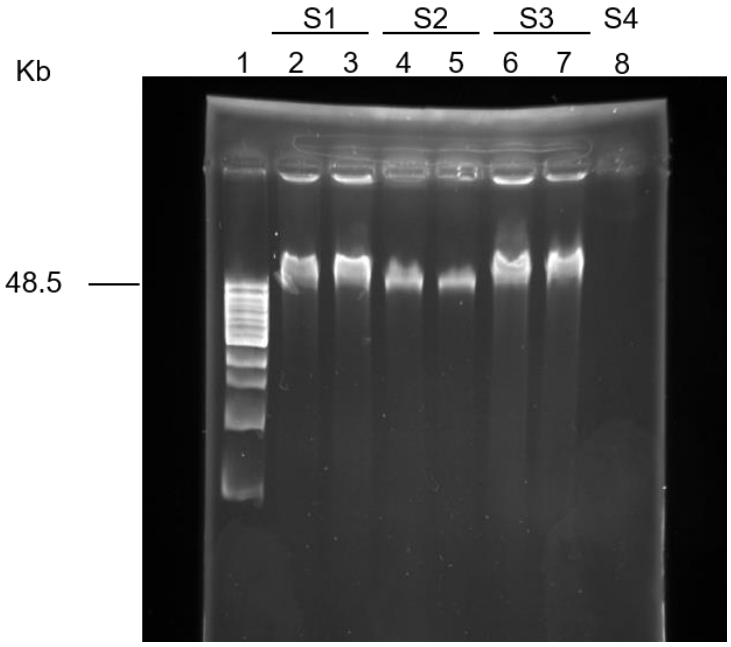


Figure 3.5. Evaluating chloroform and ultrafiltration treatment. (A) Workflow used to evaluate chloroform (route E), 300K and 1,000K MWCO filters (route G), followed by nucleases and proteinase K/SDS treatment and VLP DNA isolation. (B) DNA samples recovered from this protocol were assessed on 1% agarose gel. Lane1: 1 kb extended DNA ladder; Lane 2 and 3: duplicate aliquots (S1) resulted from chloroform treatment; Lane 4 and 5: duplicate aliquots (S2) resulted from 1,000K MWCO treatment; Lane 6 and 7: duplicate aliquots (S3) resulted from 300K MWCO treatment; Lane 8: a control without treatment.

Table 3.6. Quantity assessment of VLP DNA yields by Qubit (n = 1)

Treatments (Route)	DNA concentration (ng/µl)	DNA yield (ng)
chloroform (E)	2.2	89.6
	1.3	52.4
1,000K MWCO ultrafiltration (G)	2.2	87.6
	2.2	87.6
300K MWCO ultrafiltration (G)	10.1	404.0
	9.4	374.0
w/o treatment	1.9	74.0

(2) Evaluating GTC/2-ME Treatment

To investigate if guanidinium thiocyanate and 2-mercaptoethanol have an adverse impact on microbial/viral genomic DNA, genomic DNA isolated from *Bacteroides thetaiotaomicron* was treated with GTC/2-ME solution, followed by passage over a ZR genomic DNA Clean & Concentrator^{TM}-25 column. The recovery of gDNA without treatment was around 90% (~5,400 ng), while the recovery after treatment with GTC/2-ME was around 60% (~3,600 ng), with a reduction in yield of around 1,800 ng (**Table 3.7**).

Table 3.7. Quality and quantity assessment of the reference bacterial genomic DNA determined by Nanodrop (n = 1)

Treatment	$OD_{260/280}$	DNA concentration (ng/µl)	Total DNA amounts (ng)
with GTC/2-ME	1.96	61.5	3,639.0
w/o GTC/2-ME	1.90	90.5	5,430.0

3.4.3. Evaluating DNA Extraction Methods

To optimise faecal VLP DNA extraction, three published methods and three commercial viral genomic DNA isolation kits were selected and compared side by side. The quality and quantity of faecal VLP DNA isolated from ~5 g of fresh faecal samples of one healthy donor using selected methods or kits were measured by Nanodrop. **Table 3.8** showed that conventional phenol/chloroform approach obtains higher amounts of faecal VLP DNA (~4.7 µg) compared to other approaches. GTC-P/C/I and SDS-P/C/I recovered ~1,400 ng and ~600 ng of total VLP DNA, respectively, with the concentration of DNA being too low for NGS. Moreover, the quality and quantity of VLP DNA isolated by commercial kits were inadequate for NGS. Hence, we chose the conventional P/C/I method to extract faecal VLP DNA in combination with SDS/proteinase K treatment for sample decontamination and viral capsid disruption.

Table 3.8. Quality and quantity assessment of DNA samples determined by Nanodrop

Method/Kit (Route)	DNA concentration (ng/µl)	Total DNA (ng)	$OD_{260/280}$
P/C/I (D)	94.6	~4,700.0	1.82
GTC-P/C/I (E)	5.5	~1,375.0	1.77
SDS-P/C/I (H)	10.7	588.5	1.80
PowerViral environmental RNA/DNA isolation kit (I)	3.5	192.5	1.55
ZR viral DNA/RNA kit (J)	15.5	775.0	1.65
Phage DNA isolation kit (K)	10.0	750.0	1.67

3.4.4. Evaluating Stool Sample Size

Based on the protocol development to date, I inferred that to obtain sufficient amounts of total VLP DNA for NGS, larger stool sample sizes were needed. To verify if 10 g or more faeces improved VLP DNA yields, faecal samples of >20 g were collected and divided into 0.5 g, 1.5 g, 3 g, 5 g, 10 g and 15 g aliquots, from which VLPs were isolated using dual filtration-based protocol (0.8 µm and 0.45 µm) and DNA was extracted. The result indicated that there is an increasing trend in total DNA recovery between 0.5 g and 3 g, followed by a lower DNA recovery from samples of 3 g to 15 g (**Figure 3.6**). Therefore, 3-4 g of faecal material was optimal for each purification column to maximise DNA yield and I ultimately used >20 grams of faecal samples in total to obtain sufficient amounts of VLP DNA for NGS.

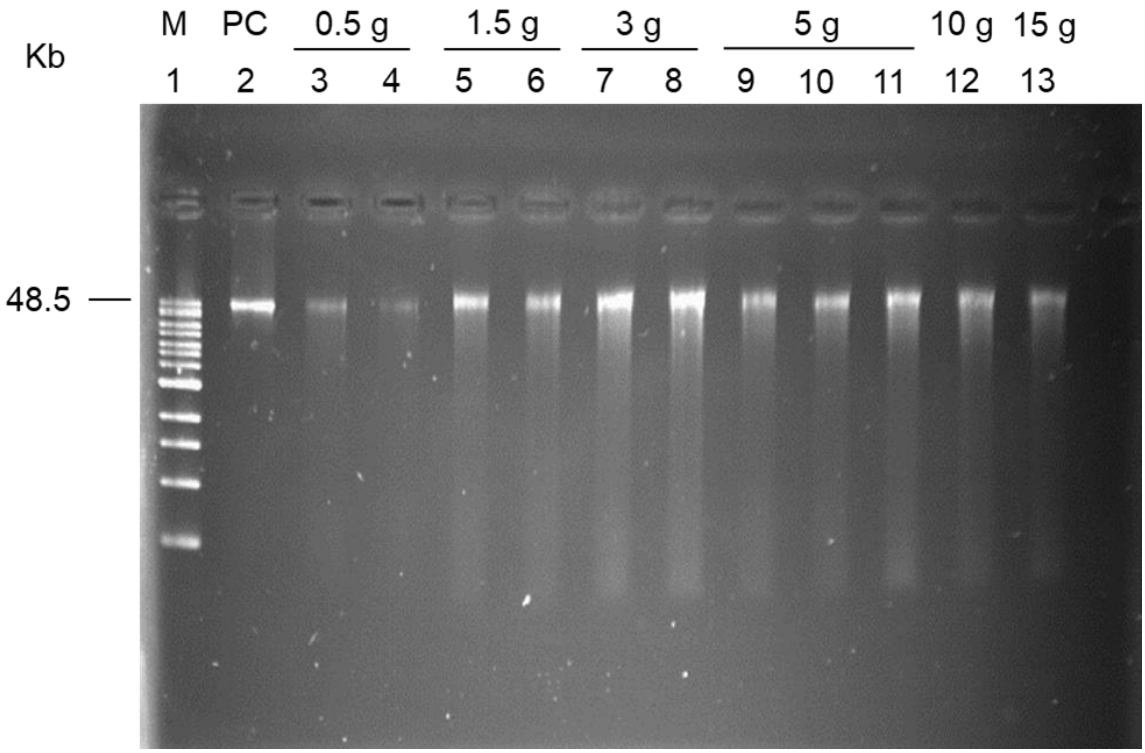


Figure 3.6. Impact of stool sample size on DNA recovery visualised by gel electrophoresis (n = 1). Lane 1: 1 kb extended DNA ladder (with the max size 48.5 kb); Lane 2: positive control (Bt. gDNA with ~100 ng in total); Lane 3 and 4: DNA from 0.5 g faeces; Lane 5 and 6: DNA from 1.5 g faeces; Lane 7 and 8: DNA from 3 g faeces; Lane 9 to 11: DNA from 5 g faeces; Lane 12: DNA from 10 g faeces; Lane 13: DNA from 15 g faeces.

3.4.5. Protocol Finalisation

Finally, I applied the optimised method to isolate VLPs and total VLP DNA from the faeces of three donors for NGS and shotgun metagenomic sequencing (**Figure 3.7 and Table 3.9**). From the analysis of DNA quality and quantity, ~4.5 µg of VLP DNA was obtained from donor 1 (~47.0 g), ~2.5 µg of VLP DNA from donor 2 (~37.5 g) and ~3.2 µg of VLP DNA from donor 3 (~37.0 g) (**Table 3.9**). Moreover, all three DNA samples were of high quality based on OD260/280 ratios. **Figure 3.7** showed that all of these samples had a major DNA product above or close to 48.5 kb with low levels of DNA smearing and small DNA/RNA fragments in donor 1 sample. Interestingly, there was a secondary DNA product in donor 2 at around 8 kb. In donor 3, DNA smearing along with distinct small DNA and/or RNA fragments were seen. Moreover, no significant protein contaminations were observed among these DNA samples with OD260/280 ratios reaching 1.8. The finalised protocol (Route B and Route F) was illustrated in **Figure 3.8**.

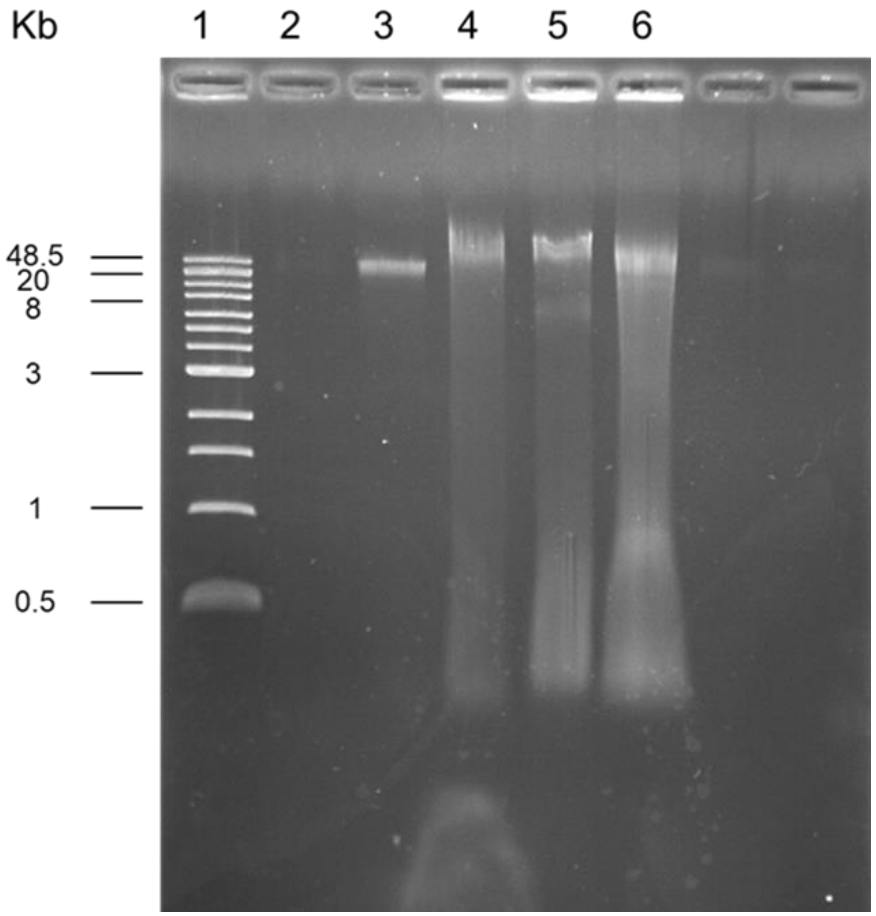


Figure 3.7. The quality of three DNA samples visualised by 1% gel electrophoresis. Lane 1: 1 kb extended DNA ladder (with the max size 48.5 kb); Lane 2: negative control (buffer only); Lane 3: positive control (Bt. gDNA with ~50 ng in total); Lane 4: DNA from donor 1 (~268.5 ng in total); Lane 5: DNA from donor 2 (~175 ng in total); Lane 6: DNA from donor 3 (~232 ng in total).

Table 3.9. Quality and quantity assessment of three faecal VLP DNA samples determined by Nanodrop and Qubit

Donor sample	Sample weight (g)	DNA concentration (ng/µl)	Total DNA (ng)	$OD_{260/280}$	DNA yield (ng/g faeces)
1	47.0	53.7	4,457.1	1.8	94.8
2	37.5	35.0	2,520.0	1.8	67.2
3	37.0	46.4	3,201.6	1.8	86.5

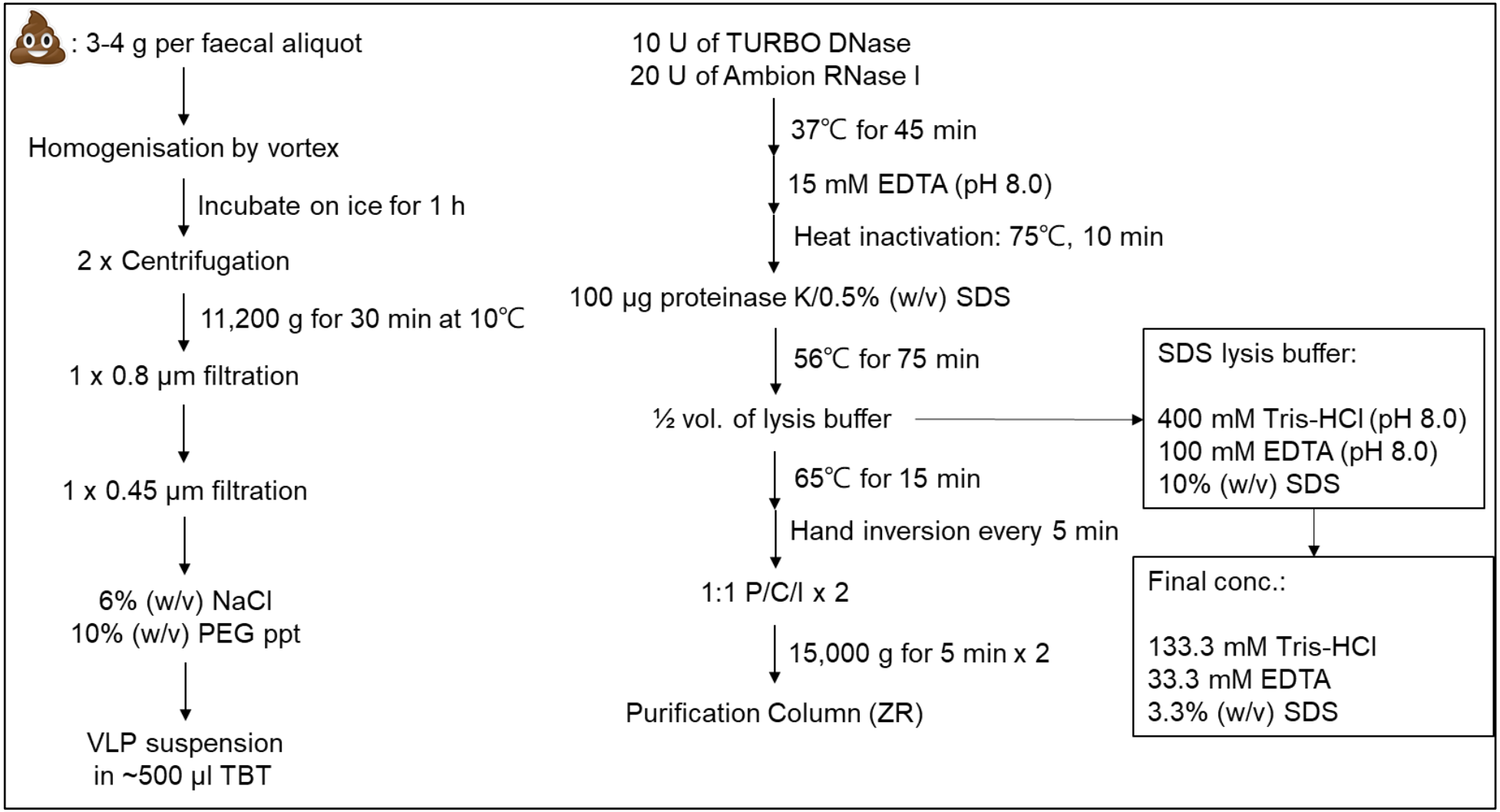

: 3-4 g per faecal aliquot
Homogenisation by vortex
Incubate on ice for 1 h
2 x Centrifugation
11,200 g for 30 min at 10℃
1 x 0.8 µm filtration
1 x 0.45 µm filtration
6% (w/v) NaCl
10% (w/v) PEG ppt
VLP suspension
in ~500 µl TBT
10 U of TURBO DNase
20 U of Ambion RNase I
37℃ for 45 min
15 mM EDTA (pH 8.0)
Heat inactivation: 75℃, 10 min
100 µg proteinase K/0.5% (w/v) SDS
56℃ for 75 min
½ vol. of lysis buffer
65℃ for 15 min
Hand inversion every 5 min
1:1 P/C/I x 2
15,000 g for 5 min x 2
Purification Column (ZR)
SDS lysis buffer:
400 mM Tris-HCl (pH 8.0)
100 mM EDTA (pH 8.0)
10% (w/v) SDS
Final conc.:
133.3 mM Tris-HCl
33.3 mM EDTA
3.3% (w/v) SDS

Figure 3.8. Workflow of the optimised VLP isolation and VLP DNA extraction protocol. 3-4 g of faeces were homogenised in sterile TBT buffer by vortexing. The faecal homogenates were centrifuged at 11,200 x g for 30 minutes at 10°C, followed by a second centrifugation under the same conditions. Supernatants were then filtered sequentially through 0.8 µm and 0.45 µm cartridge filters. NaCl (final concentration 6%, w/v) was added to faecal filtrates (FFs), followed by addition of PEG 8,000 (final concentration 10%, w/v). The samples were then left at 4°C for at least 16 hours. PEG-precipitated VLPs were harvested by centrifugation at 4,500 x g for 1 hour at 4°C. Supernatants were removed and the phage-containing pellets resuspended in ~500 µl of TBT buffer. VLP suspensions were treated with 10 U of TURBO™ DNase and 20 U of Ambion RNase I at 37°C for 45 minutes, followed by addition of 15 mM EDTA (pH 8.0) and heat inactivation at 75°C for 10 minutes. Afterwards, 100 µg of proteinase K and 0.5% (w/v) SDS were added to the samples and incubated at 56°C for 75 minutes, followed by adding 1/2 volume of 10% (w/v) SDS lysis buffer and incubating at 65°C for 15 minutes. An equal volume of P/C/I was added to the VLP lysate and mixed thoroughly by vigorously vortexing for 30 seconds, followed by centrifugation at 15,000 x g for 5 minutes at 20°C, which was repeated once more. The resulting aqueous phase was then passed over a ZR genomic DNA Clean & Concentrator™-25 column with the viral DNA eluted in 50-70 µl of elusion buffer (low EDTA TE buffer, pH 8.0). All DNA aliquots were then pooled and concentrated using a vacuum concentrator for approximately 1 hour until the total volume was reduced to 60-100 µl. Concentrated VLP DNA was stored at -70°C.

3.5. Discussion

The recent advent of NGS technologies (Margulies et al., 2005) together with improved bioinformatic tools, pipelines as well as expanded viral databases for viral reads assembly, detection, annotation and mapping, have enabled more detailed studies of the intestinal virome (Shkoporov et al., 2019, Norman et al., 2015, Minot et al., 2013, Reyes et al., 2010, Breitbart et al., 2003). However, there is a need for rigorously developed standardised protocols to ensure reproducible and reliable data. In this chapter, I set out to establish a simple, reliable and standardised method for faecal VLPs isolation and VLP DNA extraction. The optimised protocol (**Figure 3.8**) is comprised of (1) homogenisation of faecal samples in TBT buffer by vortexing without the use of bead-beating, followed by incubation on ice to facilitate the release of VLPs from solid materials; (2) partition of crude faecal matter, dietary debris and virions/VLPs by two-round of high-speed centrifugation; (3) sequential filtration using 0.8 μm and 0.45 μm filter; (4) PEG precipitation; (5) DNase and RNase treatment; (6) proteinase K digestion; (7) viral capsid lysis with SDS lysis buffer; (8) P/C/I extraction; (9) DNA purification using silica-based spin columns and (10) DNA concentration using a vacuum-based condenser.

3.5.1. Homogenisation

After reviewing published methods, I adopted the methods described by Hoyles (2014) and Shkoporov (2018) for further optimisation. To begin with, a homogenous suspension of VLPs is important for their isolation and purification from most environmental samples. Some protocols use tissue homogenisers such as Stomacher (Seward), Minilys system (Precellys) or PowerGen 125 system (Conceicao-Neto et al., 2015, Hoyles et al., 2014, Thurber et al., 2009), while others use benchtop vortex system (Shkoporov et al., 2018b, Norman et al., 2015) or vigorous shaking by hand (Reyes et al., 2010, Breitbart et al., 2008, Breitbart et al., 2003). A recent report demonstrated that the use of large-sized beads (e.g. 2.8 mm) leads to severe loss of certain viruses and increases bacterial contamination. However, this phenomenon was not observed in using small-sized beads (e.g. 0.1 mm) (Conceicao-Neto et al., 2015). Thus, to prevent VLP damage, I homogenise stool samples using vortex and hand-shaking.

3.5.2. Centrifugation

Centrifugation is used to remove human and microbial cells, dietary debris and large faecal matter. Too low or too high g-forces and longer centrifugal period may result in severe loss of VLPs (Conceicao-Neto et al., 2015), hence the need to assess centrifugal conditions (Conceicao-Neto et al., 2015, Castro-Mejia et al., 2015). In my data, around 83% of infective

Bf phages are recovered after high-speed centrifugation. Given that giant viruses and/or viral aggregates may also be lost within subsequent dual filtration and that high-speed centrifugation can effectively remove the vast majority of bacteria and large debris, I made a trade-off between the bias resulting from the loss of giant viruses and/or some clusters of VLPs and the benefits of reducing bacterial contamination as well as removing solid materials by conducting two rounds of high-speed centrifugation in this protocol.

3.5.3. Filtration

A similar compromise is made in the subsequent filtration step. A combination of 0.8 μm and 0.45 μm filters rather than 0.8 μm filter alone were used to minimise bacterial contamination accepting that some VLPs may be lost. In the spiking-and-recovery assays, approximately 85% of infective Bf phages are recovered after 0.8 μm filtration. However, a pronounced reduction in phage recovery is seen after serial 0.8 μm and 0.45 μm filtration, of between 30% and 40%. Conceição-Neto and colleagues (2015) assessed three types of low-protein-binding filter membrane materials alongside three different filter sizes, including a 0.8 μm centrifugal (PES), a 0.8 μm polycarbonate (PC), a 0.45 μm centrifugal (PVDF) or a 0.22 μm centrifugal filter (PVDF). Their findings suggested that the vast majority of bacteria in a bacterial mock-community are efficiently removed with the 0.8 μm PES, 0.45 μm and 0.22 μm filters, but only the 0.8 μm PC filter has a low efficiency for the removal of bacteria, ranging from 50% to 99.2%. On the other hand, more than 90% of viruses in a viral mock-community (e.g. herpesvirus and mimivirus) were lost during 0.22 μm filtration, consistent with the results of serial 0.45 μm and 0.22 μm filtration assessed by TEM. I therefore do not include 0.22 μm filtration in this protocol. To minimise bacterial contamination and retain viruses during filtration, I used a combination of 0.8 μm and 0.45 μm (PES) filtration to the finalised protocol for virome study. A variety of faecal virions/VLPs are seen in 0.8 μm-filtered samples including giant *Siphoviridae*-like phages with a very long tail over 1,000 nm in length, long filamentous *Inoviridae*-like VLPs with spherical structures on the end of the virion as well as several types of *Myoviridae*-like VLPs. It is not surprising some bacteria are also observed. Moreover, low abundance and diversity of VLPs are seen in dual-filtered samples. Comparing the results of 0.8 μm-filtered samples with those obtained by dual filtration (0.8 μm and 0.45 μm), noticeable variations are seen in the abundance and diversity of faecal VLPs across these three healthy donors, consistent with human intestinal viromes being highly diverse and individually specific with little overlap between individuals (Shkoporov et al., 2019, Manrique et al., 2016).

3.5.4. PEG Enrichment

PEG has been used to concentrate and precipitate viral particles from various environmental samples in combination with the presence of monovalent salt such as NaCl (Lewis and Metcalf, 1988, Albertsson and Frick, 1960, Colombet et al., 2007). In my spiking-and-recovery assays, up to 91.7% of spiked infective Bf phages are recovered after PEG precipitation. The spiking-and-recovery trial (0.8 µm filtration) showing high recovery rates has been repeated three times using faecal samples from three independent healthy donors.

Only 30-40% of spiked phages retain their infectivity after sequential filtration (0.8 µm and 0.45 µm) and PEG precipitation. The reduction in viral infectivity is most likely due to disruption of intact phages after 0.45 µm filtration as shown by TEM. In addition, SYBR Gold staining and epifluorescence microscopy shows that viral capsid recovery is high. As EFM cannot distinguish intact virions from detached phage heads which would not be infective and detectable by plaque assay, this data indicated that the majority of intact VPs and/or detached viral capsids can be recovered from faeces using VLP isolation described. Collectively, these findings demonstrate that while the VLP isolation protocol I have developed is efficient and well-suited to recovering viral capsids for downstream molecular-based virome analysis, it is less well-suited to recovering fully intact, infective phages.

3.5.5. Sample Decontamination and VLP DNA Recovery

The yields of faecal VLP DNA vary from sample to sample. My initial test revealed that although VLP DNA concentration and its total yield increases with larger faecal samples of between 10 g and 20 g, the integrity and quality of VLP DNA is poor and not optimal for NGS (data not shown). Using large amounts of faeces to isolate VLP DNA is likely to be less efficient at removing contaminants such as human or microbial cells and cell debris, dietary debris, mucins, polysaccharides, bile salts, lipids as well as proteins that interfere with the recovery of VLP DNA and lower the quality of DNA. On the basis of previous studies, 4-5 g of faeces yields between 500 ng and 1,800 ng of VLP DNA, which accounts for 2-5% of total DNA (Shkoporov et al., 2018b, Hoyles et al., 2014, Arumugam et al., 2011, Reyes et al., 2010, Thurber et al., 2009). However, such small samples are unlikely to produce sufficient viral DNA without the need for PCR- or MDA-based amplification, particularly for viruses/phages with small genomes (assuming an average size of 50 kb per phage genome) (Carding et al., 2017, Hatfull, 2008). One gram of faeces or mL of faecal filtrates may contain 10^8 to 10^{10} of VLPs based on microscopic enumeration (Hoyles et al., 2014, Lepage et al., 2008). Theoretically, ~10^{10} VLPs with an average genome size of 50 kb equate to around 1 µg of DNA. This suggests that 1 µg of phage DNA can be recovered

from >10 g of faeces containing ~10^9 VLPs/g faeces. In addition, EFM-based correlation analysis (described in **Chapter 5**) reveals that more faeces used broadly equate to more faecal VLPs and VLP DNA recovered. Therefore, I ultimately divided >20 g of faeces into 3-4 g aliquots prior to proceeding and then pooled recovered DNA for sequencing to maximise the recovery of VLP DNA.

I also considered incorporating additional steps in the protocol to purify VLP DNA and reduce contaminants. Typically, chloroform has been used to remove remaining human and bacterial cells. However, chloroform can destabilise and degrade chloroform-sensitive viruses (e.g. rotavirus and polyomavirus) and enveloped viruses (e.g. coronavirus and mimivirus) (Conceicao-Neto et al., 2015). Some bacteriophages belonging to the *Corticoviridae* (dsDNA; non-enveloped prokaryotic virus), *Plasmaviridae* (dsDNA; enveloped prokaryotic virus against mycoplasma host), *Inoviridae* (ssDNA; non-enveloped prokaryotic virus) and *Cystoviridae* (dsRNA; enveloped prokaryotic virus against pseudomonas host) are known to be chloroform-sensitive (Conceicao-Neto et al., 2015, Thurber et al., 2009). Alternatively, CsCl density gradient ultracentrifugation has been used to purity VLP samples for TEM imaging and *in vivo* studies. However, it is laborious, lacks reproducibility and often fails to recover enveloped viruses as well as those of atypical densities (Kleiner et al., 2015, Castro-Mejia et al., 2015). Also, some viruses and bacteriophages are sensitive to CsCl, including *Guttaviridae* (dsDNA; enveloped archaeal virus), *Nanoviridae* (ssDNA; non-enveloped plant virus), and *Orthomyxoviridae* (ssRNA; enveloped eukaryotic virus) (Thurber et al., 2009). Non-chemical based purification methodologies include centrifugal ultrafiltration (Minot et al., 2013). The use of filters with appropriate pore-sizes of the average molecular weight of proteins and viral particles (e.g. 100-1,000 kDa MWCO) can theoretically remove sugars, carbohydrates, proteins and other small molecules, as well as mediate buffer exchange with the retention of VLPs on the membrane. My findings implies that both 300K MWCO and 1,000K MWCO filters are efficient at removing small contamination, such as proteins, sugars, lipids and salts, and in increasing VLP DNA yields. However, the total VLP DNA yields from ultrafiltration are insufficient for NGS. As a result, I discounted the use of chloroform, CsCl and centrifugal ultrafiltration in the protocol for sequence-based virome analysis.

Moreover, I considered the issue of contaminating bacterial DNA and/or RNA which can originate from the faecal samples, the laboratory, and/or the columns, reagents and isolation/extraction kits (Sauvage et al., 2016, Zoll et al., 2015, Salter et al., 2014). I assessed enzyme-based approaches to reduce and eliminate free, non-viral capsid-protected nucleic acids including human and bacterial DNA/RNA, and used GTC/2-ME and/or SDS and proteinase K to destroy nucleases and lyse viral capsids. However, GTC/2-ME treatment produces low DNA yields which may be explained in part by adverse effect

of GTC/2-ME on genomic DNA recovery. A previous literature supports my findings and states that GTC may have adverse effects on high-molecular weight DNA from soil and sediment samples with a reduction in total DNA recovery (Miller et al., 1999). SDS and proteinase K are therefore used to remove protein contaminants and disrupt viral capsids in this protocol.

Having recovered viral capsid DNA, we evaluated DNA extraction and purification methods including conventional phenol/chloroform extraction (Sambrook and Russell, 2001), GTC/2-ME-based P/C/I extraction (Shkoporov et al., 2018b, Murphy et al., 2013) and SDS-based P/C/I extraction (Miller et al., 1999) as well as four commercial isolation/purification kits. Traditional P/C/I produces the highest yield of VLP DNA which is also of high purity. A recent report has evaluated commercial silica-based kits for viral nucleic acid extraction and purification from environmental and faecal samples (MO BIO PowerViral environmental RNA/DNA isolation kit and Zymo Research viral DNA/RNA kit). My evaluation of commercial column-based DNA purification kits shows that these kits recover low amounts and low quality of VLP DNA, consistent with the findings from similar studies (Zheng et al., 2019, Milani et al., 2018). This may be attributable to contaminants such as residual faecal debris, mucins, proteins, lipids and/or polysaccharides within viral lysates that cause columns to clog and to reduce binding affinity of the membrane. These issues may be addressed by using small aliquots of starting material (e.g. 3-4 g faeces). I also noted that some DNA samples are viscous resulting from dietary polysaccharides and/or mucins. To address this issue, CTAB could be further incorporated into the protocol to precipitate and disassociate polysaccharides from VLP DNA in high ionic strength NaCl solutions to reduce viscosity and improve DNA purity and yields (Thurber et al., 2009).

3.5.6. Contaminants in Isolation Kits

Recent reports have highlighted the presence of contaminants in reagents and isolation/purification kits on viral and microbiome metagenomic studies (Sauvage et al., 2016, Zoll et al., 2015, Salter et al., 2014). Genetically engineered nucleases, proteases and polymerases are routinely produced using bacteria-based protein expression systems and the sequences related to the expression vectors have been identified in virome datasets (Zoll et al., 2015). Silica-based columns used in commercial kits may introduce parvovirus-like, circoviruses/densoviruses and iridoviruses sequences into VLP-derived metagenomes (Sauvage et al., 2016, Zoll et al., 2015, Salter et al., 2014). Whilst avoiding the use of such reagents and kits may be preferable, their use is often unavoidable. Hence, it is better to include appropriate control samples in sequencing-based analysis to distinguish and exclude such contaminant-associated sequences (Carding et al., 2017).

3.6. Summary

The final optimised protocol has been applied to the isolation of VLPs and viral DNA from human faecal samples from three independent healthy donors. The recovered DNA has been used to generate PCR-amplified and non-amplified metagenomic DNA libraries for Illumina-based sequencing described in **Chapter 4**.

4. Investigating PCR Amplification Bias from Illumina Sequencing Libraries of the Human Intestinal Viromes

4.1. Introduction

4.1.1. Bias in Random Amplification

To date, the continuing development of metagenomics, bioinformatics tools as well as high-throughput NGS or TGS technologies have enabled the identification of novel viruses and phages, thereby making virome studies possible (Virgin, 2014). However, due to the use of diverse isolation protocols and analytical methodologies, putative bias in every step during sample preparation can impact the outcome of virus detection (Parras-Moltó et al., 2018). One of the principal sources of bias potentially comes from random amplification during sequencing library preparation (Kallies et al., 2019, Parras-Moltó et al., 2018, Aird et al., 2011). Three methods of random amplification are commonly used to enrich VLP DNA of low quantity for virome study, including sequence-independent, single-primer amplification (SISPA, or originally called random PCR) (Djikeng et al., 2008, Froussard, 1992), linker amplification shotgun libraries (LASL) (Breitbart et al., 2002) and multiple displacement amplification (MDA) (Angly et al., 2006, Dean et al., 2001).

The main drawback of random amplification is that each method potentially has its own bias which may affect the relative abundance and diversity of viruses. For example, SISPA replies on pseudo-degenerate oligonucleotides as a primer, composed of 6-12 random oligonucleotides at its 3'-end and around 20 defined nucleotides at the 5'-end, and can detect both DNA and RNA viral sequences (Djikeng et al., 2008, Froussard, 1992). However, amplification bias in SISPA may lead to an uneven read distribution across the target genomes in comparison with non-amplified samples. Some read-uncovered genetic regions tend to be overrepresented and therefore, the result may shift toward artificially dominant viral genomes (Karlsson et al., 2013). In addition, this bias can make SISPA less sensitive for detecting low-abundant viruses (Karlsson et al., 2013). LASL replies on ligating linkers/adapters to the ends of fragmented DNA molecules prior to PCR amplification (Breitbart et al., 2002). However, its efficiency may be affected by using DNA templates of low-quantity of pico- or nanogram amounts (Solonenko et al., 2013, Duhaime et al., 2012). Also, the bias associated with extreme GC content inherent in PCR amplification is likely present in LASL method (Duhaime et al., 2012). Moreover, LASL protocol may be less sensitive to detect ssDNA viruses (Szekely and Breitbart, 2016), although recently modified LASL protocols have been reported to improve the recovery of ssDNA viruses (Roux et al., 2016). MDA is not a PCR-based amplification method. Instead, it relies on random priming to DNA under isothermal conditions with very few DNA templates, random hexamer primers

and phi29 polymerase (Dean et al., 2001). Although MDA has been widely used for virome studies and benefits from converting single-stranded cDNA into a double-stranded form during reverse transcription (Shkoporov et al., 2018b), several strong biases have also been reported for phi29 amplification, including chimera formation (Lasken and Stockwell, 2007), discontinuous amplification of linear dsDNA (Zhang et al., 2006a) and preferential amplification for short circular ssDNA molecules (Kim et al., 2008).

A recent study systematically evaluated biases in the process of VLP enrichment and in different amplification protocols including SISPA and MDA in human saliva viromes using qPCR and HTS technology (Parras-Moltó et al., 2018). In this study, low-speed centrifugation followed by different pore-sized filtration or iodixanol cushion for enrichment (i.e. 0.22 μm, 0.45 μm, iodixanol cushion only or 0.22+0.45+iodixanol cushion) were assessed. Biases introduced by SISPA and MDA protocols were identified in human saliva viromes with SISPA tending to generate uneven read coverage, and also, sources of biases in MDA may be altered depending on DNA sample size. For example, MDA tended to introduce stochastic bias using picogram (e.g. 10 pg) amounts of input DNA, whereas it shifted to systematic bias with nanogram (e.g. 1 ng) amounts of DNA. Although increasing sample size for MDA is likely to introduce systematic bias, thereby leading to over-amplification of circular ssDNA viral genomes and under-amplification of extreme GC-related regions, the findings indicated that random amplification influences the relative abundance of the human saliva viromes but may only have a minor impact on inter-subject beta diversity (Parras-Moltó et al., 2018). Recently, updated protocols using alternative methods for library construction were reported to minimise bias for circular ssDNA virus detection (Roux et al., 2016, Zhong et al., 2015, Gansauge and Meyer, 2013). A similar virome study evaluated three random amplification protocols to identify potential bias using environmental samples (Kallies et al., 2019). Their findings revealed that LASL protocol displays the lowest level of read redundancy and the highest viral diversity, whereas MDA tended to generate higher abundance of viral contigs with the longest contig size, suggesting that adopting a combination strategy of LASL and MDA methods can be considered to enrich VLP DNA (Kallies et al., 2019).

4.1.2. Comparison of LASL and Non-Amplified Shotgun Library (NASL)

For LASL, the standard workflow of Illumina sequencing library preparation involves multiple steps including PCR amplification using universal primer sets to reduce amplification bias and primer dimers. However, bias cannot be completely overcome while introducing PCR into library preparation. To minimise amplification bias, particularly in the extremes of high or low G+C content, Kozarewa and colleagues developed a method for Illumina sequencing without using PCR to amplify adapter/linker-ligated DNA templates (Kozarewa et al., 2009). Unlike PCR-based protocol, fragmented, end-repaired and 3' adenine (A)-tailed DNA template is generated, followed by partially ligating complementary adapters to fragmented input DNA, which contain additional short sequences that can facilitate DNA hybridisation to the surface of the flow cell without incorporating PCR step (**Figure 4.1**) (Kozarewa et al., 2009). In another similar study, Aird and colleagues also investigated the sources of base-composition bias and made an attempt to develop a modified protocol by optimising PCR conditions to minimise amplification bias in Illumina sequencing library (Aird et al., 2011). Their finding revealed that PCR amplification step during library preparation is the principal source of bias (Aird et al., 2011).

