Investigating the Human Gut Virome in ME/CFS Patients and the Impact of Faecal Microbiota Transplantation

Rik Haagmans

Submitted for the degree of Doctor of Philosophy

University of East Anglia

Quadram Institute

July 2024

Abstract

Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) is a debilitating disease characterised by chronic fatigue resulting from emotional, physical or mental exertion. Its cause is unknown and there is no cure. Viral infections are commonly associated with ME/CFS onset and most patients have gastrointestinal (GI) disturbances, with GI microbial dysbiosis being highly associated with disease. Microbiota replacement therapy (faecal microbiota transplantation, FMT) is a promising treatment option. However, while the GI virome plays a significant role in GI microbiome homeostasis and in health, it's role in ME/CFS and the effect of FMT are unclear. This thesis project aimed to investigate this through a metagenomic and TaqMan array card-based analysis of the GI virome of participants in a Phase II placebo-controlled clinical trial, the Comeback Study (NCT03691987), investigating safety, acceptability and efficacy of FMT in 80 mild to severe ME/CFS patients.

First, to determine bias and reproducibility of the viral metagenomics methods, an epifluorescence microscopy method was optimised, and an image analysis workflow was developed. This enabled the reproducible quantification of single stranded and double stranded RNA and DNA prokaryotic and eukaryotic viruses and production of a mock virus community.

Next, the mock virus community was used to spike faecal samples to assess methodological biases and reproducibility, comparing a whole transcriptome amplification kit (WTA2) and sequence-independent single primer amplification (SISPA). This showed that assemblies ofWTA2 libraries had higher contiguity and diversity, whereas SISPA gave more consistent abundance estimates.

Finally, TaqMan analysis of the Comeback Study showed no indication of increased prevalence of GI pathogens in ME/CFS. However, viral metagenomics showed that FMT increased virus diversity in recipients virome, with donor viruses persisting and with virome similarity to the donor increasing, with distinct levels of engraftment between donors . This suggests that FMT produces a healthier virome in ME/CF patients, depending on the donor.

Acknowledgements

This thesis could not have been made without the support of numerous colleagues, collaborators, friends, and family. I would like to first express my deepest gratitude to my secondary supervisor Prof. Dr Simon Carding for sharing his knowledge, advice, and support during this project. I am also deeply thankful to my primary supervisor Dr Penny Powell, for her advice and encouragements.

I am incredibly grateful for Invest in ME Research and the University of East Anglia for funding my research project, as well as extensions to the project. I am also thankful for the Turing travel grant I was awarded by the UEA to conduct part of the virome analysis of the Comeback Study in Tromsø, Norway.

Many colleagues at the Quadram Institute contributed to this project. Dr Emily Jones assisted in setting up the nanoparticle tracking experiments using the ZetaView, and Dr Ernie Hsieh helped me get started with the epifluorescence experiments. Dr Evelien Adriaenssens provided the phages that were used in the mock community. Dr Oliver Charity conducted the mock community spiking experiments and provided valuable discussions on data analysis. Virome experts Evelien and Dr Ryan Cook gave advice on the mock community experiment and Comeback Study virome analysis. The Quadram Institute sequencing team, David Baker, Rhiannon Evans, Dr Molly Millar, and Dr Tristan Seecharan performed sequencing of mock community experiment samples and produced sequencing libraries for the Comeback Study. Our resident statistician George Savva provided valuable insights into the mock community data and the statistical analysis of the Comeback Study virome data. Dr Andrea Telatin, Dr Sumeet Tiwari and Dr Leonardo de Oliveira Martins of the Quadram Institute bioinformatics team provided valuable discussions and insights, and Andrea was of immense help with the virome analysis of the Comeback Study.

This thesis would also not exist without the work of many collaborators. I want to express my gratitude to the lovely people in Tromsø and Harstad working on the Comeback Study, particularly Dr Rasmus Goll, Dr Peter Johnsen, and Linn Skjevling for warmly welcoming me in Tromsø and Harstad and for setting up a fantastic clinical trial and sharing their knowledge with me. And thank you to Ingrid and Hilde, for their help with processing and shipping the many Comeback Study samples. Also, a huge thank you to Dr Martin Curran at the NHS Addenbrooke's hospital in Cambridge for taking the time to teach me on using the TaqMan arrays, validate the TaqMan array cards and running the Comeback Study samples in between his daily work. Several animal viruses were a crucial part of the mock community and were kindly provided by Dr Edward Mee at the National Institute for Biological Standards and

Control, Dr Joe Brownlie at the Royal Veterinary College London, and Dr James Stewart at the University of Liverpool.

A special thanks goes out to the patients in the Comeback Study and the AI-ME/CFS study that provided valuable samples for this research.

And let's not forget the Carding lab group! Massive thanks to Adnan, Aimee, Anne, Catherine, Emily, Fiona, Katharine, Regis, Rokas, Rosie, Steve, and Sonia, for the many fun lunch time conversations, coffee breaks and group meetings. In particular, I would like to thank Dr Adnan Tariq, Dr Fiona Newberry, and Dr Katharine Seton, for showing me the ropes at the beginning of my project.

While Norwich has become my home, moving here was not easy, particularly during the early days of the COVID pandemic. I am thankful for the friendship of my housemates, Kath, Ben, Jack, Georgina, Gemma, Maria, Maria, Jordan, Georgina, and Rosie, during the COVID-19 lockdown periods and after. I am grateful, too, for my Dutch friends on the continent. My "IBB" and "Hmmurica" friends, who always have their door open form me whenever I'm back home. For Bart, Robin, and Joop, for the online gaming sessions, and for my friends in Norwich, for the many coffee breaks, climbing sessions, board game nights, and pathfinder sessions: Kath, Ari, Javi, Dani, Sonia, Stuart, Miriam, Maria, Rokas, Gintare, Mike, Giedrius, Jennifer, Victor, and Georgie.

I will be forever grateful for my parents Hélène and Tom, who are always there for me and who have worked hard to support me and my sisters, and for my sisters Yvonne and Michelle, for their emotional support. And finally, a big, big thank you to my girlfriend Antía, for her incredible patience and support and, last but not least, Antía's Bernese Gaia, for the many laughs and Sunday afternoon walks in Mousehold Heath.

Table of contents

List of figures

List of tables

Abbreviations

1994-CDC/Fukuda	1994 CDC case criteria by Fukuda *et al.* 1994
aEPEC	atypical EPEC
AFQ	antibiotics and food supplement questionnaire
AGE	acute gastroenteritis
AI-ME/CFS	Autoimmunity in ME/CFS study
ASD	autism spectrum disorder
b	base
B19	Parvovirus B19
BC	Bray-Curtis
BL	blank control
BLAST	Basic Local Alignment Search Tool
bp	basepair
BVDV-1	bovine viral diarrhea virus 1
CCC	Canadian consensus criteria
CD	Crohn's disease
CDI	*Clostridium difficile* infection
cDNA	complementary DNA
CFS	chronic fatigue syndrome
cGAS	cyclic guanosine monophosphate–adenosine monophosphate synthase
CLR	centred log ratio
CMOS	complementary metal-oxide-semiconductor
COVID-19	coronavirus disease 2019
CREI	carbapenem-resistant *Enterobacteriaceae* infection
CV	coefficient of variation
DAPI	4′,6-diamidino-2-phenylindole
DC	dendritic cells
Det7	*Kuttervirus Det7*
DMEM	Dulbecco's modified Eagle medium
DNA	deoxyribonucleic acid
dsDNA	double stranded deoxyribonucleic acid
DSQ	DePaul Questionnaire
dsRNA	double stranded ribonucleic acid
dUTPase	deoxyuridine 5′-triphosphate nucleotidohydrolase
DWQ	disease worsening question
EAEC	enteroaggregative *Escherichia coli*
EBV	Epstein-Barr virus
EBV-IM	EBV infectious mononucleosis
EDTA	ethylenediaminetetraacetic acid
EFM	epifluorescence microscopy
EPEC	enteropathogenic *E. coli*
ERV	endogenous retroviruses
ETEC	enterotoxigenic *E. coli*
FCM	flow cytometry
FFQ	Food Frequency Questionnaire
FMT	faecal microbiota transplantation
FSS	Fatigue Severity Scale
GA	glutaraldehyde

GCP	Good Clinical Practice
GDPR	General Data Protection Regulation
GI	gastrointestinal
GI-TAC	gastrointestinal TaqMan array card
HADS	Hospital Anxiety Depression Scale
HCMV	human cytomegalovirus
HERV	human endogenous retroviruses
HHC	healthy household control
HHV	human herpesvirus
HI	spiked with high-concentration MC
HIV-1	human immunodeficiencyvirus 1
HSV	herpes simplex virus
IBD	inflammatory bowel disease
IBS	irritable bowel syndrome
IC	internal control
ICH	International Council for Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use
ICTV	International Committee on Taxonomy of Viruses
IFN	interferon
IoM	Institute of Medicine (currently National Academy of Medicine)
IQR	inter-quartile range
KSHV	Kaposi's sarcoma-associated herpesvirus
LASL	linker-amplified shotgun library
LB	Luria-Bertani
LO	spiked with low-concentration MC
MAMP	microbe-associated molecular pattern
MC	mock community
MDA	multiple displacement amplification
ME	myalgic encephalomyelitis
ME-ICC	ME international consensus criteria
MEM	minimal essential medium Eagle
MGB	major groove binder
MHV-68	murid gammaherpesvirus 68
MIUViG	minimum information about an uncultivated virus genome
NCBI	National Center for Biotechnology Information
NFQ	non-fluorescent quencher
NICE	National Institut for Health and Care Excellence
NO	not spiked
NR	NCBI non-redundant protein database
NRES	National Research Ethics Service
NT	NCBI non-redundant nucleotide database
NTA	nanoparticle tracking analysis
OD	optical density
P22	*Lederbergvirus P22*
PA	plaque assay
PAIS	post-acute infectious syndromes
PBMC	peripheral blood mononuclear cells
PBS	phosphate-buffered saline
PCA	principal component analysis

pEAF	*E. coli* adherence factor plasmid
PEG	polyethylene glycol
PEM	post-exertional malaise
PES	polyethersulfone
PFU	plaque forming units
PPR	particles-to-PFU ratio
PRAEQ	patient-reported adverse event questionnaire
PRR	pathogen recognition receptor
Qb	*Qubevirus durum*
QIB	Quadram Institute Biosciences
qPCR	quantitative polymerase chain reaction
RBANS	Repeatable Battery for the Assessment of Neuropsychological Status
rCDI	recurrent *C. difficile* infection
RCT	randomised clinical trial
Rn	normalised reporter fluorescence
RNA	ribonucleic acid
ROX	carboxyrhodamine
RPKM	reads per kilobase per million reads
RPM	revolutions per minute
RT	reverse transcription
RV-A	simian rotavirus SA/11
SAP	shrimp alkaline phosphatase
SaV	sapovirus
SCFA	short-chain fatty acid
SD	standard deviation
SEID	systemic exercise intolerance disorder
SF-36	MOS 36-Item Short Form Survey
SINV	Sindbis virus AR339
SISPA	sequence-independent single primer amplification
SPRI	solid phase reversible immobilisation
ssDNA	single stranded deoxyribonucleic acid
ssRNA	single stranded ribonucleic acid
T0	baseline timepoint
T3	3-month follow-up timepoint
T5	*Tequintavirus T5*
TAC	TaqMan array card
TE	tris-EDTA
TEM	transmission electron microscopy
tEPEC	typical EPEC
TLR	Toll-like receptor
UHGV	unified human gut virome database
UNN	University Hospital of North Norway
VLP	virus-like particle
vOTU	virus operational taxonomic unit
VZV	Varicella-Zoster virus
WHO	World Health Organisation
WTA2	Whole Transcriptome Amplification kit
ΔRn	baseline-corrected normalized reporter fluorescence

CHAPTER 1: GENERAL INTRODUCTION: ME/CFS AND THE GI VIROME

1.1 ME/CFS and the role of viruses

Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) is a debilitating condition in which patients suffer from fatigue, post-exertional malaise (PEM), and a variety of other symptoms (Bateman *et al.*, 2021). The quality of life of patients as well as their partners and family is significantly affected (Vyas *et al.*, 2022). Up to 25% of patients are severely affected (Pendergrast *et al.*, 2016), with patients left house- or bed bound, requiring permanent care and in some cases enteral feeding (Sommerfelt, Schei and Angelsen, 2023). While ME/CFS can develop at any age, ME/CFS develops more often around 10 – 19 years of age and 30 – 39 years of age (Bakken *et al.*, 2014). More than 60% of patients are female (Bakken *et al.*, 2014; Valdez *et al.*, 2019). Pre-COVID-19 pandemic 65 million cased were estimated globally (Hanson and Germain, 2020). However, up to 50% of those with Long COVID may develop ME/CFS (Annesley *et al.*, 2024), resulting in a doubling of global ME/CFS cases (Wong and Weitzer, 2021). The economic cost of ME/CFS in the United States alone is up to US$362 billion (Mirin, Dimmock and Jason, 2022) with up to 75% of patients unemployed (Castro-Marrero *et al.*, 2019; Unger *et al.*, 2024). There is currently no effective treatment for ME/CFS, with patient care limited to symptom management and treatment of comorbidities (Bateman *et al.*, 2021; Toogood *et al.*, 2021). While infections, chemical exposure, and major life events and trauma have been associated with the onset of ME/CFS (Chu *et al.*, 2019), disease aetiology is unknown.

ME/CFS symptomatology is heterogeneous (Brurberg *et al.*, 2014; Chu *et al.*, 2017; Lim and Son, 2020) with various immune system, metabolic, gastrointestinal (GI), nervous system, and endocrine system abnormalities been documented (Missailidis, Annesley and Fisher, 2019). The core symptoms are severe fatigue after minimal physical and mental exertion that is unresolved by resting and significantly inhibits regular activity, with worsening of symptoms after minimal exertion referred to as PEM, unrefreshing sleep, cognitive impairment, and orthostatic intolerance (Institute of Medicine, 2015; Lim and Son, 2020; NICE, 2021). Additional symptoms include sore throat, muscle and joint pains, difficulty regulating body temperature, neurosensory sensitivities, and GI disturbances (Carruthers *et al.*, 2011).

1.1.1 Diagnosis based on case definitions

The origins of the term ME/CFS dates back to the first half of the 20th century, during which several outbreaks of unknown disease occurred across the world (Acheson, 1959). While

poliomyelitis was initially suspected, this was later ruled out (Acheson, 1959). Symptoms included head, neck and muscle aches, swollen and painful lymph nodes and cognitive impairment, with the illness ultimately being named ME (Acheson, 1959). As the aetiology of ME is unknown and no diagnostic laboratory tests are available, case definitions were developed based on symptomatic criteria for ME/CFS diagnosis. During this period, case definitions for CFS were developed in parallel to diagnose patients in similar disease outbreaks (Holmes *et al.*, 1988; Fukuda *et al.*, 1994) that focused more on cognitive impairment and fatigue, while ME definitions focused more on muscle disturbance and neuro-autonomic symptoms. Whether ME and CFS describes the same illness has been disputed (Jason *et al.*, 2016; Twisk, 2018), with the term ME/CFS now generally used to describe the disease (Jason and Johnson, 2020).

Due to the heterogeneity of the patient population, the lack of biomarkers, and disputes over what does and does not constitute ME/CFS, a plethora of more than 25 case definitions have been developed with a variety of criteria (Lim and Son, 2020). Case definitions differ in the specific criteria and thus with the group of patients that meet the criteria of the various case definitions (Brurberg *et al.*, 2014; Chu *et al.*, 2017; Lim *et al.*, 2020). Nonetheless, most case definitions require the presence of the core symptoms of fatigue, cognitive impairment, PEM, sleep disorder, and orthostatic intolerance (Lim and Son, 2020).