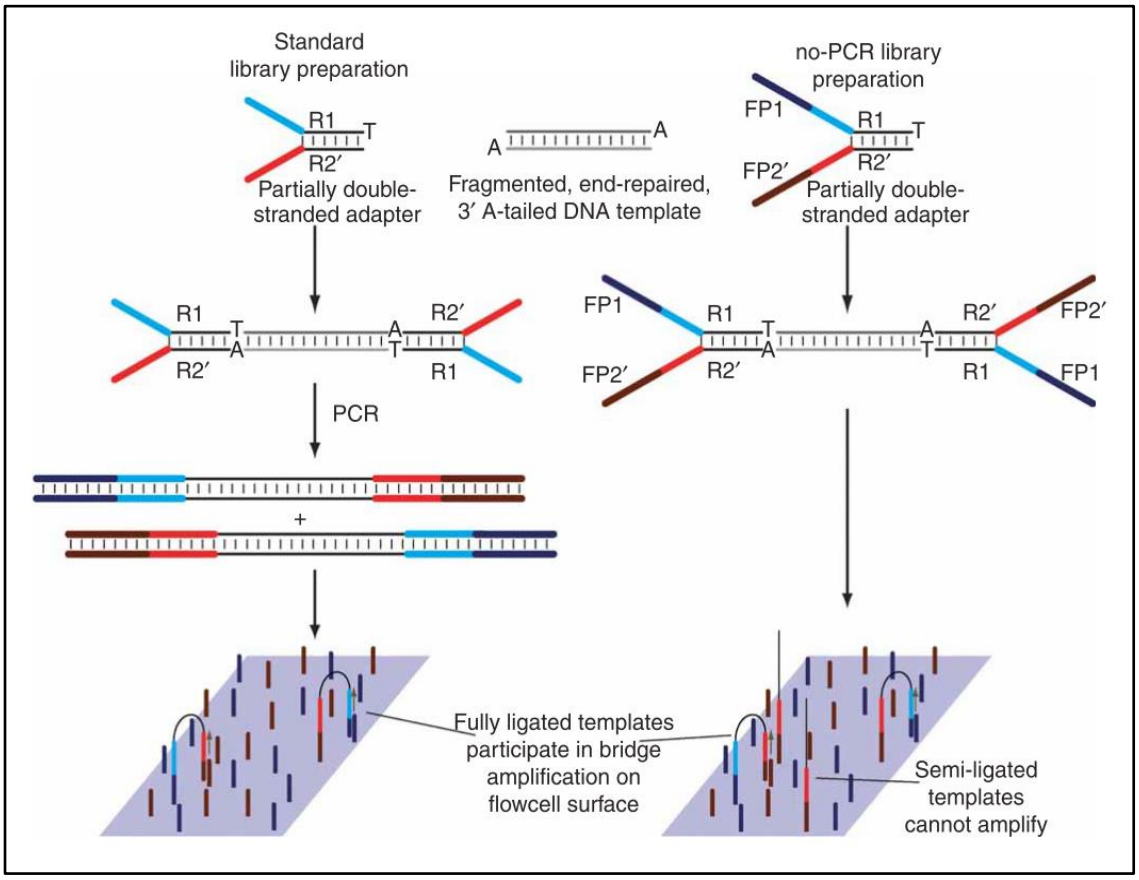


Figure 4.1. PCR and non-PCR library preparations. In both of the standard LASL protocol (left panel) and PCR-free NASL protocol (right panel), DNA is firstly fragmented, end-repaired and A-tailed on 3'-end, followed by partially ligating complementary adapters (R1 and R2') onto the DNA template. Unlike standard protocols, non-PCR adapters include additional short sequences (FP1 and FP2') that can facilitate DNA hybridisation to the surface of the flow cell without the retention of PCR step. Due to only FP2' being reverse-complementary to the oligonucleotides on the surface of the flow cell, all non-PCR templates can be hybridised to the flow cell in the same orientation. Illumina cluster amplification can therefore amplify and enrich template strands having a different adapter on either end. Image has been taken from (Kozarewa et al., 2009), with permission from Springer Nature (Copyright © 2009).

Thus, in this study, I evaluated the methods of amplification-based and non-amplified sequencing library preparations and investigated if amplification leads to bias in virome-derive PCR datasets and therefore impacts the relative abundance and diversity of intestinal viromes.

4.2. Aim

In this collaborative study, I aimed to develop a bioinformatics pipeline for viromic analysis in cooperation with Dr Mohammad Adnan Tariq (Carding group, QIB), Dr Evelien Adriaenssens (research leader, Gut Microbes & Health, QIB), Dr Andrea Telatin (QIB bioinformatics team) and Dr Rebecca Ansorge (QIB bioinformatics team). In parallel, to investigate if PCR amplification during sequencing library preparation introduces significant bias present in VLP-enriched intestinal viral metagenomes compared with a non-amplified sequencing library method. The finalised bioinformatics pipeline was then applied to an initial analysis of ME/CFS patient and SHHC samples.

4.3. Study Design

Three purified VLP DNA samples isolated from faeces of independent healthy donors were aliquoted for the preparations of PCR-based and non-PCR libraries for Illumina sequencing to develop an optimised bioinformatics pipeline for VLP-enriched virome analysis and to identify PCR-associated biases in Illumina sequencing datasets (**Figure 4.2**).

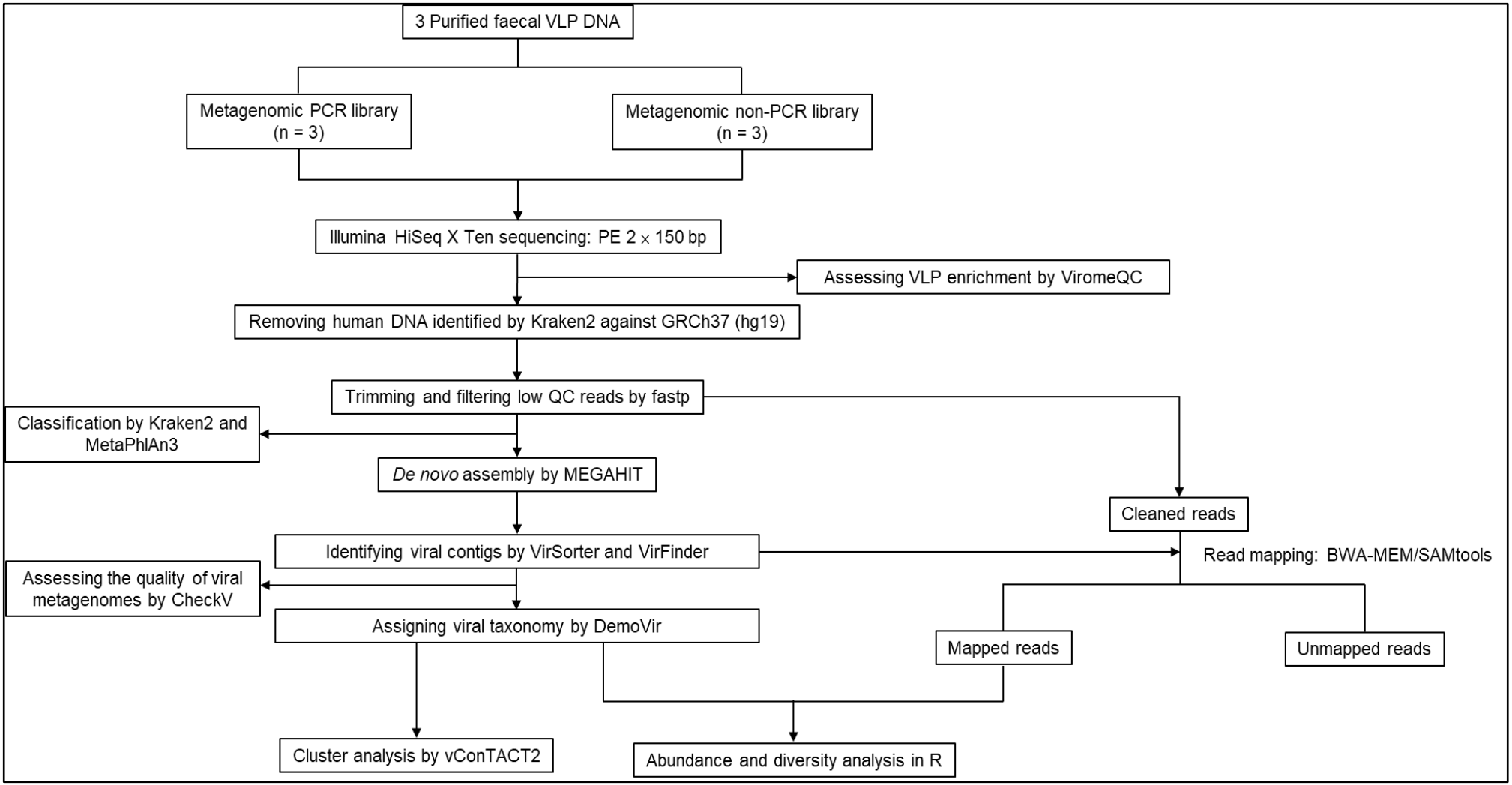

3 Purified faecal VLP DNA
Metagenomic PCR library (n = 3)
Metagenomic non-PCR library (n = 3)
Illumina HiSeq X Ten sequencing: PE 2 × 150 bp
Assessing VLP enrichment by ViromeQC
Removing human DNA identified by Kraken2 against GRCh37 (hg19)
Trimming and filtering low QC reads by fastp
Classification by Kraken2 and MetaPhlAn3
De novo assembly by MEGAHIT
Cleaned reads
Identifying viral contigs by VirSorter and VirFinder
Read mapping: BWA-MEM/SAMtools
Assessing the quality of viral metagenomes by CheckV
Assigning viral taxonomy by DemoVir
Mapped reads
Unmapped reads
Cluster analysis by vConTACT2
Abundance and diversity analysis in R

Figure 4.2. Overview of bioinformatic pipeline for cross-comparative virome study. Three faecal VLP DNA samples were aliquoted for PCR-based and non-amplified libraries for Illumina sequencing. Human DNA sequences identified by kraken 2 against GRC37 (Genome Reference Consortium; hg19) was removed, followed by trimming and filtering low quality reads using fastp prior to *de novo* assembly. In parallel, both Kraken 2 and MetaPhlAn 3 were used to taxonomically classify cleaned reads. MEGAHIT was used to assemble Illumina short reads into longer contigs, followed by detecting potential viral candidates using VirSorter and VirFinder. BWA/SAMtools were then used for reads mapping, followed by analysing mapped viral reads and visualising the outcome of PCR and non-PCR datasets in relative abundance, alpha and beta diversity. In parallel, the qualities of both pooled non-redundant viral contigs from PCR and non-PCR datasets were assessed using CheckV, respectively. In parallel, DemoVir was used to assign taxonomy to these viral contigs, followed by performing cluster analysis to compare UViG similarity between both datasets using vConTACT 2.0. For those unknown viral sequences which cannot be identified and annotated by DemoVir, vConTACT 2.0 was used to infer their taxonomy at the family level. In addition, ViromeQC was used to evaluate my VLP isolation protocol, based on the extent of VLP enrichment.

4.4. Results

4.4.1. Quality and Quantity of Raw Sequencing Output

Using my optimised protocols for isolating faecal VLPs and VLP DNA (**Chapter 3**), three DNA samples of high quality and quantity were aliquoted for library preparation and sequencing. With different strategies for sequencing library preparation using LASL and NASL protocols, a total of six shotgun metagenomic DNA libraries (PCR_1-3 and non-PCR_1-3 datasets) were sequenced on Illumina HiSeq X Ten to generate a total of 54,202,065 cleaned reads from unamplified libraries (18,067,355 ± 1,914,183.5, mean ± S.D.; n = 3) (**Table 4.1**) and 62,693,258 cleaned reads from PCR libraries (20,897,753 ± 3,471,445.7, mean ± S.D.; n = 3) (**Table 4.2**). Overall, all sequencing datasets reached the minimal Phred score of 30 (Q30) over 90%, having high sequencing depth with 32-48 million paired-end reads generated per sample. Generally, in total PCR library datasets produced more raw reads than those of non-PCR library datasets (62,974,104 reads in PCR vs. 56,085,444 reads in non-PCR).

Table 4.1. Quality and quantity of non-PCR datasets

Sample	# of raw reads	# of cleaned reads	Q30 (%)	%GC
1	18,647,093	18,072,149	93.62	36.79
2	16,540,559	16,150,779	90.91	46.71
3	20,897,792	19,979,137	90.72	39.42

Table 4.2. Quality and quantity of PCR datasets

Sample	# of raw reads	# of cleaned reads	Q30 (%)	%GC
1	17,618,631	17,547,247	90.29	36.05
2	20,715,201	20,667,351	92.47	46.89
3	24,640,272	24,478,660	92.22	38.48

4.4.2. Evaluating VLP enrichment

To evaluate the extent of faecal VLP enrichment during the isolation procedure, ViromeQC was used to computationally calculate the abundance of microbial markers and generate an enrichment score for each virome dataset by aligning to the selected microbial markers, compared to default non-enriched metagenomes as a baseline. **Table 4.3** showed that VLP enrichment of PCR-based virome datasets was relatively high, ranging from 4.16 to 11.04, which means the viromes in the amplified datasets were 11.04- (PCR-1), 4.16- (PCR-2) and 5.13-fold (PCR-3) enriched, respectively, while the enrichment scores of non-amplified virome datasets were relatively low, ranging from 0.6 to 1.17.

Table 4.3. Evaluation of VLP enrichment

Dataset	Reads	Reads_HQ	SSU rRNA alignment rate	LSU rRNA alignment rate	Bacterial markers alignment rate	Total enrichment score
non-PCR-1	37,294,186	37,140,565	0.032813179	0.88413033	0.088235061	0.599761277
non-PCR-2	33,081,118	32,889,577	0.13604614	0.453991853	0.257458465	1.168010247
non-PCR-3	41,795,584	41,435,311	0.097902004	0.694588729	0.164224663	0.763426059
PCR-1	35,237,262	34,998,689	0.0221094	0.040795814	0.063479521	11.04123903
PCR-2	41,430,402	41,300,288	0.059411208	0.108040409	0.166175597	4.161744495
PCR-3	49,280,544	49,006,610	0.048191458	0.088392158	0.096301295	5.130665822

HQ: high quality; SSU-rRNA: small subunit rRNA gene; LSU-rRNA: Large subunit rRNA gene

4.4.3. Quality and Quantity of Genome Assembly

Cleaned reads of each dataset were individually assembled into contigs yielding 280,041 contigs from non-PCR datasets and 208,529 contigs from PCR datasets, respectively. Across both library datasets, sample 1 produced 173,739 contigs, sample 2 produced the maximum yields, having 234,574 contigs, and sample 3 had the minimum output, with 80,257 contigs in total (**Table 4.4**).

Of these, non-PCR datasets yielded 50,707 contigs of length ≥1 kb across the samples, while PCR datasets produced 36,848 contigs of length ≥1 kb. Across both library datasets, sample 1 generated 32,420 contigs of length ≥1 kb, sample 2 generated 44,366 contigs of length ≥1 kb and sample 3 generated 10,769 contigs of length ≥1 kb (**Table 4.4**).

In addition, **Table 4.4** showed that in non-PCR datasets, the mean of N50 per sample was around 2,313.3 ± 285.9 bp (mean ± S.D., n = 3), while in PCR datasets, the mean of N50 per sample was around 3,858.3 ± 2,774.7 bp (mean ± S.D., n = 3).

Table 4.4. Statistics of MEGAHIT assemblies

	non-PCR library			PCR library		
Sample	**# of total contigs**	**# of contigs ≥1 kb**	**N50 (bp)**	**# of total contigs**	**# of contigs ≥1 kb**	**N50 (bp)**
1	100,493	17,951	2,019	73,246	14,469	2,011
2	131,354	26,414	2,331	103,220	17,952	2,515
3	48,194	6,342	2,590	32,063	4,427	7,049

4.4.4. Identifying Putative Viruses

To predict and identify putative uncultivated virus genomes (UViGs), VirSorter and VirFinder were used. Based on my strategy, these putative viral candidates sorted as VirSorter categories 1-6, including all linear and circular completed viruses (i.e. categories 1-3) and proviruses/prophages (i.e. categories 4-6), and predicted by VirFinder with a sorting condition score of ≥0.7 and p <0.05 considered viral (**Table 4.5**).

Table 4.5 showed that in total 701 putative UViG contigs in PCR datasets and 694 of those in non-PCR datasets were detected by VirSorter across the samples, respectively. Of these, 244 putative UViGs were present in PCR-1 dataset, 269 of which were found in PCR-2 dataset and 188 were in PCR-3 dataset. Also, 297 of UViGs were detected from non-PCR-1 dataset, 221 of which were in non-PCR-2 dataset and 176 were in non-PCR-3 dataset, respectively.

In total 27,832 putative UViG contigs in the PCR datasets and 34,418 putative UViG contigs in the non-PCR datasets were predicted by VirFinder, respectively. Of these, 8,609 putative UViGs were detected from PCR-1 dataset, 15,228 of which were found in PCR-2 dataset and 3,995 were in PCR-3 dataset Also, 12,392 of putative UViGs were found in non-PCR-1 dataset, 16,494 of which were in non-PCR-2 dataset and 5,532 were in non-PCR-3 dataset, respectively (**Table 4.5**).

By pooling UViGs together and removing redundancy across the samples, we found that the number of non-redundant UViGs of >1 kb was 17,898 and 19,591 in PCR datasets and non-PCR datasets, respectively (**Appendix 3**). On the other hand, 57,912 non-redundant UViGs <1kb were in the PCR datasets with 68,116 non-redundant UViG <1 kb in the non-PCR datasets, respectively (**Appendix 4**). Collectively, the number of short viral genomic contigs (<1kb) exceeded that of longer viral contigs in both datasets.

Table 4.5. Putative viruses/proviruses detected by VirSorter and VirFinder

Dataset	# of VirSorter-detected UViGs/UpViGs (categories 1-6)	# of VirFinder-detected UViGs/UpViGs (score ≥0.7 and p <0.05)
PCR-1	244	8,609
non-PCR-1	297	12,392
PCR-2	269	15,228
non-PCR-2	221	16,494
PCR-3	188	3,995
non-PCR-3	176	5,532

UViGs: uncultivated virus genomes; UpViGs: uncultivated provirus genomes

Based on my sorting criteria, **Figure 4.3** showed that in PCR-1 dataset, 87 shared UViG sequences considered as potential true-positive viruses were identified by both tools. Also, 144 shared UViGs were found in non-PCR-1 dataset, 122 of which were in PCR-2 dataset, 105 were in non-PCR-2 dataset, 100 and 97 of which were in PCR-3 and non-PCR-3 datasets, respectively. Venn diagrams illustrate the number of potential true-positive UViGs identified by both tools (i.e. the overlaps) and the unique UViGs individually detected by each tool, respectively.

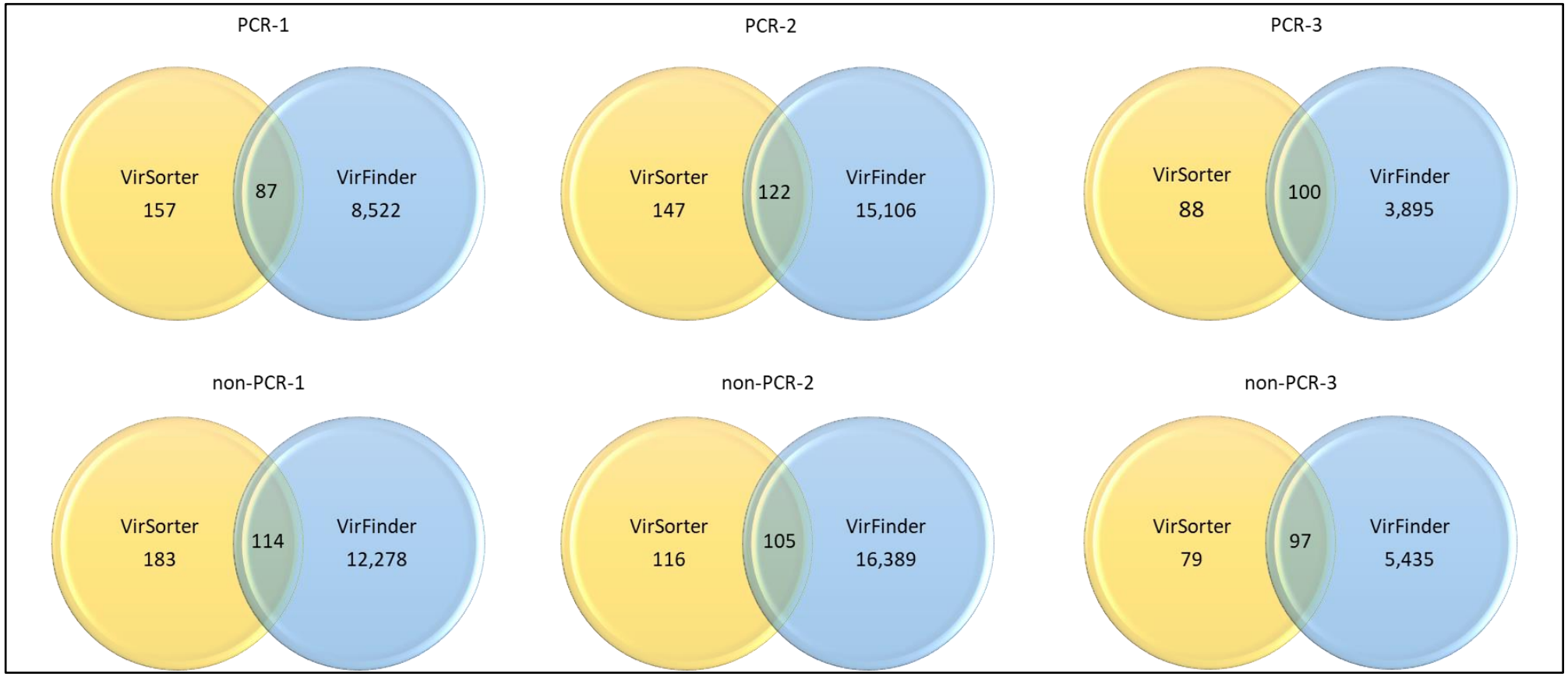


Figure 4.3. Unique and shared UViGs/UpViGs detected by VirSorter and VirFinder. Venn diagrams representing the six datasets show the numbers of unique and overlapped viral contigs identified by the two tools.

4.4.5. Read Mapping

After *de novo* assembly, filtered reads from each dataset were independently mapped against the total contigs of each dataset to the reference genome sequences. Part I of **Table 4.6** showed that cleaned reads of between 32.9 and 47.3 million from the virome-derived library datasets were individually mapped to the reference genomes, with high mapping rates reaching between 96% and 98.7% for each.

After identifying viruses, cleaned reads from each dataset were then independently mapped to the pooled non-redundant UViG/UpViG (uncultivated provirus genomes) sequences to the reference genomes. Part II of **Table 4.6** indicated that in PCR-1 and non-PCR-1 datasets, 90.1% and 87.9% of cleaned viral reads were mapped to the reference viral genomes, respectively, with mapped reads of between 30 and 32 million. In PCR-2 and non-PCR-2 datasets, around 80.7% and 72.9% of cleaned viral reads were mapped, respectively, with mapped reads of between 23 and 33 million. Also, in PCR-3 and non-PCR-3 datasets, there were 42.2 and 33.5 million of cleaned viral reads mapped to the reference viral genomes, with the mapping rates reaching up to 88.2% and 83.6%, respectively.

Overall, the PCR-3 and non-PCR-3 datasets yielded the maximum numbers of mapped viral reads, compared with other datasets of sample 1 and 2, respectively, while the PCR-1 and non-PCR-1 datasets had higher mapping rates against reference viral genomes, reaching up to 90%. In addition, non-PCR 2 datasets yielded the fewest number of mapped viral reads and both PCR-2/non-PCR-2 datasets had the minimum mapping rates to reference viral genomes, compared with other datasets.

Table 4.6. Statistics of mapped and unmapped reads from amplified and non-amplified library datasets

		PCR-1	non-PCR-1	PCR-2	non-PCR-2	PCR-3	non-PCR-3
Part I.	# of filtered reads (paired-ended)	33,872,924	36,122,668	40,749,416	32,048,918	47,904,786	40,102,922
	# of reads mapped to total contigs	32,853,847	35,035,801	39,642,037	30,762,051	47,299,101	39,089,614
	Mapping rate to total contigs (%)	97.0%	97.0%	97.3%	96.0%	98.7%	97.5%
	# of unmapped reads	1,019,077	1,086,867	1,107,379	1,286,867	605,685	1,013,308
	Unmapping rate (%)	3.0%	3.0%	2.7%	4.0%	1.3%	2.5%
Part II.	# of reads mapped to UViGs/UpViGs	30,533,123	31,754,308	32,885,103	23,357,458	42,241,639	33,515,173
	Mapping rate to UViGs/UpViGs (%)	90.1%	87.9%	80.7%	72.9%	88.2%	83.6%
	# of unmapped reads	3,339,801	4,368,360	7,864,313	8,691,460	5,663,147	6,587,749
	Unmapping rate (%)	9.9%	12.1%	19.3%	27.1%	11.8%	16.4%

4.4.6. Comparative Study Between Virome-Derived PCR and non-PCR Datasets

4.4.6.1. Relative Abundance Analysis

To investigate if PCR amplification impacts VLP-enriched intestinal viromes, virome-derived datasets were compared in terms of relative abundance at the family level. Non-redundant UViGs >1kb from both PCR and non-PCR datasets were used for further comparisons. **Figure 4.4** displayed the top 25 UViG sequences. **Figure 4.4.A** indicated that the intestinal virome profiles were considerably different and unique in each individual sample. Importantly, from the top 25 UViG data, I noted that there were some differences in their relative abundances. In sample 1, a virus (contig s18v3_PF_45906) assigned to a new family of *Bacteroides* phages was only seen in non-PCR-1 dataset rather than in PCR-1. Also, there were slight differences in the abundance of other viruses found in sample 1, comparing PCR to non-PCR dataset, such as the UViG contigs s18v1−PF_87044 (unassigned family), s18v1−PF_72494 (unassigned family), s18v1−PF_72285 (*Myoviridae*), s18v1−PF_47329 (unassigned family), s18v1_7483 (unassigned family), s18v1_39930 (unassigned family) and s18v1_17930 (*Myoviridae*).

In sample 2, no additional viruses were seen in either PCR or non-PCR datasets among the top 25 UViGs. However, there were also slight or moderate differences in other viruses identified in both datasets, including the contigs s18v2−PF_39665 (*Siphoviridae*), s18v2−PF_119664 (assigned to new_family_VC_442), s18v2_29464 (*Siphoviridae*), s18v2_12164 (*Siphoviridae*), s18v1−PF_40598 (*Siphoviridae*), s18v1−PF_17802 (assigned to a new family VC_442), s18v1_31420 (*Salasmaviridae*, a new proposed family belonging to the order *Caudovirales*) and s18v1_14784 (unassigned family).

In sample 3, a virus (contig s18v2_29464) assigned to *Siphoviridae* family and a virus (contig s18v1−PF_17802) assigned to a new family VC_442 were only seen in non-PCR-3 dataset rather than in PCR-3. Moreover, there were also slight or moderate differences in other viruses found in both datasets, including contigs s18v3−PF_48099 (*Siphoviridae*), s18v3−PF_4641 (*Podoviridae*), s18v3−PF_45906 (assigned to a new family of *Bacteroides* phages), s18v3−PF_40610 (unassigned family), s18v3−PF_11139 (*Podoviridae*), s18v3_9193 (*Podoviridae*) and s18v3_4677 (unassigned family). Overall, from the top 25 viral genomic contigs, the most abundant UViGs across the samples were Siphoviruses, followed by Podoviruses only seen in sample 3.

Figure 4.4.B showed top 25 UViGs grouped together across all datasets. Overall, the top 1 to 9 UViGs seen in each corresponding dataset could not be assigned appropriate taxonomy at the family level by neither DemoVir nor vConTACT 2.0. Of those assigned,

Siphoviridae was the most abundant virus in all datasets, followed by *Salasmaviridae* family seen in both sample 1 and 2, and *Podoviridae* family seen in sample 3. *Myoviridae* family was the least abundant and was only seen in sample 1, with no differences in relative abundance between both datasets.

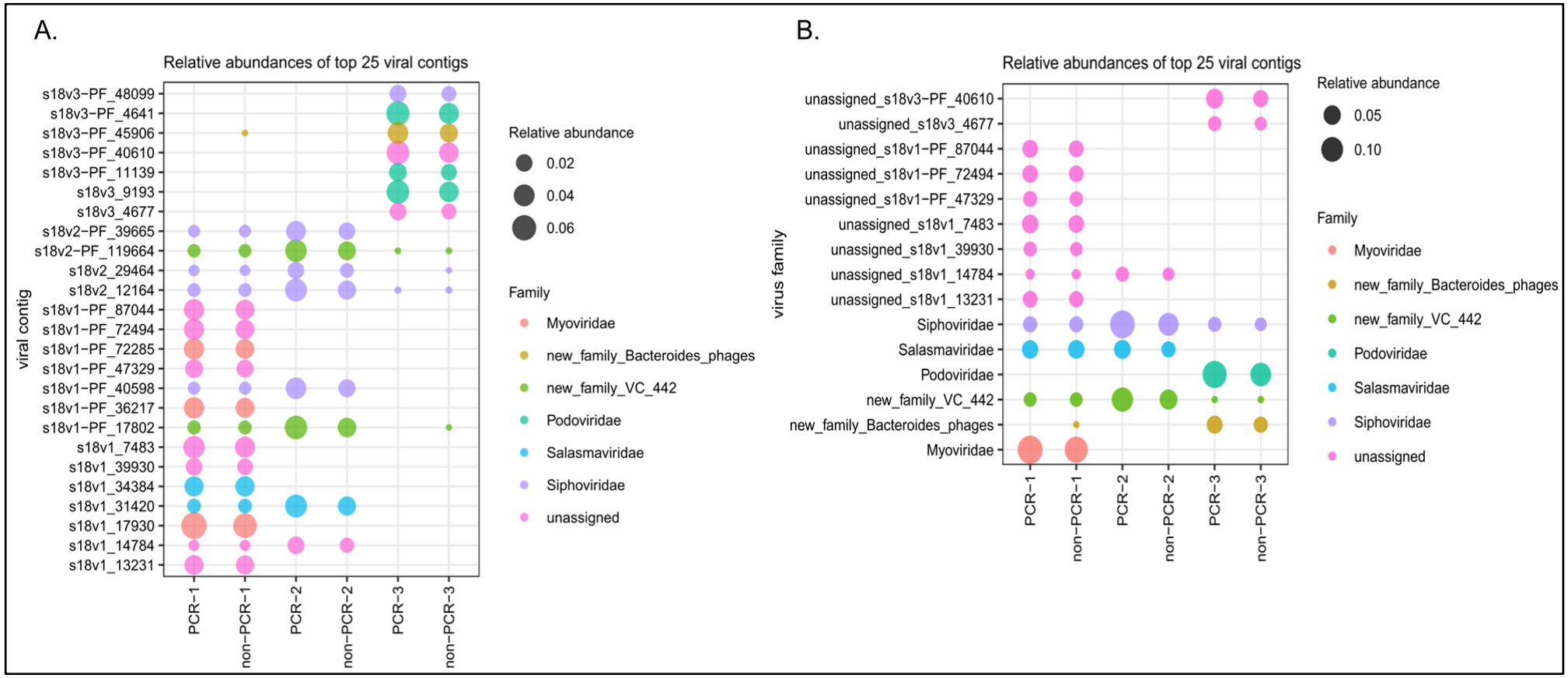


Figure 4.4. Relative abundance of top 25 viral sequences from the PCR and non-PCR datasets. (A) A bubble plot showing the relative abundance of the top 25 viral contigs with assigned and unassigned taxonomies. (B) A bubble plot shows top 25 viral contigs grouped together across the datasets. Those viral contigs that could not be classified and assigned to appropriate taxonomies by neither DemoVir nor vConTACT 2.0 were labelled as “unassigned”. The size of bubble represents the relative abundance expressed as percentage (e.g. 0.02 means 2%) and each assigned or unassigned family group was separated in colour.

4.4.6.2. Alpha Diversity Analysis

To compare intra-subject differences in sample richness and alpha diversity, all UViG/UpViG sequences based on normalised count matrices were indicated by “observed richness”, “Chao1”, “Shannon” and “Simpson” indices (**Figure 4.5**).

For observed richness (**Figure 4.5.A**), the actual numbers of UViGs/UpViGs in non-PCR datasets (7,760 in non-PCR-1, 7,581 in non-PCR-2 and 4,250 in non-PCR-3) were higher than those in PCR datasets (6,749 in PCR-1, 7,354 in PCR-2 and 3,795 in PCR-3). For Chao1 estimator (**Figure 4.5.B**), the result of richness estimates in sample 1 were inverted (10,133.1 ± 176.9 in PCR-1 vs. 9,527.2 ± 98.6 in non-PCR-1, mean ± S.E.M.), while the Chao1 estimates of non-PCR-2 and non-PCR-3 datasets were higher than those of PCR-2 and PCR-3 datasets. However, after rarefaction (**Appendix 5**), the Chao1 estimate of non-PCR-1 (7,315.3 ± 60.0, mean ± S.E.M.) was higher than that of PCR-1 (6,540.0 ± 67.0, mean ± S.E.M.), having similar results to unrarefied observed richness.

For Shannon index estimations of both richness and evenness (**Figure 4.5.C**), the estimates of non-PCR datasets (approximately 6.19 in non-PCR-1, 7.39 in non-PCR-2 and 6.53 in non-PCR-3) were higher than those of PCR datasets (approximately 5.71 in PCR-1, 6.30 in PCR-2 and 5.95 in PCR-3). For Simpson index considering the evenness (**Figure 4.5.D**), the estimates of non-PCR datasets (approximately 0.989 in non-PCR-1, 0.995 in non-PCR-2 and 0.993 in non-PCR-3) were generally higher than those of PCR datasets (approximately 0.985 in PCR-1, 0.986 in PCR-2 and 0.986 in PCR-3).

Collectively, the richness and alpha diversity of the faecal viromes in non-PCR datasets were generally higher than those in PCR datasets, except the result of Chao1 estimator in unrarefied PCR/non-PCR datasets of the sample 1.

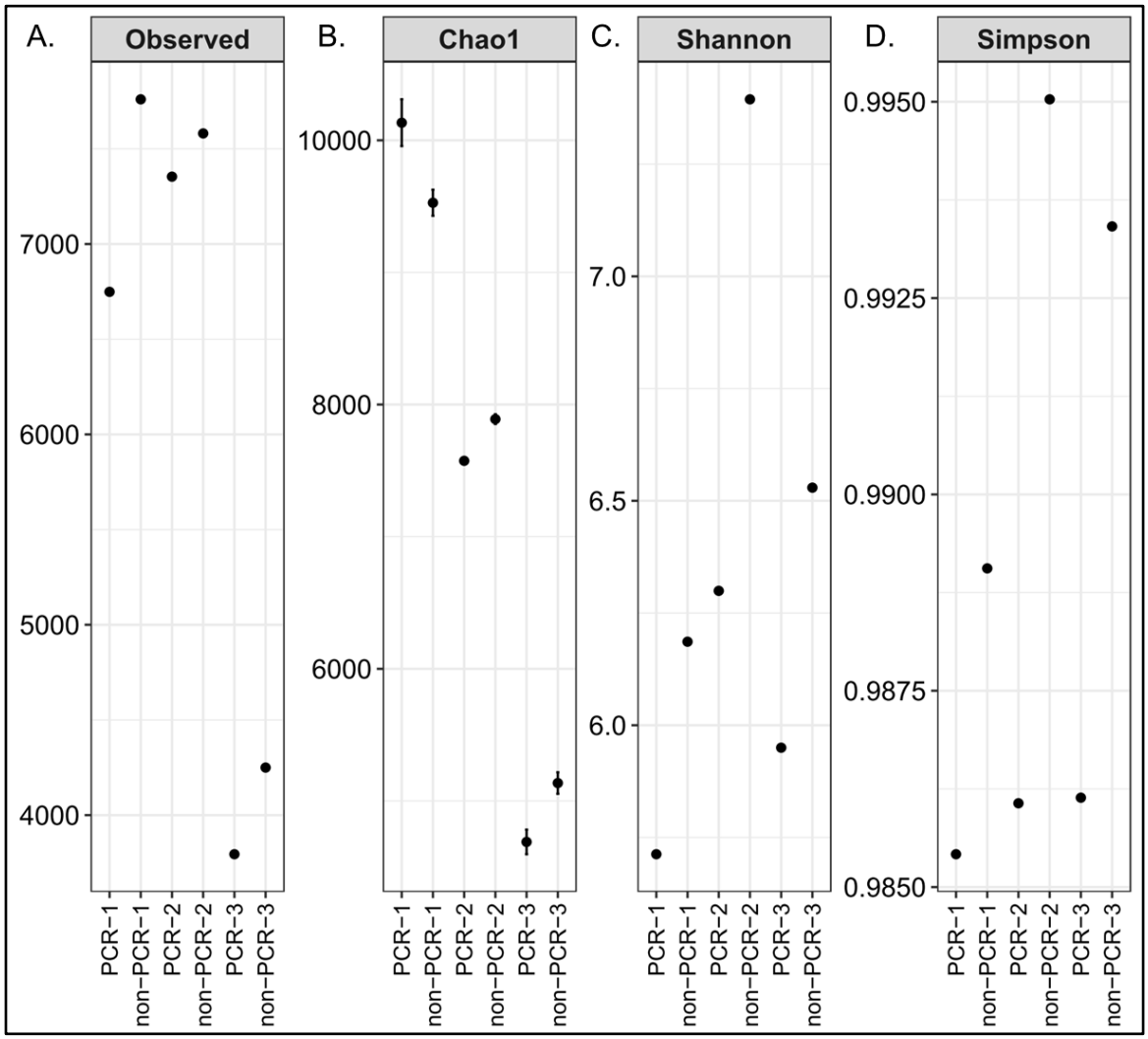


Figure 4.5. Estimation of alpha diversity from virome-derived PCR and non-PCR datasets. (A) The number of UViGs directly observed from a normalised count matrix. (B) Estimation of richness using Chao1 index for six virome-derived library datasets. (C) Estimation of Shannon index for the virome datasets. (D) Estimation of Simpson index for the virome datasets. The estimated values of unfiltered UViGs (>1kb) are indicated with black dots and the error bars of Chao1 represent standard error of the mean (S.E.M.).

4.4.6.3. Beta Diversity Analysis

To evaluate the impact of amplification in the whole viral community, we compare inter-subject differences in the beta diversity by computing the distance matrices of Bray-Curtis dissimilarities and Jensen-Shannon divergence (JSD) among all datasets (**Figure 4.6**). For the Bray-Curtis-based PCoA (**Figure 4.6.A**), over 95% of dissimilarities were captured from these datasets. The ordination plot showed that the locations between sample clusters were farther than those between PCR and non-PCR datasets in each sample. For each sample cluster, although PCR dataset localised closer to non-PCR dataset, their locations were still different, showing similar results to JSD-based PCoA with having around 100% of dissimilarities being statistically explained (**Figure 4.6.B**).