The first definition of ME was published by Ramsay in 1986 (Ramsay, 1986), describing principal symptoms of patients during several outbreaks of ME: muscle fatiguability, cognitive dysfunction and impaired circulation (Sunnquist *et al.*, 2017). The earliest definition for CFS was published shortly after to define the symptoms of chronic fatigue following Epstein-Barr virus (EBV) infection (Holmes *et al.*, 1988). In 1994, the United States Center for Disease Control published a revised definition (1994-CDC/Fukuda) to improve differentiation of people with CFS from those with other psychological or psychiatric conditions also associated with fatigue (Fukuda *et al.*, 1994). The 1994-CDC/Fukuda criteria is still widely used (Lim and Son, 2020). However, following criticism that the 1994 CDC/Fukuda criteria are too permissive, select an overly heterogeneous patient population, and do not include core symptoms like PEM and sleep disturbance, the Canadian Consensus Criteria (CCC) were developed (Carruthers *et al.*, 2003). This was the first case definition that combined ME and CFS and was developed primarily for clinical, rather than research purposes. The ME International Consensus Criteria (ME-ICC) were subsequently developed in 2011 to improve on the CCC (Carruthers *et al.*, 2011). Following a comprehensive literature review, the Unites States National Academy of Medicine (formerly the Institute of Medicine) recommended the adoption of the term Systemic Exercise Intolerance Disorder (SEID) and was accompanied by a new case definition (IoM) (Institute of Medicine, 2015). In the same year, two other case

definitions were published that each provided empirically-derived criteria (Jason *et al.*, 2015; Maes, 2015). In Europe, the 1994-CDC/Fukuda criteria and the CCC are the most widely used case definitions for diagnosis (Strand *et al.*, 2019). In the UK, the National Institute for Health and Care Excellence (NICE) recently released revised guidelines for diagnosis of ME/CFS (NICE, 2021).

1.1.2 Case definitions select distinct patient populations

As each case definition has a distinct set of selection criteria, some case definitions are more stringent than others and as a result, prevalence estimates vary depending on case definitions (Nacul *et al.*, 2011; Lim *et al.*, 2020). For instance, the CCC and ME-ICC are more restrictive than the 1994-CDC/Fukuda criteria (Jason *et al.*, 2014; Chu *et al.*, 2017; Vaes *et al.*, 2023). While nearly all patients that meet the CCC or ME-ICC also meet the 1994-CDC/Fukuda criteria, only 89% and 74% of those meeting the 1994-CDC/Fukuda criteria also meet the CCC and ME-ICC, respectively (Chu *et al.*, 2017). Additionally, assessment methods affect the prevalence estimates, with assessment through a questionnaire-based interview without medical tests producing a 5 to 10-fold higher prevalence estimate than assessment by a physician (Johnston *et al.*, 2013; Lim *et al.*, 2020). Pre-COVID-19 prevalence was estimated at 0.2% in the UK (Nacul *et al.*, 2011), 0.1-2.2% in Europe (Estévez-López *et al.*, 2020), 0.1% in Australia (Orji *et al.*, 2022), and 0.42% in adults in the US (Jason *et al.*, 1999). Sources of uncertainty include difficulties for patients to obtain a diagnosis and limited interactions with their primary care providers due to a lack of treatment options and being confronted with stigma and disbelief in the existence of ME/CFS by their care provider (Orji *et al.*, 2022). Noting high heterogeneity among prevalence studies due to differences in the case definitions used and the method of assessment, a recent meta-analysis found a global prevalence of 0.86% (Lim *et al.*, 2020). Post-COVID this estimate is likely to increase and to double due to individuals with Long COVID developing ME/CFS (Komaroff and Bateman, 2021).

Due to differences in selection criteria, patient populations diagnosed with different case definitions also exhibit different symptoms. While one of the goals of the CCC and ME-ICC was to reduce heterogeneity, it is not clear if this has been achieved (Brurberg *et al.*, 2014; Chu *et al.*, 2017). Nonetheless, patients meeting the ME-ICC have the highest severity, but also the highest prevalence of psychiatric disorders (Brurberg *et al.*, 2014). To further increase homogeneity of patient cohorts in research, patient subgrouping has been investigated to find groups of patients that are more likely to have a shared aetiology (Jason *et al.*, 2005; Collin *et al.*, 2016; White, 2019). For instance, patients can be selected based on disease onset (Domingues *et al.*, 2021), gene expression patterns (Kerr *et al.*, 2008), metabolic characteristics (Hoel *et al.*, 2021), symptoms and peak oxygen consumption (Lacasa *et al.*, 2023)the occurrence of PEM (Maes, Twisk and Johnson, 2012), or GI disturbances (Maes *et al.*, 2014;

Vaes *et al.*, 2023). Disease duration is also a contributing factor to heterogeneity, as both cytokine profiles (Hornig *et al.*, 2015) and GI dysbiosis have been found to be affected by it (He *et al.*, 2023; Xiong *et al.*, 2023).

1.1.3 Pathophysiology of ME/CFS

A large variety of physiological disturbances are associated with ME/CFS, with biomedical research describing neurological, metabolic, immunological, cardiopulmonary, muscular, cognitive, and GI changes (Komaroff and Lipkin, 2023). In the immune system, the most consistent findings are a reduction of NK cell cytotoxicity (Hardcastle *et al.*, 2014; Eaton-Fitch *et al.*, 2019) and an increase in the number of circulating NK cells (Hardcastle *et al.*, 2014; Rivas *et al.*, 2018). Presence of serum auto-antibodies have also been reported in ME/CFS, for example against the β2-adrenergic receptor (Gravelsina *et al.*, 2022) and the selenium transporter selenoprotein P that is involved in thyroid hormone metabolism (Sun *et al.*, 2023). Cytokine levels have also been extensively investigated and studies have frequently found differences between patients and healthy controls, although studies disagree on the changes to specific cytokines (Blundell *et al.*, 2015; Corbitt *et al.*, 2019; Strawbridge *et al.*, 2019; Yang *et al.*, 2019). Neurological changes in ME/CFS include an increase in microglia and astrocyte activation and pro-inflammatory cytokine profiles, reduced white and grey matter (Thapaliya *et al.*, 2022) and brain stem volume (Thapaliya *et al.*, 2023), and cognitive deficits, including reduced processing speed, attention, and short and long term memory, and increase reaction time (Aoun Sebaiti *et al.*, 2022). Sleep disturbances have also been noted (Mohamed *et al.*, 2023), with the addition of serum from ME/CFS patients with sleep disturbances to mouse fibroblasts *in vitro* interfering with their circadian rhythm (Wei *et al.*, 2023). Further, metabolic dysfunctions include changes in fatty acid and amino acid metabolism (Naviaux *et al.*, 2016; Hoel *et al.*, 2021) have also been described. Cardiopulmonary defects include reduced exercise capacity, particularly of repeated exercise (Stevens *et al.*, 2018) which induces PEM. Exercise-induced PEM is associated with changes to fatty acid and amino acid metabolism (Glass *et al.*, 2023). Accumulating evidence suggests genetic and epigenetic alterations linked to immune system, neurological, metabolic, and cardiovascular abnormalities are also present in ME/CFS patients (Apostolou and Rosén, 2024).

1.1.3.1 GI disturbances in ME/CFS

Up to 65% of people with ME/CFS experience GI disturbances including nausea, diarrhoea, constipation, abdominal pain, and bloating (Maes *et al.*, 2014; Tschopp *et al.*, 2023) and 28% of patients described a GI-related infectious trigger for their illness (Johnston, Staines and Marshall-Gradisnik, 2016). Irritable bowel syndrome (IBS) comorbidity of 17 – 92% is high in ME/CFS compared to its general prevalence of 10 – 20% (Chu *et al.*, 2019). Inflammation as a result of an increase in GI permeability, also called “leaky gut”, has been hypothesized to

contribute to ME/CFS aetiology (Navaneetharaja *et al.*, 2016; Morris *et al.*, 2019), supported by evidence of increased bacterial translocation (Maes, Mihaylova and Leunis, 2007; Maes and Leunis, 2008; Maes *et al.*, 2012, 2014; Navaneetharaja *et al.*, 2016; F. Martín *et al.*, 2023). Recent analysis of the serum antibody repertoire of ME/CFS patients has shown a relative increase in antibodies against bacterial flagellins, particularly from the genus *Lachnospiracae*, compared to healthy controls (Vogl *et al.*, 2022). Interestingly, increased antibodies against bacterial flagellins are also found in patients with Crohn's disease (CD) (Alexander *et al.*, 2021; Bourgonje *et al.*, 2023) which increases the risk for ME/CFS (Tsai *et al.*, 2019). A study in five severe ME/CFS patients and healthy household controls has provided evidence for a reduced humoral response to homologous and heterologous GI bacteria, indicative of immune insufficiency (Seton *et al.*, 2023).

Several studies have shown changes in the GI microbiome and specifically the prokaryome in ME/CFS (Frémont *et al.*, 2013; Giloteaux *et al.*, 2016; Giloteaux, Hanson and Keller, 2016; Nagy-Szakal *et al.*, 2017; Kitami *et al.*, 2020; Lupo *et al.*, 2021; Guo *et al.*, 2023; He *et al.*, 2023; Xiong *et al.*, 2023). However, the changes described are inconsistent across studies with differences in methodology, sample size and patient cohorts most likely contributing to the variance (Du Preez *et al.*, 2018; König *et al.*, 2022). Still, several studies found that GI bacterial diversity in ME/CFS patients is marked by lower species diversity, with reduced species evenness and richness, and increased heterogeneity between people with ME/CFS compared to heterogeneity between healthy controls (Giloteaux *et al.*, 2016; Nagy-Szakal *et al.*, 2017; Guo *et al.*, 2023; Xiong *et al.*, 2023). Additionally, IBS is a determining factor in ME/CFS microbiome composition, with ME/CFS patients with and without IBS having distinct prokaryomes (Nagy-Szakal *et al.*, 2017; Guo *et al.*, 2023). The most consistent findings are a reduction in *Firmicutes* bacteria, and an increase in *Bacteroides*, although results are not always statistically significant (Frémont *et al.*, 2013; Shukla *et al.*, 2015; Giloteaux *et al.*, 2016; Nagy-Szakal *et al.*, 2017; Lupo *et al.*, 2021; Xiong *et al.*, 2023). Disease duration was found to be a determining factor in dysbiosis, with the highest degree of dysbiosis found in patients with <3 – 4 years of disease (Kitami *et al.*, 2020; Xiong *et al.*, 2023), with a *Firmicutes*/*Bacteroides* ratio comparable to healthy controls seen in ME/CFS patients with a disease duration of >10 years (Xiong *et al.*, 2023). Two large-scale multi-omics studies identified a reduction in short chain fatty acid (SCFA)-producing bacteria in association with a reduction of plasma butyrate levels (Guo *et al.*, 2023; Xiong *et al.*, 2023). These studies corroborated earlier findings of reduced levels of the SCFA-producing *Faecalibacterium prausnitzii* (Giloteaux *et al.*, 2016; Nagy-Szakal *et al.*, 2017; Kitami *et al.*, 2020), which has anti-inflammatory properties and contributes to intestinal homeostasis (R. Martín *et al.*, 2023). Interestingly, *F. prausnitzii* was also found to be reduced in Long COVID (Liu, Mak, *et al.*, 2022). One study found an inverse correlation between fatigue

and levels of SCFA in IBS patients (El-Salhy *et al.*, 2021), suggestive of a possible role for SCFA or SCFA-producing bacteria in fatigue (König *et al.*, 2022). Changes in the GI eukaryotic microbiome have so far not been detected in ME/CFS (Mandarano *et al.*, 2018).

1.1.4 The role of viral infections in ME/CFS

The pathophysiology of ME/CFS suggests activation of a "sickness response" that have evolved to induce behavioural and metabolic changes that conserve energy, which is triggered by an infectious agent and possibly other non-infectious stressors such as trauma, emotional and physical stress and chemical exposures (VanElzakker, 2013; Komaroff and Lipkin, 2023). Post-acute infectious syndromes (PAIS) have been proposed to arise through one or a combination of mechanisms involving persistent infection, possibly in deep tissues where detection is difficult leading to auto-antibody production as a result of molecular mimicry, and dysbiosis resulting in inflammation and reactivation of latent virus infections (Choutka *et al.*, 2022) with all of these features being implicated in ME/CFS. Outbreaks of ME/CFS, some of which were associated with non-polio enteroviruses, suggest an infectious viral trigger in at least a subset of ME/CFS patients (Underhill, 2015). Patients often report a virus infection preceding ME/CFS, with infectious onset self-reported in 60.3% to 77% of cases (Naess *et al.*, 2010; Chu *et al.*, 2019; Jason, Yoo and Bhatia, 2022; Bretherick *et al.*, 2023; Tschopp *et al.*, 2023). While it should be noted that in one study only 68% of EBV infectious mononucleosis (EBV-IM), 51% of COVID infections, and 26% of other infections that were self-reported could be confirmed by laboratory testing (Bretherick *et al.*, 2023). This discrepancy between self-reported and laboratory-confirmed infectious onset is at least in part explained by the fact that infections typically appear initially to be self-limiting and laboratory confirmation is not deemed necessary (Bateman *et al.*, 2021). Recent analysis of Taiwanese health records suggests an adjusted hazard ratio of 1.5 for developing ME/CFS following infection with Varicella-Zoster virus (VZV), *Mycobacterium tuberculosis*, *Escherichia coli*, *Candida* spp., *Salmonella* spp., *Staphylococcus aureus*, or influenza virus (Chang *et al.*, 2023). Those infected with influenza A during the 2009 influenza A pandemic had a hazard ratio of 2.04 for developing ME/CFS (Magnus *et al.*, 2015), with risk of developing ME/CFS following EBV-IM and VZV infection increased (Katz *et al.*, 2009; Tsai *et al.*, 2014). While the pre-COVID-19 pandemic prevalence of ME/CFS in the general population is generally estimated to be 0.86% (Lim *et al.*, 2020), 11% of those infected with EBV, *Coxiella burnetiid* (Q fever), or Ross River virus (Hickie *et al.*, 2006), and 8% - 23% of people with EBV-IM (Jason *et al.*, 2021), and 5% of those with *Giardia lamblia* infection (Naess *et al.*, 2012) developed ME/CFS, showing an increased risk for ME/CFS associated with a variety of infectious pathogens. In Long COVID, which is triggered by SARS-CoV-2, 43% and 58% meet the SEID and CCC case definition criteria for ME/CFS, respectively (Bonilla *et al.*, 2023; Jason and Dorri, 2023).

Several viruses have been associated with ME/CFS including herpesviruses, parvovirus and enteroviruses. Herpesviruses are enveloped double-stranded DNA (dsDNA) viruses with human herpesviruses (HHVs) consisting of herpes simplex virus (HSV)-1 and HSV-2, VZV, EBV, human cytomegalovirus (HCMV), HHV-6A, HHV-6B, HHV-7 and Kaposi's sarcoma-associated herpesvirus (KSHV). HHVs establish latent infections for example in B cells and neuronal cells, that can cause persistent inflammation as seen in ME/CFS (Sepúlveda *et al.*, 2019). Additionally, HHV infection of endothelial cells has been proposed to cause or contribute to ME/CFS pathophysiology (Nunes, Kell and Pretorius, 2024). Of the nine HHVs, EBV, HHV-6, and HHV-7 have been extensively investigated in ME/CFS. Serological investigations have found increased prevalence of EBV and HHV-6 in sera (Ablashi *et al.*, 2000; Gravelsina *et al.*, 2022; Rasa-Dzelzkaleja *et al.*, 2023) and HHV-7 in saliva of ME/CFS patients (Lee *et al.*, 2021), although these findings have not been replicated in other studies (Chapenko *et al.*, 2006; Frémont *et al.*, 2009; Cameron *et al.*, 2010; Oakes *et al.*, 2013; Blomberg *et al.*, 2019; Cliff *et al.*, 2019). However, the high seroprevalence of HHVs in the general population makes it difficult to detect significant differences between HHVs in ME/CFS and healthy controls, with viral load not necessarily correlating with antibody titres or prevalence (Rasa *et al.*, 2018; Ariza, 2021). Molecular techniques that detect HHVs in blood fractions and peripheral blood mononuclear cells (PBMCs) have shown increased activation of persistent HHV-6 and HHV-7 (Chapenko *et al.*, 2012; Shikova *et al.*, 2020; Rasa-Dzelzkaleja *et al.*, 2023), as well as increased rates of HHV-6 and HHV-7 co-infection (Chapenko *et al.*, 2006; Rasa-Dzelzkaleja *et al.*, 2023). Moreover, severity of ME/CFS was associated with increased HHV-6 and HHV-7 gene levels (Lee *et al.*, 2021; Gravelsina *et al.*, 2022; Rasa-Dzelzkaleja *et al.*, 2023). Testing of ME/CFS antibody repertoires against an array of EBV epitopes found increased numbers against EBV protein EBNA6 (Loebel *et al.*, 2017) and EBNA4 when sub-grouped by infectious onset (Sepúlveda *et al.*, 2022), with the ME/CFS EBV antibody repertoire being distinct from that of healthy controls (Fonseca *et al.*, 2024). Antibodies against HHV deoxyuridine 5′-triphosphate nucleotidohydrolase (dUTPase) have also been detected in ME/CFS, along with increased antibodies against EBV (Lerner *et al.*, 2012), HHV-6, and VZV dUTPase (Halpin *et al.*, 2017). EBV dUTPase is excreted in exosomes and induces pro-inflammatory cytokine production in dendritic cells (DCs) (Ariza *et al.*, 2013). It has been linked to inflammatory cytokine production in the brain and increases in blood-brain-barrier permeability (Williams *et al.*, 2019). HHV-6 has been found in brain biopsies of ME/CFS patients but not healthy controls (Kasimir *et al.*, 2022). Both EBV and HHV-6 dUTPases induce activin A and interleukin-21 production by DCs, which induces T follicular helper cell differentiation and increases marginal zone and germinal centre B cell numbers, possibly pointing to a mechanism of immune system dysfunction (Cox *et al.*, 2022). Additionally, activation of HHV-6 induces changes in

mitochondria morphology leading to the production of extracellular factors that can transfer this phenotype to non-infected cells (Schreiner *et al.*, 2020). Altogether, these lines of study point to an involvement of HHVs in ME/CFS, particularly EBV, HHV-6 and HHV-7, as confirmed by two recent meta-analyses (Mozhgani *et al.*, 2021; Hwang *et al.*, 2023). This is further supported by several studies investigating the use of the nucleoside analogue antiviral drugs valacyclovir and valganciclovir, which have positive results in ME/CFS patients with high EBV and HHV-6 antibody titres (Kogelnik *et al.*, 2006; Lerner *et al.*, 2007; Watt *et al.*, 2012; Montoya *et al.*, 2013).

Parvovirus B19 (B19) is a single-stranded DNA (ssDNA) virus most commonly associated with the "fifth disease" (erythema infectiosum) in children and infects adults, with 60% of women and 30% of men experiencing arthralgia or arthritis that can persists for months to years in 20% of women (Reno, Cox and Powell, 2022). B19 primarily infects bone marrow erythroid progenitor cells but also endothelial cells and B cells in which replication is absent (Zakrzewska *et al.*, 2023). Persistent infection is often established, with integrated B19 genomes detected in both erythroid progenitor cells and PBMCs (Kerr, 2005; Janovitz *et al.*, 2017). Several case reports and studies have suggested an increased probability of developing ME/CFS following B19 infection (Kerr *et al.*, 2002; Seishima *et al.*, 2008), particularly in those experiencing high levels of psychological stress at the time of infection (Kerr and Mattey, 2008). No difference in B19 antibody prevalence between ME/CFS and healthy controls has been found (Kato *et al.*, 2009; Zhang *et al.*, 2010) although B19 antigen was detected more frequently in gastric and duodenal biopsies of ME/CFS patients (Frémont *et al.*, 2009). There was also no difference in the prevalence of B19 DNA in PBMCs of ME/CFS discordant twins (Koelle *et al.*, 2002), although evidence suggests an increase in B19 active infection and co-infection of B19 with HHV-6 and HHV-7 (Kerr *et al.*, 2010; Rasa-Dzelzkaleja *et al.*, 2023).

As the first known outbreaks of ME/CFS were thought to be caused by enteroviruses, there is a long history of association of this group of viruses with ME/CFS (Acheson, 1959). Enteroviruses are a highly diverse group of non-enveloped single-stranded RNA (ssRNA) viruses that infect a wide range of tissues in humans, including GI, respiratory, nervous system and muscle (Tapparel *et al.*, 2013). One study showed neutralising antibodies against enteroviruses in 50% of ME/CFS patients and enterovirus VP1 antigen in 81% of patients, although healthy controls were not included for comparison (Chia and Chia, 2008). A case report showed enterovirus infection in three patients preceded ME/CFS disease onset, with detectable enterovirus RNA and protein in GI biopsies years after the initial infection, suggesting persistent enterovirus infection (Chia *et al.*, 2010). Overall, the association of enterovirus through serological and molecular tests is inconclusive (Rasa *et al.*, 2018). It has been proposed that persistent enterovirus infections may play a role in ME/CFS, and that the

high genetic and antigenic diversity of enteroviruses, as well as the tropism diversity has prevented consistent detection of enterovirus in ME/CFS due to inappropriate selection of antibodies, PCR primers, and/or tissue samples (O'Neal and Hanson, 2021).

Endogenous retroviruses (ERVs) have also been implicated in ME/CFS pathogenesis. These are ancient remnants of retroviral infections in germline that have been acquired over millions of years (Grandi and Tramontano, 2018) and are estimated to make up around 8% of the human genome (International Human Genome Sequencing Consortium, 2001). They can act as regulatory elements by mobilization into a functional gene, providing enhancer functions, production of non-coding RNAs, or activate pathogen recognition receptors (PRRs) through the production of HERV nucleic acids (Kassiotis and Stoye, 2016; Grandi and Tramontano, 2018). Over time, some ERVs have evolved to contribute to essential biological processes such as placentation (Lavialle *et al.*, 2013) and myelin production (Ghosh *et al.*, 2024). ERVs have immunomodulatory properties as they, for example, contribute to antimicrobial immunity in the skin of mice (Lima-Junior *et al.*, 2021) and are associated with immune aging and senescence (Liu *et al.*, 2023). Transposable elements are genomic elements that have the ability to copy or cut and paste themselves into new regions of the genome, including ERVs, and their expression is increased during viral infection in humans and mice in a variety of cell types (Macchietto, Langlois and Shen, 2020) with the enterovirus Coxsackie B virus and EBV upregulating HERV expression in PBMCs (Dechaumes *et al.*, 2020; Apostolou and Rosén, 2024). ERVs and other transposable elements are under tight epigenetic control, with epigenetic changes in ME/CFS related to increased transcription of transposable elements. Indeed, HERV proteins are detected in GI biopsies of ME/CFS patients but not healthy controls (De Meirleir *et al.*, 2013), and HERV transcription in PBMCs (Rodrigues *et al.*, 2019) and saliva anti-HERV antibodies are elevated in ME/CFS (Apostolou *et al.*, 2022). Interestingly, the saliva antibody response was more pronounced in female than male ME/CFS patients (Apostolou *et al.*, 2022). On the other hand, another study found no difference in HERV expression in PBMCs between ME/CFS patients and healthy controls (Oakes *et al.*, 2013). HERV expression is also upregulated in related diseases, including fibromyalgia (Ovejero *et al.*, 2020) and Long COVID (Gimenez-Orenga *et al.*, 2022), with HERV expression correlating with pro-inflammatory cytokine expression and disease severity in acute COVID-19 (Balestrieri *et al.*, 2021).

Consistent and compelling evidence of HHV, parvovirus, enterovirus, or HERV involvement in the development of ME/CFS is still lacking. A recent meta-analysis found a statistically significant association of HHV-7, B19, enterovirus, coxsackie B virus, and borna disease virus with ME/CFS, although the latter association was based on a single data set (Hwang *et al.*, 2023). If these viruses are involved in ME/CFS, one unanswered question is whether they are the primary cause, or if a different infection or unrelated stressor leads to reactivation of latent

infections, with numerous articles proposing theories on how viral infections lead to the development of ME/CFS (Dowsett *et al.*, 1990; Chia and Chia, 2003; Bansal *et al.*, 2012; Blomberg *et al.*, 2018; Proal and Marshall, 2018; Nacul *et al.*, 2020; Ariza, 2021; O'Neal and Hanson, 2021; Hanson, 2023; Nunes, Kell and Pretorius, 2024; Walitt *et al.*, 2024). It is likely that various infections and other stressors cause disturbances that lead to ME/CFS (Proal and Marshall, 2018; Choutka *et al.*, 2022; Annesley *et al.*, 2024), and that infections themselves either persist or cause reactivation of latent infections, possibly leading to inflammation and/or production of auto-antibodies and disease exacerbation (Choutka *et al.*, 2022). Generally, it is thought that a viral infection leads to immune system and neurological disturbances that cause disruption of metabolic homeostasis (Nacul *et al.*, 2020) or neuromuscular changes that lead to deconditioning (Walitt *et al.*, 2024). It could be that this then leads to persistent infection of the patient (persistent infection hypothesis), or if the infection is the trigger in ME/CFS, this is then maintained after the infection has been cleared (hit-and-run hypothesis) (Günther *et al.*, 2019).

1.1.5 The role of the GI microbiota in ME/CFS

GI dysbiosis has been implicated in the aetiology of ME/CFS, including a role for GI viruses (Newberry *et al.*, 2018). While current evidence is insufficient to support a causative role, the role of GI microbiota (viruses, bacteria, archaea and eukaryotic microbes) in dysbiosis acting via the microbiota-gut-brain axis warrants further research (Stallmach *et al.*, 2024). GI dysbiosis is associated with ME/CFS and PAIS, and GI infection by eukaryotic viruses might play a role. There is evidence of HHV, parvovirus and enterovirus antigens and nucleic acid in ME/CFS GI biopsies (Chia and Chia, 2008; Frémont *et al.*, 2009; Chia *et al.*, 2010). In other diseases associated with GI dysbiosis there is also evidence of increased GI prevalence and abundance of eukaryotic viruses. For example, increased abundance of *Picobirnaviridae* associated with Graft versus Host Disease (Legoff *et al.*, 2017), and *Hepeviridae* and *Hepadnaviridae* associated with Inflammatory Bowel Disease (IBD) (Ungaro *et al.*, 2019). The human GI tract can is reservoir for human viruses, including HIV-1 and SARS-CoV-2 (Neurath, Uberla and Ng, 2021), with epithelial tuft cells forming a reservoir for persistent norovirus infection in the mouse GI tract (Wilen *et al.*, 2018).

Whether GI dysbiosis in these cases is caused by eukaryotic virus infection or vice versa, is unclear. The GI microbiome plays a role in GI immune homeostasis and changes in the GI prokaryome have been associated with increased susceptibility to infection by GI pathogens (Woods Acevedo and Pfeiffer, 2021). The prevalence of dysbiosis in ME/CFS patients (Guo *et al.*, 2023; Xiong *et al.*, 2023), the high incidence of IBS in ME/CFS (Johnston, Staines and Marshall-Gradisnik, 2016; Chu *et al.*, 2019; Tschopp *et al.*, 2023), and a possible post-infectious aetiology of both ME/CFS and IBS (Navaneetharaja *et al.*, 2016; Barbara *et al.*, 2019) indicate a

possible causative role of eukaryotic viruses. For instance, infection with SARS-CoV-2 leads to changes in GI microbiota composition that correlate with disease severity, including a reduction of *F. prausnitzii* in those with severe COVID-19 and those who develop Long COVID (Yeoh *et al.*, 2021; Liu, Mak, *et al.*, 2022; Schult *et al.*, 2022; Su *et al.*, 2023). On the other hand, at time of SARS-CoV-2 infection the microbiome composition of those who go on to develop Long COVID was distinct from those who recovered and from healthy controls, while the microbiota of recovered COVID-19 patients were more similar to healthy controls (Liu, Mak, *et al.*, 2022). Additionally, in EBV-IM, pre-existing GI disturbances increase the likelihood of developing ME/CFS (Jason *et al.*, 2022).

A range of interactions take place between GI viruses and bacteria and the human host (Virgin, 2014; Pfeiffer and Virgin, 2016). Maturation of the immune system (Woods Acevedo and Pfeiffer, 2021), humoral immunity (Zhang *et al.*, 2020), and tonic activation of innate immune responses are dependent on GI bacteria (Lima-Junior *et al.*, 2021; Erttmann *et al.*, 2022; Wirusanti, Baldridge and Harris, 2022). In mice, GI microbiota play an important role in protection against rotavirus, norovirus, hepatitis B virus, and murid herpesvirus-68 infection (Yaron *et al.*, 2020; Gozalbo-Rovira *et al.*, 2021; W. Guo *et al.*, 2021; Santiso-Bellón *et al.*, 2022). On the other hand, GI bacteria also promote eukaryotic viral infection by stabilising viral particles and their attachment to the target cells, and promote viral persistence (Berger and Mainou, 2018; Roth, Grau and Karst, 2019; Bhar *et al.*, 2022). In turn, eukaryotic viruses themselves can also affect the prokaryome and production of bacterial extracellular vesicles that promote infectivity (Mosby *et al.*, 2022). Conversely, GI viruses affect microbiota composition, as bacterial (phages) and archaeal viruses can affect microbial diversity, metabolism, and transfer of antimicrobial resistance genes, with potential implications for human health (Chevallereau *et al.*, 2022). GI viruses also interact with the human host directly. Viruses encode microbe-associated molecular patterns (MAMPs) that are recognised by host cell PRRs, including Toll-like receptors (TLRs) that are membrane-bound or cytosolic (Clinton *et al.*, 2022). For instance, endosomal membrane-bound TLR3 and TLR7 recognise double-stranded RNA (dsRNA) and ssRNA, respectively, while cytosolic cyclic GMP-AMP synthase (cGAS) recognises cytosolic DNA (Metzger, Krug and Eisenächer, 2018). Upon viral infection, production of type I and type III interferons (IFN) is activated (Odendall *et al.*, 2014; McNab *et al.*, 2015; Thaiss *et al.*, 2016). Subsequently, IFN-stimulated genes are transcribed and pro-inflammatory cytokines are produced (Iwasaki and Pillai, 2014). Type III IFNs play an important role in intestinal antiviral immunity (Kotenko *et al.*, 2003; Pott *et al.*, 2011; Selvakumar *et al.*, 2017), but also protects against damaging inflammation in intestinal and lung epithelium (Broggi *et al.*, 2017; Galani *et al.*, 2017). Interactions between viruses and the human immune system are not limited to eukaryotic viruses, since phages can rescue

immunodeficiency in germ-free mice via TLR9 signalling (Gogokhia *et al.*, 2019), inhibit antibacterial immunity via TLR3-dependent inhibition of phagocytosis (Sweere *et al.*, 2019), and elicit antibody responses (Żaczek *et al.*, 2016; Majewska *et al.*, 2019). Immunodeficiency in germ-free mice is also rescued by persistent infection by murine norovirus (Kernbauer, Ding and Cadwell, 2014) and by several other eukaryotic viruses (Liu *et al.*, 2019; Dallari *et al.*, 2021). These include HERVs, whose expression is upregulated in GI biopsies of people living with HIV-1 (Dopkins *et al.*, 2024), and HERV sequences have been found in the faecal viral extracts of Type 1 diabetes patients (Cinek *et al.*, 2021). Taken together, this illustrates the complexity of interactions between GI viruses and bacteria that affect human immunity and health and that can contribute to ME/CFS pathology.

Modulation of the GI microbiota to improve ME/CFS has been investigated. Probiotic and antibiotic treatments have been undertaken with one randomised, placebo-controlled clinical trial (RCT) finding a reduction in anxiety (Rao *et al.*, 2009), although this is not a core symptom of ME/CFS (König *et al.*, 2022). Another RCT found a reduction in blood inflammatory markers but did not assess ME/CFS symptoms (Groeger *et al.*, 2013). As dysbiosis affects the composition of both viruses and bacteria the wholesale replacement of GI microbiota may be a treatment of ME/CFS using faecal microbiota transplantation (FMT). FMT is being considered for a wide variety of dysbiosis-associated diseases, including liver disease, IBS, IBD, obesity and metabolic disorder, arthritis, cancer, and autism although studies to date have produced mixed results (Biazzo and Deidda, 2022) It has proven most successful in treating recurrent *Clostridium difficile* infections (rCDI), in which it has success rate of over 90% (Allegretti *et al.*, 2019). The determinants for FMT success are still unclear with, for example, the requirement of engraftment of donor bacteria in treating rCDI not been true for all diseases (Podlesny *et al.*, 2022). It is also not clear which factors in the donor microbiota determine strain engraftment and which donor strains lead to improvement or how many treatments are necessary, with different diseases most likely having different requirements (Biazzo and Deidda, 2022).