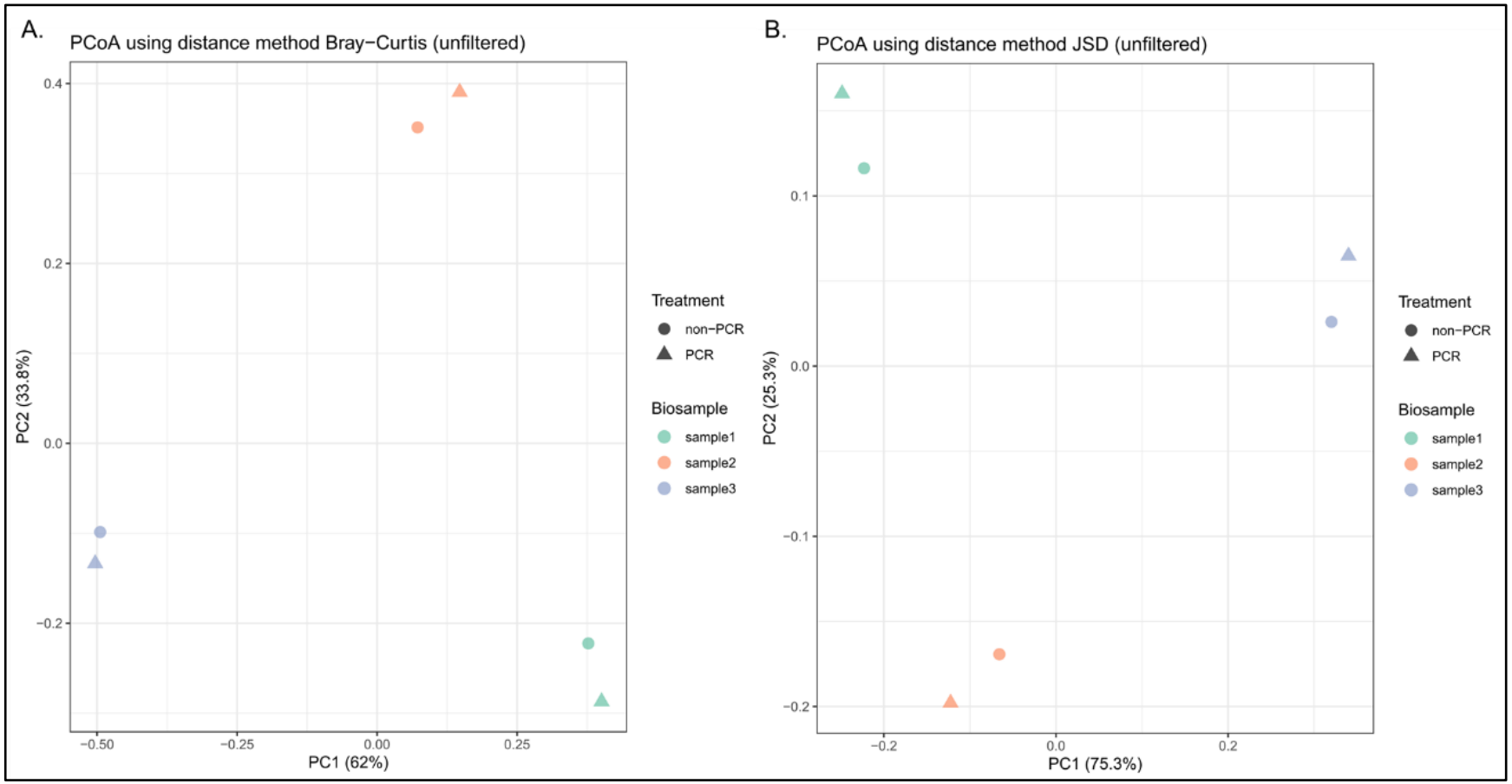
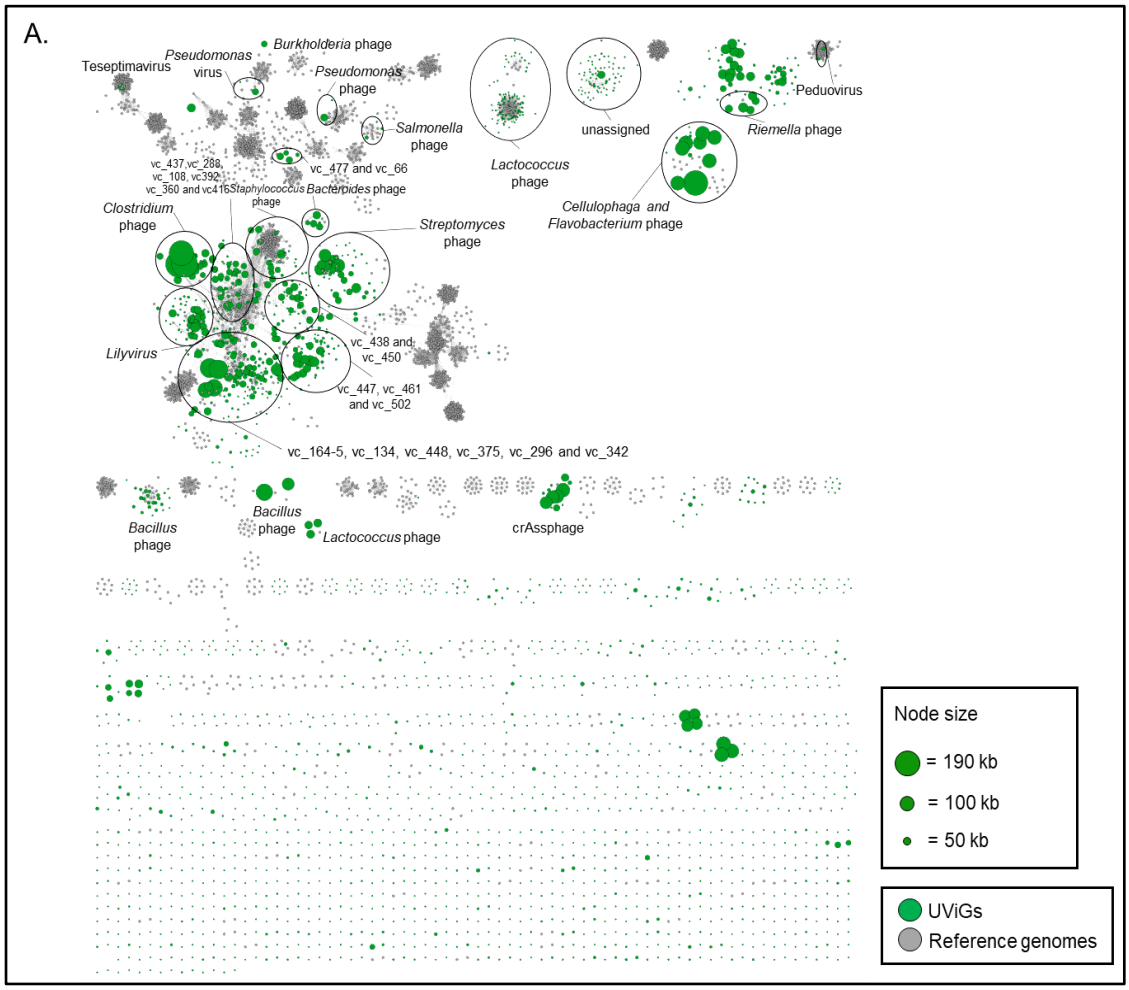


Figure 4.6. Ordination analysis of faecal viromes from unfiltered PCR and non-PCR datasets. PCoA plots based on Bray-Curtis distances (A) and Jensen-Shannon divergence (B) were used to interpret the faecal virome-derived datasets (unfiltered data). The circle represents the non-PCR dataset and the triangle represents the PCR dataset. Each sample was separately displayed in different colours.

4.4.6.4. Cluster Analysis: Sequence Similarity Networks

To further investigate the similarity of UViG sequences between PCR and non-PCR datasets, a genome-based network analysis was carried out on grouping UViGs into viral clusters (VCs) based on shared amino acid homology (**Figure 4.7**). In total, 1,132 VCs were identified in PCR datasets (**Figure 4.7.A**) and 763 VCs in non-PCR datasets (**Figure 4.7.B**), respectively. Overall, **Figure 4.7.A** showed that some nodes of UViGs (particularly larger-sized contigs) were connected to form a major network against reference genomes and were grouped into nine VCs assigned to potentially representative viruses including *Clostridium* phage, *Streptomyces* phage, *Lilyvirus*, *Staphylococcus* phage, *Bacteroides* phage and newly proposed taxonomies such as vc_438 and vc_450. Moreover, a majority of UViGs formed orphan VCs with unassigned or potential representatives, such as *Lactococcus* phage, *Cellulophaga* and *Flavobacterium* phages, *Riemella* phage, *Peduovirus*, *Bacillus* phage and crAssphage, showing either no or few connections to each other.

Figure 4.7.B showed that some nodes of UViGs were connected to form a major network and were grouped into around eleven VCs assigned to potential representatives, including *Streptomyces* phage, *Staphylococcus* phage, *Bacteroides* phage, *Streptococcus* phage, *Clostridium* phage, *Pseudomonas* phage and some newly proposed taxonomies such as vc_432, vc_444,vc_448-9 and vc_421-2. Moreover, similar to the PCR datasets, there were many orphan VCs in non-PCR datasets with unassigned or potentially assigned representatives, such as *Cellulophaga* phage, *Flavobacterium* phage, *Lactococcus* phage, *Riemella* phage, *Burkholderia* phage, *Bacillus* phage as well as crAssphage, showing either no or few connections to each other.


A.
Teseptimavirus
Pseudomonas virus
Burkholderia phage
Pseudomonas phage
Salmonella phage
vc_437,vc_288, vc_108, vc392, vc_360 and vc416
Staphylococcus phage
vc_477 and vc_66
Bacteroides phage
Clostridium phage
Streptomyces phage
Lilyvirus
vc_438 and vc_450
vc_447, vc_461 and vc_502
vc_164-5, vc_134, vc_448, vc_375, vc_296 and vc_342
Bacillus phage
Bacillus phage
Lactococcus phage
crAssphage
Lactococcus phage
unassigned
Peduovirus
Riemella phage
Cellulophaga and Flavobacterium phage
Node size
= 190 kb
= 100 kb
= 50 kb
UViGs
Reference genomes

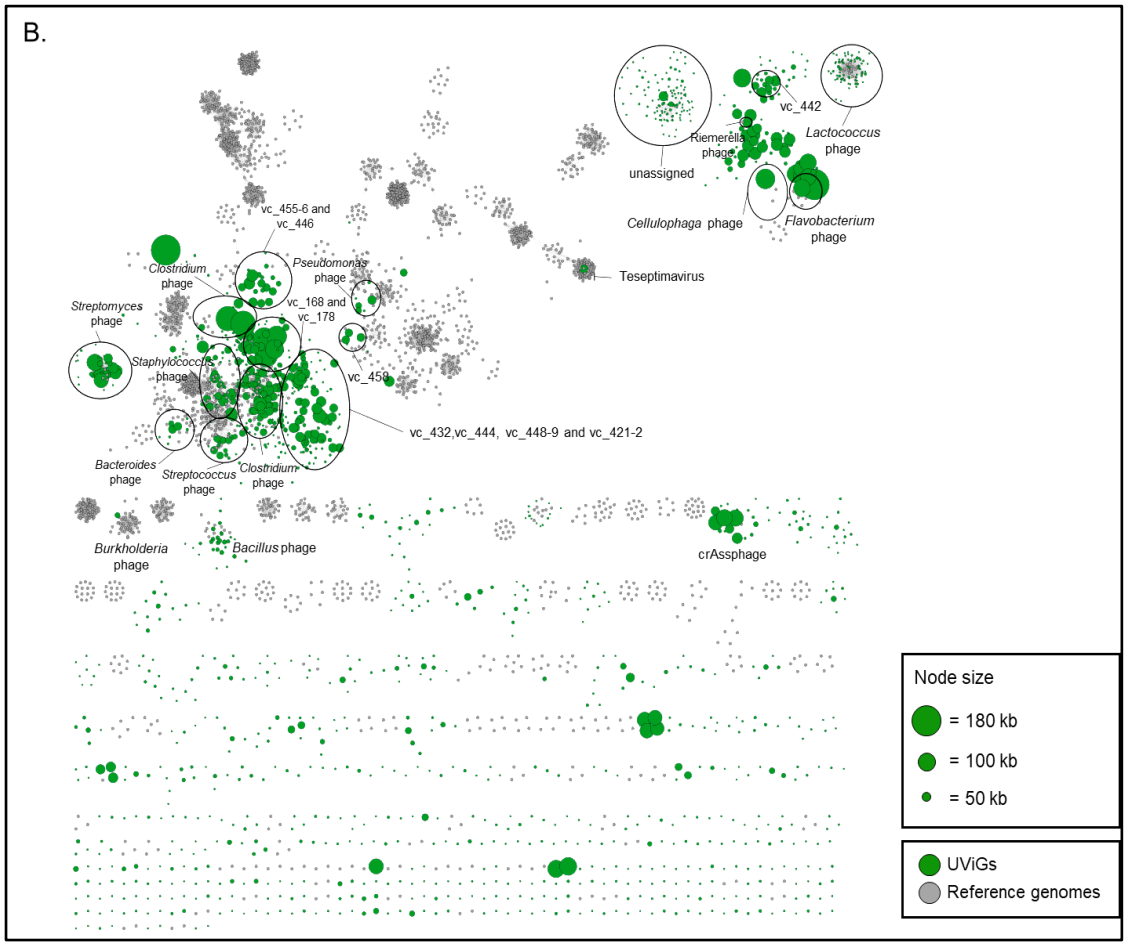

B.
vc_455-6 and vc_446
Clostridium phage
Pseudomonas phage
Streptomyces phage
vc_168 and vc_178
Staphylococcus phage
vc_450
vc_432,vc_444, vc_448-9 and vc_421-2
Bacteroides phage
Streptococcus phage
Clostridium phage
Burkholderia phage
Bacillus phage
crAssphage
vc_442
Riemerella phage
Lactococcus phage
unassigned
Cellulophaga phage
Flavobacterium phage
Teseptimavirus
Node size
= 180 kb
= 100 kb
= 50 kb
UViGs
Reference genomes

Figure 4.7. Cluster analysis displaying viral sequence similarity for virome-derived (A) PCR and (B) non-PCR datasets. Nodes (green circles) represent UViGs and edges (grey lines) represent the strength of the relationships between the genomes. UViGs were grouped into a viral cluster with shared sequence similarity based on amino acid homology against reference genomes (grey circles).

4.5. Discussion

In this chapter, I first set out to develop and optimise a bioinformatics pipeline for intestinal viral metagenomic analysis. and investigated the extent of PCR amplification bias from sequencing library preparation in my virome dataset by comparing PCR-based (LASL) to non-PCR (NASL) method of library preparation using enzyme-based NEBNext Ultra II methodology.

In human virome studies, combining VLP and viral nucleic acid enrichment with random amplification approaches is often required for sequence-based analysis in human sample such as faeces, which account for approximately 5.8% of total DNA in the whole human GIT microbial community (Shkoporov et al., 2018b, Hoyles et al., 2014, Arumugam et al., 2011). Three methods of random amplification (Djikeng et al., 2008, Angly et al., 2006, Breitbart et al., 2002, Dean et al., 2001, Froussard, 1992) to enrich viral DNA and obtain sufficient yields for sequencing library preparation are used in current virome studies. However, each method has its own benefits and limitations which can introduce different levels of bias in sequence dataset. Recently, some studies have investigated the amplification bias in three random amplification methods and multiple sequencing library preparation kits (Sato et al., 2019, Kallies et al., 2019, Parras-Moltó et al., 2018, Aird et al., 2011), but how the amplification bias impacts the human intestinal/faecal virome is unclear.

4.5.1. Assessment of Viral Metagenomic Sequences

Overall, the number of reads produced from all PCR datasets is generally higher than those produced from all non-PCR datasets, indicating that PCR effectively enriches viral DNA when the amount of input DNA reaches or exceeds ~1 µg. However, smaller amounts of viral DNA are often seen in virome studies, particularly for those using clinical samples. Low amounts of viral DNA input (e.g. a few nanogram amounts or less) is likely to cause bias when using PCR-based LASL protocol (Parras-Moltó et al., 2018, Solonenko et al., 2013, Duhaime et al., 2012). This drawback sometimes may make the LASL method unavailable due to sample limitations.

The data also shows that there is no extreme GC content in my DNA samples, with an average %GC of between 36% and 47%, although PCR amplification can introduce inherent bias particularly for those genomes with having extremely high- or low- GC content regions. This can lead to under- or over-amplification in these genomic regions (Parras-Moltó et al., 2018, Solonenko et al., 2013, Duhaime et al., 2012, Kozarewa et al., 2009). More recently, eight commercial sequencing library preparation kits including PCR and PCR-free treatment were investigated to identify amplification bias in bacterial genomes and metagenomes

(Sato et al., 2019). The extent of sequencing bias is associated with %GC content of genomes with stronger bias resulting from genomes with lower GC content and in using tagmentation-based Nextera XT methodology (Sato et al., 2019). Sato and colleagues also noted that PCR amplification contributes bias in other PCR-based library preparation kits, as noted by others (Aird et al., 2011, Kozarewa et al., 2009). To overcome this issue, Kozarewa and colleagues developed the amplification-free method (NASL) by omitting PCR steps from sequencing library preparation (Kozarewa et al., 2009). However, omitting PCR amplification steps means that relatively large amounts of input viral DNA is required for library preparation and for Illumina sequencing, and therefore, it is often impractical for many types of samples such as clinical samples (Aird et al., 2011). In addition, biases are likely to be introduced from elsewhere during sequencing process and therefore, it may be very difficult to completely remove biases from Illumina sequence datasets (Aird et al., 2011). Hence, although omitting PCR steps during library preparation is ideal to reduce bias, it is often not possible.

To evaluate the extent of VLP enrichment using the optimised protocol for faecal VLP and VLP DNA isolation, ViromeQC, was used to computationally determine eukaryotic/microbial contamination and the extent of viral enrichment, compared to non-enriched metagenomes as a baseline (Zolfo et al., 2019). High enrichment scores represent high level of VLP enrichment. The extent of VLP enrichment in all PCR datasets is relatively high, reaching 11-fold enrichment in sample 1, while the enrichment scores are low in all non-PCR datasets. Importantly, ViromeQC was originally designed for enriched viral DNA samples by amplification-based library preparation using standard protocols. It is therefore less well-suited for amplification-free samples using an NASL-based method. Despite enrichment of the non-amplified samples being relatively low, UViGs can still be detected in my analysis, as seen in a recent study (Tisza et al., 2020).

4.5.2. *De novo* Genome Assembly

In agreement with my findings, Sutton and colleagues have noted that with increased sequencing depths, MEGAHIT often has a lower average N50 length but it can generate longer contigs and has the longest contig length, compared with other assemblers (Sutton et al., 2019). Due to its performance in assembly sensitivity and low-coverage genomes in a single contig (Sutton et al., 2019, Roux et al., 2017), (meta)SPAdes is commonly used for the assembly of Illumina short reads in many virome studies (Tisza et al., 2020, Gregory et al., 2020, Shkoporov et al., 2018b, Parras-Moltó et al., 2018). However, MEGAHIT can also be used to characterise viral metagenomes, depending on the user's requirements. For example, MEGAHIT has higher performance in genome recovery, number of contigs obtained, detection of longest contigs with fewer chimeras (particularly ≥10 kb), and

assembly sensitivity and accuracy (Sutton et al., 2019, Roux et al., 2017). Moreover, our bioinformatics colleague noted that (meta)SPAdes sometimes fails in task completion before finishing assembly due to insufficient resources (i.e. RAM usage; unpublished observations), conflicting with the findings in a recent study (Sutton et al., 2019). We therefore selected the MEGAHIT assembler with default parameters as it has a higher success rate for completing genome assemblies in my virome study.

4.5.3. Identification of UViGs

To further predict and identify putative viral genomic contigs, two bioinformatics tools for virus identification were used based on the detection of viral hallmark genes (VirSorter) and *k*-mer distribution (VirFinder). VirSorter is applied to detect viral contigs and predict virus classification against two reference databases, the NCBI RefSeq (or called “RefSeqABVir”, the “RefSeq Archaea and Bacteria Viruses”) and an extended viral sequence database called “Viromes” collected from diverse environmental samples (Roux et al., 2015). VirFinder is a *k*-mer based programme via machine learning approach to identify and distinguish viruses from host sequences, and build a scoring system to predict viral sequences based on viral signatures (Ren et al., 2017). With my sorting strategy to maximise identification of potential viruses/proviruses, all putative viral candidates were identified based on meeting the VirSorter categories 1-6 including all linear and circular completed viruses (i.e. categories 1-3) and prophages (i.e. categories 4-6) with a score of ≥0.7 and p <0.05 by VirFinder, as described in a recent virome study (Gregory et al., 2020). Of these UViGs from both PCR and non-PCR datasets, shared UViGs detected by both tools can be considered to be true-positive viruses.

From the results of CheckV, the vast majority of small UViGs (<1 kb) are likely derived from VirFinder rather than VirSorter, having lower quality and fewer completed viral genomes based on estimation of genome completeness. Also, these small UViGs have difficulty in covering a significant number of viral/proviral hallmark genes. Therefore, sorting contig length at >2 kb or >5 kb would be recommended in further virome studies. On the other hand, there are 42 UViGs from PCR datasets and 49 UViGs from non-PCR datasets showing various levels of host contamination, however, when I examined the CheckV data, I noted that all of those from both PCR and non-PCR datasets have also been identified as “provirus” with several viral/proviral hallmark genes detected by CheckV. Moreover, some of these have medium- or high-levels of genome completeness, indicating that these are likely to be circular. This is evidence that my protocols and viromics pipeline can enrich sufficient sequences of viruses/proviruses (UViGs/UpViGs) for further virome investigation.

In read analysis, I individually mapped total cleaned reads to total genomic contigs from each dataset and mapped viral reads against the pool, non-redundant UViGs/UpViGs (>1kb) as the reference genomes. High mapping rate to total reference genomes can be seen in each dataset. However, there are lower mapping rates for viral reads against pool, non-redundant UViGs, with a wider range of mapping rates for each dataset, between 72% and 90%. We can infer that some paired-end viral reads have likely lost one of their single reads that may have been filtered out due to low quality, and therefore, are not successfully aligned to viral reference genomes using BWA-MEM (i.e. paired-end mode) and not incorporated during read mapping.

4.5.4. Amplification Bias in Virome-Derived Datasets

The analysis of relative abundance of the top 25 UViGs reveals that misrepresentation of relative abundance of viruses may occur after PCR amplification. Overall, most of the top 25 UViGs can be seen in both sample sets. However, a virus (contig s18v3_PF_45906) assigned to a new family of *Bacteroides* phages in PCR-1 dataset, a virus (contig s18v2_29464) assigned to *Siphoviridae* family and a virus (contig s18v1−PF_17802) assigned to a new family VC_442 found in PCR-3 dataset are underestimated, having slight or moderate differences in the top 25 UViGs, compared with non-PCR-1 and non-PCR-3 datasets, respectively. Of these differences, the relative abundance of UViGs in all three PCR datasets is generally higher than those in all non-PCR datasets, indicating that PCR amplification (LASL) influences viral representation causing over- or under-estimation for many viral taxa. This agrees with a recent virome study evaluating SISPA and MDA-based random amplification on the human saliva virome (Parras-Moltó et al., 2018). Moreover, the top 1 to 9 UViGs seen in each dataset cannot be assigned, suggesting that these are likely to be novel strains, but further investigation is required. Several published non-redundant viral reference databases, such as NCBI RefSeq (O'Leary et al., 2016), the human gut virome database (GVD) (Gregory et al., 2020), the integrated microbiome genome/virus system (IMG/VR) (Roux et al., 2020, Paez-Espino et al., 2017) and/or a new reference viral database (RVDB) (Goodacre et al., 2018), need to be considered in further studies, which can potentially shed light on “viral dark matter”. Of those assigned, *Siphoviridae* is the most abundant virus seen in all datasets, agreeing with my TEM analysis, followed by *Salasmaviridae* family in both sample 1 and 2. Also, *Podoviridae* family is only seen in sample 3 and *Myoviridae* family is only seen in sample 1, suggesting high individuality in the human intestinal/faecal viromes, as noted by others (Shkoporov et al., 2019, Shkoporov et al., 2018b).

For direct richness observation (i.e. actual counts of the UViGs >1 kb) and estimation of Shannon and Simpson indices, we noted that the richness and alpha diversity of UViGs in

non-PCR datasets are higher than those in PCR datasets, suggesting that PCR amplification affects the distributions of the human intestinal/faecal viromes and is likely to lower their richness and alpha diversity. For Chao1 estimation, however, a conflicting trend is seen in PCR-1 and non-PCR-1 datasets, whereas the same trend of higher richness and alpha diversity of UViGs in non-PCR than in PCR datasets is seen after rarefaction. This finding implies that sample 1 may encompass more low-abundant UViG taxa than sample 2 and 3, thereby leading to an incorrect tendency, demonstrating that Chao1 is less well-suited for cases with low abundant species, agreeing with previous findings (Haegeman et al., 2013).

For beta diversity analysis, both ordination plots based on Bray-Curtis dissimilarities and Jensen-Shannon divergence show that the locations between sample clusters are farther than those between PCR and non-PCR datasets in each sample, implying that the major difference in inter-subject beta diversity is driven by high specificity of the intestinal/faecal viromes in each subject (Shkoporov et al., 2019, Shkoporov et al., 2018b). Moreover, there are differences in the locations between PCR and non-PCR datasets, suggesting that PCR amplification may have a minor effect on inter-subject beta diversity in comparison with non-amplified datasets, consistent with the findings of a recent study (Parras-Moltó et al., 2018).

Finally, to investigate the similarity of UViGs between PCR and non-PCR conditions, cluster analysis was performed to group UViGs into diverse viral clusters (VCs) based on genome similarity and to investigate the strength of relationships between VCs based on their amino acid homology. Although all UViGs are grouped into diverse VCs, there is no significant differences between PCR and non-PCR datasets, sharing UViG sequences in both. This can be presumably explained by the fact that both datasets are derived from the same DNA samples and some UViGs may be conserved across the samples, such as crAssphage that tends to be conserved over time in the whole human population, particularly in healthy individuals (Shkoporov et al., 2019). Similarly, previous studies also showed that high virome stability was observed in a single healthy adult and in a cohort study of twins (Minot et al., 2013, Reyes et al., 2010). In addition, many small or very small orphan VCs can be seen in both datasets, displaying either no or few connections to each other. Most of these small orphan VCs may result from short or very short viral contigs (e.g. <1 kb), suggesting that sorting contig length to >2 kb or >5 kb would be necessary to reduce network complexity in future studies. Selecting UViGs >10 kb has also been considered in a recent virome study (Gregory et al., 2020).

Based on the aims of this chapter, it can be concluded that PCR amplification introduces significant biases that impact the relative abundance and diversity of the VLP-enriched intestinal/faecal viromes.

4.6. Summary

In this chapter, I have investigated if PCR amplification leads to bias in VLP-enriched intestinal/faecal virome datasets by comparing PCR-based LASL to NASL methods for Illumina sequencing library preparation. We also evaluated the extent of amplification bias by considering several aspects, including the composition of viral sequence datasets, relative abundance, richness, alpha and beta diversity, and UViG sequence similarity networks. Collectively, I found that PCR amplification introduces bias and can have a major impact on viral relative abundance, thereby leading to misrepresentation. For alpha diversity, viruses from non-amplified samples generally show higher richness and diversity than those from PCR-amplified samples. For beta diversity, although the major difference between subjects is likely driven by high specificity and individuality of the intestinal/faecal viromes, amplification bias has a minor effect on the beta diversity between PCR and non-PCR datasets. Moreover, in an initial analysis of the comparisons of UViG similarity between PCR and non-PCR experiments, I found that there is no significant difference with shared VCs and UViG sequences in both datasets. Based on these findings, we recommend that whenever possible, amplification-free methodologies (NASL) should be considered minimising sequence amplification bias in virome studies.

5. Enumeration and Characterisation of Faecal VLPs in Severe ME/CFS Patients and Same Household Healthy Controls

5.1. Introduction

5.1.1. Background

Quantifying the abundance of viruses and virus-like particles (VLPs) in environments such as seawater has been used to understand the potential impact of viruses on their hosts and the changes in virus-host dynamics in biology and microbial ecosystem (Wilhelm and Suttle, 1999, Weinbauer and Suttle, 1997, Suttle, 1994). To date, two main approaches have been utilised to enumerate viruses in environmental samples: transmission electron microscopy (TEM) (Wommack et al., 1992, Proctor and Fuhrman, 1990, Bergh et al., 1989) and epifluorescence microscopy (EFM) (Hara et al., 1991, Suttle et al., 1990, Coleman et al., 1981, Porter and Feig, 1980, Coleman, 1980). TEM was first applied to enumeration of bacteria and viruses in aquatic samples (e.g. seawater) (Proctor and Fuhrman, 1990, Bergh et al., 1989). However, the drawbacks of TEM analysis, including time and expense, restrict its use for routine counting of viruses (Hennes and Suttle, 1995, Hara et al., 1991). The number of viruses and VLPs are also potentially underestimated by TEM, with 43-66% of VLPs being observed and enumerated by TEM (Weinbauer and Suttle, 1997, Hennes and Suttle, 1995). As a result, researchers have relied on EFM analysis to estimate the abundance of viruses and bacteria since the 1980s (Azam et al., 1983, Coleman, 1980, Bergh et al., 1989). More recently, EFM combined with fluorescent-based staining methods has been used to identify and enumerate viruses in diverse environmental samples, from oceans to human faeces and tissue specimens (Hoyles et al., 2014, Lepage et al., 2008, Chen et al., 2001, Noble and Fuhrman, 1998).

5.1.2. Development of Fluorescent Dye Staining

DAPI (4',6-diamidino-2-phenylindole) (Porter and Feig, 1980, Coleman, 1980), Yo-Pro-1 (Weinbauer and Suttle, 1997, Hennes and Suttle, 1995), SYBR Green I (Noble and Fuhrman, 1998) as well as SYBR Gold (Tuma et al., 1999), have been extensively used for staining and detecting double- and/or single-stranded nucleic acids in samples. To date, these have been extensively applied to the identification and enumeration of bacteria and viruses in various environmental samples.

DAPI, a cyanine-based fluorochrome which specifically binds to double-stranded DNA and forms a DNA-dye complex, displays a fluorescent signal greater than DAPI-unbound molecules. It was first used to stain and enumerate viral particles by EFM (Sieburth et al.,

1988). However, DAPI was limited to direct visual observation by EFM due to its poor signal intensity (Noble and Fuhrman, 1998, Hennes and Suttle, 1995) and was replaced by Yo-Pro-1 (Weinbauer and Suttle, 1997, Hennes and Suttle, 1995). Yo-Pro-1 displays green fluorescence at λ = 510 nm after excitation with blue light when bound to DNA and RNA molecules, while unbound dye displays low levels of background fluorescence (Hennes and Suttle, 1995, Hirons et al., 1994). Compared with DAPI, Yo-Pro-1-stained viral particles in water samples exhibit brighter and more stable green-fluorescent signal intensity without bleaching than those stained with DAPI, and are easily observed by microscopy. However, samples have to be incubated in the dark at ambient temperature for up to 48 hours to ensure complete staining (Hennes and Suttle, 1995). Bacteria may grow and multiply during this long incubation period, thereby challenging VLPs identification and enumeration. A modified protocol was then developed by microwaving the Yo-Pro-1-stained samples to enhance efficiency of dye diffusion into viral capsids, thereby shortening staining and fixation time (Xenopoulos and Bird, 1997). Moreover, cell debris, nucleic acids, detritus particles and bacterial cells stained with Yo-Pro-1 are typically larger sized and are irregular shaped with more yellow fluorescence than true viruses, making it easier to distinguish non-viral particles from the Yo-Pro-1-labelled viral particles (Hennes and Suttle, 1995). Weinbauer and Suttle (1997) found that Yo-Pro-1-labelled virus counts are between 1.5 and 6.2 times higher than by TEM, consistent with the results in the similar study (Hennes and Suttle, 1995). TEM-based virus counts are more likely therefore to be underestimated.

Noble and Fuhrman (1998) recommended SYBR Green I (or SYBR I) as an effective stain for viruses particularly after concentrating on 0.02 µm pore-size Anodisc inorganic (Al_2O_3) filter membrane prior to staining. SYBR Green I is a bright, stable fluorescent dye suitable for nucleic acids staining and microbial/viral counting without significant bias or staining of detritus. This and other studies (Noble and Fuhrman, 1998, Weinbauer and Suttle, 1997, Hennes and Suttle, 1995) also revealed that SYBR Green I-stained viral counts are around 1.3 times higher than TEM-based counts. These authors speculated that viral particles may be lost when excessive uranyl acetate (UA) is wicked away from the filter grids, and may be masked by other larger, condensed and darkly stained particles. Moreover, filamentous viruses would be more readily detected by EFM and SYBR Green I staining than by TEM. SYBR Green I has been reported to stain both double- and single-stranded DNA and RNA, indicating that RNA viruses can also be enumerated. The fluorescence intensity of SYBR Green I is similar to Yo-Pro-1. However, SYBR Green I appears to bleach faster than Yo-Pro-1 (Bettarel et al., 2000). To overcome this, SYBR Gold was developed (Tuma et al., 1999).

Due to its proprietary non-symmetrical cyanine structure, SYBR gold has two fluorescence excitation peaks when bound to DNA and RNA, at ~300 nm and at ~495 nm. It is more

sensitive than ethidium bromide and SYBR Green I for broadly detecting double- and single-stranded DNA and RNA by intercalating between the bases of nucleic acids and exhibiting ~1,000-fold fluorescent signal upon binding to nucleic acids. This enables VLP numbers to be more accurately determined due to the dye's ability to penetrate viral capsid to bind nucleic acids (Choi et al., 2013, Armitage, 2005, Mosier-Boss et al., 2003, Tuma et al., 1999). Chen and colleagues (2001) as well as Shibata and colleagues (2006) have evaluated and refined the protocol of Noble and Fuhrman (1998) for rapid staining and precisely estimating bacteria and/or viruses in aquatic samples by EFM, based on SYBR Gold having a more stable fluorescence intensity than SYBR Green I.

Recently, SYBR Gold staining has been applied to various human samples such as faeces and tissue biopsies (Hoyles et al., 2014, Lepage et al., 2008). Hoyles and colleagues (2014) established a protocol for VLP isolation, examining and enumerating VLPs in faecal filtrates (FFs) collected from six healthy individuals using TEM and SYBR Gold-based EFM. The number of VLPs detected ranged from 2.4×10^8 to 1.12×10^9 VLP/ml FFs (equivalent to 1.2×10^9 to 5.58×10^9 VLP/g faeces). By comparison, Lepage et al (2008) used mucosal samples from 14 healthy donors and 19 Crohn's disease patients to detect an average of 1.2×10^9 VLPs per biopsy. Of note, they revealed that more VLPs were observed in patients (mean 2.9×10^9 VLP/biopsy) than in healthy individuals (mean 1.2×10^8 VLPs/biopsy).

In this study, SYBR Gold staining (Budinoff et al., 2011) was used to enumerate VLPs in faecal samples.

5.2. Aim

To evaluate methods of VLP enumeration for determinations of VLPs in faecal samples from severe ME/CFS patients and from healthy individuals living in the same household (same household healthy controls; SHHC).

5.3. Study Design

Faecal VLPs were isolated from nine patients and eight SHHC for detection and enumeration using EFM analysis as outlined in **Figure 5.1**. Briefly, PEG-VLP suspensions (**Chapter 2**) were stained with SYBR Gold dye, followed by VLP immobilisation on 0.02 µm pore-size Anodisc filter membrane and VLP visualised by EFM. VLP counts were then determined using ImageJ and by manual counting as well as digital image analysis (DIA). The formula for VLP enumeration (Budinoff et al., 2011) was as described in **Chapter 2**.

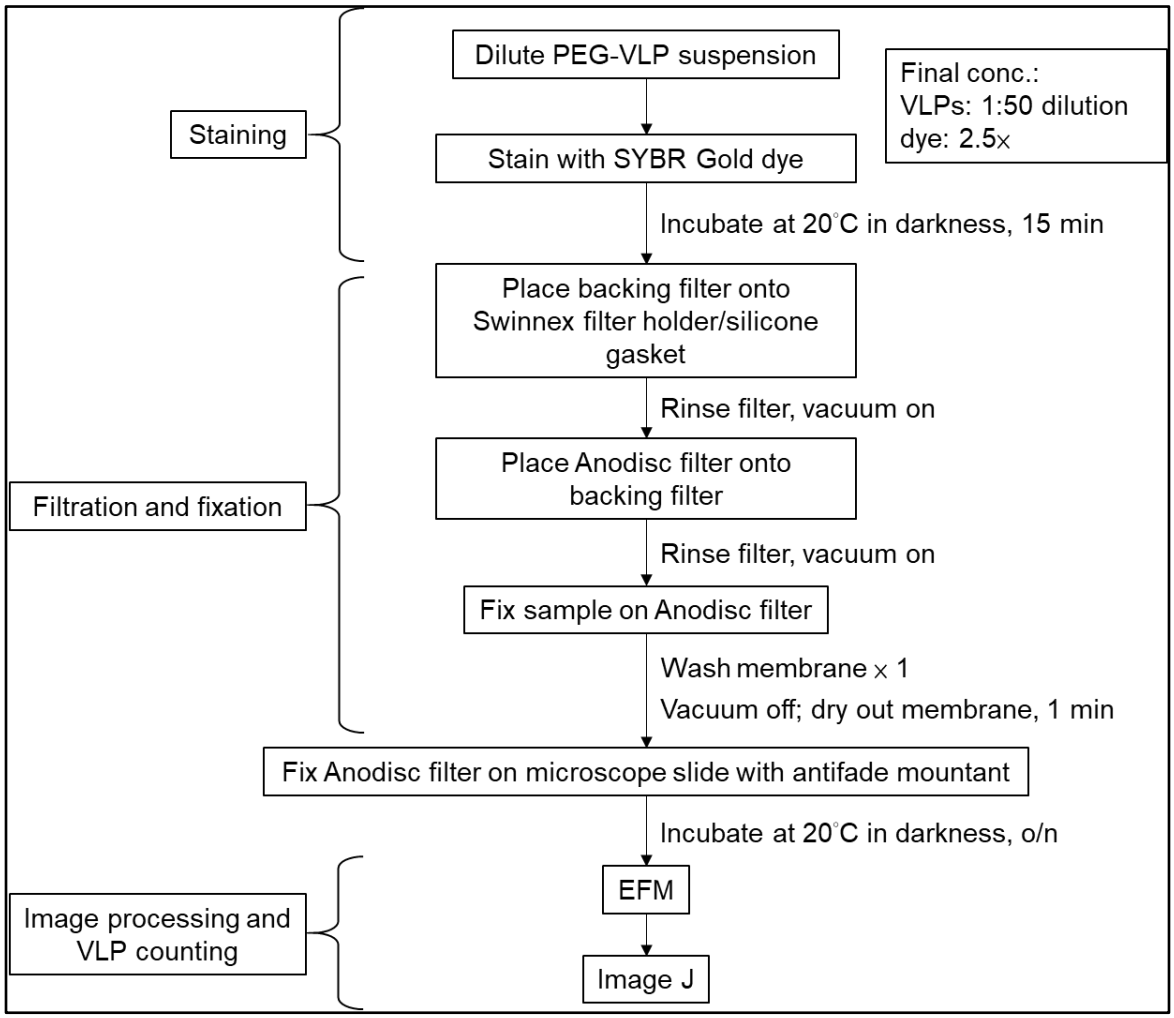


Figure 5.1. Workflow of SYBR Gold staining and EFM analysis. The workflow is composed of VLP staining, sample filtration and fixation, EFM imaging, image post-processing and VLP enumeration by manual counting by eye.

5.4. Results

5.4.1. Evaluating VLP Enumeration by Manual and Automated Counting

Initially, the threshold for particle size sorting in ImageJ was optimised to ensure that the size and number of the SYBR Gold-stained VLPs are accurate and to avoid over- or underestimation. For this, the reference Bf phage ΦB124-14 was used and particle sizes were determined between 0 and 0.2 μm^2 with ImageJ, compared to direct manual counting by eye (**Figure 5.2** and **Table 5.1**).

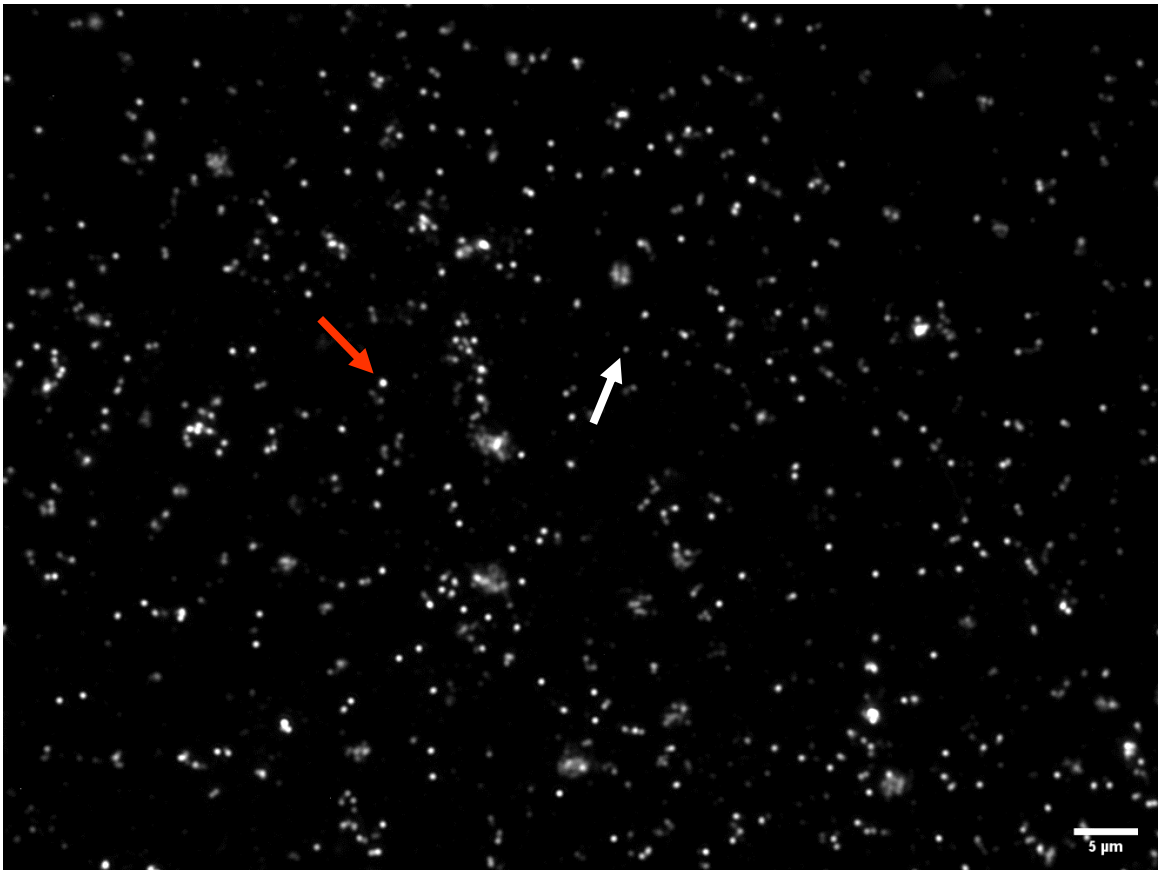


Figure 5.2. Fluorescence micrograph of SYBR Gold-stained Bf phage ΦB124-14. SYBR Gold-stained viral particles (VPs) were identified as "pinpricks" (white arrow) with larger dots of intense fluorescence representing VP aggregates (red arrow). The scale bar represents 5 µm under 1,000X magnification.