Interest in FMT as a treatment for ME/CFS grew after publication of a retrospective study in which ME/CFS patients, most of whom had IBS, received a transcolonic or rectal infusion of a culture of 13 common faecal bacteria (Borody, Nowak and Finlayson, 2012). In this study, 70% of patients responded to the treatment after 4 weeks. ME/CFS did not return in 58% of patients contacted 15 – 20 years later. Two studies have published results on FMT today. One compared FMT to a therapy consisting of management of diet and lifestyle and prescription of pre- and probiotics. It found that of the patients receiving 10 treatments from 10 different donors 17 out of 21 achieved at least 60% improvement, and improvement was significantly higher than the other interventions (Kenyon, Coe and Izadi, 2019). However, an RCT in 5 ME/CFS patients found no improvement after FMT (Salonen *et al.*, 2023). More research has been conducted in

IBS, with nine FMT-RCTs to date (Halkjær *et al.*, 2018; Johnsen *et al.*, 2018; Aroniadis *et al.*, 2019; Holster *et al.*, 2019; Huang *et al.*, 2019; El-Salhy *et al.*, 2020; Lahtinen *et al.*, 2020; Holvoet *et al.*, 2021; Q. Guo *et al.*, 2021). Meta-analysis of seven RCTs showed that while the response rate for FMT was higher than placebo, the difference was not significant due imprecision in reporting, risk of bias and inconsistent methodology (Rodrigues *et al.*, 2023). However, efficacy seems to depend on the method of delivery, as FMTs reached a significant improvement of IBS over placebo when only including studies that provided FMT *via* gastroscopy and nasojejunal tube (Wu, Lv and Wang, 2022; Rodrigues *et al.*, 2023). Nevertheless, quality of life score was significantly increased after FMT versus placebo, regardless of delivery method (Wu, Lv and Wang, 2022). Several studies have shown that FMT in IBS leads to a change in the faecal microbiome (Körner and Lorentz, 2023), and that patient microbiomes became more similar to their donors in the responders, but not in non-responders (Goll *et al.*, 2020). One study found an inverse correlation in IBS patients between fatigue and levels of SCFA (El-Salhy *et al.*, 2021), which suggests a role for SCFA or SCFA-producing bacteria in fatigue (König *et al.*, 2022). This finding supports the idea that the reduced capacity for SCFA metabolism in ME/CFS GI microbiota and reduced levels of SCFA in the ME/CFS GI tract are related to fatigue (Guo *et al.*, 2023; Xiong *et al.*, 2023). Given the diverse interactions between viruses, bacteria and the human host discussed above, it is likely viruses are a determinant of FMT success or failure (Lam *et al.*, 2022). For example, transfer of phages during FMT was associated with treatment success in rCDI (Draper *et al.*, 2018; Zuo *et al.*, 2018), and rCDI has previously been successfully treated using faecal viral filtrate transplantation (Ott *et al.*, 2017). However, much is still unclear about the role of GI viruses in FMT and ME/CFS.

1.2 The GI virome

The GI microbiome contains a vast number of phages, estimated between 10^8 and 10^{10} phage particles per gram of faeces (Lepage *et al.*, 2008; Kim *et al.*, 2011; Hoyles *et al.*, 2014; Shkoporov *et al.*, 2018), as well as some eukaryotic viruses. Current estimates suggest there are between 1 and 10 virus particles for every bacteria in the GI-tract (Reyes *et al.*, 2010; Shkoporov and Hill, 2019). This collection of viruses (the virome) has received less attention in scientific research compared to the bacterial component of the microbiome, but nonetheless plays a key role in shaping the GI microbiome and the human host (Cao *et al.*, 2022). Most viruses in the GI tract are DNA phages that infect the resident prokaryotic microbiota. The most abundant and prevalent phages in the human GI tract are tailed phages of the class *Caudoviricetes*, of which for example the laboratory strains T4 and T5 are a member, and microviruses of the class *Malgrandaviricetes*, of which phage phiX174 is one of the most widely used laboratory phages (Liang and Bushman, 2021). *Caudoviricetes* typically have icosahedral capsid of 45 – 185 nm in

diameter (Dion, Oechslin and Moineau, 2020), and can have up to 200 nm long tails through which the genome is injected into the host cell (Letellier *et al.*, 2004). Their linear dsDNA genomes can be as small as 11,500 bp and up to 500,000 bp (Hatfull and Hendrix, 2011). Phages in the class *Malgrandaviricetes*, of which *Microviridae* is the only family, are viruses with small icosahedral capsids with a diameter of ~ 25 nm (Doore and Fane, 2016) containing a circular single-stranded ssDNA genome of 3,000 – 8,000 bases (Kirchberger, Martinez and Ochman, 2022). Another class of ssDNA phages, less prevalent in the human GI tract are the *Faserviricetes*, of which *Inoviridae* is the largest family, which consists of filamentous phages of 6 – 10 nm in diameter and 600 – 2,500 nm in length, with 5,500 – 10,600 base genomes (Knezevic, Adriaenssens and ICTV Report Consortium, 2021). Filamentous phage M13 is a frequently used lab strain of this family. Phages have various replication strategies including a lytic replication cycle in which phages attach to the host cell and inject their genome into the host cell, after which the host genome is replicated, and virus products are produced and assembled into new virions. During this cycle the host cell is ultimately lysed by endolysins to release virions. Another strategy is a lysogenic replication cycle, in which after injection of the virus genome, the genome is first integrated into the bacterial chromosome as a prophage and the prophage is replicated during bacterial cell division (Mirzaei and Maurice, 2017). The prophage can be induced by external cues, including GI inflammation (Diard *et al.*, 2017), to enter into the lytic cycle with replication of the genome and production of viral products to build new virions that are released through cell lysis (Chevallereau *et al.*, 2022). Virulent phages (also referred to as lytic phages) are obligately lytic replicating, whereas temperate phages have the capacity to integrate their genomes and enter into lytic or lysogenic replication cycles depending on environmental cues (Chevallereau *et al.*, 2022). As the lysogenic replication cycle depends on host cell fitness, temperate phages often carry genes that can give a competitive advantage to the host cell and contribute to horizontal gene transfer (Mirzaei and Maurice, 2017).

Eukaryotic viruses have a low abundance in the GI tract in adults compared to phages (Liang and Bushman, 2021). A large portion of the GI eukaryotic viruses are likely diet-derived, with many viruses belonging to *Alpha*-, and *Betaflexiviridae* families of the plant and fungal ssRNA virus class *Alsuvircetes*, and the plant ssDNA virus family *Geminiviridae* in the class *Arfiviricetes*. Human-infecting viruses are also found in the GI tract and can be roughly divided into disease- and non-disease associated viruses. Disease associated viruses include are associated with gastroenteritis including those in the genera *Rotavirus*, *Norovirus*, *Sapovirus*, *Astrovirus*, *Adenovirus*, and *Enterovirus* (do Socorro Fôro Ramos *et al.*, 2021). Disease-associated viruses have a high prevalence and abundance in humans in the first years of life, but prevalence steadily decreases and in adults are only found sporadically during symptomatic or

asymptomatic infections (Lim *et al.*, 2015; Beller and Matthijnssens, 2019; Walters *et al.*, 2023). Non-disease associated human viruses include ssDNA viruses of the *Annelloviridae* and *Redondoviridae* families, which have a high prevalence and diversity in humans , although their roles in human health and disease is unclear (Gregory *et al.*, 2019; Taylor *et al.*, 2022). In adults, the GI virome is individually unique and relatively stable over time (Gregory *et al.*, 2019; Shkoporov *et al.*, 2019).

1.2.1 Analysing the GI virome

Techniques for virome analysis include culture-based methods that rely on isolating viruses from biological samples, molecular techniques like quantitative (q)PCR to detect viral sequences, and sequencing (Trubl *et al.*, 2020). Isolation of viruses relies on the availability of suitable host cells for propagation in cell culture with molecular techniques relying on pre-existing knowledge of specific viral sequence. Sequencing techniques, on the other hand, enable sequence-independent investigation of viromes.

1.2.1.1 Sequence-targeted analysis of the virome

In clinical settings, a fast turnaround time and low costs are an important factor for virome analysis and diagnosis. In these settings, detection of individual viruses can be achieved using qPCR and reverse transcription (RT)-PCR assays that target specific pathogen sequences. Assay array systems such as BioFire and Luminex allow syndromic testing, in which a range of pathogens can be tested in a single array (Greatorex *et al.*, 2014). TaqMan array cards (TACs) offer a similar platform, with the added advantage of providing customisable assays (Kodani *et al.*, 2011), which has enabled the development of GI pathogen TACs (Liu *et al.*, 2013; Agoti *et al.*, 2022).

1.2.1.2 Sequence-independent analysis of the virome

While various virus clades share genes such as the HK97 major capsid protein shared by phages and herpesviruses of the realm *Duplodnaviria*, the RNA-dependent RNA polymerase gene shared by RNA viruses in the realm *Riboviria*, and the double jelly roll capsid protein shared by *Varidnaviria*, there is no universally conserved viral gene (Koonin, Krupovic and Agol, 2021). A comprehensive virome analysis therefore requires sequence-independent techniques with GI virome research usually relying on metagenomic sequencing. The GI virome can change in composition along the GI tract, as well as from the lumen to the mucosal surface, and thus, faeces are not a perfect reflection of the GI virome (Shkoporov *et al.*, 2022; Yan *et al.*, 2023). Nonetheless, analysis of the faecal virome is a convenient and non-invasive means of sampling the GI virome, and is therefore used most frequently (Gregory *et al.*, 2020).

Metagenomic virome analysis can be performed on total faecal DNA and/or RNA (bulk), or on nucleic acid from virus-like particles (VLP)s isolated from faeces using techniques like filtration, ultrafiltration, polyethylene glycol precipitation, and caesium chloride gradient

purification (Kleiner, Hooper and Duerkop, 2015; Mirzaei *et al.*, 2021). Of note, while the term VLP has been used to describe non-replicating virus particles in, for example, vaccine development settings, VLP here refers to particles in the size range of viruses, that contain nucleic acid, and may or may not be infectious (Hyman, Trubl and Abedon, 2021). Sequencing of VLP-enriched nucleic acid extracts reduces the amount of bacterial and human host nucleic acid, while bulk metagenomic sequences includes prophages and thus enables analysis of non-replicating integrated temperate phages (Li *et al.*, 2022).

Various biases can occur during the process from sampling to nucleic acid extraction that affect the final virome composition. For instance, storage time and temperature, VLP extraction methods, VLP nucleic acid amplification by PCR, and operator bias affect the results, and batch-to-batch variation needs to be considered as well (Shkoporov *et al.*, 2018; Hsieh *et al.*, 2021). To account for methodological variability, mock communities (MC) of artificially constructed and defined microbial populations can be used (Knight *et al.*, 2018; Amos *et al.*, 2020). Additionally, laboratory equipment and reagents are a frequent source of contaminants, and blanks help control for contaminants (Boers, Jansen and Hays, 2019).

Virome research has historically focused on DNA viruses as these were assumed to represent the majority of phages and because RT is typically required to sequence RNA. With computational advances and reducing costs of sequencing, the number of known virus sequences has grown exponentially in the last decade (Call, Nayfach and Kyrpides, 2021). However, the amount of RNA phages in the environment may have been greatly underestimated (Handley and Virgin, 2019). Importantly, the majority of eukaryotic viruses are RNA viruses (Koonin, Krupovic and Agol, 2021). Inclusion of RNA viruses is therefore important, particularly for human (GI) virome research. Recently, several meta-transcriptomic studies have expanded the number of known RNA viruses, contributing to a more than 5-fold increase in RNA viral diversity (Callanan *et al.*, 2020; Edgar *et al.*, 2022; Neri *et al.*, 2022) and the discovery of new RNA viral clades (Janowski *et al.*, 2017). While some studies perform separate RNA and DNA virome analyses, several protocols have been published that enable combined sequencing of RNA and DNA viromes using whole transcriptome-based approaches (Conceição-Neto *et al.*, 2015; L. Li *et al.*, 2015; Kramná and Cinek, 2018; Shkoporov *et al.*, 2018).

Several sequencing platforms are widely used to obtain viral sequences. Most commonly, high-throughput short read sequencing platforms provide individual sequences of 100 – 300 bases, like those produced by Illumina (Smith *et al.*, 2022). While their low cost and high accuracy makes these platforms an attractive option, low-complexity genomic sequences are difficult to resolve due to the short reads. Pacific Biosciences and Oxford Nanopore have produced long read sequencing platforms that can generate up to 50 kilobase (kb) and 4.2 megabase (Mb)

sequences, but with reduced accuracy and the need for larger amounts of input nucleic acid (Mirzaei *et al.*, 2021; Smith *et al.*, 2022). Nonetheless, Illumina-based short read sequencing remains the gold standard for metagenomic virome sequencing (Cook *et al.*, 2024).

Following sequencing, reads are processed to remove low-quality reads, trim adapter sequences, and trim low-quality sequences and sequencing artifacts and a frequently program is fastp, which includes all of these steps in one program (Chen *et al.*, 2018; Chen, 2023). Once high-quality reads have been filtered, reads can be mapped directly against reference genomes to determine the origin of the read and calculate relative abundances of the reference genome. This can be done using alignment-based tools like bowtie2 (Langmead and Salzberg, 2012; Langmead *et al.*, 2019), Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990; Camacho *et al.*, 2009), or MMseqs2 (Steinegger and Söding, 2017; Mirdita, Steinegger and Söding, 2019). Alternatively, Kraken2 determines abundances by mapping reads to reference genomes by scoring the last common ancestor of sequential *k*-mers, stretches of *k* nucleotides, in the read (Wood, Lu and Langmead, 2019). A Kraken 2-based pipeline was recently published that is optimised for the analysis of viromes (Pinto *et al.*, 2023). However, the drawback of direct read mapping approaches is that these all rely on reference sequences, and a high fraction of viral reads remain unclassified (Shkoporov and Hill, 2019).

Rather than direct virome profiling of sequencing reads, the current approach for virome analysis depends on *de novo* assembly of reads into larger contiguous sequences (contigs). Various tools are available, the most frequently used tools being MEGAHIT (D. Li *et al.*, 2015; Li *et al.*, 2016) and SPAdes (Bankevich *et al.*, 2012), with metagenomic- and viral metagenomic-optimised versions available for the latter (Nurk *et al.*, 2017; Antipov *et al.*, 2020). MEGAHIT and SPAdes-based assemblers often outperform other assemblers in benchmarking studies, with (meta)SPAdes having increased accuracy and contiguity (Roux *et al.*, 2017; Sutton *et al.*, 2019), while MEGAHIT produces fewer chimeric sequences (Roux *et al.*, 2017). MEGAHIT typically is faster and has lower memory requirements than (meta)SPAdes, and a comparison of three large datasets did not find substantial differences in assembly quality between the two (D. Li *et al.*, 2015; Nayfach, Páez-Espino, *et al.*, 2021).

Following assembly, contigs can be analysed to determine the origin of sequences. Many virus mining tools exist that work by predicting protein sequences and analysing viral protein sequences using Hidden Markov model searches, nucleotide *k*-mer frequencies and GC content, and detecting terminal repeats and inverted terminal repeats (Li *et al.*, 2022). These tools can be used to filter potential viral sequences for further analysis using CheckV, which produces a score determining the confidence of a viral origin of the sequence and an estimate of the genome completeness of the sequence through removal of host sequences and detection of

viral genes and terminal repeats (Nayfach, Camargo, *et al.*, 2021). CheckV produces a final quality score of Complete (100% completeness), High Quality (>90% completeness), Medium Quality (>50% completeness), Low Quality (≤50% completeness), and Undetermined in cases where no confident assessment can be made, according to Minimum Information about an Uncultivated Virus Genome (MIUViG) standards (Roux *et al.*, 2019). Depending on the research question, putative viral sequences can be filtered by various criteria like genome completeness and length. For instance, those interested in *Caudoviricetes* phages remove sequences <5 kb in length, since these phages have large genomes. However, many viruses have much smaller genomes and a lower cut-off is required to preserve these viruses.

Finally, viral sequences can be assigned a taxonomy using alignment-based tools to align nucleotide or predicted protein sequences against reference databases, for example using BLAST or MMseqs2 (Camacho *et al.*, 2009; Steinegger and Söding, 2017). While many large virome datasets have been published recently that have vastly increased the number of virus sequences (Callanan *et al.*, 2020; Gregory *et al.*, 2020; Camarillo-Guerrero *et al.*, 2021; Nayfach, Camargo, *et al.*, 2021; Neri *et al.*, 2022), assigning a taxonomy is challenging as only 14.2 – 56.6% of sequences can be assigned a family (Li *et al.*, 2022). Instead, viral sequences can be clustered *de novo* based on genome sequence features such as shared genes, for example using the tool vConTACT2 (Jang *et al.*, 2019).

1.3 The Comeback Study

Studies involving FMT in ME/CFS can provide valuable insights into possible treatment avenues of ME/CFS, as well as into whether GI dysbiosis plays a causative role in ME/CFS (Stallmach *et al.*, 2024). The Comeback Study is a phase II, randomised, double-blinded, placebo controlled clinical trial in which the efficacy and safety of FMT is investigated (Skjevling *et al.*, 2024). The trial was conducted in Harstad, by the University Hospital of North Norway (UNN). The principal investigator of the study is Peter H. Johnsen, co-principal investigator is Rasmus Goll, and coordinating investigator was Linn Christin Kallbekken Skjevling. A total of 80 mild to severe ME/CFS patients, as determined by the ME-ICC, were recruited. Treatment consisted of a faecal transplant delivered via enema, with patients randomly allocated to an active transplant from a healthy donor, or a placebo transplant prepared from the patient's own faecal sample. Baseline and follow-up measurements at three and twelve months include assessment of symptom severity, GI symptoms, and collection of faecal, urine and blood samples. The longitudinal and interventional nature of The Comeback Study offers a unique opportunity for studying the GI microbiome in ME/CFS, and its role in FMT. Virome analysis was led by the Quadram Institute, with VLP enrichment and nucleic acid extraction performed in Harstad, and sequencing and analysis performed at the Quadram

Institute. UNN will conduct a prokaryome analysis, which will enable integrating changes in the virome to changes in the prokaryome and improvements in the patients.

As explained in Chapter 2, methods for quantifying diverse VLPs are currently lacking. An accurate method for the quantification of MC members is required to correlate virus input levels to the virus levels as observed in the metagenomic data. A virus MC can be used to assess bias and reproducibility of metagenomic methods, and to serve as an internal control for virome analysis. This will be valuable for analysis of the virome of The Comeback Study participants, as studying changes in the virome requires knowledge of the methodological variability in order to differentiate between methodological variability and biological variability and change. This will also help improve our understanding of the GI virome in ME/CFS and its role in FMT. Additionally, TACs could provide a cost-effective means of high-throughput targeted detection of microbes with potential clinical use beyond determining prevalence, which warrants evaluation of this platform in the setting of this clinical trial.

1.4 Hypothesis and aims

FMT from healthy donors induces changes in the composition of the GI virome in ME/CFS patients, leading to engraftment of GI donor viruses.

The aims of this project are:

1. To develop a method to accurately quantify diverse virus particle titres, including non-infectious particles.
2. To construct an MC for use as a reference standard and evaluate the bias and reproducibility of metagenomic sequencing approaches using the MC reference standard.
3. To analyse the GI virome of ME/CFS patients and donors participating in The Comeback Study to determine the changes in the ME/CFS patient virome following FMT.
4. To determine the presence of common GI pathogens and the effect of FMT in ME/CFS patients using TACs and evaluate the use of TACs for large scale studies.

CHAPTER 2: DEFINING A REFERENCE STANDARD FOR METAGNOMIC VIROME ANALYSIS

2.1 Introduction

The human GI tract contains a wide variety of viruses, from prokaryotic viruses (phages) that predate on the resident microbes, to eukaryotic viruses that originate from the diet or infect human cells (Liang and Bushman, 2021). Viruses have ssRNA, dsRNA, ssDNA, or dsDNA genomes and enveloped or non-enveloped virions in a wide range of sizes. A comprehensive analysis of the GI virome therefore requires methods that can capture all these viruses (Conceição-Neto *et al.*, 2015). An MC consisting of a collection of representative viruses can be used as a reference standard to estimate the accuracy of next generation sequencing-based metagenomics (Knight *et al.*, 2018; Boers, Jansen and Hays, 2019). This helps to evaluate methodological bias and quantify metagenomic results (Roux *et al.*, 2016). Accurate quantification of individual viruses is a crucial step in the construction of such a mock community with several methods being used.

2.1.1 Virus quantification methods

Plaque assays (PA) determine the number of infectious viruses, defined as plaque forming units (PFU), in a sample. However, not all virus particles are infectious, and the particles-to-PFU ratio (PPR) can range from 1 for bacterial viruses (phages) to over 10,000 for some eukaryotic viruses (Klasse, 2015). The PPR for Varicella-zoster virus has been reported to be 40,000 (Carpenter, Henderson and Grose, 2009). qPCR can also be used to quantify the number of genome copies and is not affected by the PPRs but requires individual assays and primers to be developed and calibrated for each virus.

There are also virus-independent approaches to quantifying viruses. Transmission electron microscopy (TEM) and other techniques that rely on the optical detection of particles through fluorescence or light scattering are widely used in viral ecology to enumerate virus particles in water samples (Turzynski *et al.*, 2021).

TEM was first used for environmental samples, and has been superseded by epifluorescence microscopy (EFM) due to lower costs and preparation time (Turzynski *et al.*, 2021). For EFM, particles are stained with a fluorescent dye to visualise and enumerate by fluorescence microscopy (Budinoff *et al.*, 2011). A similar staining process has been adapted for flow cytometry (FCM) (Brussaard, Marie and Bratbak, 2000) and more recently, nanoparticle tracking analysis (NTA) has been used, in which scattered or fluorescent light from particles is

detected to determine the speed of Brownian motion and calculate the hydrodynamic radius of the particle (Kramberger *et al.*, 2012).

For EFM, virus particles are deposited onto a transparent filter membrane by vacuum pressure, stained using a nucleic acid-binding fluorescent dye, and counted by fluorescence microscopy (Hara, Terauchi and Koike, 1991). Samples were originally stained with 4′,6-diamidino-2-phenylindole (DAPI), but this was later replaced by Yo-Pro-1, SYBR Green, and SYBR Gold dyes (Noble and Fuhrman, 1998). The original 15 nm Nuclepore membrane has been substituted for 20 nm Al_2O_3 filter membranes (Noble and Fuhrman, 1998). The use of 13 mm diameter filters reduced sample volume, and pre-staining the sample before filtration accelerated the process (Budinoff *et al.*, 2011). Similar staining protocols have been developed and optimised for FCM (Brussaard, Marie and Bratbak, 2000). Virus detection is affected by the choice of dilution media, e.g., deionised water, Tris-EDTA or phage culture media (Brussaard, 2004; Ortmann and Suttle, 2009; Budinoff *et al.*, 2011; Holmfeldt *et al.*, 2012; Hoyles *et al.*, 2014), and by fixation of samples (Chen *et al.*, 2001; Wen, Ortmann and Suttle, 2004; Ortmann and Suttle, 2009), although results have been inconsistent and an optimal, standardised protocol has not yet been established.

Despite methodological variations, several studies have reported a good correlation between EFM and TEM measurements, supporting the accuracy of EFM (Noble and Fuhrman, 1998; Chen *et al.*, 2001; Brum and Sullivan, 2015). Still, there are sources of bias in environmental samples, notably extracellular vesicles, and gene transfer agents (virus-like particles produced by bacteria) that contain or bind nucleic acid and appear as virus particles under EFM (Forterre *et al.*, 2013; Soler *et al.*, 2015). Additionally, SYBR dyes form autofluorescent colloids at high concentrations in aqueous solutions and can be mistaken for virus particles in FCM and thus potentially EFM (Dlusskaya *et al.*, 2021). Small viruses with ssDNA or a RNA genomes are detectable in EFM, but not FCM, as these viruses, being smaller with shorter and often single-stranded genomes, bind less dye and are less fluorescent (Tomaru and Nagasaki, 2007; Holmfeldt *et al.*, 2012). Due to large difference in brightness between small and large viruses, small ssDNA and RNA viruses can be missed in environmental samples, although measurements of single viruses is possible (Holmfeldt *et al.*, 2012). As <1% of EVs in environmental samples are detectable by EFM, EVs are unlikely to affect measurements of environmental samples (Biller *et al.*, 2017), although this is less clear for virus stocks, particularly small viruses for which greater sensitivity needed. Thus, quantification of virus stocks through EFM will require inclusion of appropriate controls to control for any host cell-derived products and media components that might appear as virus particles.

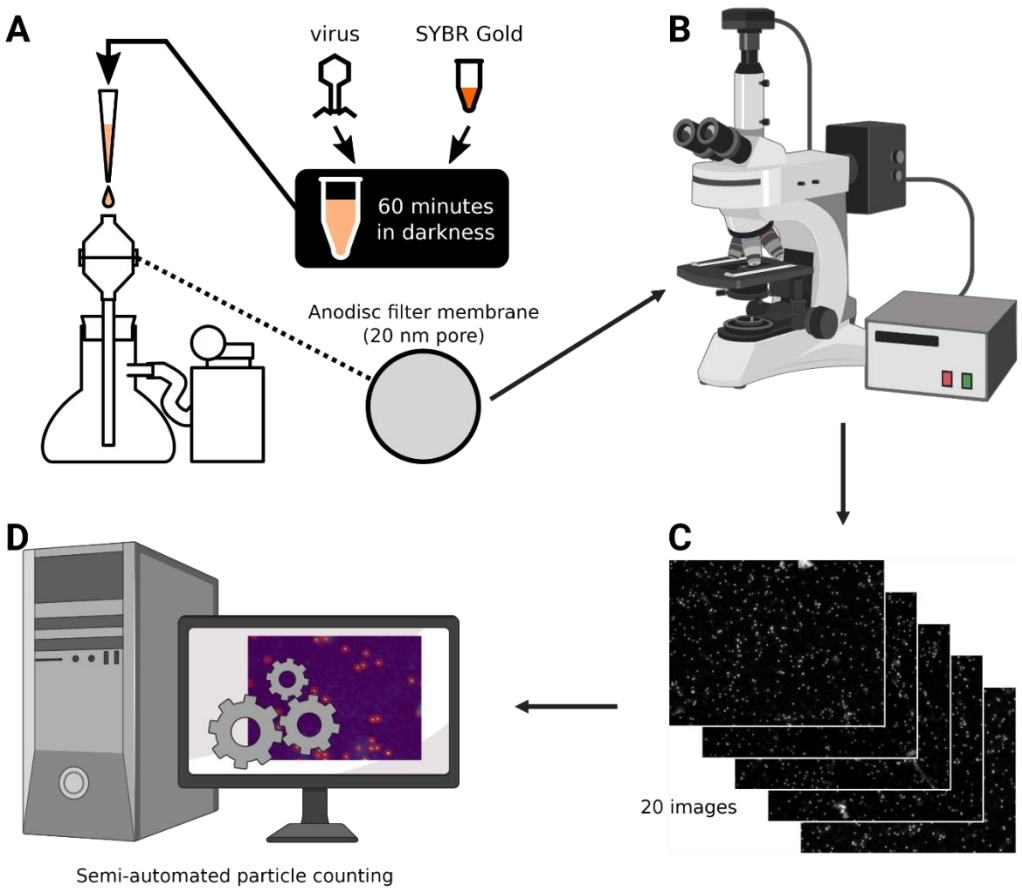

Figure 2.1: Titration of virus stocks. Epifluorescence microscopy was used to titre nucleic acid-containing virus particles in virus stocks. A: To identify nucleic acid, virus stocks were stained using SYBR Gold and incubated for 60 minutes in the dark at 21°C. A 20 nm-pore size Anodisc filter membrane is placed in a Swinnex filter holder and mounted onto a filtration flask connected to a vacuum pump. B: The transparent filter is imaged using a fluorescence microscope. C: Up to 20 views are taken from a single filter disc. D: The images are analysed using a custom Python 3.5 script which loads an instance of ImageJ. Figure created with BioRender.com

Nanoparticle tracking analysis is an alternative to EFM which works by measuring the Brownian motion of a particle to calculate the hydrodynamic diameter to estimate a particle's size (Heider and Metzner, 2014). NTA offers an attractive alternative to EFM due to fast sample preparation and measurement times. NTA has been used to determine phage and eukaryotic virus titres (Anderson *et al.*, 2011; Kramberger *et al.*, 2012). A recent study compared EFM, FCM and NTA and qPCR capabilities on several viruses (Kaletta *et al.*, 2020). NTA was the most accurate measure compared to qPCR. EFM systematically underperformed in this assessment, although a low concentration of SYBR Gold was used and membranes were stained after

Table 2.1: Mock community viruses.

Virus		Host	Genome		Virion	
Name	**Species**	**Species**	**Type**	**Size[n] (bp)**	**Enveloped**	**Size (nm)**
T5[1]	*Tequintavirus T5*	*E. coli*	dsDNA	121,750		90[c], 160[t]
M13	*Inovirus M13*	*E. coli*	ssDNA	6,407		6.5[d], 860[l]
P22[1]	*Lederbergvirus P22*	*S. tyhphimurium*	dsDNA	41,724		60
Det7[1]	*Kuttervirus Det7*	*S. tyhphimurium*	dsDNA	157,498		90[c], 110[t]
Qb[1]	*Qubevirus durum*	*E. coli*	ssRNA	4,215		26
MHV-68	*Rhadinovirus muridgamma4*	Mouse	dsDNA	119,451	Yes	220
BVDV-1[1]	*Pestivirus bovis*	Cow	ssRNA	12,513	Yes	50
RV-A[1]	*Rotavirus A*	Simian	dsRNA	18,550		80

c: capsid diameter, d: filament diameter, env: envelope, l: filament length, n: NCBI reference genome length, t: tail length, 1: (Hulo *et al.*, 2011), 2: (Flynn *et al.*, 2003), 3: (Liu and Zhou, 2007)

filtration, which requires higher concentrations for small viruses (Holmfeldt *et al.*, 2012). Additionally, the EFM measurement were below the PA titre in their study, suggesting a suboptimal methodology.

2.1.2 Specific aims

The main aim of this chapter is to define a reference standard for virus metagenomic sequencing. For this, an accurate and reproducible virus quantification pipeline was developed that is suitable for diverse viruses. The specific aims of this chapter are:

- To optimise a SYBR Gold staining technique using EFM and NTA with the representative large dsDNA phage T5 and the representative small ssRNA phage Qb (Table 2.1).
- To establish a semi-automated image analysis pipeline to enable reproducible particle counting (Fig. 2.1).
- To determine the titres of the ssDNA filamentous phage M13, dsDNA phages T5, Det7, and P22, and eukaryotic dsDNA virus murid gammaherpesvirus-68 (MHV-68), ssRNA virus bovine viral diarrhoea virus (BVDV-1), and the dsRNA virus simian rotavirus SA/11 (RV-A) (Table 2.1) using EFM.

2.2 Materials and methods

2.2.1 Phage stocks

Phage stocks were kindly provided by Dr Evelien Adriaenssens (QIB, Norwich, UK). All LB broth and agar used for phage culture and experiments was supplemented with 10 mM $CaCl_2$. T5 was cultured on *Escherichia coli* strain MG1655, Det7 and P22 on *S. tyhpimurium* strain LT2, Qb on *E. coli* DSM5210, and M13 on *E. coli* strain M13. *E. coli* phage phiX174 (*Sinsheimervirus phiX174*) was grown on *E. coli* strain phiX and used for calculating the effective filter area of Anodisc membranes (see below). Phage working stocks were grown from primary freezer

stocks. Primary freezer stocks were stored at -70°C and working stocks were stored at 4°C. Host strains were streaked onto LB agar plates and grown overnight at 37 °C. A single colony was picked and cultured overnight in 10 ml LB broth in an orbital shaking incubator at 37 °C and 200 RPM. Subcultures were produced by inoculating 10 ml LB broth with 100 µl overnight culture and incubation in an orbital shaking incubator at 37 °C and 200 RPM until an OD_{600} of 0.4 – 0.5 was reached. T5, Det7 and P22 were then grown by the plate culture method. For each phage, 100 µl primary stock was incubated with 200 µl of the respective host subculture for 5 minutes at room temperature (21 °C). The mixture was then added to 5 ml semi-solid (0.8%) agar overlay and poured onto LB agar plates. For each host, a negative control plate was made for which only the host was added to the overlay. After 20 minutes the plates were incubated at 37 °C for 16 – 20 hours incubation. Then, the plates were incubated with 5 ml sterile PBS on an orbital shaker for 1 hour. The overlay was chopped into small pieces and the mix of overlay and PBS was transferred into a 15 ml Falcon tube. The tube was centrifuged at 3,200 × *g* for 10 minutes at 21°C and the supernatant was filtered through a 0.22 µm PES syringe filter to sterilise the phage stock.