Table 5.1. VP counts determined by ImageJ or by direct manual counting

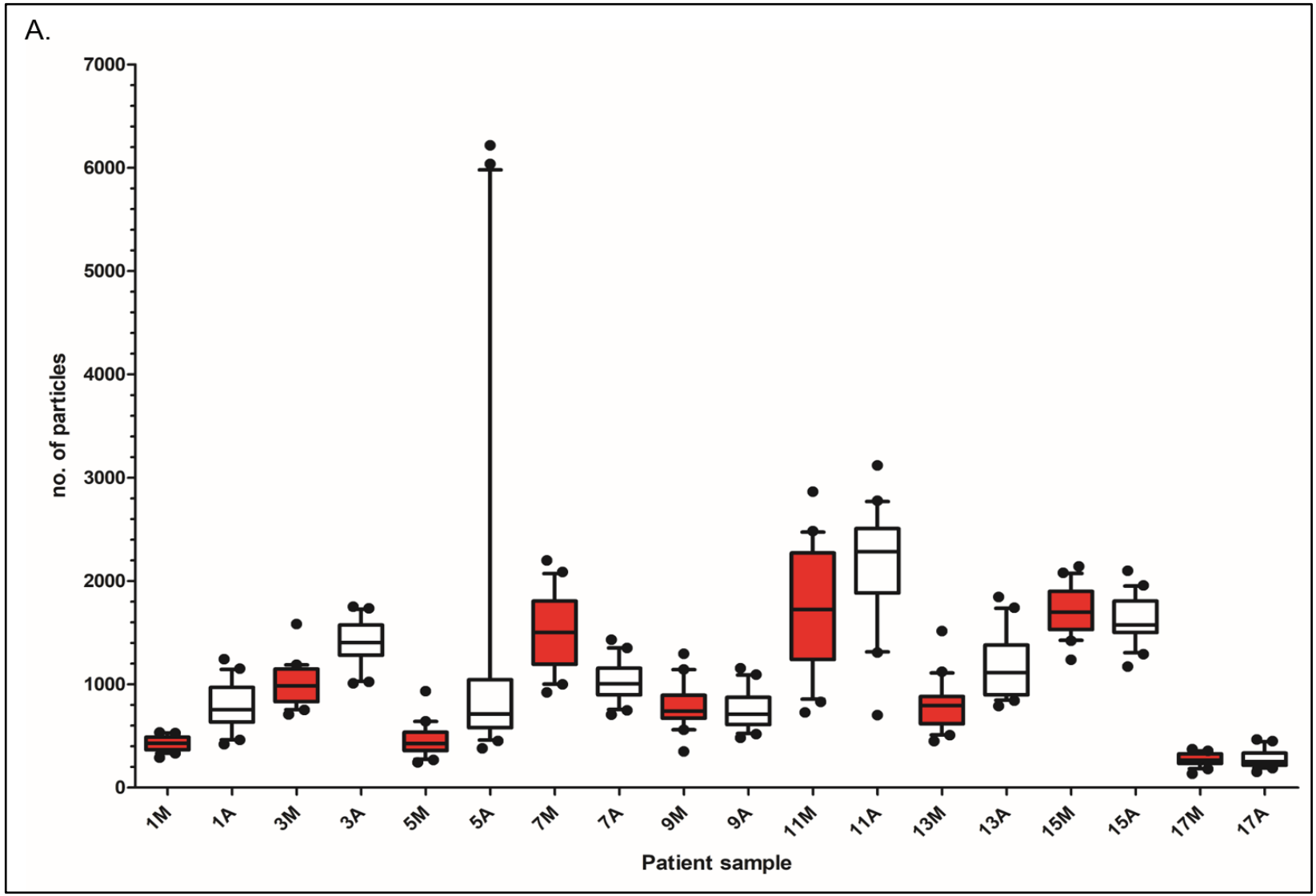
Particle size (μm^2)	# of VPs (ImageJ automated counting)
0-infinity	791
0-100	791
0-10	791
0-7	790
0-5	789
0-4	785
0-3	779
0-2	766
0-1	702
0-0.9	689
0-0.8	664
0-0.7	633
0-0.6	590
0-0.5	538
0-0.4	418
0-0.3	392
0-0.2	**342**
0-0.1	242
0-0.05	177
0-0.01	99
0-0.005	81
Direct manual counting	**361**

VP: viral particles

Using automated counting and a one-size-fits-all sorting condition, we noted that the electronic images contained background and electronic noise, the level of which varied from sample to sample. To confirm the accuracy of VLP enumeration, we adjusted the conditions and thresholds in ImageJ for each image and manually checked by eye the accuracy of particle size, shape, number and distribution, and then compared these to the original images.

Figure 5.3 showed that VLP numbers varied and fluctuated widely in the patient and SHHC samples by both manual and automated counting. Automated counting gave higher counts in comparison with corresponding manual counting, for example, in sample 1 (manual vs. automated counting: 428 ± 72.2 vs. 788.5 ± 226, mean ± S.D.; n = 20), sample 2 (531.7 ± 95.2 vs. 845.9 ± 104.7), sample 3 (1,004.5 ± 206.8 vs. 1,395 ± 219.5), sample 5 (462.4 ±

155.3 vs. 1,501.1 ± 1911.1), sample 6 (231.8 ± 92.1 vs. 492.6 ± 230.9), sample 10 (815.1 ± 116.5 vs. 1363.7 ± 504.1), sample 11 (1735.6 ± 581.8 vs. 2148.1 ± 568.2), sample 13 (797.6 ± 239.8 vs. 1175.3 ± 320.8) and sample 16 (701.9 ± 279.3 vs. 833.3 ± 182.6).


A.
7000
6000
5000
4000
3000
2000
1000
0
no. of particles
1M
1A
3M
3A
5M
5A
7M
7A
9M
9A
11M
11A
13M
13A
15M
15A
17M
17A
Patient sample

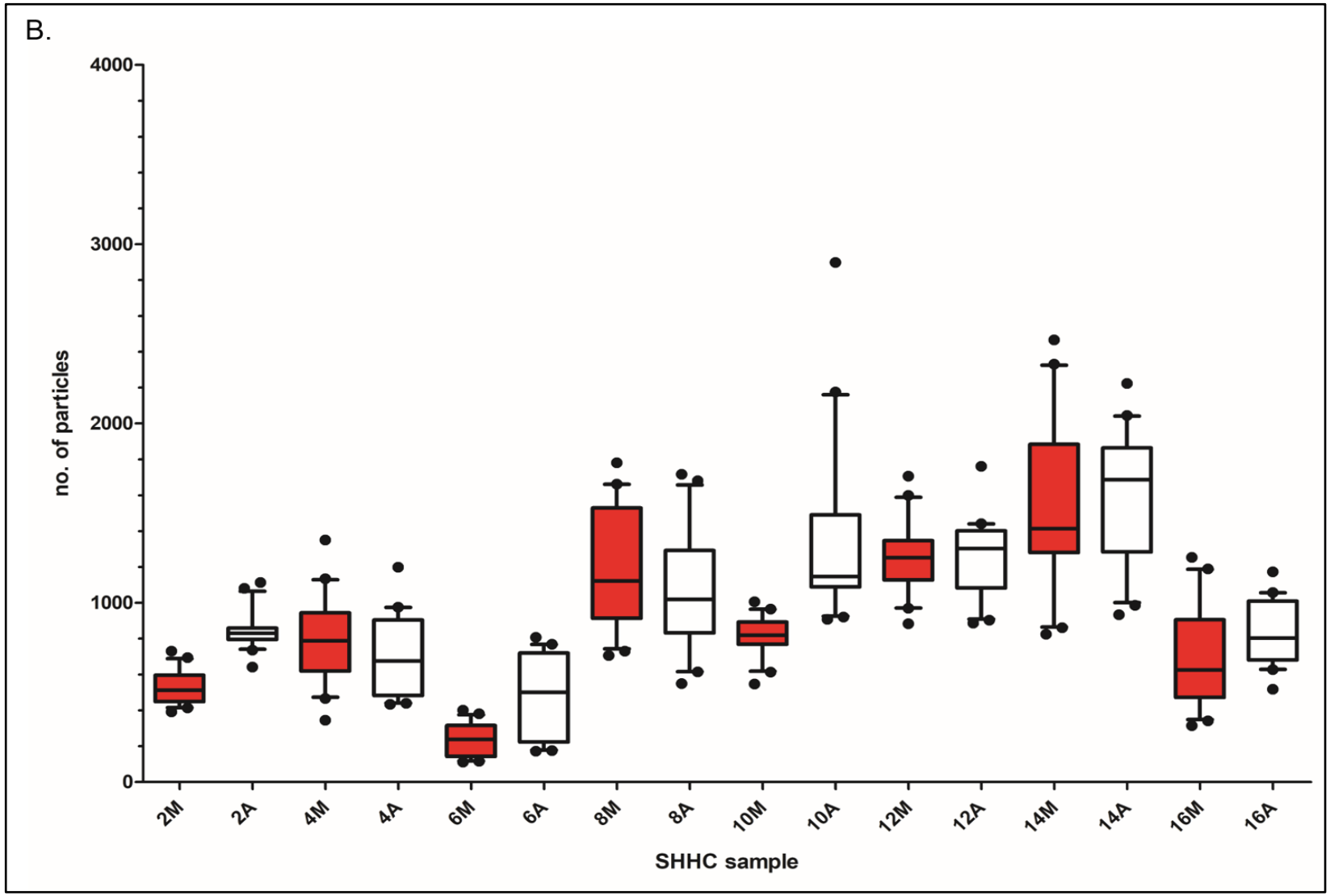

B.
4000
3000
2000
1000
0
no. of particles
2M
2A
4M
4A
6M
6A
8M
8A
10M
10A
12M
12A
14M
14A
16M
16A
SHHC sample

Figure 5.3. Box whisker plot (10-90 percentiles) of manual- and automated VLP counts in patient (A) and SHHC (B) samples. Manual counting-based DIA method (M, red) and automated counts determined using identical settings in ImageJ for all samples (A, white) based upon 20 images per sample taken by EFM-connected cooled CCD camera. Points below and above the whiskers represent extreme values.

Based on these findings, we adopted manual counting for VLP enumeration in patient and SHHC samples (**Figure 5.4 and Table 5.2**). **Figure 5.4** showed an example of fluorescence micrographs of faecal VLPs collected from a patient (sample 11). Across all 17 samples, faecal VLPs ranged from 6.2×10^7 to 4.6×10^8 VLP/ml ($2.5 \times 10^8 \pm 1.3 \times 10^8$ VLP/ml, mean ± S.D.), equivalent to 3.9×10^8 to 2.9×10^9 VLP/g faeces ($1.5 \times 10^9 \pm 8.1 \times 10^8$ VLP/g faeces) (**Table 5.2**).

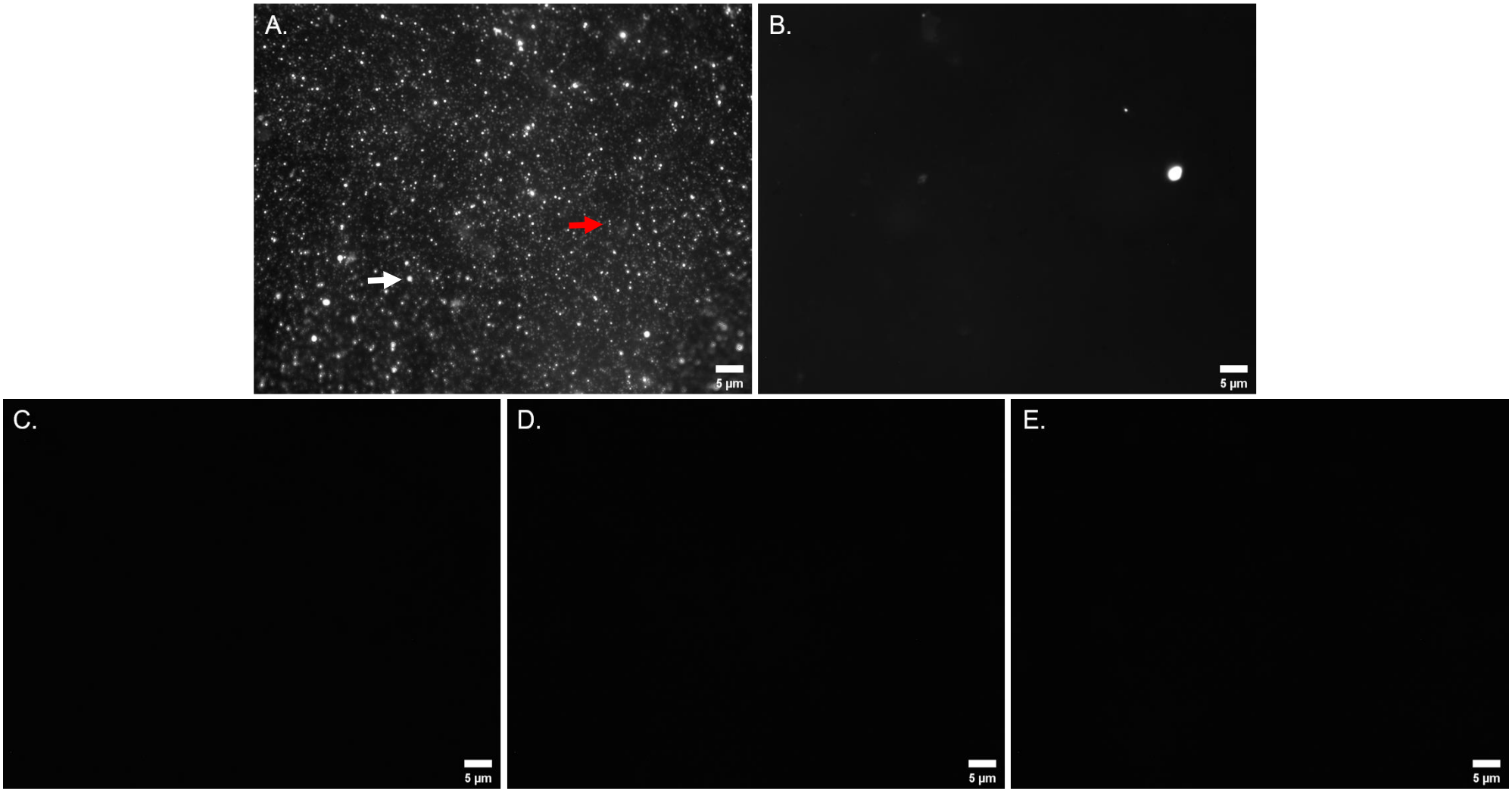

A.
5 µm
B.
5 µm
C.
5 µm
D.
5 µm
E.
5 µm

Figure 5.4. Fluorescence micrograph of faecal VLPs. (A) Faecal VLPs collected from a severe ME/CFS patient. The red arrow represents "pinprick" dots indicative of VLPs; the white arrow represents brighter dots of intense fluorescence signal representing aggregated VLPs. (B) An example of a suspected bacterial contamination. (C-E) Unstained VLP suspension (C), dye incubated with sterile TBT buffer (D) and dye incubated with nuclease-free water (E) were included as controls to monitor contamination. The scale bars = 5 μm.

Table 5.2. Summary of VLP count and VLP DNA yield from faecal samples of ME/CFS patient and SHHC

Sample	Health status	Sample dry weight (g)	VLP/ml	VLP/g faeces	Total DNA (ng)	DNA yield (ng/g faeces)
1	PT	22.25	1.1×10^8	7.1×10^8	2,000.0	89.9
2	SHHC	21.92	1.4×10^8	8.8×10^8	6,120.0	279.2
3	PT	44.83	2.7×10^8	1.7×10^9	2,745.0	61.2
4	SHHC	37.29	2.1×10^8	1.3×10^9	3,852.0	103.3
5	PT	20.77	1.2×10^8	7.7×10^8	383.5	18.5
6	SHHC	19.42	6.2×10^7	3.9×10^8	3,657.0	188.3
7	PT	20.0	4.0×10^8	2.5×10^9	615.0	30.8
8	SHHC	20.29	3.2×10^8	2.0×10^9	3,733.2	184.0
9	PT	50.0	2.1×10^8	1.3×10^9	600.0	12.0
10	SHHC	50.0	2.2×10^8	1.4×10^9	1,842.0	36.8
11	PT	41.21	4.6×10^8	2.9×10^9	1,027.0	24.9
12	SHHC	38.17	3.3×10^8	2.1×10^9	945.0	24.8
13	PT	16.37	2.1×10^8	1.3×10^9	488.0	29.8
14	SHHC	17.42	4.1×10^8	2.6×10^9	5,592.0	321.0
15	PT	40.16	4.6×10^8	2.8×10^9	1,738.0	43.3
16	SHHC	31.65	1.9×10^8	1.2×10^9	2,806.0	88.7
17	PT	33.62	7.2×10^7	4.5×10^8	4,670.0	138.9

PT: severe ME/CFS patient; SHHC: same household healthy control

5.4.2. Characterising Faecal VLPs by TEM

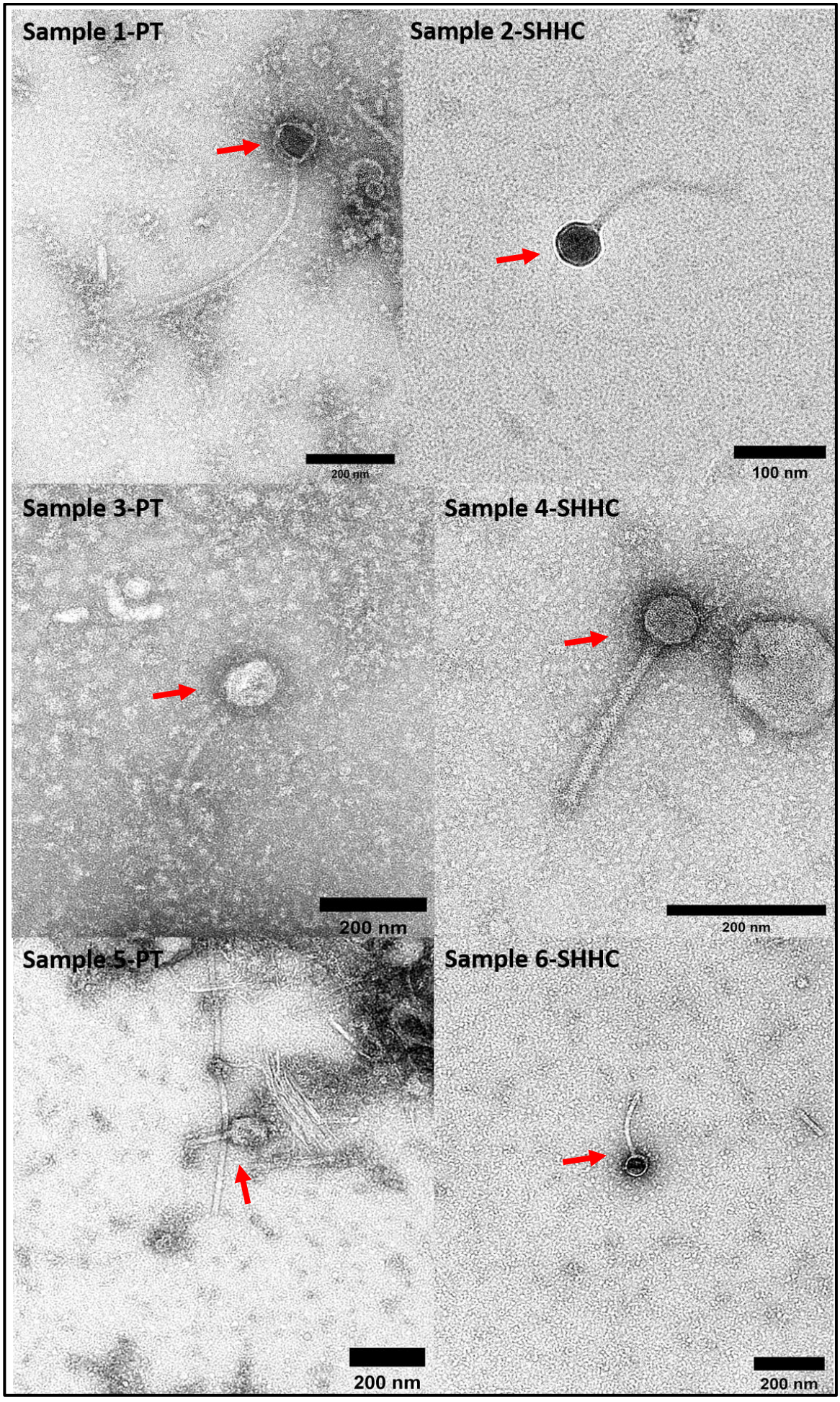
VLPs collected from faecal filtrates (FFs) of patient and SHHC samples were morphologically characterised by TEM analysis taken by myself and Dr Catherine Booth (**Figure 5.5** and **Appendix 6**). Dimensions and predicted viral classification based on morphology were summarised in **Table 5.3**. In sample 1, the majority of detected VLPs appeared to be double-stranded DNA bacteriophages, belonging to the order *Caudovirales* with three families: *Sipho*-like VLPs of average size 412.8 ± 289.4 nm (mean ± S.D., n = 10), *Podo*-like VLPs of average size 117.5 ± 17.7 nm (n = 2) and *Myo*-like phage structures of average size 172.3 ± 17.1 nm (n = 5). In sample 2, *Sipho*-like VLPs with an average size 294.9 ± 94.7 nm (n = 3) and *Myo*-like VLPs with an average size 191.2 ± 11.8 nm (n = 3) were seen. In sample 3, *Sipho*-like VLPs of average size 413.3 ± 256.4 nm (n = 4) and a *Myo*-like VLP (~155.6 nm) were seen. In sample 4, a *Sipho*-like VLP (~312.5 nm) and a *Myo*-like VLP (~218.5 nm) were detected.

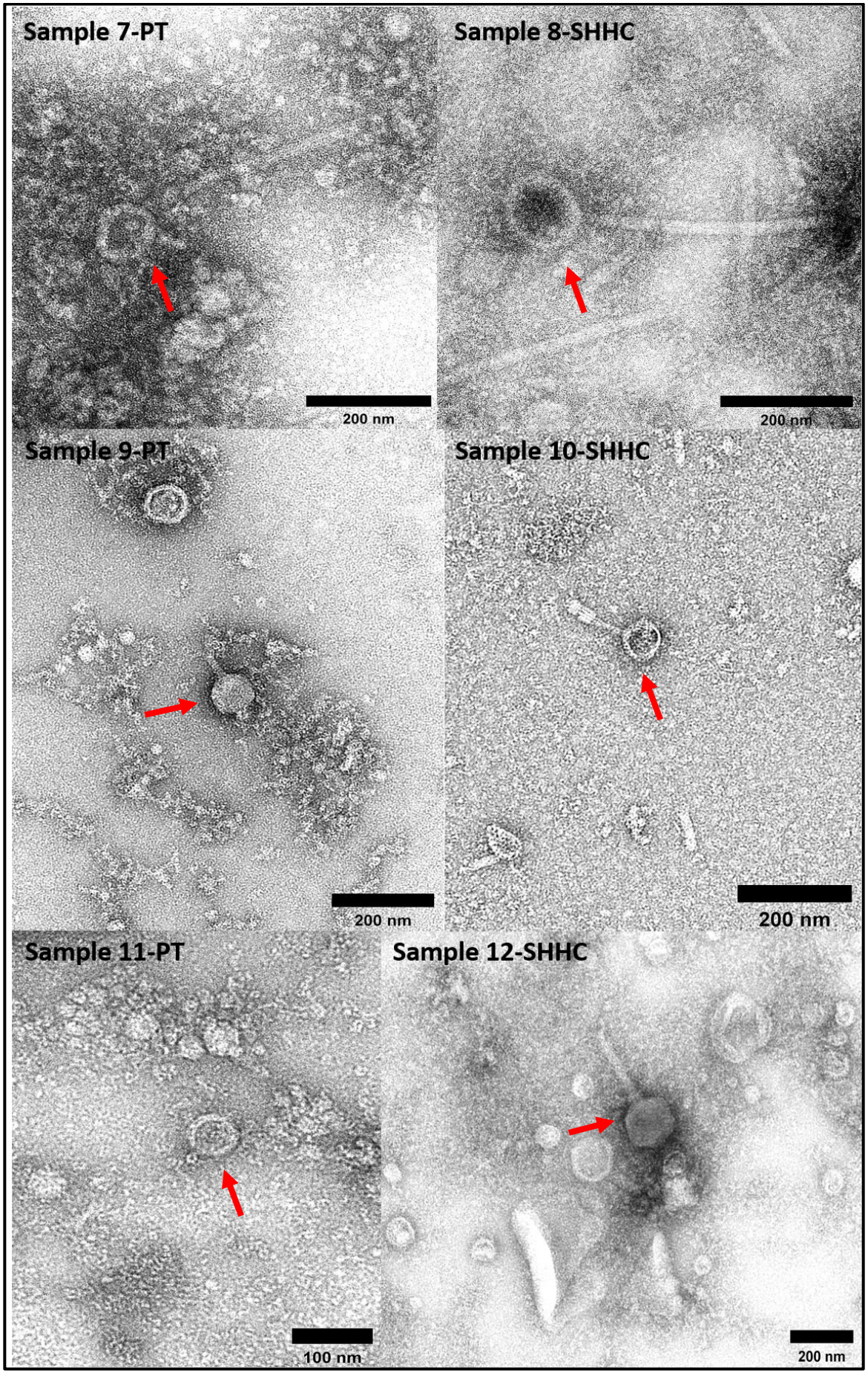
In sample 5, a *Myo*-like VLP (~200 nm) was found in addition to a number of detached phage heads and tails. In sample 6, *Sipho*-like VLPs of 344.5 ± 124.5 nm (n = 3) and a giant *Podo*-like structure or a detached viral capsid (~240 nm) were observed. In sample 7, three intact viral-like structures were seen, including two *Sipho*-like VLPs of 416.4 ± 23.6 nm (n = 2) and one *Myo*-like VLP (~77 nm). In sample 8, seven intact *Sipho*-like VLPs of average size 344.3 ± 107.3 nm (n = 7) were seen. Also, two intact *Myo*-like phage structures of average size 161.2 ± 54.9 nm (n = 2) were seen in sample 9 and a *Myo*-like VLP (~175 nm) as well as a *Podo*-like structure (~100 nm) were detected in sample 10. Very few intact VLPs were seen in sample 11, except for a *Podo*-like structure (~90 nm). Two intact *Sipho*-like VLPs of 350.6 ± 70.7 nm (n = 2) and detached viral structures or non-tail virions were observed in sample 12.

In addition, an intact *Myo*-like structure (~200 nm) were seen in sample 13 with a *Sipho*-like structure (~392.8 nm) and a *Sipho*-like structure (~228.6 nm) in sample 14 and 15, respectively. A *Sipho*-like VLP (~416.6 nm), a *Myo*-like VLP (~185.7 nm) and a *Podo*-like VLP (~69.5 nm) were present in sample 16 with ten intact viral-like structures observed in sample 17, including *Sipho*-like VLPs of 243.4 ± 62.1 nm (n = 4), *Myo*-like VLPs of 189.3 ± 7.3 nm (n = 4) and *Podo*-like VLPs of 106.3 ± 8.8 nm (n = 2).

Collectively, the TEM data indicated that the size, number and diversity of faecal VLPs varied from sample to sample, with variable numbers of detached phage heads and tails, non-tail virions and/or filamentous-like viruses seen in all samples. Across these samples, the most common phage morphology observed was *Siphoviridae* (57.4%), followed by *Myoviridae* (30.9%) and *Podoviridae* (11.8%). In patient samples, the most prevalent phage

was *Siphoviridae* (51.2%), followed by *Myoviridae* (36.6%) and *Podoviridae* (12.2%). A similar distribution was seen amongst SHHC samples with *Siphoviridae* (66.7%) dominating, followed by *Myoviridae* (22.2%) and *Podoviridae* (11.1%) (**Table 5.3**). Interestingly, lysed bacteria were occasionally observed with virions being “released” from the damaged bacterial body (**Appendix 6; Figure A14**).


Sample 1-PT
200 nm
Sample 2-SHHC
100 nm
Sample 3-PT
200 nm
Sample 4-SHHC
200 nm
Sample 5-PT
200 nm
Sample 6-SHHC
200 nm


Sample 7-PT
200 nm
Sample 8-SHHC
200 nm
Sample 9-PT
200 nm
Sample 10-SHHC
200 nm
Sample 11-PT
100 nm
Sample 12-SHHC
200 nm

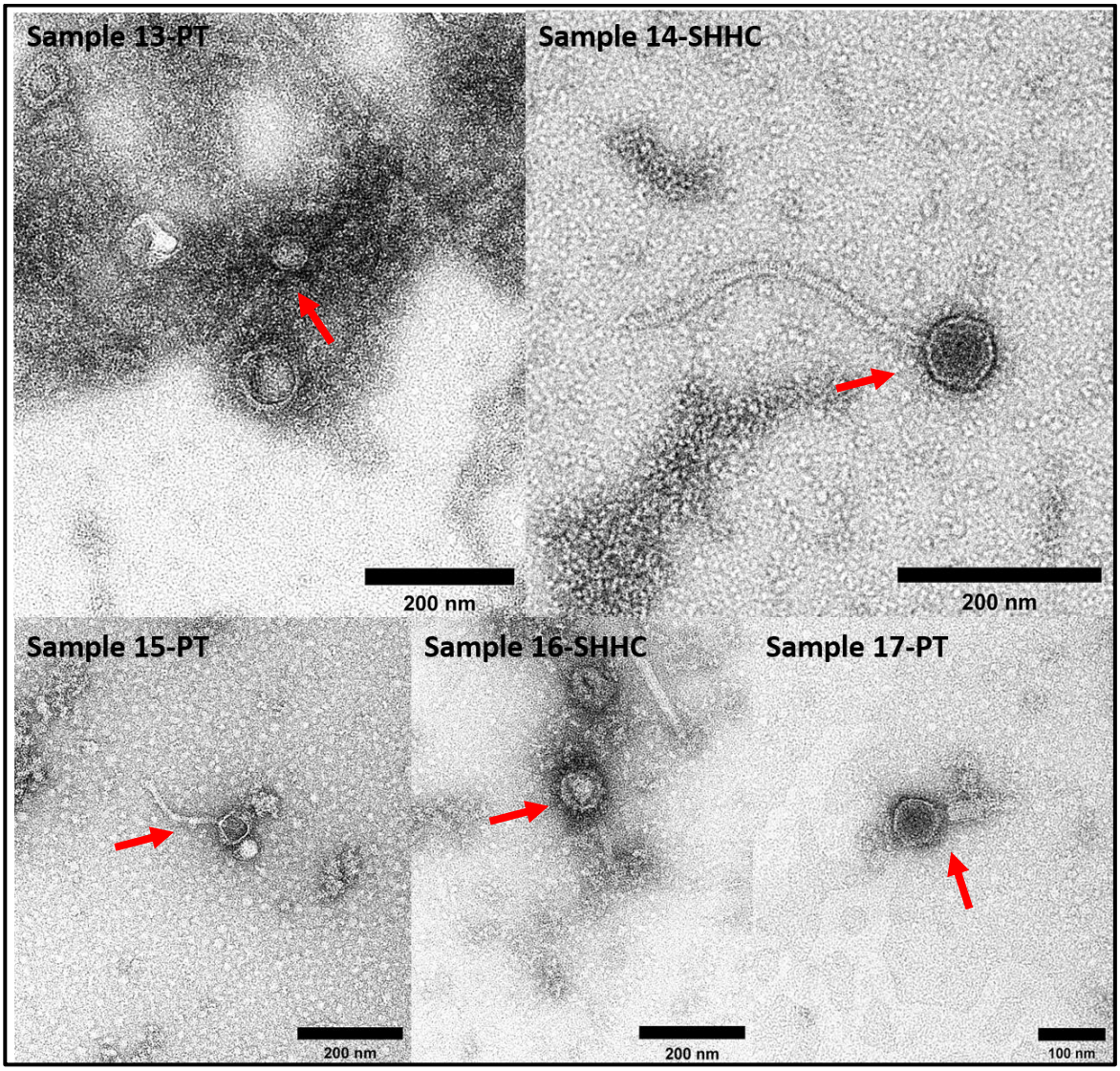


Figure 5.5. Examples of transmission electron micrographs of faecal VLPs collected from patient and SHHC samples. The red arrows indicate VLPs and the scale bars shown represent 100-200 nm in length. PT: severe ME/CFS patient; SHHC: same household healthy control.

Table 5.3. Dimensions and morphological classifications of faecal VLPs detected by TEM in ME/CFS patient and SHHC samples

Sample	Head (nm)	Tail (nm)	Total size (nm)	Classification to family
1	80.0	500.0	580.0	*Siphoviridae*
	80.0	50.0	130.0	*Podoviridae*
	64.3	100.0	164.3	*Siphoviridae*
	80.0	120.0	200.0	*Myoviridae*
	100.0	400.0	500.0	*Siphoviridae*
	90.0	77.0	167.0	*Myoviridae*
	66.7	160.0	226.7	*Siphoviridae*
	63.0	343.0	406.0	*Siphoviridae*
	91.0	1055.0	1146.0	*Siphoviridae*
	48.0	166.7	215.0	*Siphoviridae*
	69.0	187.5	256.5	*Siphoviridae*
	66.7	110.0	176.7	*Myoviridae*
	60.0	220.0	286.0	*Siphoviridae*
	60.0	100.0	160.0	*Myoviridae*
	68.0	90.0	158.0	*Myoviridae*
	84.0	263.0	347.0	*Siphoviridae*
	75.0	30.0	105.0	*Podoviridae*
2	53.3	133.0	186.3	*Siphoviridae*
	50.0	150.0	200.0	*Myoviridae*
	92.3	246.0	338.3	*Siphoviridae*
	77.8	100.0	177.8	*Myoviridae*
	60.0	300.0	360.0	*Siphoviridae*

	50.0	145.8	195.8	*Myoviridae*
3	107.7	246.0	353.7	*Siphoviridae*
	70.0	180.0	250.0	*Siphoviridae*
	136.4	655.0	791.4	*Siphoviridae*
	55.6	100.0	155.6	*Myoviridae*
	70.6	187.5	258.1	*Siphoviridae*
4	62.5	250.0	312.5	*Siphoviridae*
	106.0	112.5	218.5	*Myoviridae*
5	100.0	100.0	200.0	*Myoviridae*
6	60.0	200.0	260.0	*Siphoviridae*
	143.0	143.0	286.0	*Siphoviridae*
	112.5	375.0	487.5	*Siphoviridae*
	200.0	40.0	240.0	*Podoviridae*
7	100.0	333.0	433.0	*Siphoviridae*
	66.7	333.0	399.7	*Siphoviridae*
	36.5	40.5	77.0	*Myoviridae*
8	112.5	375.0	487.5	*Siphoviridae*
	71.4	315.0	386.4	*Siphoviridae*
	90.9	145.5	236.4	*Siphoviridae*
	111.1	322.2	433.3	*Siphoviridae*
	100.0	277.8	377.8	*Siphoviridae*
	80.0	220.0	300.0	*Siphoviridae*
	55.6	133.3	188.9	*Siphoviridae*
9	66.7	55.6	122.3	*Myoviridae*
	100.0	100.0	200.0	*Myoviridae*

10	62.5	112.5	175.0	*Myoviridae*
	57.1	42.9	100.0	*Podoviridae*
11	50.0	40.0	90.0	*Podoviridae*
12	155.6	245.0	400.6	*Siphoviridae*
	55.6	245.0	300.6	*Siphoviridae*
13	50.0	150.0	200.0	*Myoviridae*
14	71.4	321.4	392.8	*Siphoviridae*
15	61.9	166.7	228.6	*Siphoviridae*
16	85.7	100.0	185.7	*Myoviridae*
	83.3	333.3	416.6	*Siphoviridae*
	56.5	13.0	69.5	*Podoviridae*
17	69.2	115.4	184.6	*Myoviridae*
	75.0	37.5	112.5	*Podoviridae*
	63.6	181.8	245.4	*Siphoviridae*
	72.7	27.3	100.0	*Podoviridae*
	75.0	110.0	185.0	*Myoviridae*
	75.0	112.5	187.5	*Myoviridae*
	75.0	250.0	325.0	*Siphoviridae*
	68.2	160.0	228.2	*Siphoviridae*
	75.0	125.0	200.0	*Myoviridae*
	75.0	100.0	175.0	*Siphoviridae*

5.4.3. Correlation Analysis Between VLP Counts, Sample Weights and DNA Yields

By epifluorescence microscopy, VLP numbers (per gram frozen faeces) in patient and SHHC samples varied, ranging from 3.9 x 10^8 to 2.9 x 10^9 VLP/g faeces, with total VLP DNA ranging from 383.5 to 6,120.0 ng, equating to a range of 12.0 to 321.0 ng of VLP DNA per gram faeces (**Table 5.2**). Among these 17 samples, the average yields of VLP DNA was 2,518.5 ng ± 1,832.9 ng (mean ± S.D.), while the average VLP counts was 1.5 x 10^9 ± 8.1 x 10^8 VLP/g faeces (mean ± S.D.).

A direct comparison between patient and SHHC samples showed that EFM-based VLP counts and stool weights were positively correlated with a correlation coefficient: r = 0.18 (**Figure 5.6**) consistent with more faeces used equating to more faecal VLPs recovered.

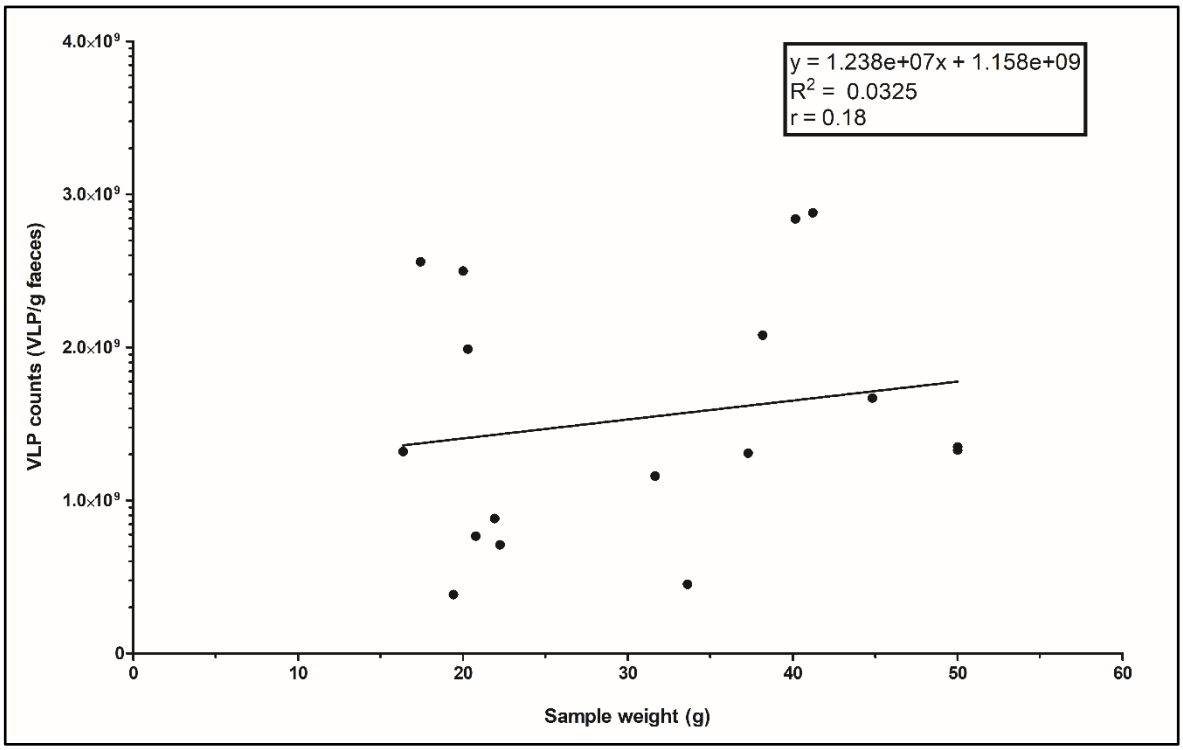


Figure 5.6. Linear regression and correlation analysis between sample weights and VLP counts in ME/CFS patient and SHHC samples. The solid line corresponds to a linear regression: y =1.238E+07x + 1.158E+09 (r = 0.18; n = 17).

Next, we investigated the relationships between viral counts, stool weights and VLP DNA yields in patient and SHHC samples. A direct comparison of the patient group (n = 9) showed a weak positive correlation between stool weights and VLP DNA yields (r = 0.258) (**Figure 5.7.A**), whereas VLPs per gram faeces and VLP DNA yields were negatively correlated (r = -0.347) (**Figure 5.7.B**). In the SHHC group (n = 8) stool weights and VLP DNA yields (r = -0.732) were negatively correlated as were VLP counts (VLPs/g faeces) and total VLP DNA (r = -0.059) (**Figure 5.8.A** and **B**).

Moreover, comparing faecal VLP counts in all samples showed that there was no significant difference between patient and SHHC samples: VLPs in patient samples were $1.6 \times 10^9 \pm 9.3 \times 10^8$ VLP/g faeces (mean ± S.D., n = 9), compared to $1.5 \times 10^9 \pm 7.1 \times 10^8$ VLP/g faeces (mean ± S.D., n = 8) for SHHC samples (**Table 5.2** and **Figure 5.9.A**). However, in comparing VLP DNA yields, VLP DNA yields from SHHCs were significantly higher than those in severe patients (p <0.05): the average VLP DNA yield in patient samples was 1,585.2 ± 1,411.0 ng (mean ± S.D.), while for SHHC samples it was 3,568.4 ± 1738.8 ng (mean ± S.D.) (**Table 5.2** and **Figure 5.9.B**).