For phage M13 and Qb, stock was produced by the liquid culture method as follows. Host strain subculture at an OD_{600} of 0.4 – 0.5 was inoculated with 100 µl primary phage stock and incubated overnight in an orbital shaking incubator at 37 °C and 200 RPM. After 16 – 20 hours incubation, the lysate was centrifuged at 3,200 × *g* for 10 minutes and the supernatant was filtered through a 0.22 µm PES syringe filter to sterilise the stock. Stocks were stored at 4 °C.

2.2.2 Animal virus stocks

RV-A strain Simian rotavirus SA11 was kindly provided by Dr Edward Mee (National Institute for Biological Standards and Control, Potters Bar, UK). BVDV-1 strain Bovine viral diarrhoea virus Ky1203nc is a field isolate of BVDV-1 that was kindly provided by Dr J Brownlie (Royal Veterinary College, London, UK) (Howard, Brownlie and Clarke, 1987). MHV-68 strain Murid herpesvirus-68 was kindly provided by Dr James Stewart (University of Liverpool, Liverpool, UK).

2.2.3 Plaque assays

Phage PFU titres were determined by PA. Host strain subcultures were produced from 100 µl of overnight culture in 10 ml LB broth supplemented with 10 mM $CaCl_2$ in an orbital shaking incubator at 37 °C and 200 RPM, until an OD_{600} of 0.4 – 0.5 was reached. Phages were serially diluted 10-fold in PBS and 100 µl of each was dilution and 200 µl host subculture were added to 5 ml semi-solid (0.8%) agar, mixed and poured onto LB agar plates. Plates were incubated overnight at 37°C and plaques were counted. A minimum of three plates was used to calculate the final PFU titre of each stock.

2.2.4 Nanoparticle tracking analysis

A ZetaView PMX120 (Particle Metrix, Ammersee, Germany) nanoparticle tracking analysis device fitted with a 488 nm laser was used to measure fluorescently labelled T5 phage stock. The ZetaView can detect particles as small as 25 – 40 nm, depending on the particles refractive index, up to 1 µm in size. Scattered and fluorescent light from the 488 nm laser is detected by a CMOS camera through a 10× objective, and a 500 nm cut-off filter can be used to allow only fluorescent light to pass through. Samples were loaded manually into the ZetaView measurement cell using a 1 ml syringe. Before each experiment, the ZetaView was calibrated using 5 ml of 100 nm polystyrene calibration beads as supplied by the manufacturer, diluted 1:250,000 in deionised water. To find a dilution factor that results in the optimal number of 50 – 200 particles per field of view, samples were diluted 1:5,000, 1:10,000, and 1:20,000 and then analysed. Samples were then loaded into the ZetaView and two rounds of measurements were taken. The first round was taken with the light filter in place to observe fluorescent particles and the second round was taken without light filter to observe both fluorescence and scattering. During each round of measurement, two cycles were recorded at all 11 positions of the measurement cell at 30 frames per second. Between samples, the measurement cell was flushed three times with 10 ml deionised water. For samples diluted in PBS, PBS was used to flush the system before loading and between samples.

2.2.5 Epifluorescence microscopy

To determine virus stock titres by epifluorescence microscopy a protocol adapted from Holmfeldt *et al.* and Hoyles *et al.* was used. Briefly, a 13 mm Al_2O_3 Anodisc 0.02 µm filter membrane (Whatman, ref. 6809-7003) was placed in a Swinnex filter holder (Millipore, Darmstadt, Germany, ref. SX0001300) fitted onto a glass tube protruding through a rubber stopper into a Büchner flask (Fig. 2.1). The flask was connected to a Millivac Maxi vacuum pump (Millipore, Darmstadt, Germany, ref. SD1P014M04). After the pump was switched on, 400 µl deionised water or PBS was added to the filter holder to confirm the inlet was sealed. The stained sample was added gradually using a 1 ml micropipette. The filter was then washed with 400 µl of deionised water or PBS and any remaining liquid was aspirated for one minute. The filter was then placed sample-side-up onto Whatman filter paper (Whatman, Marlborough, MA, USA, ref. 1004-055) to dry for 5 minutes. A 7.5 µl drop of Fluoromount G (Invitrogen, Waltham, MA, USA, ref. 00-4958-02) was placed on a glass microscopy slide (VWR, Leicestershire, UK, ref. 631-0117). The filter was placed on top, a 7.5 µl drop of Fluoromount G was added onto the filter and covered with a glass cover slip (VWR, ref. 631-0125). The slide was then placed in the dark at 21 °C for at least 2 hours to allow the mountant to set.

The slides were imaged on an Axio Imager.M2 upright fluorescence microscope (Zeiss, Cambourne, UK). Samples were illuminated using a HAL 100 illuminator with a quartz collector

(Zeiss, Cambourne, UK, ref. 423000-9901-000) and a 65HE Alexa 488 filter. For titre measurements, a 100X EC Plan-Neofluar oil immersion objective (Zeiss, Cambourne, UK, ref. 420496-990-000) was used. Images were captured on an ICX 285 CCD monochrome camera (Sony, Surrey, UK). Between 15 and 20 images were taken of each filter at random locations.

2.2.6 Effective filter area calculation

The outer rim of the filter disc was covered by the filter holder gasket, which reduced the effective area of the filter disc. The effective area was measured using a 1:10 diluted sample of phage phiX174 stock, which stained very brightly. Six images were taken of the outer rim of the filter disc using a 10X EC Plan-Neofluar objective at several points the along edge of the filter. In ImageJ 1.52p, the distance from the filter edge to the sample was measured by drawing a perpendicular line from the filter edge to the sample edge. The average distance was 0.8 mm. By subtracting this from the filter radius (6.5 mm), the effective filter area was calculated to be 102.1 mm^2:

$$A_{effective} = \pi\,(6.5 - 0.8)^2 = 102.1\ mm^2$$

2.2.7 Staining time course experiment

For the staining time course, T5 stock was diluted 1:250 in deionised water and stained in 2.5X and 25X SYBR Gold (Thermo Fisher Scientific, Loughborough, UK, ref. S11494) for 5, 15, 30 and 60 minutes at 21°C. A blank consisting of only deionised water was stained in parallel. Samples were further diluted in deionised water to a final dilution of 1:5,000, loaded into the ZetaView and analysed as described above.

2.2.8 Stock dilution experiment

To determine the effect of stock dilution on staining efficiency, phage T5 stock was diluted 1:100, 1:300 and 1:1,000 in deionised water and 1:1,000 in PBS and incubated with 2.5X SYBR gold for 30 minutes at 21°C. The sample was then further diluted to a final dilution of the stock of 1:5,000 and analysed in the ZetaView as detailed in above. Another 1:5,000 dilution of the same sample was then filtered through an Anodisc filter membrane for EFM imaging. The filter disc was placed on the filter holder outlet, 400 µl of deionised water was added and the pump was turned on. Then, the inlet was mounted, and 1 ml of deionised water was added to confirm a good seal. The sample was then added to the inlet chamber and filtered. The filter was washed with 1 ml deionised water, remaining liquid was aspirated for an additional minute, and then filter was then dried on Whatman filter paper for 2 minutes. A 20 µl drop of Fluoromount G was then added to a microscopy slide, the filter was placed on top and another 20 µl drop of Fluoromount G was added. The filter was then covered with a microscopy slide and left at 21 °C in the dark to harden overnight. Due to a pipetting error, data for dilutions 1:300 and 1:1,000 in deionised water for the first replicate were excluded.

The filter membranes were imaged as described above. Images were analysed in ImageJ version 2.0.0-rc69/1.53c in batch mode using an ImageJ macro. For each image, the background was subtracted using a rolling ball size of 50 pixels. Image contrast was then enhanced by normalised to saturation of 0.1% of all pixels. Local maxima were then detected with a prominence threshold of 20,000 and the resulting points were saved to a text file. The particle locations were then marked on the image and the image was saved for later inspection.

2.2.9 Fixation and staining media optimisation

To determine the effect of buffers and cell culture media on stock EFM, Tris-EDTA (TE) buffer (Sigma, Darmstadt, Germany, ref. 93283-100ML) and Opti-MEM without phenol red (Gibco, Waltham, MA, USA, ref. 11058021) were used to dilute T5 and Qb stocks. Additionally, a fixation method using glutaraldehyde was adapted from Ortmann and Suttle (2009) and Budinoff et al. (2011), combining SYBR Gold pre-staining with fixation, to determine its effect on particle staining. Phage T5 stock was diluted 1:1000, while phage Qb stock was diluted 1:4000 in deionised water, TE buffer and Opti-MEM with 25X SYBR Gold and with or without 0.5% glutaraldehyde. Samples were incubated on ice for 1 hour and then processed as described in the section "Epifluorescence microscopy". Images of T5 were processed using the same macro as described above. For Phage Qb, images were processed in batch mode in ImageJ using a slightly adapted macro, in which a gaussian blur filter with σ = 2 was applied after contrast enhancement and before detection of local maxima to reduce noise.

2.2.10 Fixation and staining of eukaryotic viruses

BVDV-1, MHV-68 and RV-A stocks were resuspended using a 23G 32mm hypodermic needle (Terumo, Surrey, UK, ref. TUAN-2332R) and a 1 ml syringe to prevent particle clumping. Stocks were diluted 1:2 and stained with a final concentration for 25X SYBR Gold and 0.5% glutaraldehyde for 1 hour on ice in the dark. Samples were then processed in a microbiological safety cabinet as described above.

2.2.11 EFM-based virus stock titres

For the final stock titre measurements of the phages, stock titres were first determined by plaque assay. Stocks were then diluted to approximately 10^7 PFU/ml and stained with 25X SYBR Gold for 1 hour at 21°C in the dark. Four replicates were performed for T5 and Det7, five for P22 and three for M13. Negative controls and a blank were included for all phages. For BVDV-1, MHV-68 and RV-A, no plaque assay titres were available, and stocks were diluted 10-fold three times in PBS and each dilution was incubated with 25X SYBR Gold for 1 hour on ice. In addition to SYBR Gold, BVDV-1 was also incubated with 0.5% glutaraldehyde. For eukaryotic viruses, only blank samples were included, consisting of DMEM and MEM cell culture media.

2.2.12 Image analysis for final titre measurements

For increased automation, consistency, and ease of data analysis, an image analysis library and workflow were developed in Python 3.8.10 using Jupyter Notebook 6.4.5. First, sample metadata was entered into a cell in the workflow Notebook. Then, using the pyimagej 1.0.2 package, an instance of ImageJ version 2.1.0/1.53c was started. ImageJ scripts were then applied to prepare the images for particle detection. First, the background was subtracted using the "Subtract background…" function with a rolling ball radius of 50 pixels. The resulting image was saved and then displayed in the Notebook to compare the raw image. In the ImageJ instance, the contrast of the background-subtracted image was then enhanced by normalising to saturation of 0.3% of pixels. The resulting images were saved and displayed in the Notebook. To determine the appropriate threshold for particle detection, the "Find maxima…" function was then applied to all enhanced images with thresholds from 100 to 60,000 with a step of 500. The number of detected particles in each image at each threshold was then saved and the average number was plotted in the Notebook with 95%-confidence bands. This produced an inverse sigmoidal curve, with a sharply decreasing number of detections at a low threshold that plateaued over a large threshold range and then decreased sharply again. The threshold was chosen close to the beginning of the plateau. Five images were then randomly chosen and detections were marked for a small range of thresholds around the chosen threshold to fine-tune the threshold. Then, the "Find maxima…" function was applied to each of the images at the chosen threshold, the detected particles are marked on the images, saved, and displayed in the Notebook. The particle coordinates in the images are saved and loaded into the Notebook and the particle coordinates are then used to determine the particle intensities from the background-subtracted images.

For increased consistency, images of each sample and the corresponding negative control were then normalised to a fixed maximum value to enhance contrast. The maximum intensity was chosen as the mean particle intensity of the image with the lowest mean particle intensity detected in the previous step. Each background-subtracted image was then normalised using the NumPy 1.21.4 package and saved using the Pillow 8.4.0 package. The fixed-value normalised images were displayed in the Notebook and the above process repeated to determine the appropriate threshold for particle detection, detect particles, save marked images, and save and plot particle statistics. Particles with an intensity more than 1.5X inter-quartile range (IQR) above the 75-precentile were marked as outliers and excluded from subsequent titre calculation. Finally, all the particle data and sample metadata was saved for each sample and the data further processed in R 4.1.2 using the Tidyverse package version 2.0.0.

2.2.13 Data analysis

EFM and NTA sample data files were analysed using RStudio 2022.12.0 running R version 4.1.2. tidyverse package 1.3.1 was used for analysis and plotting of the data. For the EFM data files, a stock titre estimate was calculated for each image of a filter. The view size of the camera with a 100X objective was 89.53 μm by 67.08 μm. The surface area thus was 1/16,996 of the effective filter area. For each image, the diluted sample titre was then calculated by multiplying the particle count by 16,996 and dividing by the sample volume. The sample titre was then calculated by taking the average of the diluted sample titres calculated for each of the images and the stock titre calculated by multiplying by the dilution factor of the stock.

2.3 Results

Due to their relatively low fluorescence compared to tailed phages like T5, observation of ssDNA and RNA viruses through EFM is more sensitive to staining parameters like buffer and dye (Tomaru and Nagasaki, 2007; Holmfeldt *et al.*, 2012). To measure a wide variety of viruses, optimal staining conditions were investigated in this chapter to maximise particle brightness. Initial experiments were conducted using NTA which is quicker compared to EFM. These initial measurements were done using representative large and small viruses, T5 and Qb respectively.

2.3.1 Incubation time does not improve particle fluorescence of T5

T5, incubated at various timepoints and SYBR Gold concentrations, was analysed using NTA. In the scattered light measurement, a peak was visible in the 50 – 200 nm range (Fig. 2.2A). Since some *Caudoviricetes* phages, including phage T5, have long tails, their hydrodynamic sphere might not accurately reflect their physical size and the NTA-estimated particle size might therefore not be accurate. The capsid of phage T5 is 90 nm in diameter and the tail is 160 nm in length (Hulo *et al.*, 2011), so particles 50 – 200 nm in diameter were assumed to be virus particles. For T5 stained in 2.5X SYBR Gold, no fluorescent peak was visible above background level up to 90 minutes incubation (Fig. 2.2A). Incubation with 25X SYBR Gold produced a fluorescent peak around 120 nm around half the height of the peak of the scattered light measurement, suggesting around half of the particles are detectable by fluorescence.