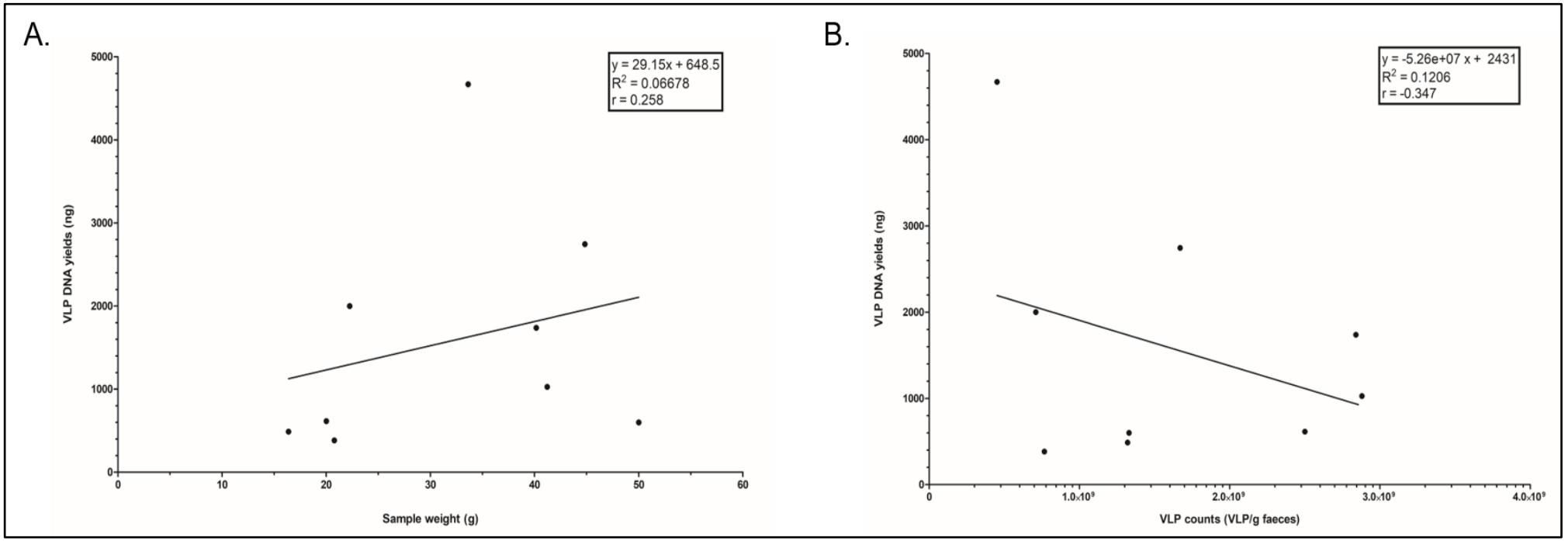


Figure 5.7. Linear regression and correlation analysis between sample weights, VLP counts, and VLP DNA yields in severe ME/CFS patients. (A) The solid line corresponds to a linear regression: y =29.15x + 648.5 (r = 0.258; n = 9), with positive correlation between stool weights and VLP DNA yields. (B) The solid line corresponds to a linear regression: y = -5.26E+07x + 2431 (r = -0.347), with negative correlation between VLP counts and VLP DNA yields.

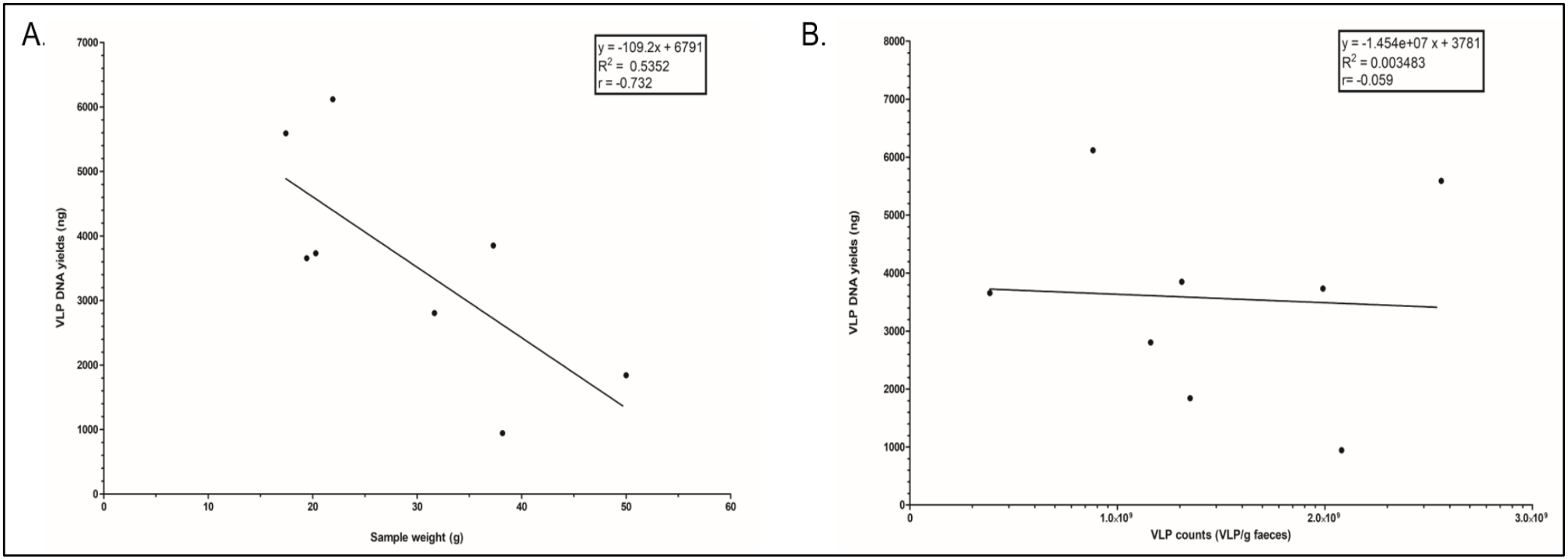


Figure 5.8. Linear regression and correlation analysis between sample weights, VLP counts and VLP DNA yields in same household healthy controls. (A) The solid line corresponds to a linear regression: y = -109.2x + 6791 (r = -0.732; n = 8), with negative correlation between stool weights and VLP DNA yields. (B) The solid line corresponds to a linear regression: y = -1.454E+07x + 3781 (r = -0.059), with negative correlation between VLP counts and VLP DNA yields.

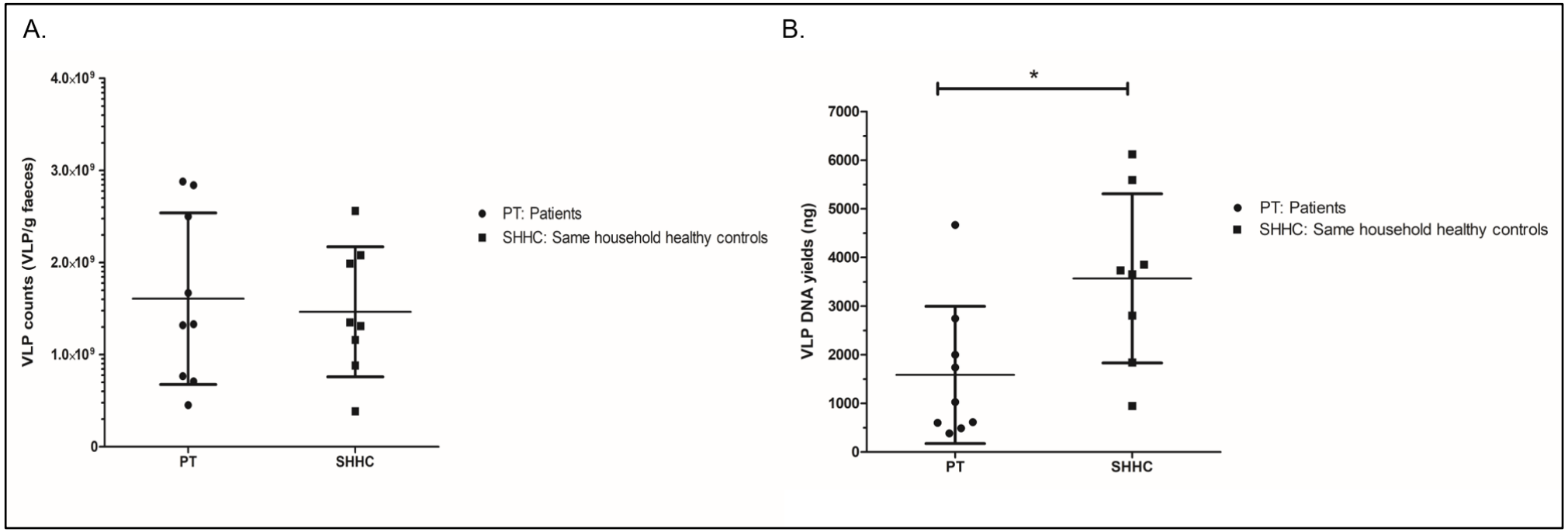


Figure 5.9. Comparisons of faecal VLP counts and faecal VLP DNA yields between ME/CFS patient and SHHC samples. (A) Scatter plot shows a comparison of VLP counts between patient (PT) and SHHC groups (mean ± S.D., no significant, two-tailed *t* test). (B) Scatter plot shows a comparison of total VLP DNA yields between PT and SHHC groups, indicating that VLP DNA yields in SHHC group are significantly higher than those in patient group (mean ± S.D., **p* <0.05, two-tailed *t* test).

5.5. Discussion

Previous studies investigating the human faecal virome have regularly used 0.22 μm filters to enrich and immobilise faecal VLPs for enumeration (Minot et al., 2013, Minot et al., 2011, Reyes et al., 2010, Breitbart et al., 2003). To improve this approach, I used dual filtration (0.8 μm and 0.45 μm) to enhance VLP recovery as part of my faecal VLP enumeration method based on the protocol described by Budinoff (2011). Using this optimised method we found that the faecal VLPs with nucleic acid-containing capsids in SHHC samples were higher than those of severe ME/CFS patients, although the variation and diversity of VLPs were seen in both sets of faecal samples.

5.5.1. Comparison of DIA-Based VLP Counts by Manual and Automated Methods

Due to their similar particle size and fluorescent intensity, it can be a challenge to distinguish small bacterial cells from fluorescence-stained viral particles and VLP clusters by EFM, and to accurately enumerate VLPs. Manual enumeration by eye is time consuming and low throughput. Thus, a fast and precise enumeration method for estimating VLP numbers in many samples is required. Digital images captured by EFM-connected cooled CCD camera can record more fluorescent particles and VLPs. Chen and colleagues (Chen et al., 2001) found a strong linear correlation between the viral counts by direct EFM and by direct image analysis (DIA). Furthermore, using water samples they found that the DIA counts exceed EFM counts on average by 1.31 fold, suggesting that some VLPs with weak fluorescent intensity may be difficult to identify by eye. In this study, I initially adopted the DIA-based method using ImageJ software to first detect and enumerate a reference phage (ΦB124-14) and optimise settings for the analysis of the ME/CFS patient and SHHC samples using an automated programme in ImageJ to count viral particles of up to 0.2 μm^2. However, the extent of background and electronic noise on digital images varies in faecal samples, which affected the accuracy of VLP counts. To verify the accuracy of VLP enumeration by DIA method, the thresholds in ImageJ for each image is optimised and we manually confirm the particle size, shape, number and distribution on photomicrographs by eye, comparing it to original raw images. ImageJ-based automated counting results in overestimations of VLP numbers in most samples, compared to the results of manual counting, demonstrating that the background and electronic noise displayed on electronic images have an impact on the accuracy of VLP detection and enumeration. However, manual counting from photomicrographs allows more faecal VLPs including those with weak staining to be identified. Manual counting-based DIA method for faecal VLP enumeration is therefore more accurate and reliable than the automated counting-based methods.

5.5.2. Quantitative Analysis of Faecal VLPs in Patient and SHHC Samples

In my findings, the number of faecal VLPs from ME/CFS and SHHC samples reaches up to ~10^9 VLPs per gram of faeces, consistent with previous studies (Hoyles et al., 2014, Lepage et al., 2008). However, due to unavoidable losses during VLP isolation and filtration, the VLP counts are conservative estimates with the true number of faecal/enteric VLPs being higher, as noted previously (Hoyles et al, 2014). Lepage *et al.* (2008) also demonstrated that by EFM there are ~10^{10} VLPs/mm^3 in human mucosal tissues.

The vast majority of faecal VLPs detected here are tailed bacteriophages belonging to the order *Caudovirales* with three families including *Siphoviridae*, *Myoviridae* as well as *Podoviridae*, of highly diverse abundance, sizes and structures represented. Overall, across all samples the most common morphotype of intact phage was *Siphoviridae*. Moreover, the largest VLP is a *Siphoviridae* of >1,000 nm in size found in sample 1 (patient) with the smallest VLP being a *Podoviridae* of approximately 70 nm in size found in sample 16 (healthy control). Most recent studies suggested that viruses of 50-100 nm in size with an average genome of 30-50 kb which includes dsDNA and ssDNA viruses most likely make up the majority of the human faecal virome (**Figure 5.10.A**) (Garmaeva et al., 2019). However, the TEM findings are not in total agreement with these faecal virome studies due to the presence of giant *Sipho*-like viruses in my samples. Their size and structure are comparable to other *Siphoviridae* members such as *Thermus thermophilus* phage P23-45 (**Figure 5.10.B**) and P74-26 isolated from hot springs, which are known for their extremely long tails (>800 nm) (Minakhin et al., 2008, Yu et al., 2006) These species have not to my knowledge been previously described in the human GIT or faeces, suggesting novel virulent strains potentially related to the human intestinal/faecal virome are present in patients and/or SHHC. Although dual filtration can remove the vast majority of faecal matter and bacterial cells, some giant VLPs may pass through the filter due to the diameters of their heads and tails being smaller than the filter pore size, but further investigation is required to confirm this. In addition, some faecal VLPs with intense fluorescent intensity may be VLP aggregates or giant viruses (i.e. Megaviruses or Jumbo phages) containing larger genomes (Devoto et al., 2019, Colson et al., 2013b).

The maximal number of VLPs detected by TEM were present in sample 1. In contrast, only one intact VLP is found in the sample 5, 11 and 13-15, respectively. To increase intact phage recovery and visibility, we used 300K MWCO filters to remove small contaminants and enrich VLPs from FFs prior to TEM analysis. We noted that this reduced the background and contaminating debris with viral structures becoming clearer in TEM images, with many detached phage heads and tails detected in most samples. This is consistent with the loss of intact viruses during storage at ultra-low temperature and/or the occurrence of intact VLP

disruption during isolation, as described previously (Steward and Culley, 2010, Sambrook and Russell, 2001).

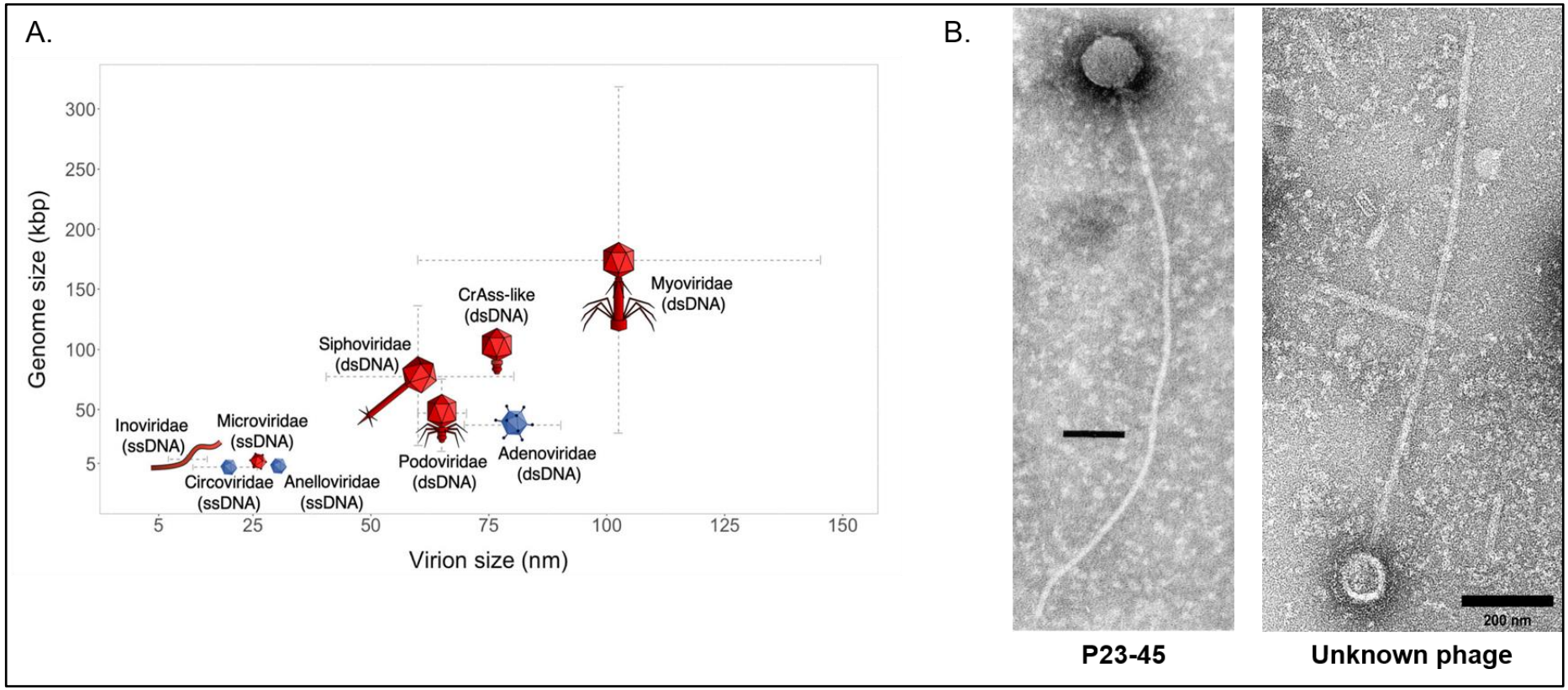


Figure 5.10. (A) Size distributions of viral genomes and the most dominant viral families in the human intestine. Prokaryotic viruses are shown in red and eukaryotic viruses are shown in blue. The most dominant viral families in the human intestine include *Inoviridae*, *Circoviridae*, *Adenoviridae*, *Microviridae*, *Podoviridae*, *Myoviridae*, *Siphoviridae*, *Anelloviridae* and crAss-like phages, with various genome and virion sizes (Garmaeva, 2019). **(B) Electron micrograph of a phage P23-45 (left) and a giant unknown Siphovirus (right) found in a severe ME/CFS patient (sample 1).** Both show the extremely long tail (>800 nm in length) and icosahedral head (~100 nm in diameter). Scale bar: 100 (left) and 200 nm (right). Electron micrograph of a phage P23-45 was reproduced from (Minakhin et al., 2008), with permission from Elsevier (Copyright © 2008).

5.5.3. Correlation Between VLP Counts, Stool Weights and DNA Yields

VLP counts and initial stool weight are positively correlated, indicating that larger sample sizes yield more VLPs. However, the data shows a weak correlation between VLP counts and sample sizes, suggesting that there may be other factors affecting VLP recovery and hence leading to VLP loss during sample preparation, as suggested previously (Hoyles et al., 2014). Hoyles and colleagues pointed out that many VLPs may be lost during centrifugation or filtration due to filter clogging. Therefore, the number of faecal VLPs determined is likely to be an underestimate. On the other hand, more VLPs isolated from faeces can theoretically yield higher amounts of VLP DNA. For the patient samples, the data identifies a weak positive correlation between initial stool sizes and VLP DNA yields, whereas the VLP counts (VLP/g faeces) and DNA yields are negatively correlated. In SHHC samples, there is a strong negative correlation between initial stool sizes and VLP DNA yields, with VLP counts (VLP/g faeces) and DNA yields also being negatively correlated. Possible explanations for these negative correlations and variable VLP recovery includes clogging of the 0.45 µm filter due to larger VLPs and/or VLP aggregates or high sample viscosity (Hoyles et al., 2014). VLP DNA may also be lost during extraction and purification steps. For example, residual DNase activity may destroy viral DNA prior to P/C/I extraction. Also, viral genomic DNA can be lost during P/C/I extraction and subsequent column-based purification. Contaminants such as salts, residual phenol and polysaccharides may also lower DNA yield. In addition, the efficiency of viral capsid lysis and viral DNA release from the capsid can influence the DNA yield. Multiple factors can therefore impact on the accuracy of VLP counts and VLP DNA yields and explain the relationships not fitting a simple linear regression model. Moreover, the small sample size used in this study has insufficient statistical power, impacting on statistical analysis, thereby leading to bias. Thus, further studies with larger sample sizes or cohorts is required.

In comparing VLP counts and VLP DNA yields between patient and SHHC samples, although total intact VLPs detected from patient samples (41 out of 68 totally) outnumber those detected from SHHC samples (27 out of 68 totally) based on TEM, there is no significant difference in EFM-based VLP counts between the groups. However, the VLP DNA yields from SHHC samples are significantly higher than those from patient samples. In considering the predominant viruses in human faeces, the theoretical amount of viral DNA in the human intestinal virome can be estimated as follows: the average number of VLPs per gram of faeces is around 10^9 (Hoyles et al., 2014, Lepage et al., 2008) and the average genome size of phages is approximately 30-50 kb (Hatfull, 2008), equating to a total amount of VLP DNA per gram of faeces of between 30×10^9 and 50×10^9 kb with an estimated mass of between 30 and 55 ng (Garmaeva et al., 2019) equating to 1 µg of VLP DNA being recovered from 18 g of faeces (approximately 55 ng/g faeces). The amount of

DNA I recovered from patient and SHHC samples ranges from 12 to 321 ng/g faeces, broadly consistent with the theoretical calculation, albeit with some samples yielding as little as 12-25 ng/g faeces as noted previously (Shkoporov and Hill, 2019, Shkoporov et al., 2018b).

5.6. Summary

In this chapter, I have adapted and optimised the DIA-based method for VLP enumeration by manual counting. Using this method I have shown that the faecal VLPs with nucleic acid-containing capsids in SHHC significantly outnumber those in severe ME/CFS patients. Variations in numbers and morphotypes of faecal VLPs have been observed amongst these samples. Although there is no significant difference in EFM-based VLP counts between patients and SHHC, total intact VLPs detected from patient samples outnumber those in SHHC samples based on TEM analysis, with the most predominant virus being *Siphoviridae*. Moreover, giant Siphoviruses were occasionally detected in the TEM findings, suggesting potential novel strains are likely present in these samples. The biological meaning of these findings is not clear and requires further investigation.

6. General Discussion

The human GIT contains a variety of commensal and/or potentially pathogenic (pathobionts) microbes including viruses. To understand the contribution the virome makes to human health and disease, there is a need to characterise it in detail including eukaryotic as well as prokaryotic viruses (i.e. bacteriophages or phages). The human virome is overlooked or ignored in most microbiome studies due to limited tools and methodologies hampering viral detection and cultivation. Its taxonomic composition therefore remains largely uncharacterised and is often referred to as “viral dark matter” (Reyes et al., 2012, Pedulla et al., 2003). Driven by the development of advanced high-throughput NGS and TGS technologies as well as up-to-date bioinformatics tools and reference viral databases, significant progress is now being made in viral metagenomics (Shkoporov and Hill, 2019, Reyes et al., 2012). Although many virome studies have begun to focus on protocol development to enrich VLPs and VLP DNA for high-throughput, in-depth metagenomic analysis (Shkoporov et al., 2018b, Kleiner et al., 2015, Conceicao-Neto et al., 2015, Castro-Mejia et al., 2015, Hoyles et al., 2014, Thurber et al., 2009), most have not thoroughly assessed the extent of contamination, the biases from isolation procedures used, random amplification to generate sequencing libraries, and the recovery efficiency of the adopted protocol.

In this thesis, I therefore first aimed to develop and optimise a protocol and a bioinformatics pipeline to characterise the composition of human intestinal/faecal virome using VLP enrichment-based methodology, systematically evaluating the efficiency of recovery/enrichment and the extent of contamination during isolation and purification, and assessing the extent of amplification bias from the use of PCR-amplified and PCR-free sequencing libraries. Optimised protocols as well as viromics pipeline have then been applied to an initial analysis of severely affected ME/CFS patients and same household healthy control (SHHC) subjects.

To investigate potential amplification bias from sequencing library preparation in virome-derived metagenomic datasets, I compared existing methods of library preparation, including PCR-based (LASL) and PCR-free (NASL) techniques. The principal benefit of random amplification approaches is to effectively enrich viral nucleic acid for high-throughput sequencing. However, the adverse effect is that PCR amplification simultaneously introduces systemic bias, particularly affecting genomes with extreme GC content and causing uneven amplified distribution of sequencing reads, thereby leading to over- or under-representation of viruses, as noted previously (Kallies et al., 2019, Parras-Moltó et al., 2018, Kim and Bae, 2011, Aird et al., 2011, Kozarewa et al., 2009). We demonstrated that PCR amplification leads to misrepresentation of viruses in their relative

abundance, likely resulting in misinterpreting the composition of viruses in human GIT, agreeing with a recent study (Parras-Moltó et al., 2018).

Recently, multiple commercial sequencing library preparation kits with PCR and PCR-free workflows for Illumina sequencing have been investigated, demonstrating that significant bias results from the genomes with low GC content seen in a particular library preparation kit (i.e. Nextera XT) (Sato et al., 2019). In addition, another bias of the PCR-based LASL method comes from the use of low amounts of input DNA which lowers its efficiency (Solonenko et al., 2013, Duhaime et al., 2012). Therefore, based on research purpose and user's requirement, selecting appropriate methodologies and kits for library preparation is important to minimise unnecessary bias (Sato et al., 2019). In this study, I used NEBNext Ultra II methodology for PCR (LASL) and non-PCR (NASL) library preparation, which has no significant GC-related sequencing bias unlike that seen using the Nextera XT technique (Sato et al., 2019). This thesis showed that PCR amplification reduces the richness and alpha diversity of viruses, while Chao1 is less well-suited for low-abundant viruses, as described previously (Haegeman et al., 2013). On the other hand, this thesis demonstrated that although high specificity of intestinal/faecal virome has been seen in each sample, agreeing with recent studies (Shkoporov et al., 2019, Shkoporov et al., 2018b), amplification also has a minor impact on virus distribution in beta diversity analysis between virome-derived PCR and non-PCR datasets, as supported by a recent study (Parras-Moltó et al., 2018).

To minimise potential PCR-associated artifacts, I first aimed to use amplification-free method for intestinal/faecal viromic analysis. However, the key limitation of amplification-free method is that larger amounts of faecal samples and DNA input are required for sequence-based analysis (Aird et al., 2011). In the process of VLP isolation and enrichment, the data showed that 0.8 µm filtration is beneficial in enhancing VLP recovery, as noted previously (Conceicao-Neto et al., 2015). However, a key drawback is that the removal of the contaminants is incomplete, including human or microbial cell debris, salts, proteins, lipids, carbohydrates, mucins and residue faecal material which can lower the efficiency of downstream enzymatic reactions and reduce VLP DNA recovery. Most current virome studies utilise 0.45 µm and/or 0.22 µm filtration to enrich VLPs and minimise contamination (Shkoporov et al., 2018b, Hoyles et al., 2014, Thurber et al., 2009). However, small pore size filters potentially introduce bias with the exclusion of the megavirome (e.g. *Mimivirus*) and may lead to VLP loss due to filter clogging (Hoyles et al., 2014). By evaluating different filter pore sizes and avoiding 0.2 µm filters, I demonstrated that a combination of 0.8 µm and 0.45 µm filtration is efficient at recovering viral capsids from a 3-4 g of stool sample, and is well-suited for sequence-based, VLP-enriched metagenomics using PCR-generated libraries. Importantly, a key advantage of my isolation/purification protocol is eliminating the

use of CsCl and chloroform to purify faecal VLPs. While CsCl is often used for EM-based analysis and *in vivo* studies, it is laborious and time-consuming, lacks reproducibility and often fails to recover enveloped viruses as well as those of atypical densities (Kleiner et al., 2015, Castro-Mejia et al., 2015). Also, some viruses and bacteriophages are sensitive to CsCl, such as *Guttaviridae*, *Nanoviridae* and *Orthomyxoviridae* (Thurber et al., 2009). Similarly, chloroform can also destabilise chloroform-sensitive viruses such as rotavirus and polyomavirus, as well as enveloped viruses such as coronavirus and mimivirus (Conceicao-Neto et al., 2015).

Based on the estimated number of faecal VLPs and VLP DNA yields described previously (Hoyles et al., 2014, Reyes et al., 2010, Lepage et al., 2008), I inferred that approximately 10^{10} VLPs with an average genome size of 50 kb theoretically equate to around 1 µg of VLP DNA, suggesting that reaching or exceeding 1 µg of VLP DNA requires large amounts of faeces (>10 grams) containing approximately 10^9 VLPs/g faeces. In the EFM-based analysis for ME/CFS and SHHC samples, I revealed that although the output of VLPs and VLP DNA varies from sample to sample, more faeces used broadly equate to more faecal VLPs and VLP DNA recovered, reaching around 10^8 to 10^9 VLPs/g faeces, as supported by previous studies (Castro-Mejia et al., 2015, Hoyles et al., 2014, Lepage et al., 2008). Thus, based on the theoretical estimation, 1 µg of VLP DNA can be recovered from around 18 g of frozen/dried faeces (i.e. approximately 55 ng/g faeces). The findings revealed that VLP DNA yields recovered from ME/CFS and SHHC samples range from 12 to 321 ng/g frozen faeces, broadly consistent with the theoretical calculation, albeit with some frozen samples yielding as little as 12-25 ng/g faeces, as noted elsewhere (Shkoporov and Hill, 2019, Shkoporov et al., 2018b). I therefore inferred that under most circumstances, small amounts of faeces (i.e. 3-5 grams or less) can yield sufficient VLPs and VLP DNA for amplification-based library preparation and shotgun metagenomic sequencing (Shkoporov et al., 2018b, Castro-Mejia et al., 2015), with larger amounts of faeces required to obtain sufficient VLPs and VLP DNA for amplification-free-based sequencing (Aird et al., 2011), particularly when amounts of sample are limiting or frozen stool samples are used. However, using large amounts of faeces to isolate VLP DNA is laborious and time-consuming, and is likely to be less efficient at removing contaminants that can interfere with VLP DNA recovery and lower VLP DNA quality.

In addition, this thesis have shown that members of the order *Caudovirales* are the most predominant viruses in human GIT, consistent with most virome studies (Shkoporov et al., 2018b, Manrique et al., 2016, Hoyles et al., 2014, Minot et al., 2013, Reyes et al., 2010, Breitbart et al., 2003). Of these, Siphoviruses were frequently observed in my TEM images consistent with the relative abundance and cluster analysis, as noted previously (Reyes et al., 2010, Breitbart et al., 2003). They are therefore likely to be the most dominant and

universal viruses in the human GIT virome. In particular, I identified giant *Sipho*-like viruses (with the tail >1,000 nm in length) by TEM. These giant, long-tailed viruses may be novel and may be virulent as giant viruses potentially have larger genomes and spaces to bring and supply essential lytic/virulent components (Shkoporov and Hill, 2019, Cui et al., 2014). Moreover, crAss-like phages were also seen in both sequencing datasets. Although the human intestinal virome is highly specific and diverse in the human population, crAssphages tend to be stable and universal over time in human populations, particularly in healthy individuals (Shkoporov et al., 2019, Minot et al., 2013, Reyes et al., 2010).

This thesis also showed that the considerable sequencing output with high sequencing depth and highly enriched intestinal/faecal virome in my shotgun metagenomic sequencing datasets facilitates in-depth analysis. While many short UViG sequences were seen in both viral metagenomic datasets, these can be removed in further virome studies to reduce complexity and to enrich true-positive UViGs using viral hallmark genes and by selecting longer contig length with the inclusion of circular genomes, as adopted in recent studies (Gregory et al., 2020, Shkoporov et al., 2018b, Parras-Moltó et al., 2018). Furthermore, several published non-redundant viral reference databases, such as NCBI RefSeq (O'Leary et al., 2016), the human gut virome database (GVD) (Gregory et al., 2020), the integrated microbiome genome/virus system (IMG/VR) (Roux et al., 2020, Paez-Espino et al., 2017) and/or a new reference viral database (RVDB) (Goodacre et al., 2018), need to be incorporated to build an in-house, non-redundant comprehensive virome database for virome studies including for ME/CFS, which can shed light on “unknown” taxonomic composition of intestinal viruses.

In conclusion, this thesis has provided robust and reliable protocols for human faecal VLP and VLP DNA isolation/purification and for faecal VLP estimation using an adapted DIA-based methodology. This thesis demonstrated the presence of PCR-associated artifacts in my intestinal/faecal virome-enriched metagenomic analysis based upon which I recommend using amplification-free-based approach for library preparation and sequencing to minimise misinterpretation in future virome studies whenever possible.

Further Work: Investigating the Human Intestinal Virome in ME/CFS

Finalised protocols and viromics pipeline have begun to be applied to the initial analysis of whole faecal and VLP-enriched metagenomes from ME/CFS and healthy control subjects. Giant viruses and VLPs have also been identified in ME/CFS and/or SHHC samples by TEM analysis but further investigation for their role and biological meaning is required.

7. References

Aaron, L. A., Burke, M. M. & Buchwald, D. 2000. Overlapping conditions among patients with chronic fatigue syndrome, fibromyalgia, and temporomandibular disorder. *Arch Intern Med,* 160**,** 221-227.

Ackermann, H. W. 1998. Tailed bacteriophages: the order caudovirales. *Adv Virus Res,* 51**,** 135-201.

Ackermann, H. W. 2007. 5500 Phages examined in the electron microscope. *Arch Virol,* 152**,** 227-243.

Ackermann, H. W. & Prangishvili, D. 2012. Prokaryote viruses studied by electron microscopy. *Arch Virol,* 157**,** 1843-1849.

Ahmed, T., Auble, D., Berkley, J. A., Black, R., Ahern, P. P., Hossain, M., Hsieh, A., Ireen, S., Arabi, M. & Gordon, J. I. 2014. An evolving perspective about the origins of childhood undernutrition and nutritional interventions that includes the gut microbiome. *Ann N Y Acad Sci,* 1332**,** 22-38.

Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C. & Gnirke, A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol,* 12**,** R18.

Aksyuk, A. A. & Rossmann, M. G. 2011. Bacteriophage assembly. *Viruses,* 3**,** 172-203.

Albertsson, P. A. & Frick, G. 1960. Partition of virus particles in a liquid two-phase system. *Biochim Biophys Acta,* 37**,** 230-237.

Alter, H. J., Mikovits, J. A., Switzer, W. M., Ruscetti, F. W., Lo, S. C., Klimas, N., Komaroff, A. L., Montoya, J. G., Bateman, L., Levine, S., Peterson, D., Levin, B., Hanson, M. R., Genfi, A., Bhat, M., Zheng, H., Wang, R., Li, B., Hung, G. C., Lee, L. L., Sameroff, S., Heneine, W., Coffin, J., Hornig, M. & Lipkin, W. I. 2012. A multicenter blinded analysis indicates no association between chronic fatigue syndrome/myalgic encephalomyelitis and either xenotropic murine leukemia virus-related virus or polytropic murine leukemia virus. *MBio,* 3**,** 1-7.

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., Chan, A. M., Haynes, M., Kelley, S., Liu, H., Mahaffy, J. M., Mueller, J. E., Nulton, J., Olson, R., Parsons, R., Rayhawk, S., Suttle, C. A. & Rohwer, F. 2006. The marine viromes of four oceanic regions. *PLoS Biol,* 4**,** e368:2121-2131.

Armitage, B. A. 2005. Cyanine dye-DNA interactions: Intercalation, groove binding, and aggregation. *DNA Binders and Related Subjects,* 253**,** 55-76.

Armstrong, C. W., Mcgregor, N. R., Lewis, D. P., Butt, H. L. & Gooley, P. R. 2017. The association of fecal microbiota and fecal, blood serum and urine metabolites in myalgic encephalomyelitis/chronic fatigue syndrome. *Metabolomics,* 13**,** 1-13.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J. M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H. B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E. G., Wang, J., Guarner, F., Pedersen, O., De Vos, W. M., Brunak, S., Dore, J., Meta, H. I. T. C., Antolin, M., Artiguenave, F., Blottiere, H. M., Almeida, M., Brechot, C., Cara, C., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G.,

Dervyn, R., Foerstner, K. U., Friss, C., Van De Guchte, M., Guedon, E., Haimet, F., Huber, W., Van Hylckama-Vlieg, J., Jamet, A., Juste, C., Kaci, G., Knol, J., Lakhdari, O., Layec, S., Le Roux, K., Maguin, E., Merieux, A., Melo Minardi, R., M'rini, C., Muller, J., Oozeer, R., Parkhill, J., Renault, P., Rescigno, M., Sanchez, N., Sunagawa, S., Torrejon, A., Turner, K., Vandemeulebrouck, G., Varela, E., Winogradsky, Y., Zeller, G., Weissenbach, J., Ehrlich, S. D. & Bork, P. 2011. Enterotypes of the human gut microbiome. *Nature,* 473**,** 174-180.

Azam, F., Fenchel, T., Field, J. G., Gray, J. S., Meyerreil, L. A. & Thingstad, F. 1983. The Ecological Role of Water-Column Microbes in the Sea. *Marine Ecology Progress Series,* 10**,** 257-263.

Baldridge, M. T., Nice, T. J., Mccune, B. T., Yokoyama, C. C., Kambal, A., Wheadon, M., Diamond, M. S., Ivanova, Y., Artyomov, M. & Virgin, H. W. 2015. Commensal microbes and interferon-lambda determine persistence of enteric murine norovirus infection. *Science,* 347**,** 266-269.

Bansal, A. S. 2016. Investigating unexplained fatigue in general practice with a particular focus on CFS/ME. *BMC Fam Pract,* 17**,** 81:1-14.

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., Salamon, P., Youle, M. & Rohwer, F. 2013. Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proc Natl Acad Sci U S A,* 110**,** 10771-10776.

Beghini, F., Mciver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Thomas, A. M., Manghi, P., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E. A. & Segata, N. 2020. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *bioRxiv***,** 2020.2011.2019.388223.

Behan, W. M., More, I. A. & Behan, P. O. 1991. Mitochondrial abnormalities in the postviral fatigue syndrome. *Acta Neuropathol,* 83**,** 61-65.

Beijerinck, M. W. 1898. Über ein Contagium vivum fluidum als Ursache der Fleckenkrankheit der Tabaksblätter. *Verhandelingen der Koninklijke Akademie van Wetenschappen,* 65**,** pp. 1-22.

Bell, D. S., Jordan, K. & Robinson, M. 2001. Thirteen-year follow-up of children and adolescents with chronic fatigue syndrome. *Pediatrics,* 107**,** 994-998.

Bergh, O., Borsheim, K. Y., Bratbak, G. & Heldal, M. 1989. High abundance of viruses found in aquatic environments. *Nature,* 340**,** 467-468.

Bettarel, Y., Sime-Ngando, T., Amblard, C. & Laveran, H. 2000. A comparison of methods for counting viruses in aquatic systems. *Appl Environ Microbiol,* 66**,** 2283-2289.

Bhutta, Z. A., Berkley, J. A., Bandsma, R. H. J., Kerac, M., Trehan, I. & Briend, A. 2017. Severe childhood malnutrition. *Nat Rev Dis Primers,* 3**,** 17067:1-44.

Biswal, N., Kleinschmidt, A. K., Spatz, H. C. & Trautner, T. A. 1967. Physical properties of the DNA of bacteriophage SP50. *Mol Gen Genet,* 100**,** 39-55.

Black, R. E., Victora, C. G., Walker, S. P., Bhutta, Z. A., Christian, P., De Onis, M., Ezzati, M., Grantham-Mcgregor, S., Katz, J., Martorell, R., Uauy, R., Maternal & Child Nutrition Study, G. 2013. Maternal and child undernutrition and overweight in low-income and middle-income countries. *Lancet,* 382**,** 427-451.

Bolduc, B., Jang, H. B., Doulcier, G., You, Z. Q., Roux, S. & Sullivan, M. B. 2017. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ,* 5**,** e3243:1-26.

Bradley, A. S., Ford, B. & Bansal, A. S. 2013. Altered functional B cell subset populations in patients with chronic fatigue syndrome compared to healthy controls. *Clin Exp Immunol,* 172**,** 73-80.

Bradley, D. E. 1967. Ultrastructure of bacteriophage and bacteriocins. *Bacteriol Rev,* 31**,** 230-314.