The number of fluorescent particles in the selected size range did not increase over time for either concentration. The total average number of particles detected increased when 25X SYBR Gold was used compared to 2.5X SYBR Gold. However, with an average of 4.04 (standard deviation (SD)=0.56, n=5) × 10^{10} and 7.84 (SD=0.69, n=5) × 10^{10} fluorescent particles/ml for 2.5X and 25X SYBR Gold, respectively, the average fraction of fluorescent particles remained the same between both concentrations (Fig. 2.2B). The highest ratio was measured after 90 minutes and 60 minutes for 2.5X and 25X SYBR Gold, respectively (Fig. 2.2C). The average ratio

of fluorescent particles over all time points was 0.46 (±0.05 SD, *n*=5) and 0.44 (±0.08 SD, *n*=5) for 2.5X and 25X SYBR

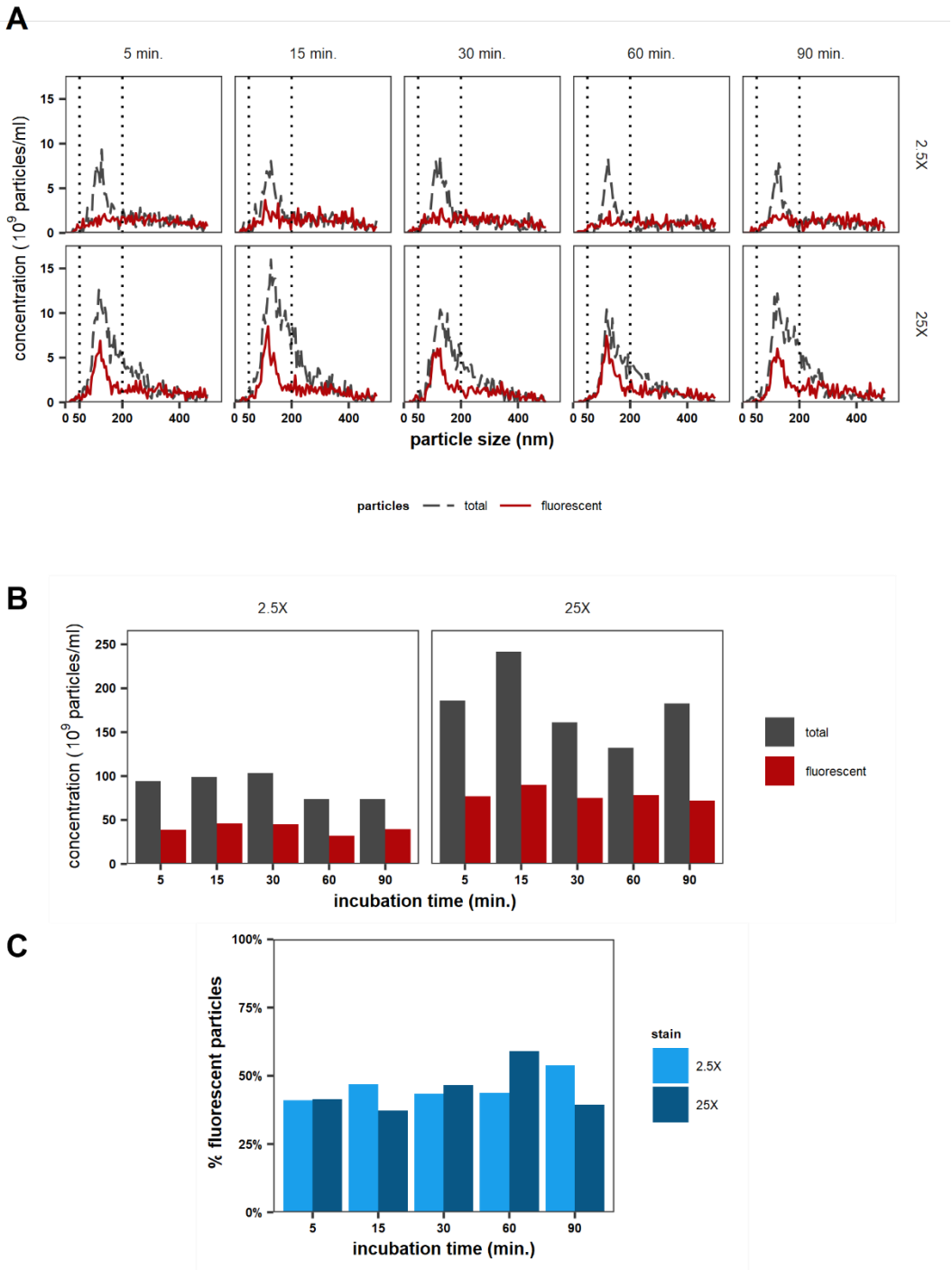

Figure 2.2: Staining time of phage T5 shows no effect of incubation time on particle detection. Phage T5 stock was stained in 2.5X and 25X SYBR Gold and analysed by EFM and nanoparticle tracking analysis (*n*=1) **A:** Size distribution of all (grey) and fluorescently labelled (red) particles. **B:** Number of particles (grey) and fluorescent particles (red) detected in the size range of phage T5. **C:** Proportion of fluorescently labelled particles in the size range of phage T5.

Gold, respectively. Together, this suggests incubation time does not affect the number or fraction of fluorescent particles, while 25X SYBR Gold produces the highest number of detected fluorescent particles. Qb was also tested but was not visible by NTA and was omitted from subsequent experiments involving NTA.

2.3.2 Staining of diluted phage stock increases particle fluorescence

Since particle fluorescence was independent of the incubation time, increasing the amount of dye per virus particle was tested next. Increasing the concentration of SYBR Gold above 25X affected NTA measurements, possibly due to colloid formation (Dlusskaya *et al.*, 2021). Instead, dilutions of T5 stock were tested to increase labelling of particles.

NTA revealed a peak of fluorescent particles with a hydrodynamic diameter of between 50 and 200 nm for all dilutions in deionised water, while no peak was visible for dilutions made in PBS (Fig. 2.3A). The fraction of fluorescent particles increased with higher dilutions. Also, the fluorescent fraction was highest for PBS although no peak was visible in the expected phage size range, which suggests the signal is background only. Additionally, the number of fluorescent particles was double the number detected by scattering, which suggests measurements in PBS were inaccurate. In deionised water, the number fluorescent particles peaked in the 50 – 200 nm size range at every dilution. The highest fraction of fluorescent particles was obtained for the 1:1,000 dilution in deionised water, with a mean of 0.79 (±0.22 SD, n=2) (Fig. 2.3B).

Particle intensity and concentration of the same phage stock dilutions were assessed by EFM. One replicate of the 1:100 dilution in deionised water and one of the 1:1,000 dilution in PBS failed due to broken filter membranes. All samples were imaged using the same laser intensity and exposure settings with particle intensity measured for the 1:1,000 dilution in deionised water (Fig. 2.3C). Concentration measurements of dilutions made in deionised water decreased from 5.25×10^{10} to 4.74×10^{10} particles/ml from 1:100 to 1:1,000, respectively (Fig. 2.3D), representing a 2.3 and 2.0-fold increase, respectively, compared to the PFU measured by plaque assay. NTA concentration measurements of the same dilutions increased with higher dilution and were higher than the EFM measurement; from 6.61×10^{10} to 1.12×10^{11} particles/ml for 1:100 and 1:1,000 dilutions, respectively. NTA requires a narrow range of 20 – 100 particles per view, and all samples were diluted to a final dilution of the stock to 1:5,000. This meant that for the stocks stained at 1:100 and 1:1,000 required a 50- and 5-fold dilution, respectively. Since each sample contained the same concentration of SYBR Gold, the final concentration of SYBR Gold was 10-fold higher for the 1:1,000 sample than the 1:100 sample. Therefore, the higher particle counts in the 1:1,000 sample might, at least in part, be due to increased SYBR Gold colloids and background noise. The 1:1,000 dilution in PBS resulted in

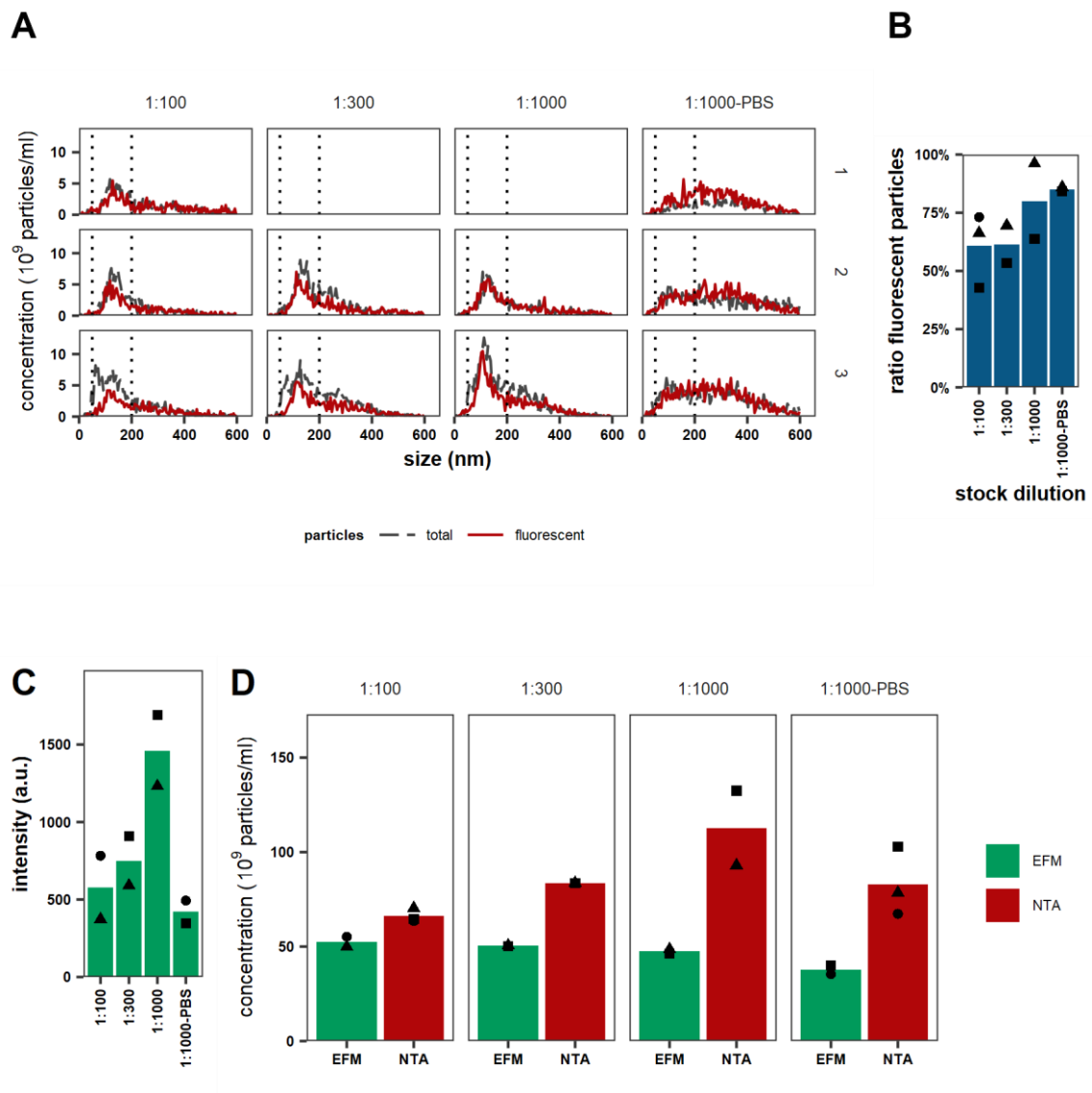

Figure 2.3: Serial dilution of phage stocks shows increased dilution enhances particle intensity and detection. Phage stock was diluted 1:100, 1:300 and 1:1000 in distilled water, and 1:1000 in PBS and stained with 25X SYBR Gold for 1h. (*n*=3) **A:** Size distribution of all particles (grey) and fluorescent particles (red) **B:** Ratio of fluorescent particles in the size range for phage T5, as measured by nanoparticle tracking analysis. Symbols mark ratios of matching replicates; the bars show the average of the replicates. **C:** Intensity of particles as measured by EFM, measured by the average value of the brightest pixel of all particles in all images from an Anodisc filter membrane. Symbols mark the average intensity of matching replicates; the bars show the average of the replicates. **D:** Stock titres as estimated by nanoparticle tracking analysis and EFM. Symbols mark the titre estimate of matching replicates; the bars show the average of the replicates.

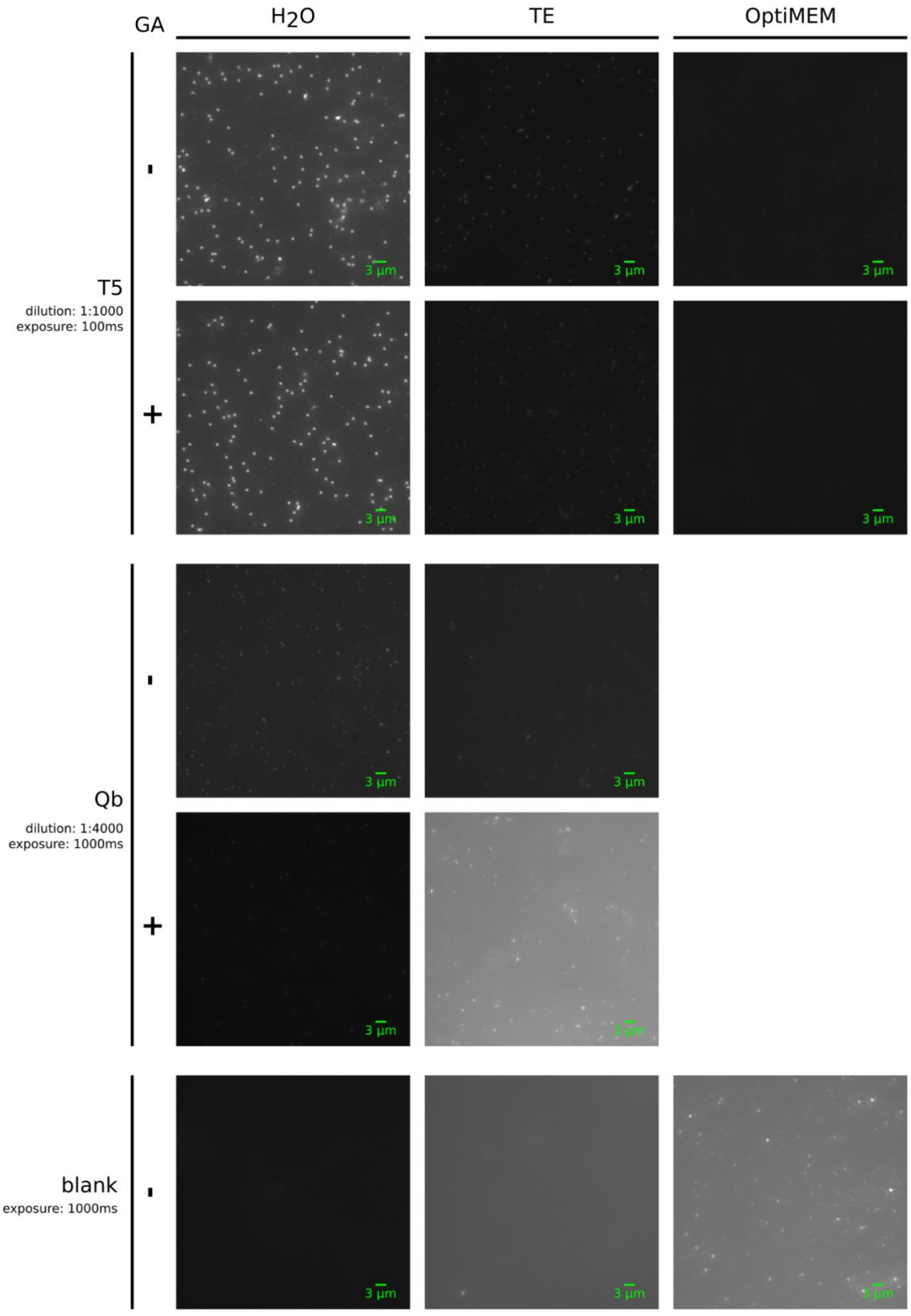

Figure 2.4: Comparison of phages T5 and Qb incubated with and without glutaraldehyde in various media. Phages T5 and Qb were diluted in H_2O, TE buffer and OptiMEM with and

without 0.5% glutaraldehyde (GA) and stained with 25X SYBR Gold. Blanks were medium only with 25X SYBR Gold. Dilution refers to the stock dilution of each phage, exposure to the microscope exposure time. Images are normalised to enhance contrast, with all images of phage T5, and all images of phage Qb and blanks, normalised to the same values.

lower intensity in EFM images, and the 1:1000 dilution in deionised water was higher than the same dilution in PBS measured both by EFM and NTA, strongly suggesting that deionised water is a better staining medium than PBS.

2.3.3 Deionised water is the optimal staining medium

Some staining protocols for EFM and FCM use TE buffer and fixatives (Brussaard, 2004; Ortmann and Suttle, 2009; Budinoff *et al.*, 2011), and eukaryotic virus stocks are produced in cell culture media. To test the effects of these reagents on particle staining for EFM, T5 and Qb were stained in deionised water, TE buffer and cell culture medium OptiMEM (n=3), with and without 0.5% glutaraldehyde (GA) (n=1). Staining with TE buffer and OptiMEM reduced the image brightness for T5 compared to deionised water, while GA had no effect (Fig. 2.4). For Qb, TE buffer lowered overall image intensity compared to deionised water (Fig. 2.4). OptiMEM results are excluded for Qb, as at 1,000 ms exposure, fluorescent particles were also visible in OptiMEM blank control samples (Fig. 2.4). These particles were also apparent in non-stained OptiMEM blanks, but not 20 nm filtered OptiMEM, suggesting that OptiMEM contains fluorescent particles. Correspondence with the manufacturer identified vitamin B as a potential source of autofluorescence, although the proprietary recipe of the medium prohibited identification of specific vitamin B compounds.