Breitbart, M., Haynes, M., Kelley, S., Angly, F., Edwards, R. A., Felts, B., Mahaffy, J. M., Mueller, J., Nulton, J., Rayhawk, S., Rodriguez-Brito, B., Salamon, P. & Rohwer, F. 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. & Rohwer, F. 2003. Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol,* 185**,** 6220-6223.

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

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

Brenchley, J. M. 2013. Mucosal immunity in human and simian immunodeficiency lentivirus infections. *Mucosal Immunol,* 6**,** 657-665.

Brenchley, J. M., Price, D. A., Schacker, T. W., Asher, T. E., Silvestri, G., Rao, S., Kazzaz, Z., Bornstein, E., Lambotte, O., Altmann, D., Blazar, B. R., Rodriguez, B., Teixeira-Johnson, L., Landay, A., Martin, J. N., Hecht, F. M., Picker, L. J., Lederman, M. M., Deeks, S. G. & Douek, D. C. 2006. Microbial translocation is a cause of systemic immune activation in chronic HIV infection. *Nat Med,* 12**,** 1365-1371.

Brown, C. T., Davis-Richardson, A. G., Giongo, A., Gano, K. A., Crabb, D. B., Mukherjee, N., Casella, G., Drew, J. C., Ilonen, J., Knip, M., Hyoty, H., Veijola, R., Simell, T., Simell, O., Neu, J., Wasserfall, C. H., Schatz, D., Atkinson, M. A. & Triplett, E. W. 2011. Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes. *PLoS One,* 6**,** e25792:1-9.

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

Budinoff, C. R., Loar, S. N., Lecleir, G. R., Wilhelm, S. W. & Buchan, A. 2011. A protocol for enumeration of aquatic viruses by epifluorescence microscopy using Anodisc 13 membranes. *BMC Microbiol,* 11:168**,** 1-6.

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. & Virgin, H. W. 2010. Virus-plus-susceptibility gene interaction determines Crohn's disease gene Atg16L1 phenotypes in intestine. *Cell,* 141**,** 1135-1145.

Cairns, R. & Hotopf, M. 2005. A systematic review describing the prognosis of chronic fatigue syndrome. *Occup Med (Lond),* 55**,** 20-31.

Carding, S. R., Davis, N. & Hoyles, L. 2017. Review article: the human intestinal virome in health and disease. *Aliment Pharmacol Ther,* 46**,** 800-815.

Carruthers, B. M., Van De Sande, M. I., De Meirleir, K. L., Klimas, N. G., Broderick, G., Mitchell, T., Staines, D., Powles, A. C., Speight, N., Vallings, R., Bateman, L., Baumgarten-Austrheim, B., Bell, D. S., Carlo-Stella, N., Chia, J., Darragh, A., Jo, D., Lewis, D., Light, A. R., Marshall-Gradisbik, S., Mena, I., Mikovits, J. A., Miwa, K., Murovska, M., Pall, M. L. & Stevens, S. 2011. Myalgic encephalomyelitis: International Consensus Criteria. *J Intern Med,* 270**,** 327-338.

Casjens, S. R. 2005. Comparative genomics and evolution of the tailed-bacteriophages. *Curr Opin Microbiol,* 8**,** 451-458.

Castro-Mejia, J. L., Muhammed, M. K., Kot, W., Neve, H., Franz, C. M., Hansen, L. H., Vogensen, F. K. & Nielsen, D. S. 2015. Optimizing protocols for extraction of bacteriophages prior to metagenomic analyses of phage communities in the human gut. *Microbiome,* 3:64**,** 1-14.

Centers for Disease Control and Prevention 2014. Revised surveillance case definition for HIV infection--United States, 2014. *MMWR Recomm Rep,* 63**,** 1-10.

Centers for Disease Control and Prevention 2019. Antibiotic resistance threats in the United States, 2019. Atlanta, Georgia, U.S.: Centers for Disease Control and Prevention**,** 1-139.

Chen, F., Lu, J. R., Binder, B. J., Liu, Y. C. & Hodson, R. E. 2001. Application of digital image analysis and flow cytometry to enumerate marine viruses stained with SYBR gold. *Appl Environ Microbiol,* 67**,** 539-545.

Chen, S., Zhou, Y., Chen, Y. & Gu, J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics,* 34**,** i884-i890.

Chia, J. K. & Chia, A. Y. 2008. Chronic fatigue syndrome is associated with chronic enterovirus infection of the stomach. *J Clin Pathol,* 61**,** 43-48.

Choi, Y., Shin, H., Lee, J. H. & Ryu, S. 2013. Identification and characterization of a novel flagellum-dependent Salmonella-infecting bacteriophage, iEPS5. *Appl Environ Microbiol,* 79**,** 4829-4837.

Cinek, O., Mazankova, K., Kramna, L., Odeh, R., Alassaf, A., Ibekwe, M. U., Ahmadov, G., Mekki, H., Abdullah, M. A., Elmahi, B. M. E., Hyoty, H. & Rainetova, P. 2018. Quantitative CrAssphage real-time PCR assay derived from data of multiple geographically distant populations. *J Med Virol,* 90**,** 767-771.

Clayton, E. W. 2015. Beyond myalgic encephalomyelitis/chronic fatigue syndrome: an IOM report on redefining an illness. *JAMA,* 313**,** 1101-1102.

Clooney, A. G., Sutton, T. D. S., Shkoporov, A. N., Holohan, R. K., Daly, K. M., O'regan, O., Ryan, F. J., Draper, L. A., Plevy, S. E., Ross, R. P. & Hill, C. 2019. Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. *Cell Host Microbe*, 26**,** 764-778.

Clowes, R. C. & Hayes, W. 1968. *Experiments in Microbial Genetics,* Oxford: Blackwell Scientific Publications.

Coleman, A. W. 1980. Enhanced Detection of Bacteria in Natural Environments by Fluorochrome Staining of DNA. *Limnology and Oceanography,* 25**,** 948-951.

Coleman, A. W., Maguire, M. J. & Coleman, J. R. 1981. Mithramycin- and 4'-6-diamidino-2-phenylindole (DAPI)-DNA staining for fluorescence microspectrophotometric

measurement of DNA in nuclei, plastids, and virus particles. *J Histochem Cytochem,* 29**,** 959-968.

Coleman, T. J., Gamble, D. R. & Taylor, K. W. 1973. Diabetes in mice after Coxsackie B 4 virus infection. *Br Med J,* 3**,** 25-27.

Colombet, J., Robin, A., Lavie, L., Bettarel, Y., Cauchie, H. M. & Sime-Ngando, T. 2007. Virioplankton 'pegylation': use of PEG (polyethylene glycol) to concentrate and purify viruses in pelagic ecosystems. *J Microbiol Methods,* 71**,** 212-219.

Colson, P., De Lamballerie, X., Yutin, N., Asgari, S., Bigot, Y., Bideshi, D. K., Cheng, X. W., Federici, B. A., Van Etten, J. L., Koonin, E. V., La Scola, B. & Raoult, D. 2013a. "Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. *Arch Virol,* 158**,** 2517-2521.

Colson, P., Fancello, L., Gimenez, G., Armougom, F., Desnues, C., Fournous, G., Yoosuf, N., Million, M., La Scola, B. & Raoult, D. 2013b. Evidence of the megavirome in humans. *J Clin Virol,* 57**,** 191-200.

Colson, P., La Scola, B., Levasseur, A., Caetano-Anolles, G. & Raoult, D. 2017. Mimivirus: leading the way in the discovery of giant viruses of amoebae. *Nat Rev Microbiol,* 15**,** 243-254.

Colson, P., Richet, H., Desnues, C., Balique, F., Moal, V., Grob, J. J., Berbis, P., Lecoq, H., Harle, J. R., Berland, Y. & Raoult, D. 2010. Pepper mild mottle virus, a plant virus associated with specific immune responses, Fever, abdominal pains, and pruritus in humans. *PLoS One,* 5**,** e10041:1-12.

Compston, N. D. 1978. An outbreak of encephalomyelitis in the Royal Free Hospital Group, London, in 1955. *Postgrad Med J,* 54**,** 722-724.

Conceicao-Neto, N., Zeller, M., Lefrere, H., De Bruyn, P., Beller, L., Deboutte, W., Yinda, C. K., Lavigne, R., Maes, P., Van Ranst, M., Heylen, E. & Matthijnssens, J. 2015. Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci Rep,* 5**,** 16532:1-14.

Cowlishaw, J. & Ginoza, W. 1970. Induction of lambda prophage by nalidixic acid. *Virology,* 41**,** 244-255.

Cui, J., Schlub, T. E. & Holmes, E. C. 2014. An allometric relationship between the genome length and virion volume of viruses. *J Virol,* 88**,** 6403-6410.

D'hérelle, F. 1917. Sur un microbe invisible antagoniste des bacilles dysentériques. *C R Acad Sci Ser D,* 165**,** 373–375.

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

De Goffau, M. C., Luopajarvi, K., Knip, M., Ilonen, J., Ruohtula, T., Harkonen, T., Orivuori, L., Hakala, S., Welling, G. W., Harmsen, H. J. & Vaarala, O. 2013. Fecal microbiota composition differs between children with beta-cell autoimmunity and those without. *Diabetes,* 62**,** 1238-1244.

De Paepe, M., Leclerc, M., Tinsley, C. R. & Petit, M. A. 2014. Bacteriophages: an underestimated role in human and animal health? *Front Cell Infect Microbiol,* 4**,** 1-11.

Dean, F. B., Nelson, J. R., Giesler, T. L. & Lasken, R. S. 2001. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res,* 11**,** 1095-1099.

Devoto, A. E., Santini, J. M., Olm, M. R., Anantharaman, K., Munk, P., Tung, J., Archie, E. A., Turnbaugh, P. J., Seed, K. D., Blekhman, R., Aarestrup, F. M., Thomas, B. C. & Banfield, J. F. 2019. Megaphages infect Prevotella and variants are widespread in gut microbiomes. *Nat Microbiol,* 4**,** 693-700.

Dillon, S. M., Lee, E. J., Kotter, C. V., Austin, G. L., Dong, Z., Hecht, D. K., Gianella, S., Siewe, B., Smith, D. M., Landay, A. L., Robertson, C. E., Frank, D. N. & Wilson, C. C. 2014. An altered intestinal mucosal microbiome in HIV-1 infection is associated with mucosal and systemic immune activation and endotoxemia. *Mucosal Immunol,* 7**,** 983-994.

Dinh, D. M., Volpe, G. E., Duffalo, C., Bhalchandra, S., Tai, A. K., Kane, A. V., Wanke, C. A. & Ward, H. D. 2015. Intestinal microbiota, microbial translocation, and systemic inflammation in chronic HIV infection. *J Infect Dis,* 211**,** 19-27.

Djikeng, A., Halpin, R., Kuzmickas, R., Depasse, J., Feldblyum, J., Sengamalay, N., Afonso, C., Zhang, X., Anderson, N. G., Ghedin, E. & Spiro, D. J. 2008. Viral genome sequencing by random priming methods. *BMC Genomics,* 9**,** 5:1-9.

Drulis-Kawa, Z., Olszak, T., Danis, K., Majkowska-Skrobek, G. & Ackermann, H. W. 2014. A giant Pseudomonas phage from Poland. *Arch Virol,* 159**,** 567-572.

Duerkop, B. A., Clements, C. V., Rollins, D., Rodrigues, J. L. & Hooper, L. V. 2012. A composite bacteriophage alters colonization by an intestinal commensal bacterium. *Proc Natl Acad Sci U S A,* 109**,** 17621-17626.

Duerkop, B. A. & Hooper, L. V. 2013. Resident viruses and their interactions with the immune system. *Nat Immunol,* 14**,** 654-659.

Duhaime, M. B., Deng, L., Poulos, B. T. & Sullivan, M. B. 2012. Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. *Environ Microbiol,* 14**,** 2526-2537.

Dutilh, B. E., Cassman, N., Mcnair, K., Sanchez, S. E., Silva, G. G., Boling, L., Barr, J. J., Speth, D. R., Seguritan, V., Aziz, R. K., Felts, B., Dinsdale, E. A., Mokili, J. L. & Edwards, R. A. 2014. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun,* 5**,** 4498:1-11.

Ebdon, J., Muniesa, M. & Taylor, H. 2007. The application of a recently isolated strain of Bacteroides (GB-124) to identify human sources of faecal pollution in a temperate river catchment. *Water Res,* 41**,** 3683-3690.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. & Turner, S. 2009. Real-time DNA sequencing from single polymerase molecules. *Science,* 323**,** 133-138.

Ewing, B. & Green, P. 1998. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research,* 8**,** 186-194.

Ewing, B., Hillier, L., Wendl, M. C. & Green, P. 1998. Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment. *Genome Research,* 8**,** 175-185.

Fauquet, C. M. 1999. TAXONOMY, CLASSIFICATION AND NOMENCLATURE OF VIRUSES. *Encyclopedia of Virology***,** 1730-1756.

Fauquet, C. M., Mayo, M. A., Maniloff, J., Desselberger, U. & Ball, L. A. 2005. The Single Stranded DNA Viruses. *In:* Fauquet, C. M., Mayo, M. A., Maniloff, J., Desselberger, U. & Ball, L. A. (eds.) *Virus Taxonomy.* San Diego: Academic Press.

Feldman, H. A. & Wang, S. S. 1961. Sensitivity of various viruses to chloroform. *Proc Soc Exp Biol Med,* 106**,** 736-738.

Foulongne, V., Sauvage, V., Hebert, C., Dereure, O., Cheval, J., Gouilh, M. A., Pariente, K., Segondy, M., Burguiere, A., Manuguerra, J. C., Caro, V. & Eloit, M. 2012. Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing. *PLoS One,* 7**,** e38499:1-11.

Foxman, E. F. & Iwasaki, A. 2011. Genome-virome interactions: examining the role of common viral infections in complex disease. *Nat Rev Microbiol,* 9**,** 254-264.

Fremont, M., Metzger, K., Rady, H., Hulstaert, J. & De Meirleir, K. 2009. Detection of herpesviruses and parvovirus B19 in gastric and intestinal mucosa of chronic fatigue syndrome patients. *In Vivo,* 23**,** 209-213.

Froussard, P. 1992. A random-PCR method (rPCR) to construct whole cDNA library from low amounts of RNA. *Nucleic Acids Res,* 20**,** 2900.

Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics,* 28**,** 3150-3152.

Fukuda, K., Straus, S. E., Hickie, I., Sharpe, M. C., Dobbins, J. G., Komaroff, A., Schluederberg, A., Jones, J. F., Lloyd, A. R., Wessely, S., Gantz, N. M., Holmes, G. P., Buchwald, D., Abbey, S., Rest, J., Levy, J. A., Jolson, H., Peterson, D. L., Vercoulen, J. H. M. M., Tirelli, U., Evengard, B., Natelson, B. H., Steele, L., Reyes, M. & Reeves, W. C. 1994. The Chronic Fatigue Syndrome - a Comprehensive Approach to Its Definition and Study. *Annals of Internal Medicine,* 121**,** 953-959.

Gansauge, M. T. & Meyer, M. 2013. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat Protoc,* 8**,** 737-748.

Garmaeva, S., Sinha, T., Kurilshikov, A., Fu, J., Wijmenga, C. & Zhernakova, A. 2019. Studying the gut virome in the metagenomic era: challenges and perspectives. *BMC Biol,* 17**,** 84:1-14.

George, B. 1869. Neurasthenia, or Nervous Exhaustion. *The Boston Medical and Surgical Journal,* 80**,** 217-221.

Germain, A., Ruppert, D., Levine, S. M. & Hanson, M. R. 2017. Metabolic profiling of a myalgic encephalomyelitis/chronic fatigue syndrome discovery cohort reveals disturbances in fatty acid and lipid metabolism. *Mol Biosyst,* 13**,** 371-379.

Ghurye, J. S., Cepeda-Espinoza, V. & Pop, M. 2016. Metagenomic Assembly: Overview, Challenges and Applications. *Yale J Biol Med,* 89**,** 353-362.

Giloteaux, L., Goodrich, J. K., Walters, W. A., Levine, S. M., Ley, R. E. & Hanson, M. R. 2016a. Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome. *Microbiome,* 4**,** 30:1-12.

Giloteaux, L., Hanson, M. R. & Keller, B. A. 2016b. A Pair of Identical Twins Discordant for Myalgic Encephalomyelitis/Chronic Fatigue Syndrome Differ in Physiological Parameters and Gut Microbiome Composition. *Am J Case Rep,* 17**,** 720-729.

Goerke, C., Koller, J. & Wolz, C. 2006. Ciprofloxacin and trimethoprim cause phage induction and virulence modulation in Staphylococcus aureus. *Antimicrob Agents Chemother,* 50**,** 171-177.

Goodacre, N., Aljanahi, A., Nandakumar, S., Mikailov, M. & Khan, A. S. 2018. A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection. *mSphere,* 3**,** e00069-18:1-18.

Goodwin, S., Mcpherson, J. D. & Mccombie, W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet,* 17**,** 333-351.

Gregory, A. C., Zablocki, O., Zayed, A. A., Howell, A., Bolduc, B. & Sullivan, M. B. 2020. The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host Microbe,* 28**,** 724-740.

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics,* 29**,** 1072-1075.

Haegeman, B., Hamelin, J., Moriarty, J., Neal, P., Dushoff, J. & Weitz, J. S. 2013. Robust estimation of microbial diversity in theory and in practice. *ISME J,* 7**,** 1092-1101.

Handley, S. A., Thackray, L. B., Zhao, G., Presti, R., Miller, A. D., Droit, L., Abbink, P., Maxfield, L. F., Kambal, A., Duan, E., Stanley, K., Kramer, J., Macri, S. C., Permar, S. R., Schmitz, J. E., Mansfield, K., Brenchley, J. M., Veazey, R. S., Stappenbeck, T. S., Wang, D., Barouch, D. H. & Virgin, H. W. 2012. Pathogenic simian immunodeficiency virus infection is associated with expansion of the enteric virome. *Cell,* 151**,** 253-266.

Hara, S., Terauchi, K. & Koike, I. 1991. Abundance of viruses in marine waters: assessment by epifluorescence and transmission electron microscopy. *Appl Environ Microbiol,* 57**,** 2731-2734.

Hatfull, G. F. 2008. Bacteriophage genomics. *Curr Opin Microbiol,* 11**,** 447-453.

Hatfull, G. F. & Hendrix, R. W. 2011. Bacteriophages and their genomes. *Curr Opin Virol,* 1**,** 298-303.

Hendrix, R. W. 2009. Jumbo bacteriophages. *Curr Top Microbiol Immunol,* 328**,** 229-240.

Hennes, K. P. & Suttle, C. A. 1995. Direct counts of viruses in natural waters and laboratory cultures by epifluorescence microscopy *Limnol. Oceanogr.,* 40**,** 1050-1055.

Hirons, G. T., Fawcett, J. J. & Crissman, H. A. 1994. TOTO and YOYO: new very bright fluorochromes for DNA content analyses by flow cytometry. *Cytometry,* 15**,** 129-140.

Holmes, G. P., Kaplan, J. E., Gantz, N. M., Komaroff, A. L., Schonberger, L. B., Straus, S. E., Jones, J. F., Dubois, R. E., Cunningham-Rundles, C., Pahwa, S. & Et Al. 1988. Chronic fatigue syndrome: a working case definition. *Ann Intern Med,* 108**,** 387-389.

Hoyles, L., Mccartney, A. L., Neve, H., Gibson, G. R., Sanderson, J. D., Heller, K. J. & Van Sinderen, D. 2014. Characterization of virus-like particles associated with the human faecal and caecal microbiota. *Res Microbiol,* 165**,** 803-812.

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W. & Hauser, L. J. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics,* 11**,** 119:1-11.

Illumina 2011. Quality scores for next-generation sequencing: Assessing sequencing accuracy using Phred quality scoring *Technical Note: Sequencing*.

ISO 10705-4:2001 Water quality -Detection and enumeration of bacteriophages - Part 4: Enumeration of bacteriophages infecting *Bacteroides fragilis*

Iwanowski, D. 1892. Über die Mosaikkrankheit der Tabakspflanze. *Bulletin Scientifique publié par l'Académie Impériale des Sciences de Saint-Pétersbourg / Nouvelle Serie III,* 35**,** pp. 67-70.

Jain, M., Olsen, H. E., Paten, B. & Akeson, M. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol,* 17**,** 239.

Jang, H. B., Bolduc, B., Zablocki, O., Kuhn, J. H., Roux, S., Adriaenssens, E. M., Brister, J. R., Kropinski, A. M., Krupovic, M., Lavigne, R., Turner, D. & Sullivan, M. B. 2019. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol,* 37**,** 632-639.

Jason, L. A., Benton, M. C., Valentine, L., Johnson, A. & Torres-Harding, S. 2008. The economic impact of ME/CFS: individual and societal costs. *Dyn Med,* 7**,** 6:1-8.

Jason, L. A., Richman, J. A., Rademaker, A. W., Jordan, K. M., Plioplys, A. V., Taylor, R. R., Mccready, W., Huang, C. F. & Plioplys, S. 1999. A community-based study of chronic fatigue syndrome. *Arch Intern Med,* 159**,** 2129-2137.

Joint United Nations Programme on Hiv/Aids 2020. UNAIDS data 2020. *In:* Joint United Nations Programme on Hiv/Aids (Unaids) (ed.). Geneva, Switzerland: Joint United Nations Programme on HIV/AIDS (UNAIDS).

Jones, M. K., Watanabe, M., Zhu, S., Graves, C. L., Keyes, L. R., Grau, K. R., Gonzalez-Hernandez, M. B., Iovine, N. M., Wobus, C. E., Vinje, J., Tibbetts, S. A., Wallet, S. M. & Karst, S. M. 2014. Enteric bacteria promote human and mouse norovirus infection of B cells. *Science,* 346**,** 755-759.

Kallies, R., Holzer, M., Brizola Toscan, R., Nunes Da Rocha, U., Anders, J., Marz, M. & Chatzinotas, A. 2019. Evaluation of Sequencing Library Preparation Protocols for Viral Metagenomic Analysis from Pristine Aquifer Groundwaters. *Viruses,* 11**,** 484:1-18.

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

Karlsson, O. E., Belak, S. & Granberg, F. 2013. The effect of preprocessing by sequence-independent, single-primer amplification (SISPA) on metagenomic detection of viruses. *Biosecur Bioterror,* 11 Suppl 1**,** S227-234.

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

Kim, K.-H. & 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.

Kim, K. H., Chang, H. W., Nam, Y. D., Roh, S. W., Kim, M. S., Sung, Y., Jeon, C. O., Oh, H. M. & Bae, J. W. 2008. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Appl Environ Microbiol,* 74**,** 5975-5985.

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

Kleiner, M., Hooper, L. V. & Duerkop, B. A. 2015. Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics,* 16**,** 7:1-15.

Klimas, N. G., Salvato, F. R., Morgan, R. & Fletcher, M. A. 1990. Immunologic abnormalities in chronic fatigue syndrome. *J Clin Microbiol,* 28**,** 1403-1410.

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M. & Turner, D. J. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods,* 6**,** 291-295.

Kramna, L., Kolarova, K., Oikarinen, S., Pursiheimo, J. P., Ilonen, J., Simell, O., Knip, M., Veijola, R., Hyoty, H. & Cinek, O. 2015. Gut virome sequencing in children with early islet autoimmunity. *Diabetes Care,* 38**,** 930-933.

Krupovic, M., Cvirkaite-Krupovic, V., Iranzo, J., Prangishvili, D. & Koonin, E. V. 2018. Viruses of archaea: Structural, functional, environmental and evolutionary genomics. *Virus Res,* 244**,** 181-193.

Kurochkina, L. P., Semenyuk, P. I., Sykilinda, N. N. & Miroshnikov, K. A. 2018. The unique two-component tail sheath of giant Pseudomonas phage PaBG. *Virology,* 515**,** 46-51.

Lane, R. J., Soteriou, B. A., Zhang, H. & Archard, L. C. 2003. Enterovirus related metabolic myopathy: a postviral fatigue syndrome. *J Neurol Neurosurg Psychiatry,* 74**,** 1382-1386.

Lasken, R. S. & Stockwell, T. B. 2007. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol,* 7**,** 19:1-11.

Lazarevic, V., Whiteson, K., Gaia, N., Gizard, Y., Hernandez, D., Farinelli, L., Osteras, M., Francois, P. & Schrenzel, J. 2012. Analysis of the salivary microbiome using culture-independent techniques. *J Clin Bioinforma,* 2**,** 4:1-8.

Lecuit, M. & Eloit, M. 2013. The human virome: new tools and concepts. *Trends Microbiol,* 21**,** 510-515.

Lepage, P., Colombet, J., Marteau, P., Sime-Ngando, T., Dore, J. & Leclerc, M. 2008. Dysbiosis in inflammatory bowel disease: a role for bacteriophages? *Gut,* 57**,** 424-425.

Lewis, G. D. & Metcalf, T. G. 1988. Polyethylene glycol precipitation for recovery of pathogenic viruses, including hepatitis A virus and human rotavirus, from oyster, water, and sediment samples. *Appl Environ Microbiol,* 54**,** 1983-1988.

Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics,* 31**,** 1674-1676.

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Genome Project Data Processing, S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-2079.

Li, S. K., Leung, R. K., Guo, H. X., Wei, J. F., Wang, J. H., Kwong, K. T., Lee, S. S., Zhang, C. & Tsui, S. K. 2012. Detection and identification of plasma bacterial and viral elements in HIV/AIDS patients in comparison to healthy adults. *Clin Microbiol Infect,* 18**,** 1126-1133.

Li, W. & Godzik, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* 22**,** 1658-1659.

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**,** 1228-1234.

Loebel, M., Strohschein, K., Giannini, C., Koelsch, U., Bauer, S., Doebis, C., Thomas, S., Unterwalder, N., Von Baehr, V., Reinke, P., Knops, M., Hanitsch, L. G., Meisel, C., Volk, H. D. & Scheibenbogen, C. 2014. Deficient EBV-specific B- and T-cell response in patients with chronic fatigue syndrome. *PLoS One,* 9**,** e85387:1-10.

Lombardi, V. C., Ruscetti, F. W., Das Gupta, J., Pfost, M. A., Hagen, K. S., Peterson, D. L., Ruscetti, S. K., Bagni, R. K., Petrow-Sadowski, C., Gold, B., Dean, M., Silverman, R. H. & Mikovits, J. A. 2009. Detection of an infectious retrovirus, XMRV, in blood cells of patients with chronic fatigue syndrome. *Science,* 326**,** 585-589.

Lorusso, L., Mikhaylova, S. V., Capelli, E., Ferrari, D., Ngonga, G. K. & Ricevuti, G. 2009. Immunological aspects of chronic fatigue syndrome. *Autoimmun Rev,* 8**,** 287-291.

Lotka, A. J. 1910. Contribution to the Theory of Periodic Reactions. *The Journal of Physical Chemistry,* 14**,** 271-274.

Luganini, A. & Gribaudo, G. 2020. Retroviruses of the Human Virobiota: The Recycling of Viral Genes and the Resulting Advantages for Human Hosts During Evolution. *Front Microbiol,* 11**,** 1140:1-8.

Lysholm, F., Wetterbom, A., Lindau, C., Darban, H., Bjerkner, A., Fahlander, K., Lindberg, A. M., Persson, B., Allander, T. & Andersson, B. 2012. Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. *PLoS One,* 7**,** e30875:1-12.

Ma, Y., You, X., Mai, G., Tokuyasu, T. & Liu, C. 2018. A human gut phage catalog correlates the gut phageome with type 2 diabetes. *Microbiome,* 6**,** 24:1-12.

MacDuff, D. A., Reese, T. A., Kimmey, J. M., Weiss, L. A., Song, C., Zhang, X., Kambal, A., Duan, E., Carrero, J. A., Boisson, B., Laplantine, E., Israel, A., Picard, C., Colonna, M., Edelson, B. T., Sibley, L. D., Stallings, C. L., Casanova, J. L., Iwai, K. & Virgin, H. W. 2015. Phenotypic complementation of genetic immunodeficiency by chronic herpesvirus infection. *Elife,* 4**,** e04494:1-20.

Manrique, P., Bolduc, B., Walk, S. T., Van Der Oost, J., De Vos, W. M. & Young, M. J. 2016. Healthy human gut phageome. *Proc Natl Acad Sci U S A,* 113**,** 10400-10405.

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

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., Mcdade, K. E., Mckenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. & Rothberg, J. M. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437**,** 376-380.

Mateen, F. J. & Black, R. E. 2013. Expansion of acute flaccid paralysis surveillance: beyond poliomyelitis. *Trop Med Int Health,* 18**,** 1421-1422.

Matos, R. C., Lapaque, N., Rigottier-Gois, L., Debarbieux, L., Meylheuc, T., Gonzalez-Zorn, B., Repoila, F., Lopes Mde, F. & Serror, P. 2013. Enterococcus faecalis prophage dynamics and contributions to pathogenic traits. *PLoS Genet,* 9**,** e1003539:1-16.

McKibbin, J. M. & Taylor, W. E. 1949. The nitrogenous constituents of the tissue lipides; the determination of sphingosine in tissue lipide extracts. *J Biol Chem,* 178**,** 29-35.

Mcmurdie, P. J. & Holmes, S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One,* 8**,** e61217:1-11.

Meals, R. W., Hauser, V. F. & Bower, A. G. 1935a. Poliomyelitis-The Los Angeles Epidemic of 1934 : Part I. *Cal West Med,* 43**,** 123-125.

Meals, R. W., Hauser, V. F. & Bower, A. G. 1935b. Poliomyelitis-The Los Angeles Epidemic of 1934: Part II. *Cal West Med,* 43**,** 215-222.

Medical Staff of the Royal Free Hospital 1957. An outbreak of encephalomyelitis in the Royal Free Hospital Group, London, in 1955. *Br Med J,* 2**,** 895-904.

Meessen-Pinard, M., Sekulovic, O. & Fortier, L. C. 2012. Evidence of in vivo prophage induction during Clostridium difficile infection. *Appl Environ Microbiol,* 78**,** 7662-7670.

Mejia-Leon, M. E., Petrosino, J. F., Ajami, N. J., Dominguez-Bello, M. G. & De La Barca, A. M. 2014. Fecal microbiota imbalance in Mexican children with type 1 diabetes. *Sci Rep,* 4**,** 3814:1-5.

Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. 2018. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics,* 34**,** i142-i150.

Mikheyev, A. S. & Tin, M. M. 2014. A first look at the Oxford Nanopore MinION sequencer. *Mol Ecol Resour,* 14**,** 1097-1102.

Milani, C., Casey, E., Lugli, G. A., Moore, R., Kaczorowska, J., Feehily, C., Mangifesta, M., Mancabelli, L., Duranti, S., Turroni, F., Bottacini, F., Mahony, J., Cotter, P. D., Mcauliffe, F. M., Van Sinderen, D. & Ventura, M. 2018. Tracing mother-infant transmission of bacteriophages by means of a novel analytical tool for shotgun metagenomic datasets: METAnnotatorX. *Microbiome,* 6**,** 145:1-16.

Miller, D. N., Bryant, J. E., Madsen, E. L. & Ghiorse, W. C. 1999. Evaluation and optimization of DNA extraction and purification procedures for soil and sediment samples. *Appl Environ Microbiol,* 65**,** 4715-4724.

Minakhin, L., Goel, M., Berdygulova, Z., Ramanculov, E., Florens, L., Glazko, G., Karamychev, V. N., Slesarev, A. I., Kozyavkin, S. A., Khromov, I., Ackermann, H. W., Washburn, M., Mushegian, A. & Severinov, K. 2008. Genome comparison and proteomic characterization of Thermus thermophilus bacteriophages P23-45 and P74-26: siphoviruses with triplex-forming sequences and the longest known tails. *J Mol Biol,* 378**,** 468-480.

Minot, S., Bryson, A., Chehoud, C., Wu, G. D., Lewis, J. D. & 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. & 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. & Bushman, F. D. 2011. The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res,* 21**,** 1616-1625.

Modi, S. R., Lee, H. H., Spina, C. S. & Collins, J. J. 2013. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature,* 499**,** 219-222.

Mohsen, M. O., Gomes, A. C., Vogel, M. & Bachmann, M. F. 2018. Interaction of Viral Capsid-Derived Virus-Like Particles (VLPs) with the Innate Immune System. Vaccines (Basel), 6**,** 1-12

Monaco, C. L., Gootenberg, D. B., Zhao, G., Handley, S. A., Ghebremichael, M. S., Lim, E. S., Lankowski, A., Baldridge, M. T., Wilen, C. B., Flagg, M., Norman, J. M., Keller, B. C., Luevano, J. M., Wang, D., Boum, Y., Martin, J. N., Hunt, P. W., Bangsberg, D. R., Siedner, M. J., Kwon, D. S. & Virgin, H. W. 2016. Altered Virome and Bacterial Microbiome in Human Immunodeficiency Virus-Associated Acquired Immunodeficiency Syndrome. *Cell Host Microbe,* 19**,** 311-322.

Morgan, C., Rose, H. M. & Moore, D. H. 1956. Structure and development of viruses observed in the electron microscope. III. Influenza virus. *J Exp Med,* 104**,** 171-182.

Morris, G., Berk, M., Galecki, P. & Maes, M. 2014. The emerging role of autoimmunity in myalgic encephalomyelitis/chronic fatigue syndrome (ME/cfs). *Mol Neurobiol,* 49**,** 741-756.

Mosier-Boss, P. A., Lieberman, S. H., Andrews, J. M., Rohwer, F. L., Wegley, L. E. & Breitbart, M. 2003. Use of fluorescently labeled phage in the detection and identification of bacterial species. *Appl Spectrosc,* 57**,** 1138-1144.

Murphy, J., Royer, B., Mahony, J., Hoyles, L., Heller, K., Neve, H., Bonestroo, M., Nauta, A. & Van Sinderen, D. 2013. Biodiversity of lactococcal bacteriophages isolated from 3 Gouda-type cheese-producing plants. *J Dairy Sci,* 96**,** 4945-4957.

Murri, M., Leiva, I., Gomez-Zumaquero, J. M., Tinahones, F. J., Cardona, F., Soriguer, F. & Queipo-Ortuno, M. I. 2013. Gut microbiota in children with type 1 diabetes differs from that in healthy children: a case-control study. *BMC Med,* 11**,** 46:1-12.

Myhill, S., Booth, N. E. & Mclaren-Howard, J. 2009. Chronic fatigue syndrome and mitochondrial dysfunction. *Int J Clin Exp Med,* 2**,** 1-16.

Nacul, L. C., Lacerda, E. M., Pheby, D., Campion, P., Molokhia, M., Fayyaz, S., Leite, J. C., Poland, F., Howe, A. & Drachler, M. L. 2011. Prevalence of myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) in three regions of England: a repeated cross-sectional study in primary care. *BMC Med,* 9**,** 91:1-12.

Naess, H., Nyland, M., Hausken, T., Follestad, I. & Nyland, H. I. 2012. Chronic fatigue syndrome after Giardia enteritis: clinical characteristics, disability and long-term sickness absence. *BMC Gastroenterol,* 12**,** 13:1-7.

Nakamura, S., Yang, C. S., Sakon, N., Ueda, M., Tougan, T., Yamashita, A., Goto, N., Takahashi, K., Yasunaga, T., Ikuta, K., Mizutani, T., Okamoto, Y., Tagami, M., Morita, R., Maeda, N., Kawai, J., Hayashizaki, Y., Nagai, Y., Horii, T., Iida, T. & Nakaya, T. 2009. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS One,* 4**,** e4219:1-8.

Nayfach, S., Camargo, A. P., Eloe-Fadrosh, E., Roux, S. & Kyrpides, N. 2020. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol,* 2020.

Nepusz, T., Yu, H. & Paccanaro, A. 2012. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods,* 9**,** 471-472.

Noble, R. T. & Fuhrman, J. A. 1998. Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. *Aquatic Microbial Ecology,* 14**,** 113-118.

Nokso-Koivisto, J., Kinnari, T. J., Lindahl, P., Hovi, T. & Pitkaranta, A. 2002. Human picornavirus and coronavirus RNA in nasopharynx of children without concurrent respiratory symptoms. *J Med Virol,* 66**,** 417-420.

Norman, J. M., Handley, S. A., Baldridge, M. T., Droit, L., Liu, C. Y., Keller, B. C., Kambal, A., Monaco, C. L., Zhao, G., Fleshner, P., Stappenbeck, T. S., Mcgovern, D. P., Keshavarzian, A., Mutlu, E. A., Sauk, J., Gevers, D., Xavier, R. J., Wang, D., Parkes, M. & Virgin, H. W. 2015. Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell,* 160**,** 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., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., Mcgarvey, K. M., Murphy, M. R., O'neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., Dicuccio, M., Kitts, P., Murphy, T. D. & Pruitt, K. D. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res,* 44**,** D733-745.

Ogilvie, L. A., Caplin, J., Dedi, C., Diston, D., Cheek, E., Bowler, L., Taylor, H., Ebdon, J. & Jones, B. V. 2012. Comparative (meta)genomic analysis and ecological profiling of human gut-specific bacteriophage phiB124-14. *PLoS One,* 7**,** e35053:1-17.

Oldstone, M. B., Nerenberg, M., Southern, P., Price, J. & Lewicki, H. 1991. Virus infection triggers insulin-dependent diabetes mellitus in a transgenic model: role of anti-self (virus) immune response. *Cell,* 65**,** 319-331.

Olsen, R. H., Siak, J. S. & Gray, R. H. 1974. Characteristics of PRD1, a plasmid-dependent broad host range DNA bacteriophage. *J Virol,* 14**,** 689-699.

Orlova, E. V. 2012. Bacteriophages and their structural organisation. *In:* Kurtböke, I. (ed.) *Bacteriophages.* Croatia: InTech.

Oulas, A., Pavloudi, C., Polymenakou, P., Pavlopoulos, G. A., Papanikolaou, N., Kotoulas, G., Arvanitidis, C. & Iliopoulos, I. 2015. Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinform Biol Insights,* 9**,** 75-88.

Paez-Espino, D., Chen, I. A., Palaniappan, K., Ratner, A., Chu, K., Szeto, E., Pillay, M., Huang, J., Markowitz, V. M., Nielsen, T., Huntemann, M., Tb, K. R., Pavlopoulos, G. A., Sullivan, M. B., Campbell, B. J., Chen, F., Mcmahon, K., Hallam, S. J., Denef, V., Cavicchioli, R., Caffrey, S. M., Streit, W. R., Webster, J., Handley, K. M., Salekdeh, G. H., Tsesmetzis, N., Setubal, J. C., Pope, P. B., Liu, W. T., Rivers, A. R., Ivanova, N. N. & Kyrpides, N. C. 2017. IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. *Nucleic Acids Res,* 45**,** D457-D465.

Paprotka, T., Delviks-Frankenberry, K. A., Cingoz, O., Martinez, A., Kung, H. J., Tepper, C. G., Hu, W. S., Fivash, M. J., Jr., Coffin, J. M. & Pathak, V. K. 2011. Recombinant origin of the retrovirus XMRV. *Science,* 333**,** 97-101.

Parras-Moltó, M., Rodríguez-Galet, A., Suárez-Rodríguez, P. & López-Bueno, A. 2018. Evaluation of bias induced by viral enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses. *Microbiome,* 6**,** 119:1-18.

Pedulla, M. L., Ford, M. E., Houtz, J. M., Karthikeyan, T., Wadsworth, C., Lewis, J. A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N. R., Brucker, W., Kumar, V., Kandasamy, J., Keenan, L., Bardarov, S., Kriakov, J., Lawrence, J. G., Jacobs, W. R., Jr., Hendrix, R. W. & Hatfull, G. F. 2003. Origins of highly mosaic mycobacteriophage genomes. *Cell,* 113**,** 171-182.

Pemberton, S. & Cox, D. L. 2014. Experiences of daily activity in chronic fatigue syndrome/myalgic encephalomyelitis (CFS/ME) and their implications for rehabilitation programmes. *Disabil Rehabil,* 36**,** 1790-1797.

Petersen, C. & Round, J. L. 2014. Defining dysbiosis and its influence on host immunity and disease. Cell Microbiol, 16**,** 1024-1033

Pfeiffer, J. K. & Virgin, H. W. 2016. Viral immunity. Transkingdom control of viral infection and immunity in the mammalian intestine. *Science,* 351**,** 239-244.

Pirtle, E. C. & Beran, G. W. 1991. Virus survival in the environment. *Rev Sci Tech,* 10**,** 733-748.

Popgeorgiev, N., Boyer, M., Fancello, L., Monteil, S., Robert, C., Rivet, R., Nappez, C., Azza, S., Chiaroni, J., Raoult, D. & Desnues, C. 2013. Marseillevirus-like virus recovered from blood donated by asymptomatic humans. *J Infect Dis,* 208**,** 1042-1050.

Porter, K. G. & Feig, Y. S. 1980. The use of DAPI for identifying and counting aquatic microflora. *Limnol. Oceanogr.,* 25**,** 943-948.

Pride, D. T., Salzman, J., Haynes, M., Rohwer, F., Davis-Long, C., White, R. A., 3rd, Loomer, P., Armitage, G. C. & Relman, D. A. 2012. Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. *ISME J,* 6**,** 915-926.

Proctor, L. M. & Fuhrman, J. A. 1990. Viral Mortality of Marine-Bacteria and Cyanobacteria. *Nature,* 343**,** 60-62.

Public Health England 2019. Annual epidemiological commentary: Gram-negative bacteraemia, MRSA bacteraemia, MSSA bacteraemia and C. difficile infections, up to and including financial year April 2018 to March 2019. *In:* Chudasama, D., Thelwall, S., Nsonwu, O., Rooney, G., Wasti, S. & Hope, R. (eds.). London, UK: Public Health England (PHE).

Purnell, S., Ebdon, J., Buck, A., Tupper, M. & Taylor, H. 2015. Bacteriophage removal in a full-scale membrane bioreactor (MBR) - Implications for wastewater reuse. *Water Res,* 73**,** 109-117.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., Wu, P., Dai, Y., Sun, X., Li, Z., Tang, A., Zhong, S., Li, X., Chen, W., Xu, R., Wang, M., Feng, Q., Gong, M., Yu, J., Zhang, Y., Zhang, M., Hansen, T., Sanchez, G., Raes, J., Falony, G., Okuda, S., Almeida, M., Lechatelier, E., Renault, P., Pons, N., Batto, J. M., Zhang, Z., Chen, H., Yang, R., Zheng, W., Li, S., Yang, H., Wang, J., Ehrlich, S. D., Nielsen, R., Pedersen, O., Kristiansen, K. & Wang, J. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature,* 490**,** 55-60.

Raoult, D. & Forterre, P. 2008. Redefining viruses: lessons from Mimivirus. *Nat Rev Microbiol,* 6**,** 315-319.

Raoult, D., La Scola, B. & Birtles, R. 2007. The discovery and characterization of Mimivirus, the largest known virus and putative pneumonia agent. *Clin Infect Dis,* 45**,** 95-102.

Rasa, S., Nora-Krukle, Z., Henning, N., Eliassen, E., Shikova, E., Harrer, T., Scheibenbogen, C., Murovska, M., Prusty, B. K. & European Network On, M. C. 2018. Chronic viral infections in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS). *J Transl Med,* 16**,** 268:1-25.

Reid, S., Chalder, T., Cleare, A., Hotopf, M. & Wessely, S. 2000. Chronic fatigue syndrome. *BMJ,* 320**,** 292-296.

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. 2017. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome,* 5**,** 69:1-20.

Reyes, A., Blanton, L. V., Cao, S., Zhao, G., Manary, M., Trehan, I., Smith, M. I., Wang, D., Virgin, H. W., Rohwer, F. & Gordon, J. I. 2015. Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. *Proc Natl Acad Sci U S A,* 112**,** 11941-11946.

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

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

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

Rhoads, A. & Au, K. F. 2015. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics,* 13**,** 278-289.

Rodrigues, L. S., Da Silva Nali, L. H., Leal, C. O. D., Sabino, E. C., Lacerda, E. M., Kingdon, C. C., Nacul, L. & Romano, C. M. 2019. HERV-K and HERV-W transcriptional activity in myalgic encephalomyelitis/chronic fatigue syndrome. *Auto Immun Highlights,* 10**,** 12:1-5.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. & Nyren, P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem,* 242**,** 84-89.

Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A. & Sullivan, M. B. 2017. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ,* 5**,** e3817:1-26.

Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. 2015. VirSorter: mining viral signal from microbial genomic data. *PeerJ,* 3**,** e985:1-20.

Roux, S., Krupovic, M., Daly, R. A., Borges, A. L., Nayfach, S., Schulz, F., Sharrar, A., Matheus Carnevali, P. B., Cheng, J. F., Ivanova, N. N., Bondy-Denomy, J., Wrighton, K. C., Woyke, T., Visel, A., Kyrpides, N. C. & Eloe-Fadrosh, E. A. 2019. Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nat Microbiol,* 4**,** 1895-1906.

Roux, S., Krupovic, M., Poulet, A., Debroas, D. & Enault, F. 2012. Evolution and diversity of the Microviridae viral family through a collection of 81 new complete genomes assembled from virome reads. *PLoS One,* 7**,** e40418:1-12.

Roux, S., Paez-Espino, D., Chen, I. A., Palaniappan, K., Ratner, A., Chu, K., Reddy, T. B. K., Nayfach, S., Schulz, F., Call, L., Neches, R. Y., Woyke, T., Ivanova, N. N., Eloe-Fadrosh, E. A. & Kyrpides, N. C. 2020. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res,* gkaa946**,** 1-12.

Roux, S., Solonenko, N. E., Dang, V. T., Poulos, B. T., Schwenck, S. M., Goldsmith, D. B., Coleman, M. L., Breitbart, M. & Sullivan, M. B. 2016. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ,* 4**,** e2777:1-17.

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J. & Walker, A. W. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol,* 12**,** 87:1-12.

Sambrook, J. & Russell, D. W. 2001. *Molecular cloning: A laboratory manual,* New York, USA, Cold Spring Harbor Laboratory Press.

Sansone, C. L., Cohen, J., Yasunaga, A., Xu, J., Osborn, G., Subramanian, H., Gold, B., Buchon, N. & Cherry, S. 2015. Microbiota-Dependent Priming of Antiviral Intestinal Immunity in Drosophila. *Cell Host Microbe,* 18**,** 571-581.

Santiago-Rodriguez, T. M., Ly, M., Bonilla, N. & Pride, D. T. 2015. The human urine virome in association with urinary tract infections. *Front Microbiol,* 6**,** 14:1-12.

Sato, M. P., Ogura, Y., Nakamura, K., Nishida, R., Gotoh, Y., Hayashi, M., Hisatsune, J., Sugai, M., Takehiko, I. & Hayashi, T. 2019. Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. *DNA Res,* 26**,** 391-398.

Sausset, R., Petit, M. A., Gaboriau-Routhiau, V. & De Paepe, M. 2020. New insights into intestinal phages. *Mucosal Immunology,* 13**,** 205-215.

Sauvage, V., Laperche, S., Cheval, J., Muth, E., Dubois, M., Boizeau, L., Hebert, C., Lionnet, F., Lefrere, J. J. & Eloit, M. 2016. Viral metagenomics applied to blood donors and recipients at high risk for blood-borne infections. *Blood Transfus,* 14**,** 400-407.

Schadt, E. E., Turner, S. & Kasarskis, A. 2010. A window into third-generation sequencing. *Human Molecular Genetics,* 19**,** R227-R240.

Schowalter, R. M., Pastrana, D. V., Pumphrey, K. A., Moyer, A. L. & Buck, C. B. 2010. Merkel cell polyomavirus and two previously unknown polyomaviruses are chronically shed from human skin. *Cell Host Microbe,* 7**,** 509-515.

Schreiner, P., Harrer, T., Scheibenbogen, C., Lamer, S., Schlosser, A., Naviaux, R. K. & Prusty, B. K. 2020. Human Herpesvirus-6 Reactivation, Mitochondrial Fragmentation, and the Coordination of Antiviral and Metabolic Phenotypes in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome. *Immunohorizons,* 4**,** 201-215.

Selva, L., Viana, D., Regev-Yochay, G., Trzcinski, K., Corpa, J. M., Lasa, I., Novick, R. P. & Penades, J. R. 2009. Killing niche competitors by remote-control bacteriophage induction. *Proc Natl Acad Sci U S A,* 106**,** 1234-1238.

Serwer, P., Hayes, S. J., Thomas, J. A. & Hardies, S. C. 2007. Propagating the missing bacteriophages: a large bacteriophage in a new class. *Virol J,* 4**,** 21:1-5.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res,* 13**,** 2498-2504.

Sharpe, M. C., Archard, L. C., Banatvala, J. E., Borysiewicz, L. K., Clare, A. W., David, A., Edwards, R. H., Hawton, K. E., Lambert, H. P., Lane, R. J. & Et Al. 1991. A report--chronic fatigue syndrome: guidelines for research. *J R Soc Med,* 84**,** 118-121.

Sheedy, J. R., Wettenhall, R. E., Scanlon, D., Gooley, P. R., Lewis, D. P., Mcgregor, N., Stapleton, D. I., Butt, H. L. & Kl, D. E. M. 2009. Increased d-lactic Acid intestinal bacteria in patients with chronic fatigue syndrome. *In Vivo,* 23**,** 621-628.

Shen, W., Le, S., Li, Y. & Hu, F. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One,* 11**,** e0163962:1-10.

Shkoporov, A. N., Clooney, A. G., Sutton, T. D. S., Ryan, F. J., Daly, K. M., Nolan, J. A., Mcdonnell, S. A., Khokhlova, E. V., Draper, L. A., Forde, A., Guerin, E., Velayudhan, V., Ross, R. P. & Hill, C. 2019. The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host Microbe,* 26**,** 527-541.

Shkoporov, A. N. & Hill, C. 2019. Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome. *Cell Host Microbe,* 25**,** 195-209.

Shkoporov, A. N., Khokhlova, E. V., Fitzgerald, C. B., Stockdale, S. R., Draper, L. A., Ross, R. P. & Hill, C. 2018a. PhiCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nat Commun,* 9**,** 4781:1-8.

Shkoporov, A. N., Ryan, F. J., Draper, L. A., Forde, A., Stockdale, S. R., Daly, K. M., Mcdonnell, S. A., Nolan, J. A., Sutton, T. D. S., Dalmasso, M., Mccann, A., Ross, R. P. & Hill, C. 2018b. Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome,* 6**,** 68:1-17.

Shukla, S. K., Cook, D., Meyer, J., Vernon, S. D., Le, T., Clevidence, D., Robertson, C. E., Schrodi, S. J., Yale, S. & Frank, D. N. 2015. Changes in Gut and Plasma Microbiome following Exercise Challenge in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS). *PLoS One,* 10**,** e0145453:1-15.

Sieburth, J. M., Johnson, P. W. & Hargraves, P. E. 1988. Ultrastructure and Ecology of Aureococcus-Anophagefferens Gen-Et-Sp-Nov (Chrysophyceae) - the Dominant Picoplankter during a Bloom in Narragansett Bay, Rhode-Island, Summer 1985. *Journal of Phycology,* 24**,** 416-425.

Siedner, M. J., Kim, J. H., Nakku, R. S., Bibangambah, P., Hemphill, L., Triant, V. A., Haberer, J. E., Martin, J. N., Mocello, A. R., Boum, Y., 2nd, Kwon, D. S., Tracy, R. P., Burdo, T., Huang, Y., Cao, H., Okello, S., Bangsberg, D. R. & Hunt, P. W. 2016. Persistent Immune Activation and Carotid Atherosclerosis in HIV-Infected Ugandans Receiving Antiretroviral Therapy. *J Infect Dis,* 213**,** 370-378.

Sigurdsson, B., Sigurjónsson, J., Sigurdsson, J. H. J., Thorkelsson, J. & Gudmundsson, K. R. 1950. A DISEASE EPIDEMIC IN ICELAND SIMULATING POLIOMYELITIS. *American Journal of Epidemiology,* 52**,** 222-238.

Smith, M. I., Yatsunenko, T., Manary, M. J., Trehan, I., Mkakosya, R., Cheng, J., Kau, A. L., Rich, S. S., Concannon, P., Mychaleckyj, J. C., Liu, J., Houpt, E., Li, J. V., Holmes, E., Nicholson, J., Knights, D., Ursell, L. K., Knight, R. & Gordon, J. I. 2013. Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. *Science,* 339**,** 548-554.

Solonenko, S. A., Ignacio-Espinoza, J. C., Alberti, A., Cruaud, C., Hallam, S., Konstantinidis, K., Tyson, G., Wincker, P. & Sullivan, M. B. 2013. Sequencing platform and library preparation choices impact viral metagenomes. *BMC Genomics,* 14**,** 320:1-12.

Stachler, E., Kelty, C., Sivaganesan, M., Li, X., Bibby, K. & Shanks, O. C. 2017. Quantitative CrAssphage PCR Assays for Human Fecal Pollution Measurement. *Environ Sci Technol,* 51**,** 9146-9154.

Steele, L., Dobbins, J. G., Fukuda, K., Reyes, M., Randall, B., Koppelman, M. & Reeves, W. C. 1998. The epidemiology of chronic fatigue in San Francisco. *Am J Med,* 105**,** 83S-90S.

Stene, L. C., Oikarinen, S., Hyoty, H., Barriga, K. J., Norris, J. M., Klingensmith, G., Hutton, J. C., Erlich, H. A., Eisenbarth, G. S. & Rewers, M. 2010. Enterovirus infection and progression from islet autoimmunity to type 1 diabetes: the Diabetes and Autoimmunity Study in the Young (DAISY). *Diabetes,* 59**,** 3174-3180.

Steward, G. F. & Culley, A. I. 2010. Extraction and purification of nucleic acids from viruses. *In:* Wilhelm, S. W., Weinbauer, M. G. & Suttle, C. A. (eds.) *Manual of Aquatic Viral Ecology.* Waco, TX 76710, USA: American Society of Limnology and Oceanography, Inc.

Stremlau, M. H., Andersen, K. G., Folarin, O. A., Grove, J. N., Odia, I., Ehiane, P. E., Omoniwa, O., Omoregie, O., Jiang, P. P., Yozwiak, N. L., Matranga, C. B., Yang, X., Gire, S. K., Winnicki, S., Tariyal, R., Schaffner, S. F., Okokhere, P. O., Okogbenin, S., Akpede, G. O., Asogun, D. A., Agbonlahor, D. E., Walker, P. J.,

Tesh, R. B., Levin, J. Z., Garry, R. F., Sabeti, P. C. & Happi, C. T. 2015. Discovery of novel rhabdoviruses in the blood of healthy individuals from West Africa. *PLoS Negl Trop Dis,* 9**,** e0003631:1-17.

Subramanian, S., Huq, S., Yatsunenko, T., Haque, R., Mahfuz, M., Alam, M. A., Benezra, A., Destefano, J., Meier, M. F., Muegge, B. D., Barratt, M. J., Vanarendonk, L. G., Zhang, Q., Province, M. A., Petri, W. A., Jr., Ahmed, T. & Gordon, J. I. 2014. Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature,* 510**,** 417-421.

Sullivan, M. B., Huang, K. H., Ignacio-Espinoza, J. C., Berlin, A. M., Kelly, L., Weigele, P. R., Defrancesco, A. S., Kern, S. E., Thompson, L. R., Young, S., Yandava, C., Fu, R., Krastins, B., Chase, M., Sarracino, D., Osburne, M. S., Henn, M. R. & Chisholm, S. W. 2010. Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environ Microbiol,* 12**,** 3035-3056.

Suttle, C. A. 1994. The significance of viruses to mortality in aquatic microbial communities. *Microb Ecol,* 28**,** 237-243.

Suttle, C. A., Chan, A. M. & Cottrell, M. T. 1990. Infection of Phytoplankton by Viruses and Reduction of Primary Productivity. *Nature,* 347**,** 467-469.

Suttle, C. A., Chan, A. M. & Cottrell, M. T. 1991. Use of ultrafiltration to isolate viruses from seawater which are pathogens of marine phytoplankton. *Appl Environ Microbiol,* 57**,** 721-726.

Sutton, T. D. S., Clooney, A. G., Ryan, F. J., Ross, R. P. & Hill, C. 2019. Choice of assembly software has a critical impact on virome characterisation. *Microbiome,* 7**,** 12:1-15.

Szekely, A. J. & Breitbart, M. 2016. Single-stranded DNA phages: from early molecular biology tools to recent revolutions in environmental microbiology. *FEMS Microbiol Lett,* 363**,** fnw027:1-9.

Tamboli, C. P., Neut, C., Desreumaux, P. & Colombel, J. F. 2004. Dysbiosis in inflammatory bowel disease. *Gut,* 53**,** 1-4.

Tartera, C., Araujo, R., Michel, T. & Jofre, J. 1992. Culture and decontamination methods affecting enumeration of phages infecting Bacteroides fragilis in sewage. *Appl Environ Microbiol,* 58**,** 2670-2673.

The Lancet 1956. A New Clinical Entity? *The Lancet,* 267**,** 789-790.

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

Tisza, M. J., Pastrana, D. V., Welch, N. L., Stewart, B., Peretti, A., Starrett, G. J., Pang, Y. S., Krishnamurthy, S. R., Pesavento, P. A., Mcdermott, D. H., Murphy, P. M., Whited, J. L., Miller, B., Brenchley, J., Rosshart, S. P., Rehermann, B., Doorbar, J., Ta'ala, B. A., Pletnikova, O., Troncoso, J. C., Resnick, S. M., Bolduc, B., Sullivan, M. B., Varsani, A., Segall, A. M. & Buck, C. B. 2020. Discovery of several thousand highly diverse circular DNA viruses. *Elife,* 9:1-26.

Tuma, R. S., Beaudet, M. P., Jin, X., Jones, L. J., Cheung, C. Y., Yue, S. & Singer, V. L. 1999. Characterization of SYBR Gold nucleic acid gel stain: a dye optimized for use with 300-nm ultraviolet transilluminators. *Anal Biochem,* 268**,** 278-288.

Twort, F. W. 1915. An Investigation on the Nature of Ultra-Microscopic Viruses. *Lancet,* 186**,** 1241-1243.

Underhill, R. A. 2015. Myalgic encephalomyelitis, chronic fatigue syndrome: An infectious disease. *Med Hypotheses,* 85**,** 765-773.

United Nations Children's Fund, World Health Organization & World Bank Group 2020. Levels and trends in child malnutrition: Key Findings of the 2020 Edition of the Joint Child Malnutrition Estimates. *In:* United Nations Children's Fund, World Health Organization & World Bank Group (eds.) 2020 ed. Geneva, Switzerland: World Health Organization.

Van't Leven, M., Zielhuis, G. A., Van Der Meer, J. W., Verbeek, A. L. & Bleijenberg, G. 2010. Fatigue and chronic fatigue syndrome-like complaints in the general population. *Eur J Public Health,* 20**,** 251-257.

Van Deusen, E. H. 1869. Observations on a Form of Nervous Prostration, (Neurasthenia,) Culminating in Insanity. *American Journal of Psychiatry,* 25**,** 445-461.

Victoria, J. G., Kapoor, A., Li, L., Blinkova, O., Slikas, B., Wang, C., Naeem, A., Zaidi, S. & Delwart, E. 2009. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *J Virol,* 83**,** 4642-4651.

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

Wang, Y., Zhu, N., Li, Y., Lu, R., Wang, H., Liu, G., Zou, X., Xie, Z. & Tan, W. 2016. Metagenomic analysis of viral genetic diversity in respiratory samples from children with severe acute respiratory infection in China. *Clin Microbiol Infect,* 22**,** 458 e451-459.

Weigle, J., Meselson, M. & Paigen, K. 1959. Density Alterations Associated with Transducing Ability in the Bacteriophage-Lambda. *J Mol Biol,* 1**,** 379-386.

Weinbauer, M. G. & Suttle, C. A. 1997. Comparison of epifluorescence and transmission electron microscopy for counting viruses in natural marine waters. *Aquatic Microbial Ecology,* 13**,** 225-232.

Wessely, S., Chalder, T., Hirsch, S., Wallace, P. & Wright, D. 1997. The prevalence and morbidity of chronic fatigue and chronic fatigue syndrome: a prospective primary care study. *Am J Public Health,* 87**,** 1449-1455.

White, D. W., Suzanne Beard, R. & Barton, E. S. 2012. Immune modulation during latent herpesvirus infection. *Immunol Rev,* 245**,** 189-208.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis.,* New York, USA, Springer-Verlag.

Wickham, H., Averick, M., Bryan, J., Chang, W., Mcgowan, L., François, R., Grolemund, G., Hayes, A., Henry, L. L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. & Yutani, H. 2019. Welcome to the Tidyverse. *Journal of Open Source Software,* 4**,** 1-6.

Wilhelm, S. W. & Suttle, C. A. 1999. Viruses and Nutrient Cycles in the Sea - Viruses play critical roles in the structure and function of aquatic food webs. *Bioscience,* 49**,** 781-788.

Williams, M. V., Cox, B., Lafuse, W. P. & Ariza, M. E. 2019. Epstein-Barr Virus dUTPase Induces Neuroinflammatory Mediators: Implications for Myalgic Encephalomyelitis/Chronic Fatigue Syndrome. *Clin Ther,* 41**,** 848-863.

Willner, D., Furlan, M., Haynes, M., Schmieder, R., Angly, F. E., Silva, J., Tammadoni, S., Nosrat, B., Conrad, D. & Rohwer, F. 2009. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS One,* 4**,** e7370:1-12.

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

Wommack, K. E., Hill, R. T., Kessel, M., Russekcohen, E. & Colwell, R. R. 1992. Distribution of Viruses in the Chesapeake Bay. *Applied and Environmental Microbiology,* 58**,** 2965-2970.

Wood, D. E., Lu, J. & Langmead, B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol,* 20**,** 257:1-13.

Wood, D. E. & Salzberg, S. L. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol,* 15**,** R46:1-12.

World Health Organization 2016. *Global report on diabetes,* Geneva, Switzerland, World Health Organization, 1-86.

World Health Organization 2019. *Classification of diabetes mellitus 2019* Geneva, Switzerland, World Health Organization, 1-37.

Wylie, K. M., Mihindukulasuriya, K. A., Sodergren, E., Weinstock, G. M. & Storch, G. A. 2012. Sequence analysis of the human virome in febrile and afebrile children. *PLoS One,* 7**,** e27735:1-10.

Xenopoulos, M. A. & Bird, D. F. 1997. Virus à la sauce Yo-Pro: Microwave enhanced staining for counting viruses by epifluorescence microscopy. *Limnology and Oceanography,* 42**,** 1648-1650.

Yamada, T., Satoh, S., Ishikawa, H., Fujiwara, A., Kawasaki, T., Fujie, M. & Ogata, H. 2010. A jumbo phage infecting the phytopathogen Ralstonia solanacearum defines a new lineage of the Myoviridae family. *Virology,* 398**,** 135-147.

Yu, M. X., Slater, M. R. & Ackermann, H. W. 2006. Isolation and characterization of Thermus bacteriophages. *Arch Virol,* 151**,** 663-679.

Yuan, Y. & Gao, M. 2017. Jumbo Bacteriophages: An Overview. *Front Microbiol,* 8**,** 403:1-9.

Yutin, N., Makarova, K. S., Gussow, A. B., Krupovic, M., Segall, A., Edwards, R. A. & Koonin, E. V. 2018. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat Microbiol,* 3**,** 38-46.

Zeltins, A. 2013. Construction and characterization of virus-like particles: a review. Mol Biotechnol, 53**,** 92-107

Zhang, K., Martiny, A. C., Reppas, N. B., Barry, K. W., Malek, J., Chisholm, S. W. & Church, G. M. 2006a. Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol,* 24**,** 680-686.

Zhang, T., Breitbart, M., Lee, W. H., Run, J. Q., Wei, C. L., Soh, S. W., Hibberd, M. L., Liu, E. T., Rohwer, F. & Ruan, Y. 2006b. RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol,* 4**,** e3:0108-0118.

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

Zhao, G., Vatanen, T., Droit, L., Park, A., Kostic, A. D., Poon, T. W., Vlamakis, H., Siljander, H., Harkonen, T., Hamalainen, A. M., Peet, A., Tillmann, V., Ilonen, J., Wang, D., Knip, M., Xavier, R. J. & Virgin, H. W. 2017. Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. *Proc Natl Acad Sci U S A,* 115**,** E6166-E6175.

Zheng, Q., Chen, Q., Xu, Y., Suttle, C. A. & Jiao, N. 2018. A Virus Infecting Marine Photoheterotrophic Alphaproteobacteria (Citromicrobium spp.) Defines a New Lineage of ssDNA Viruses. *Front Microbiol,* 9**,** 1418:1-7.

Zheng, T., Li, J., Ni, Y., Kang, K., Misiakou, M. A., Imamovic, L., Chow, B. K. C., Rode, A. A., Bytzer, P., Sommer, M. & Panagiotou, G. 2019. Mining, analyzing, and integrating viral signals from metagenomic data. *Microbiome,* 7**,** 42:1-15.

Zhong, X., Guidoni, B., Jacas, L. & Jacquet, S. 2015. Structure and diversity of ssDNA Microviridae viruses in two peri-alpine lakes (Annecy and Bourget, France). *Res Microbiol,* 166**,** 644-654.

Zolfo, M., Pinto, F., Asnicar, F., Manghi, P., Tett, A., Bushman, F. D. & Segata, N. 2019. Detecting contamination in viromes using ViromeQC. *Nat Biotechnol,* 37**,** 1408-1412.

Zoll, J., Rahamat-Langendoen, J., Ahout, I., De Jonge, M. I., Jans, J., Huijnen, M. A., Ferwerda, G., Warris, A. & Melchers, W. J. 2015. Direct multiplexed whole genome sequencing of respiratory tract samples reveals full viral genomic information. *J Clin Virol,* 66**,** 6-11.

Zuo, T., Wong, S. H., Lam, K., Lui, R., Cheung, K., Tang, W., Ching, J. Y. L., Chan, P. K. S., Chan, M. C. W., Wu, J. C. Y., Chan, F. K. L., Yu, J., Sung, J. J. Y. & Ng, S. C. 2017. Bacteriophage transfer during faecal microbiota transplantation in Clostridium difficile infection is associated with treatment outcome. *Gut,* 67**,** 634-643.

8. Appendices

Appendix 1. Letters of Ethical Approval

Faculty of Medicine and Health Sciences Research Ethics Committee

University of East Anglia

Daniel Vipond
MED

Research & Enterprise Services
West Office (Science Building)
University of East Anglia
Norwich Research Park
Norwich, NR4 7TJ

Telephone: +44 (0) 1603 591720

Email: fmh.ethics@uea.ac.uk

Web: www.uea.ac.uk/researchandenterprise

16th July 2015

Dear Daniel

Title: A role for a leaky gut and the intestinal microbiota in the pathophysiology of chronic fatigue syndrome/myalgic encephalomyelitis (CFS/ME)
Reference: 20142015 - 28

Thank you for your e-mail dated 13th July 2015 notifying us of the amendments you would like to make to your above proposal. These have been considered and we can now confirm that your amendments have been approved.

We would like to remind you that following clarification of the NRES approval for the Biorepository/tissue bank to collect tissue, it is now a requirement that researchers such as yourself collecting tissue using the NNUH biorepository process are obliged to bank some of their sample with the tissue bank. This banking will allow other researchers to potentially access and use these samples. Therefore, you will be required to bank some of the tissue that you are collecting as a result of your amendment. Mark Wilkinson will be able to offer you guidance as to how much and in what form you should bank this tissue.

Please can you ensure that any further amendments to either the protocol or documents submitted are notified to us in advance, and also that any adverse events which occur during your project are reported to the Committee.

Please can you also arrange to send us a report once your project is completed.

Yours sincerely,

p.p.

Linda Harvey
Chair Human Tissue - FMH Research Ethics Committee

Cc Simon Carding

Epsom and St Helier NHS
University Hospitals
NHS Trust

Mr Daniel Vipond
20 Church Road
Magdalen
King's Lynn
PE34 3DG

Please reply to:
Research and Development General Manager
Research and Development Department
2nd Floor Ferguson House
St Helier Hospital
Wrythe Lane
Carshalton
Surrey SM5 1AA

Tel: 020 8296 4698
Web: www.epsom-sthelier.nhs.uk
Email: yvonne.reilly@esth.nhs.uk

5 March 2015

R&D No: 020S/2014/SMED/D6
(Please quote the R&D Number in all correspondence)

Dear Mr Vipond

Re: NHS Permission

Chief Investigator:	Mr Daniel Vipond
ESTH Collaborator:	Dr Amolak Bansal
Study Title/Acronym:	A role for a leaky gut and the intestinal microbiota in the pathophysiology of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome.
REC No.	20142015-28
Sponsor:	University of East Anglia
NHS Permission Date:	5 March 2015

Thank you for submitting documentation in respect of the above-mentioned study for NHS Permission from Epsom and St Helier University Hospitals NHS Trust (ESTH).

ESTH has granted NHS Permission for the study to proceed on the basis described in the application form, protocol and supporting documentation submitted, subject to the following terms and conditions being satisfied:

1. The study is conducted in accordance with:
   - Research Governance Framework for Health and Social Care, Second Edition, April 2005.
   - Trust Policies and Procedures
   - The Data Protection Act 1998
   - NHS Confidentiality Code of Practice
   - NHS Caldicott Report and Caldicott Guardians
   - The Human Tissue Act 2004
   - Good Clinical Practice
   - Other relevant legislation released during the course of the study.

2. Members of the Research Team who wish to conduct research at ESTH and are not Trust employees must contact the R&D Office to establish appropriate contractual arrangements are in place (e.g. honorary contracts, letters of access), prior to commencement of the research study.

3. Please ensure that you submit a copy of any amendments made to this study to the R&D Office.

4. A requirement of the Research Governance Framework is that Trusts have a duty to monitor research studies. If this study is selected for monitoring, it is your responsibility, as Principal

NHS
Health Research Authority

Miss Katharine Seton
Quadram Institute Bioscience
Norwich Research Park
Colney Lane
NR4 7UA

Email: hra.approval@nhs.net

19 July 2017

Dear Miss Seton,

Letter of HRA Approval

Study title:	**Defining autoimmune aspects of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS)**
IRAS project ID:	**218545**
REC reference:	**17/LO/1102**
Sponsor	**University of East Anglia**

I am pleased to confirm that **HRA Approval** has been given for the above referenced study, on the basis described in the application form, protocol, supporting documentation and any clarifications noted in this letter.

Participation of NHS Organisations in England

The sponsor should now provide a copy of this letter to all participating NHS organisations in England.

Appendix B provides important information for sponsors and participating NHS organisations in England for arranging and confirming capacity and capability. **Please read *Appendix B* carefully**, in particular the following sections:

- *Participating NHS organisations in England* – this clarifies the types of participating organisations in the study and whether or not all organisations will be undertaking the same activities
- *Confirmation of capacity and capability* - this confirms whether or not each type of participating NHS organisation in England is expected to give formal confirmation of capacity and capability. Where formal confirmation is not expected, the section also provides details on the time limit given to participating organisations to opt out of the study, or request additional time, before their participation is assumed.
- *Allocation of responsibilities and rights are agreed and documented (4.1 of HRA assessment criteria)* - this provides detail on the form of agreement to be used in the study to confirm capacity and capability, where applicable.

Further information on funding, HR processes, and compliance with HRA criteria and standards is also provided.

It is critical that you involve both the research management function (e.g. R&D office) supporting each organisation and the local research team (where there is one) in setting up your study. Contact details

Appendix 2. Table of Modified Bansal's Diagnostic Scoring System for ME/CFS Diagnosis

Table A1. Modified Diagnostic Scoring System for ME/CFS diagnosis

Factor	Score*
Post exertional malaise beginning 12 h after physical, mental or emotional activity and lasting longer than 24 h	3
Post exertional malaise beginning 3 hours after activity and lasting longer than 24 h	2
Rapid onset of post exertional malaise lasting less than 24 h	1
Non-restorative sleep with frequent difficulty initiating and/or maintaining sleep	2
Non-restorative sleep in the absence of difficulty initiating and/or maintaining sleep	1
Reduced short term memory with word finding difficulty	1
Impaired concentration which is reduced further by external stimuli	1
Arthralgia in multiple joints with stiffness lasting longer than 1 h in the absence of swelling and inflammation	1
Hypersensitivity to light and sound present for longer than half the waking period	1

Sore throat with cervical tenderness/recurrent flu like episodes that last several hours and occur at least once per week	1
New onset of headaches of a different pattern lasting longer than 2 months	1
Myalgia exacerbated by mild exertion	1
Postural instability feeling unstable on standing, prolonged standing or sitting	1

***** 8 out of 13 points was required for ME/CFS diagnosis. This modified diagnostic scoring system adapted in this study was taken from (Bansal, 2016).

Appendix 3. Table of the Number of Non-Redundant UViGs >1 kb Per Dataset

Table A2. Statistics of non-redundant UViGs >1 kb

	PCR-1	non-PCR-1	PCR-2	non-PCR-2	PCR-3	non-PCR-3
# of UViGs*	6,749	7,760	7,354	7,581	3,795	4,250

* Pooled non-redundant UViGs >1 kb were used for the analysis of relative abundance, alpha and beta diversity.

Appendix 4. Table of the Number of Non-Redundant UViGs <1 kb Per Dataset

Table A3. Statistics of non-redundant UViGs <1 kb

	PCR-1	non-PCR-1	PCR-2	non-PCR-2	PCR-3	non-PCR-3
# of UViGs*	20,768	25,442	25,192	28,271	11,952	14,403

* UViGs <1 kb were not incorporated in the analysis of relative abundance, alpha and beta diversity.

Appendix 5. Sample Rarefaction for Alpha Diversity Analysis

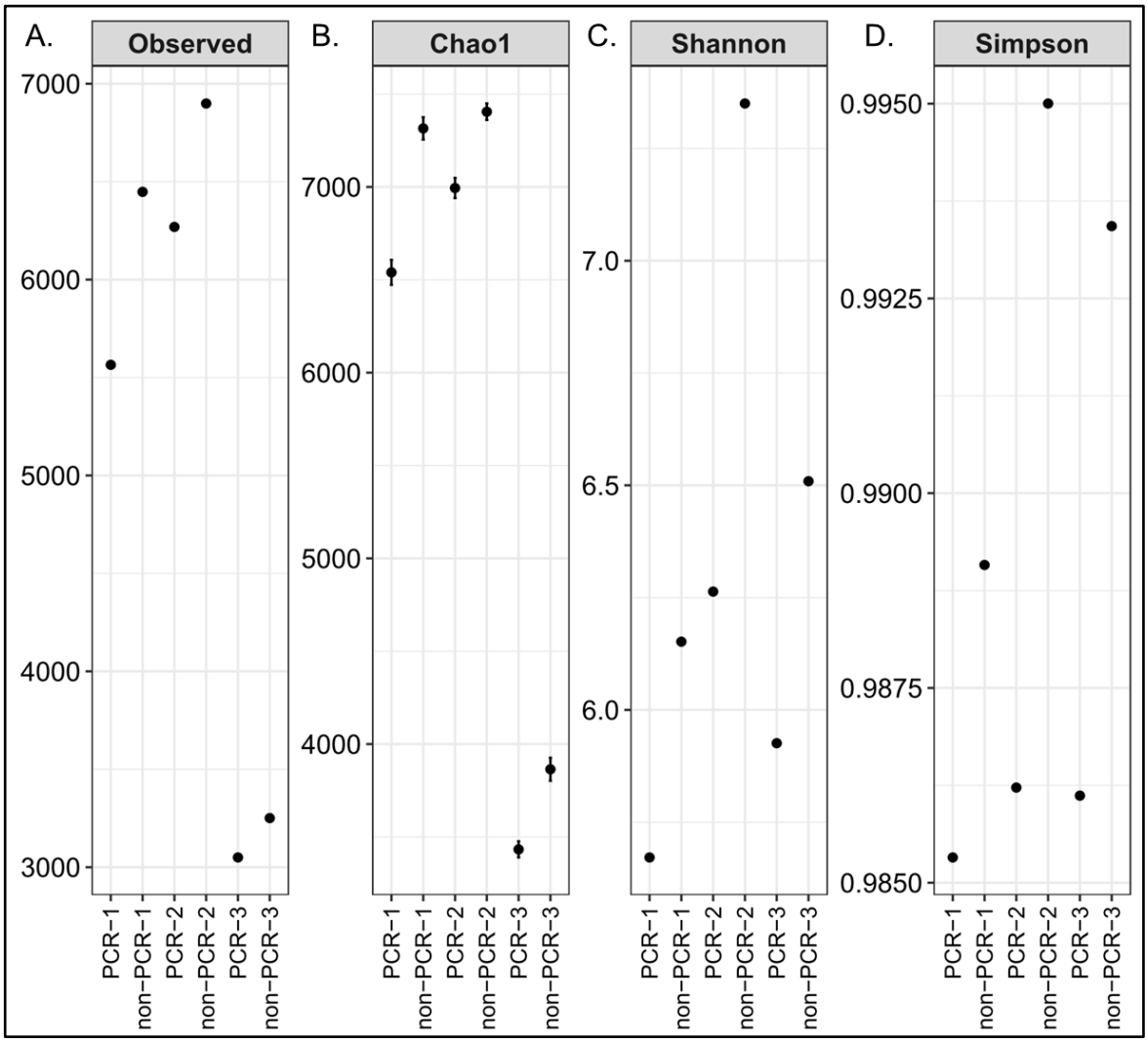
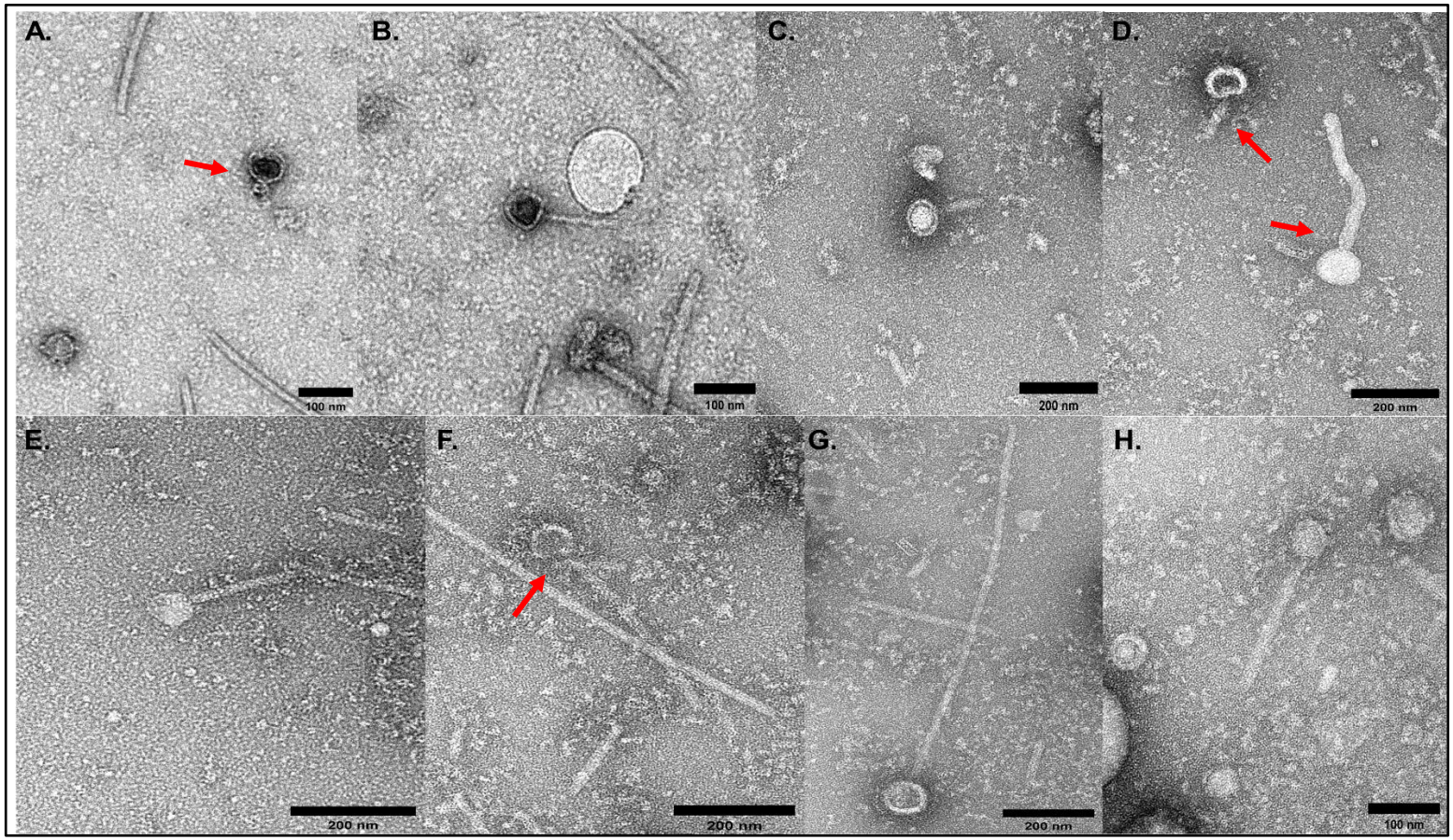


Figure A1. Estimation of alpha diversity from rarefied virome-derived PCR and non-PCR datasets. (A) The number of UViGs directly observed from a normalised count matrix. (B) Estimation of Chao1 index for six virome-derived library datasets. (C) Estimation of Shannon index for the virome datasets. (D) Estimation of Simpson index for the virome datasets. The estimated values of unfiltered UViGs (>1kb) are indicated with black dots.

Appendix 6. TEM Analysis of Faecal VLPs from ME Patient and SHHC Samples

(1) Sample 1 (patient):



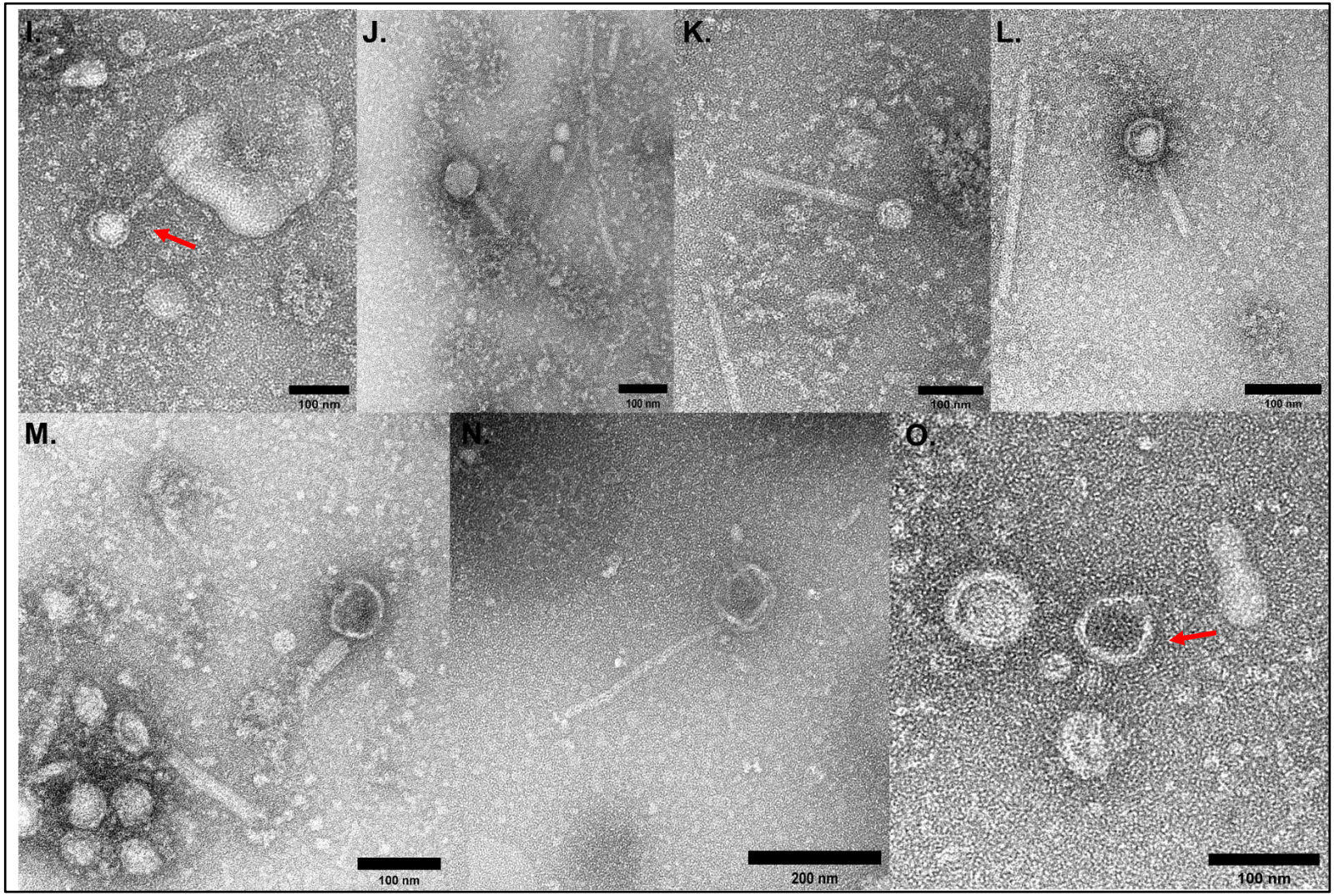

I.
100 nm
J.
100 nm
K.
100 nm
L.
100 nm
M.
100 nm
N.
200 nm
O.
100 nm

Figure A2. Transmission electron micrographs of faecal VLPs from sample 1. (A) *Podo*-like VLP (red arrow). (B) *Sipho*-like VLP. (C) *Myo*-like VLP. (D) A *Sipho*-like and A *Myo*-like VLP (red arrows). (E-I) *Sipho*-like VLPs (red arrows). (J) *Myo*-like VLP. (K) *Sipho*-like VLP. (L) *Myo*-like VLP. (M) *Myo*-like VLP with several virions. (N) *Sipho*-like VLP. (O) *Podo*-like VLP (red arrow). Scale bar: (A, B , H, I-M and O) 100 nm; (C-G and N) 200 nm.

(2) Sample 2 (SHHC):

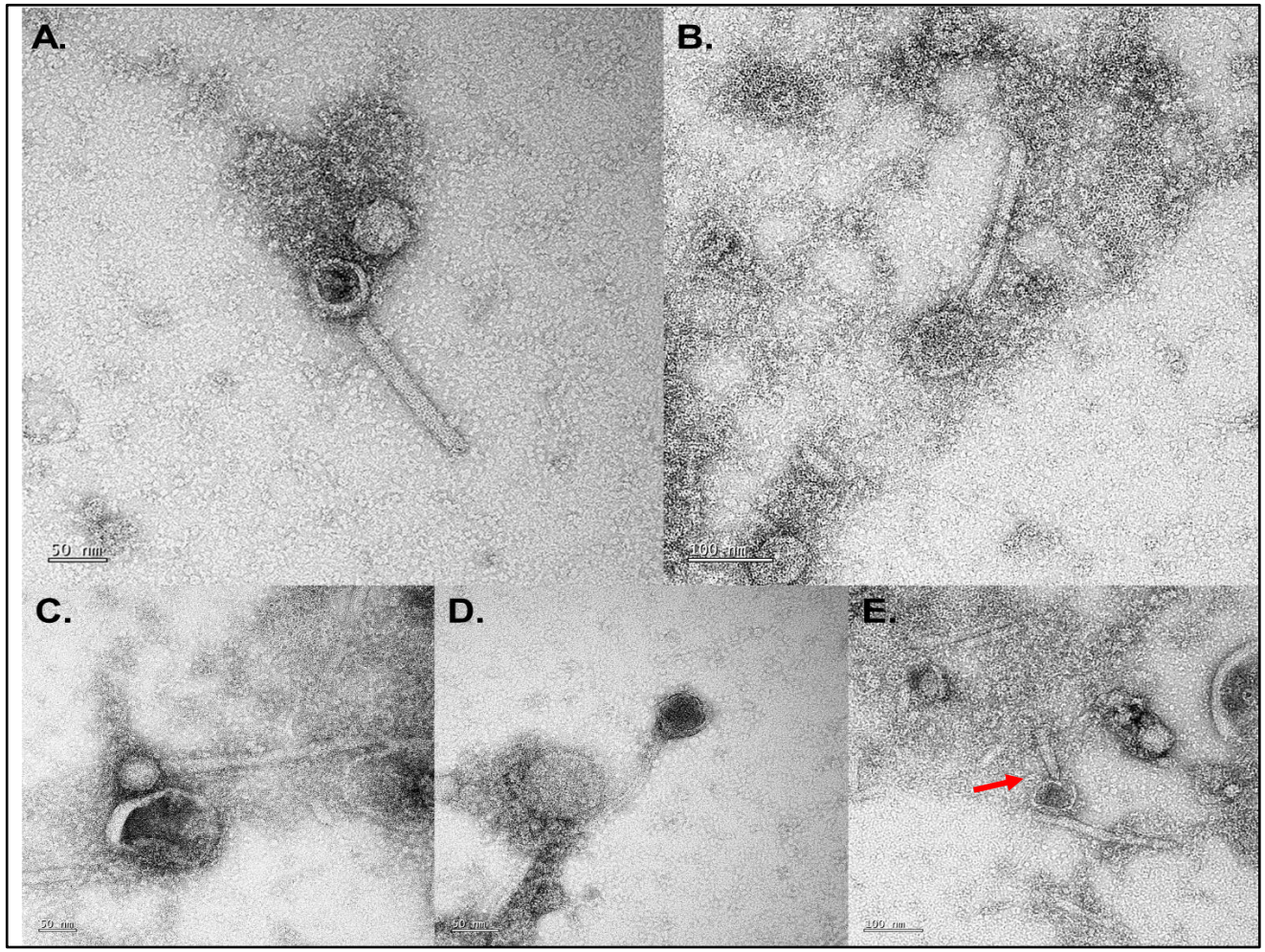


Figure A3. Transmission electron micrographs of faecal VLPs from sample 2. (A) *Myo*-like VLP. (B) *Sipho*-like VLP. (C) *Myo*-like VLP. (D) *Sipho*-like VLPs. (E) *Myo*-like VLP. Scale bar: (A, C and D) 50 nm; (B and E) 100 nm.

(3) Sample 3 (patient):

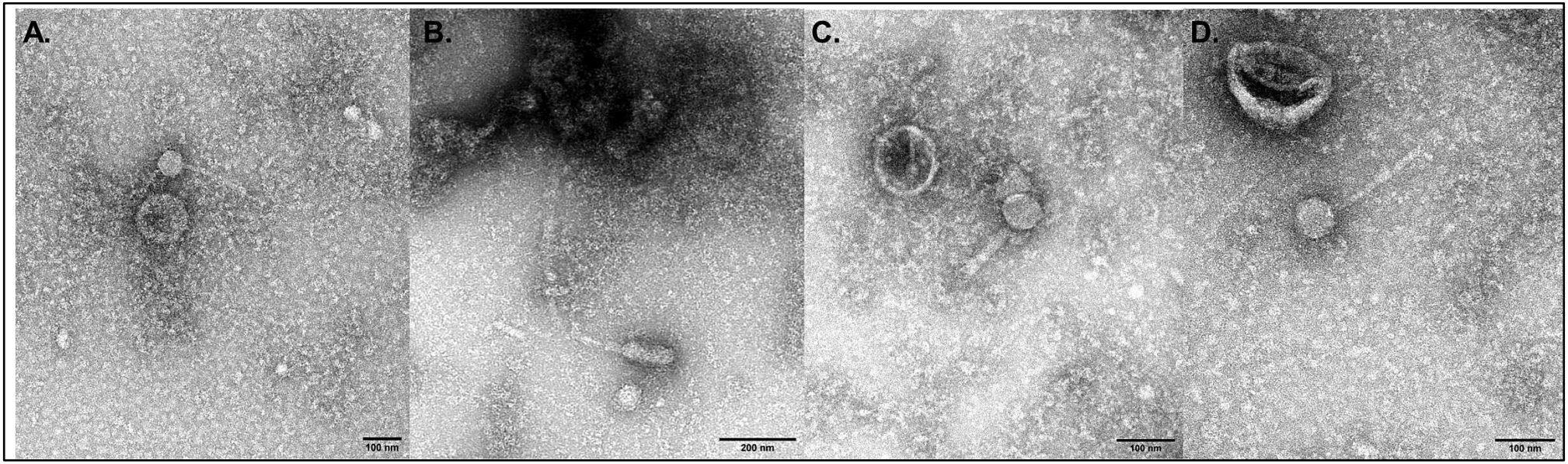


Figure A4. Transmission electron micrographs of faecal VLPs from sample 3. (A-B) *Sipho*-like VLPs. (C) *Myo*-like VLP. (D) *Sipho*-like VLP. Scale bar: (A, C and D) 100 nm; (B) 200 nm.

(4) Sample 4 (SHHC):

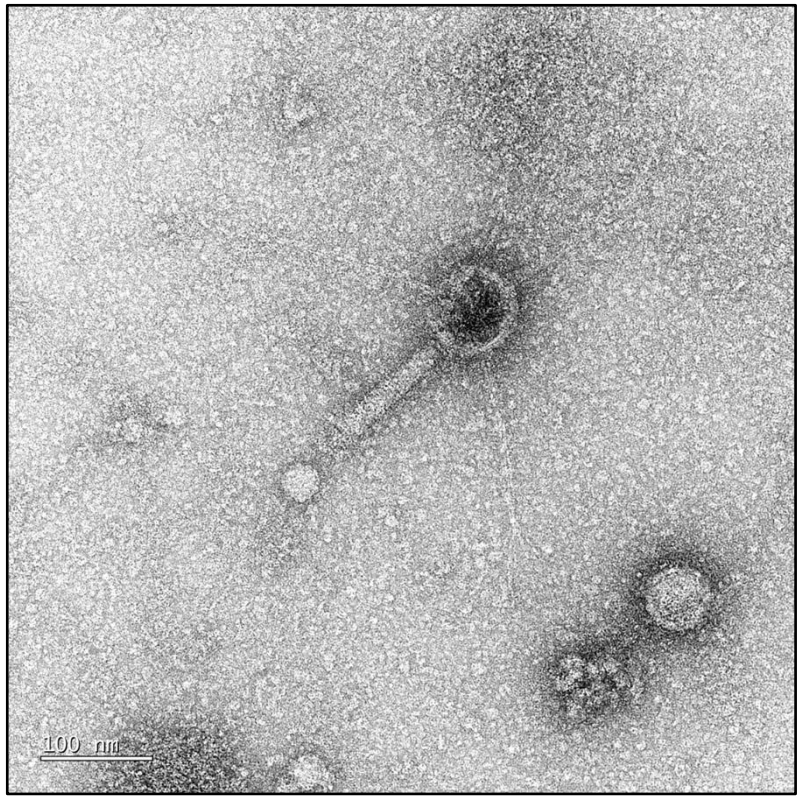


Figure A5. Transmission electron micrographs of faecal VLPs from sample 4. A *Myo*-like VLP. Scale bar: 100 nm.

(5) Sample 6 (SHHC):

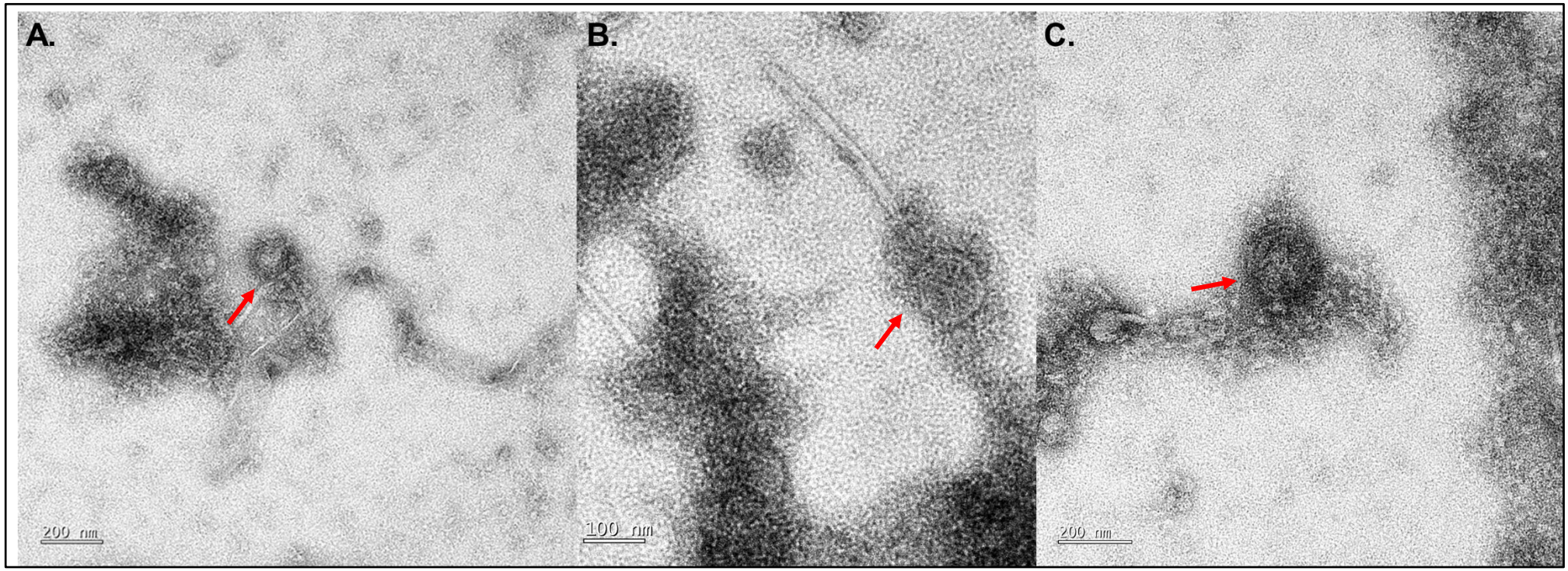


Figure A6. Transmission electron micrographs of faecal VLPs from sample 6. (A-B) *Sipho*-like VLPs (red arrows). (C) *Podo*-like VLP (red arrow). Scale bar: (A and C) 200 nm; (B) 100 nm.

(6) Sample 7 (patient):

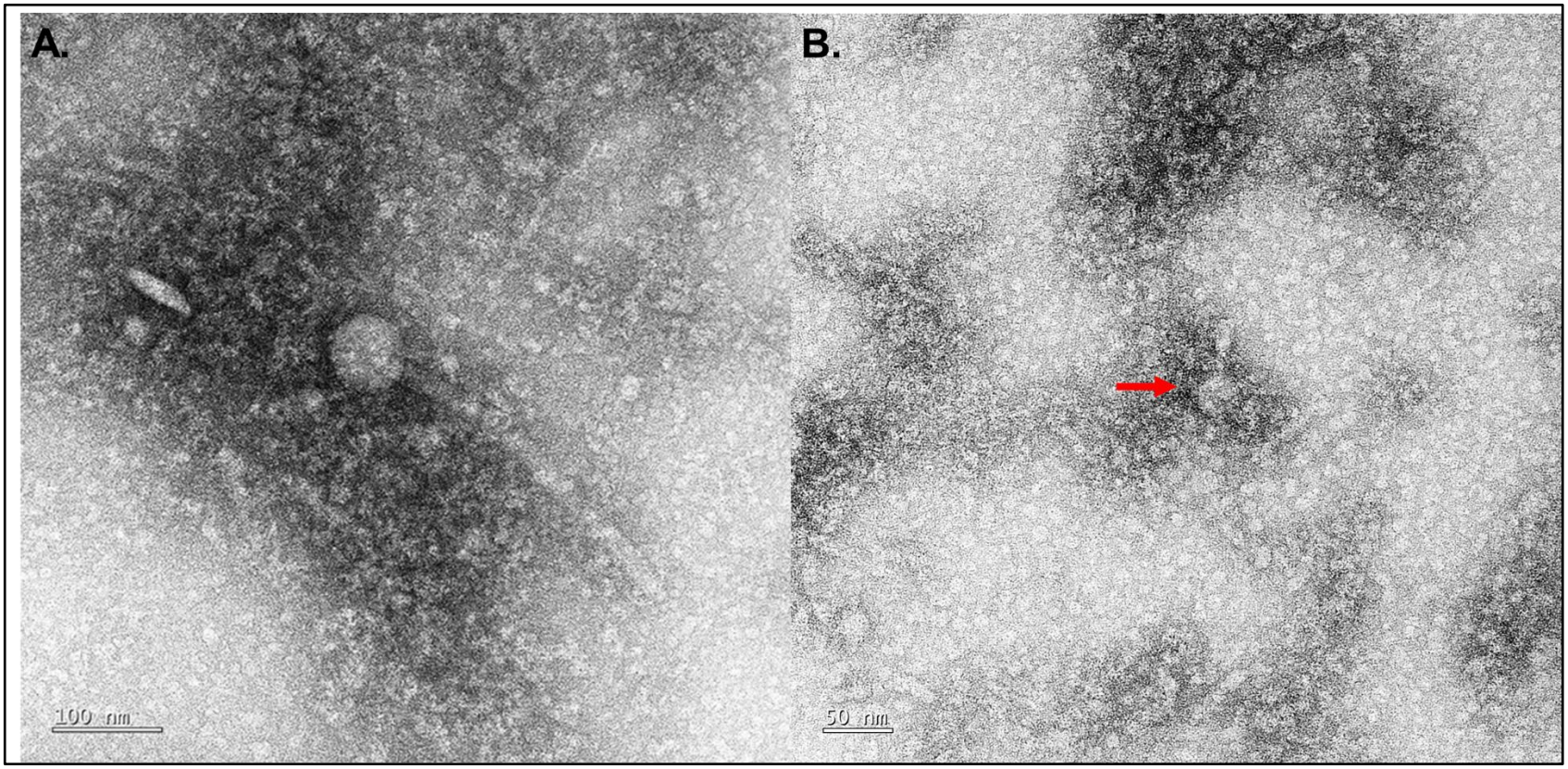


Figure A7. Transmission electron micrographs of faecal VLPs from sample 7. (A) *Sipho*-like VLP. (B) *Myo*-like VLP (red arrow). Scale bar: (A) 100 nm; (B) 50 nm.

(7) Sample 8 (SHHC):

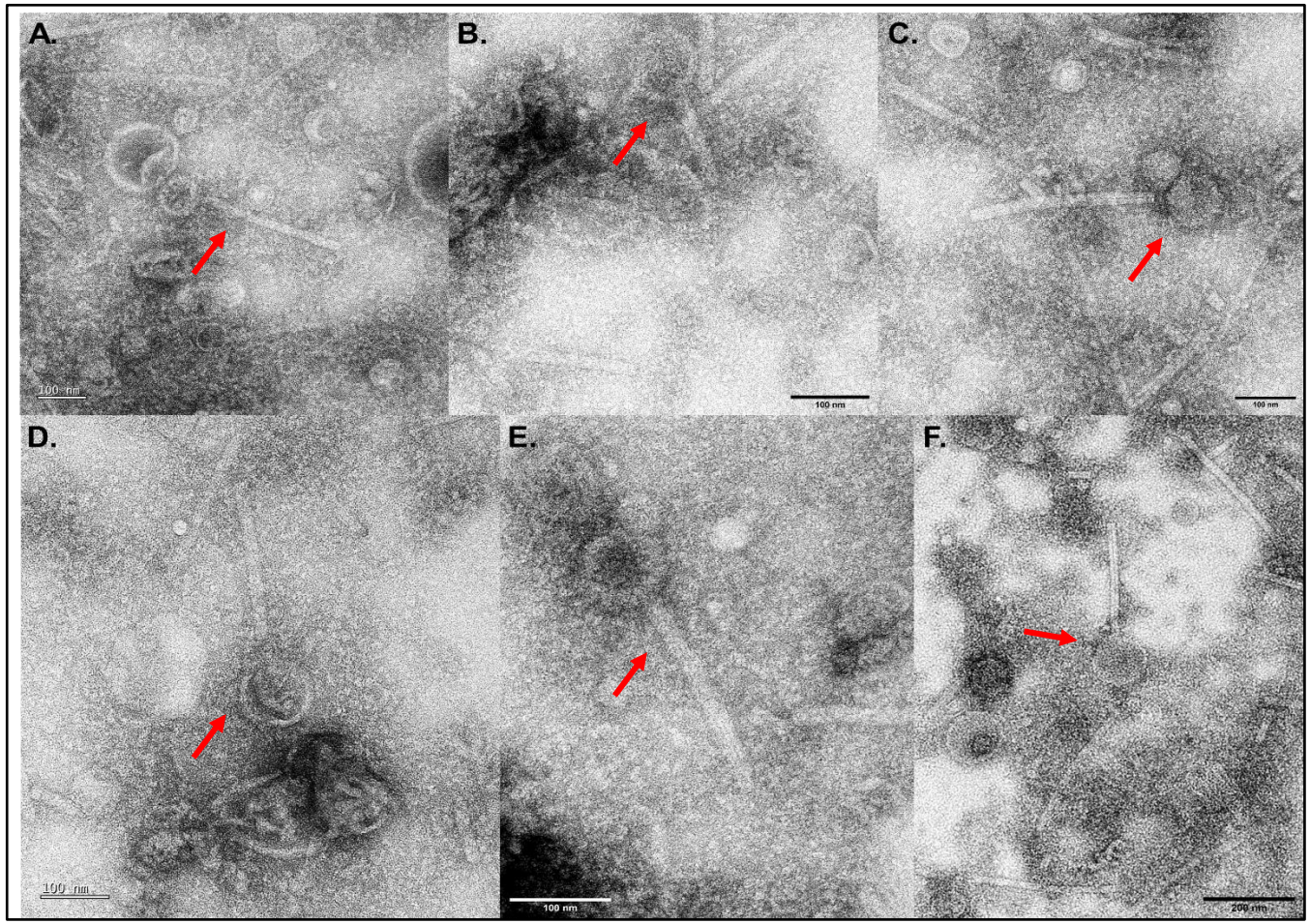


Figure A8. Transmission electron micrographs of faecal VLPs from sample 8. (A-F) *Sipho*-like VLPs (red arrows). Scale bar: (A-E) 100 nm; (F) 200 nm.

(8) Sample 9 (patient):

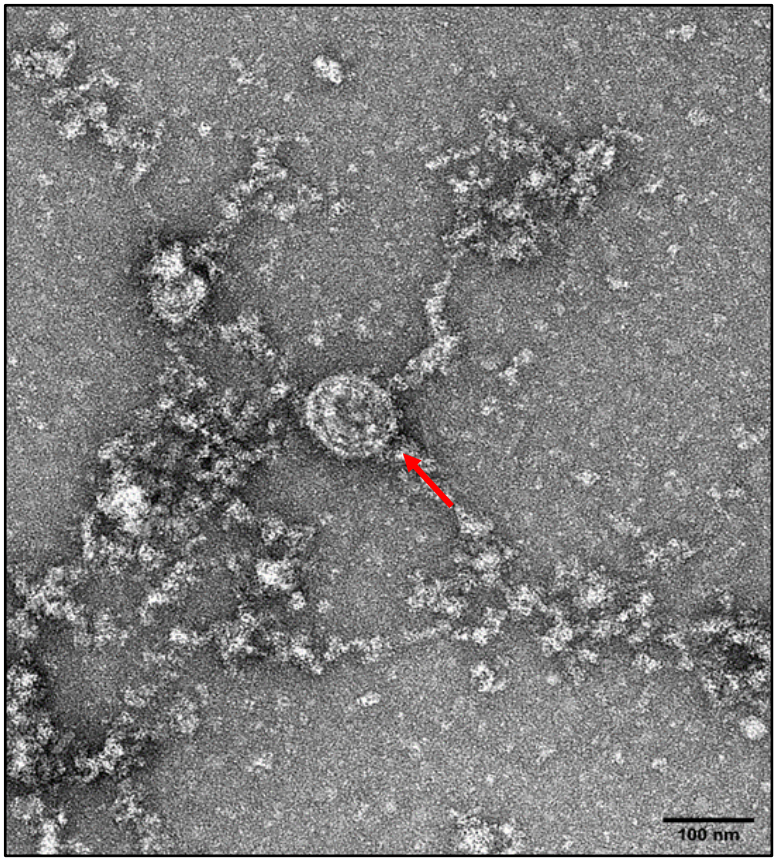


Figure A9. Transmission electron micrographs of faecal VLPs from sample 9. A *Myo*-like VLP. Scale bar: 100 nm.

(9) Sample 10 (SHHC):

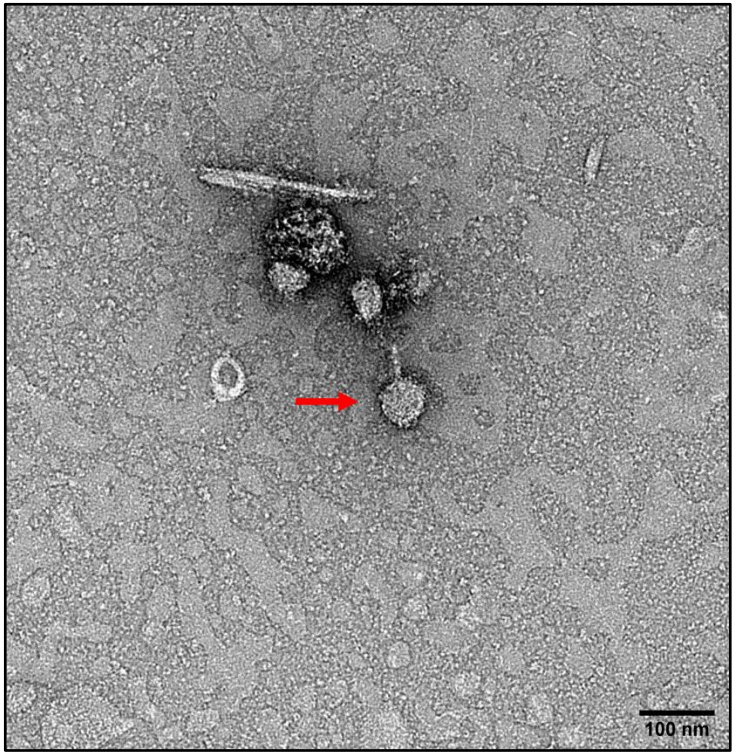


Figure A10. Transmission electron micrographs of faecal VLPs from sample 10. A *Myo*-like VLP (red arrow). Scale bar: 100 nm.

(10) Sample 12 (SHHC):

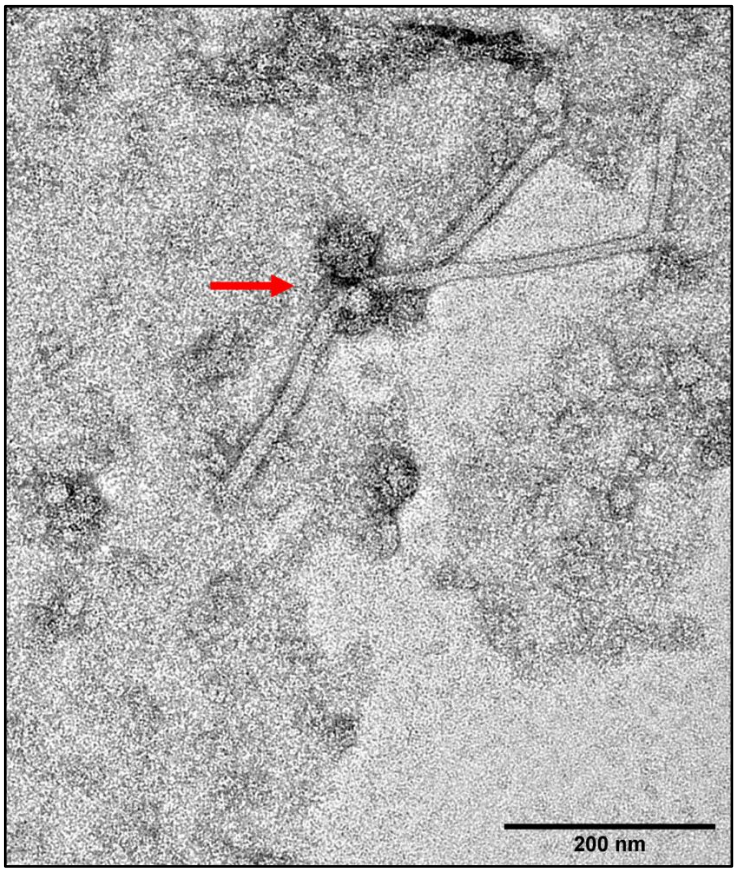


Figure A11. Transmission electron micrographs of faecal VLPs from sample 12. A *Sipho*-like VLP (red arrow). Scale bar: 200 nm.

(11) Sample 16 (SHHC):

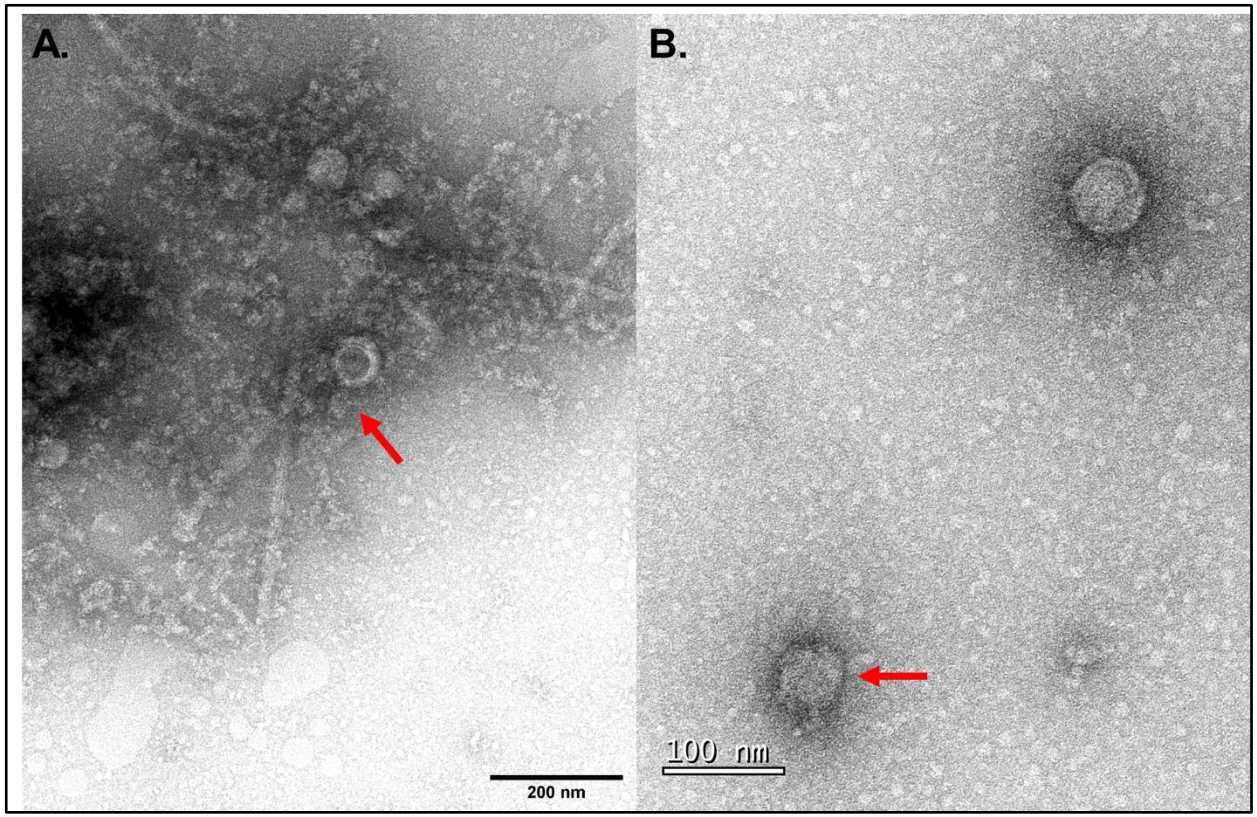


Figure A12. Transmission electron micrographs of faecal VLPs from sample 16. (A) *Sipho*-like VLP (red arrow). (B) A *Podo*-like VLP (red arrow) with a capsid-like structure. Scale bar: (A) 200 nm; (B) 100 nm.

(12) Sample 17 (patient):

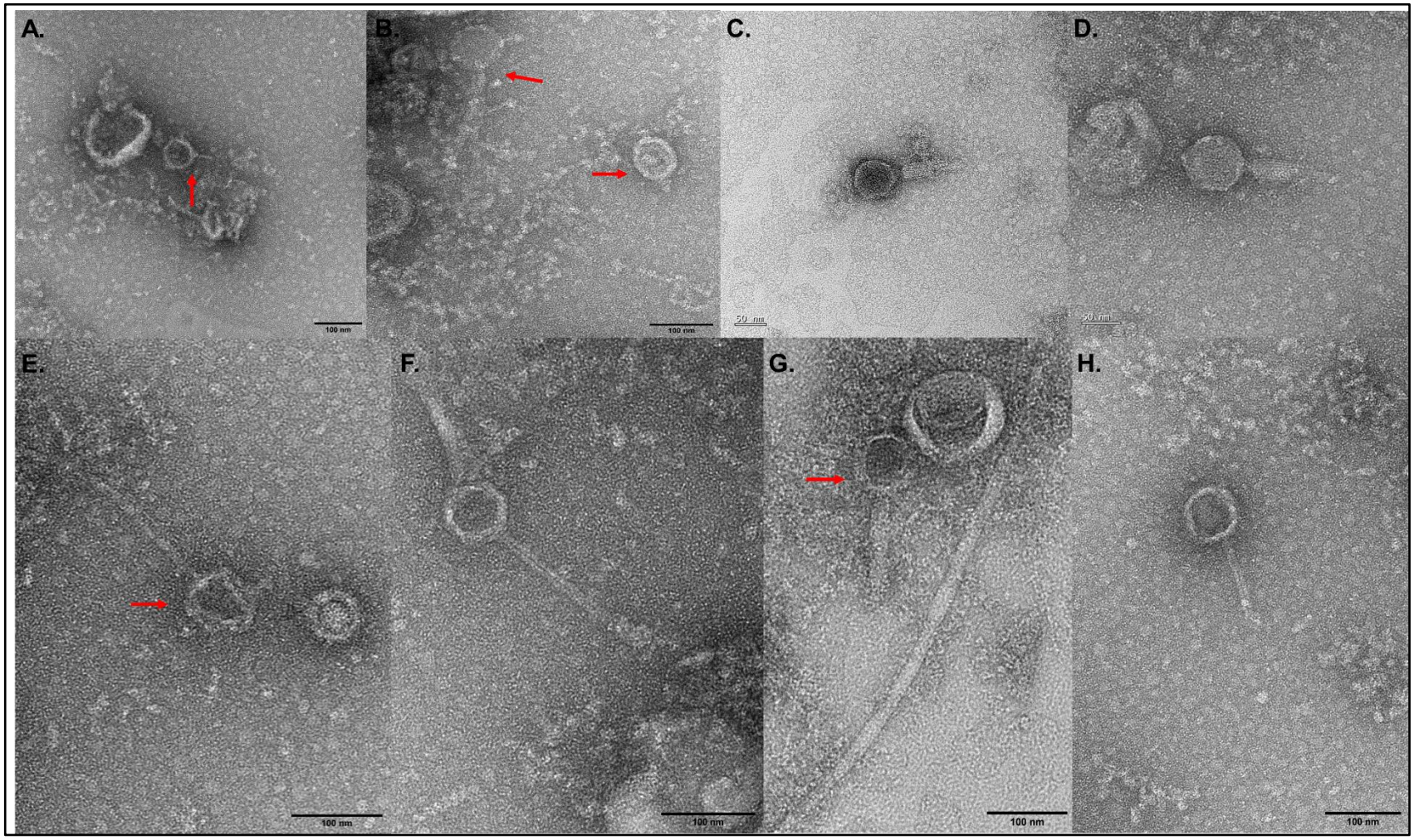


Figure A13. Transmission electron micrographs of faecal VLPs from sample 17. (A) *Podo*-like VLP (red arrow). (B) A *Sipho*-like VLP and a *Podo*-like VLP (red arrow). (C-D) *Myo*-like VLPs. (E-F) *Sipho*-like VLPs. (G) *Myo*-like VLP (red arrow). (H) *Sipho*-like VLP. Scale bar: (A-B and E-H) 100 nm; (C-D) 50 nm.

(13) Lysed bacteria observed:

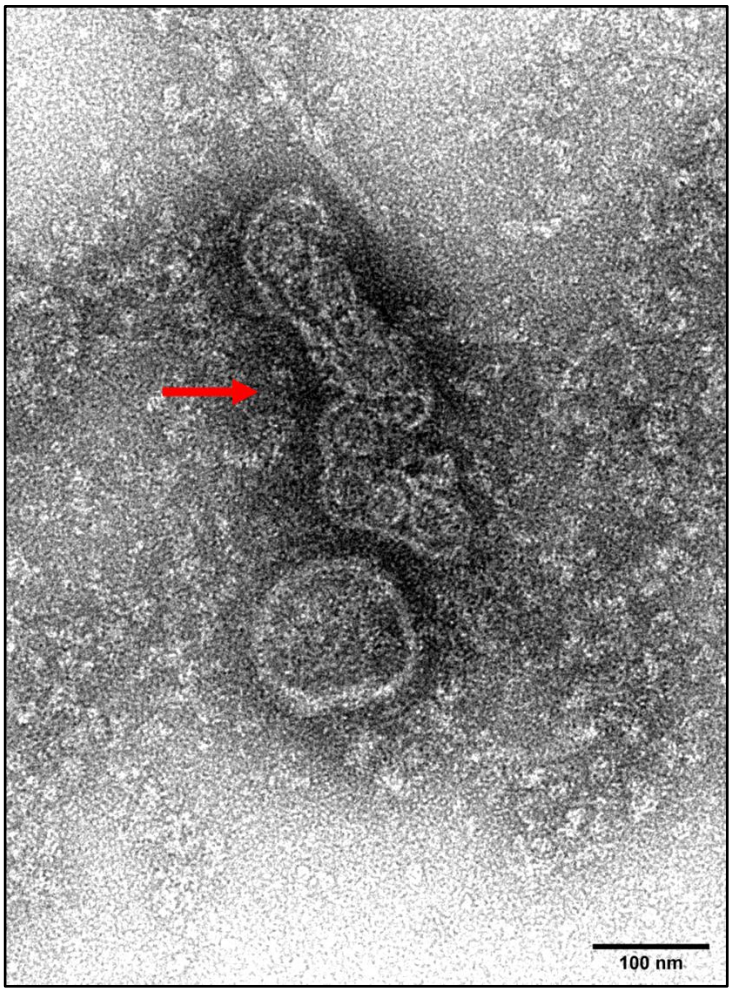


Figure A14. Transmission electron micrographs of lysed bacteria. Lysed bacteria were occasionally observed with virions being “released” from the damaged bacterial body. Scale bar: 100 nm.