Particle counts for both T5 and Qb were higher in deionised water compared to TE buffer (Fig. 2.5A). Particle intensity of T5 was highest in deionised water, lower in TE buffer and lowest in OptiMEM (Fig. 2.5B). Fixing with GA had little effect on T5 in deionised water but had reduced intensity in TE buffer and OptiMEM (Fig. 2.5B). For Qb, particle intensity was higher in deionised water and GA had reduced intensity in deionised water whereas TE buffer had no impact (Fig. 2.5B). Comparing the particle count for fixed and non-fixed samples showed an increase for T5 in all three media for the fixed samples, while Qb saw a reduction in deionised water with an increase in TE buffer, although the size particle count was much lower in TE buffer than in deionised water.

2.3.4 Fixation improves staining of BVDV-1 but not MHV-68 and RV-A

Fixation was evaluated for eukaryotic viruses BVDV-1, MHV-68, and RV-A (Fig. 2.6). While incubation in GA increased the image brightness in BVDV-1 compared to the non-fixed sample,

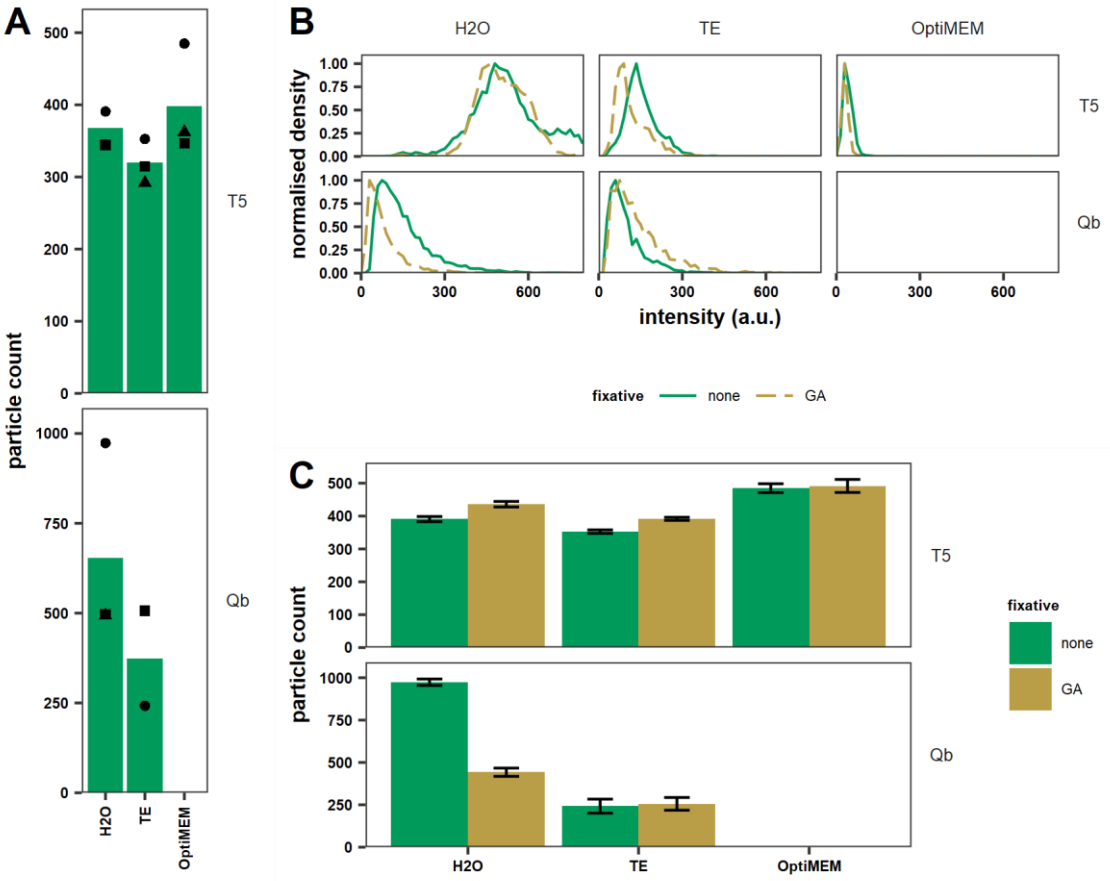

Figure 2.5: Comparison of diluents and fixative for the optimisation of phage staining. Phages T5 and Qb were diluted in distilled water, TE buffer and OptiMEM with or without 0.5% glutaraldehyde (GA). **A:** Phages T5 and Qb were diluted and stained in triplicate in H_2O, TE buffer or OptiMEM and measured by EFM (*n*=3). Symbols mark the average particle count over all images of a single replicate; bars depict the average number of particles detected across replicates. **B:** Particle intensity distribution of phages T5 and Qb in different media with (yellow) and without (green) 0.5% glutaraldehyde. Particle intensity is the value of the brightest pixel of each detected particle. The graphs depict the normalised density distribution of particle intensity (*n*=3). **C:** Particle count of phages T5 and Qb in different media with (yellow) and without (green) 0.5% glutaraldehyde. Count refers to the average number of particles detected in the images of a single filter membrane. Error bars show the standard error of the mean count over all images (*n*=1).

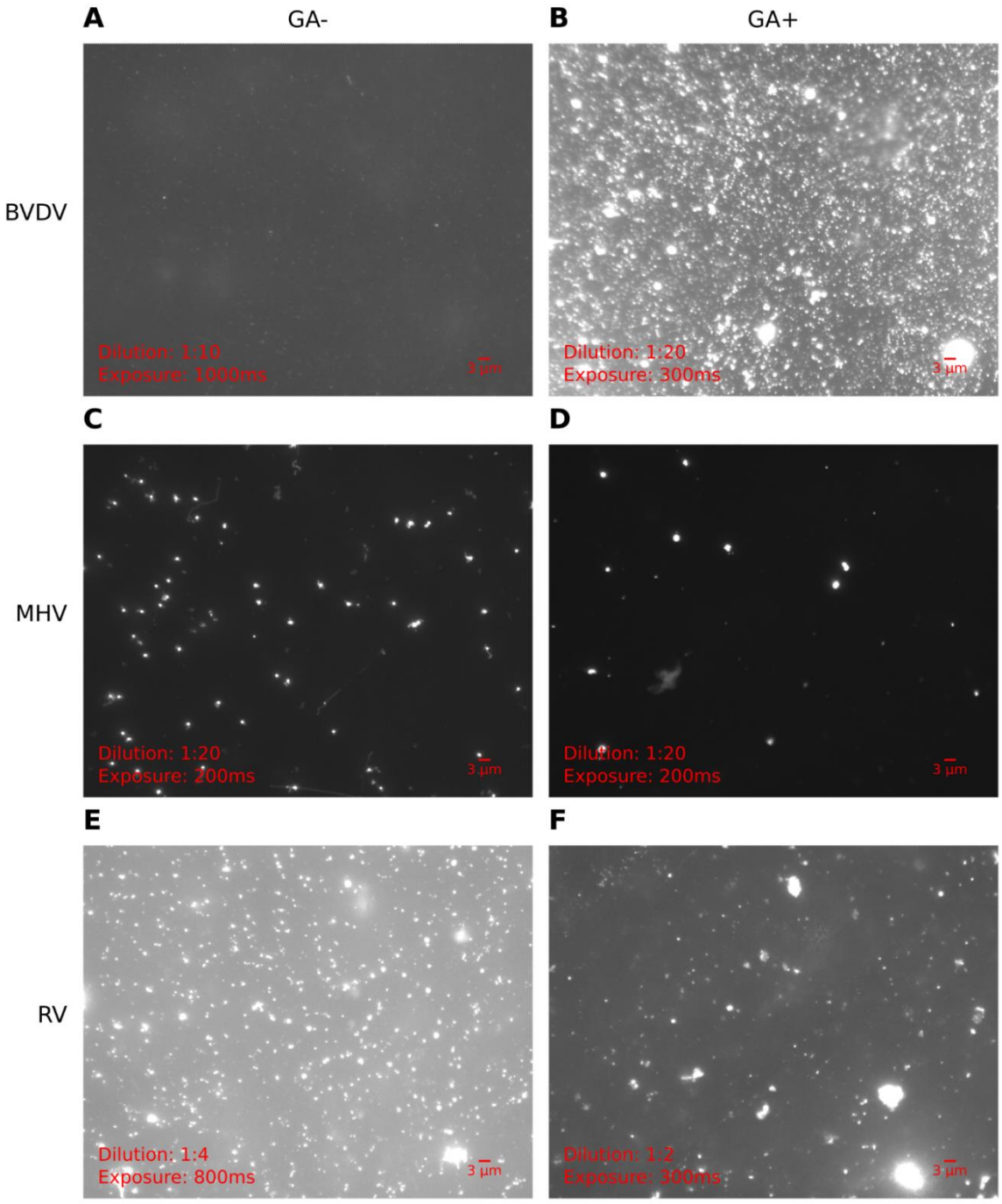

Figure 2.6: Impact of glutaraldehyde on eukaryotic virus particle staining. Dilutions in PBS of virus stocks of BVDV-1, MHV-68 and RV-A in PBS are incubated in 25X SYBR Gold with (**A**, **C**, **E** respectively) and without (**B**, **D**, **F**, respectively) 0.5% glutaraldehyde (GA) for 1 hour. Dilutions referred to in the images represent the stock dilution used for staining. Scale bar = 3 µm.

it had the opposite effect for MHV-68 and RV-A. For subsequent EFM experiments, GA was therefore only applied to BVDV-1, while all other viruses were stained without GA.

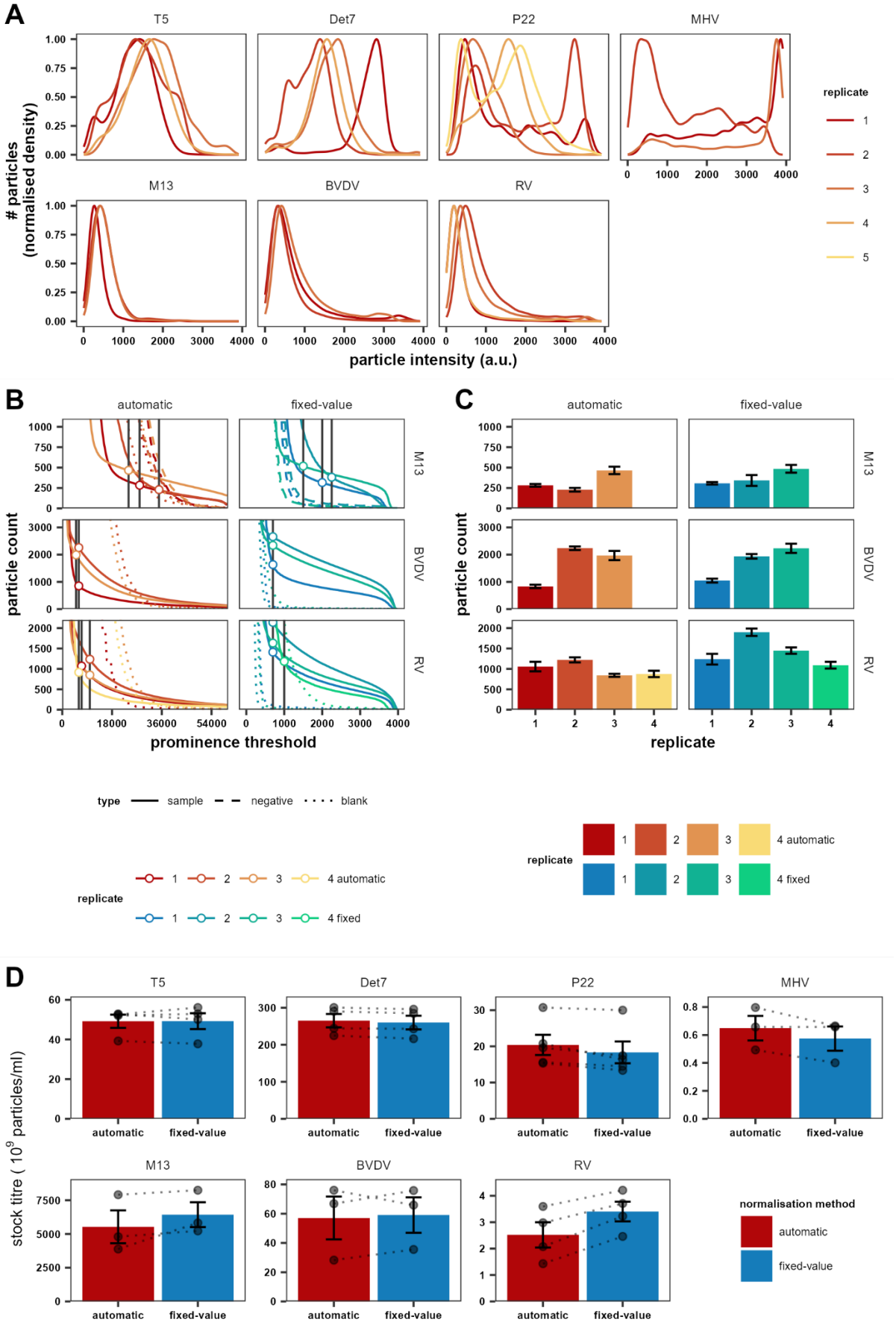
A
T5
Det7
P22
MHV
M13
BVDV
RV
particles (normalised density)
particle intensity (a.u.)
replicate
1
2
3
4
5
B
automatic
fixed-value
particle count
prominence threshold
C
replicate
type sample negative blank
replicate 1 2 3 4 automatic
replicate 1 2 3 4 fixed
D
stock titre (10^9 particles/ml)
automatic
fixed-value
normalisation method

Figure 2.7: Use of consistent normalisation thresholds enhances particle detection for non-dsDNA viruses. Using the ImageJ software, images of the same filter membrane are first normalised to saturation of 0.3% of pixels before particles are located by the "Find maxima" function. Maxima are then detected in the image to locate virus particles, and the brightness value of the Background-subtracted image is taken at each position. The 75-percentile brightness is then taken and used to apply a fixed-value normalisation on all images of the same set, by normalising the Background-subtracted images with the maximum set to that value. **A**: Gaussian smoothed density estimates of the intensity distributions of virus particles across replicates as detected in automatically normalised images. The x-axis depicts the entire dynamic range of the microscope camera. **B**: the number of detections as a function of particle intensity prominence, which is the particle intensity compared to the background in automatically normalised (auto) and fixed-value normalised (fixed-value) images. Vertical lines represent the chosen threshold and the points emphasize the particle count at the corresponding threshold. Lines show the particles detected in virus stock samples (solid lines), negative controls (dashed lines, M13 only), and blanks (dotted lines). **C**: distribution of stock titre estimates from images of replicate EFM measurements in automatically and fixed-value normalised images. Error bars represent the standard error of the mean of the particle count of images in a single EFM measurement. Colours of the graphs in **B** and bars in **C** correspond to the same replicate. **D**: Titre calculation of virus stocks, based on the means of at least three EFM measurements. Points show the estimates from individual replicates, corresponding replicates are connected by a dotted line, and error bars depict standard error of the mean.

2.3.5 Virus detection is limited by particle brightness

Following the staining optimisation experiments, seven viruses were selected for use in an MC for viral metagenomics (Chapter 3). Four dsDNA viruses were selected. The phages T5, Det7 and P22 chosen as representatives of the tailed phage families *Siphvoridae*, *Myoviridae*, and *Podoviridae* of the order *Caudovirales*. Recently, tailed phage taxonomy has been reorganised, with the order *Caudovirales* and its families being abolished. The phages T5 and Det7 are now classified into the *Demercviridae* and *Ackermannviridae* families, respectively, while P22 has not been assigned to a family. Herpesvirus MHV-68 was included as the fourth dsDNA virus with the ssDNA phage M13, ssRNA virus BVDV-1 and dsRNA virus RV-A also included.

Particle brightness was not only influenced by the staining process but also exposure time. For T5, Det7, P22, MHV-68 exposure times of 30 – 500 ms were sufficient to illuminate particles in the middle of the dynamic range of the camera. In the case of T5 and Det7, the particle

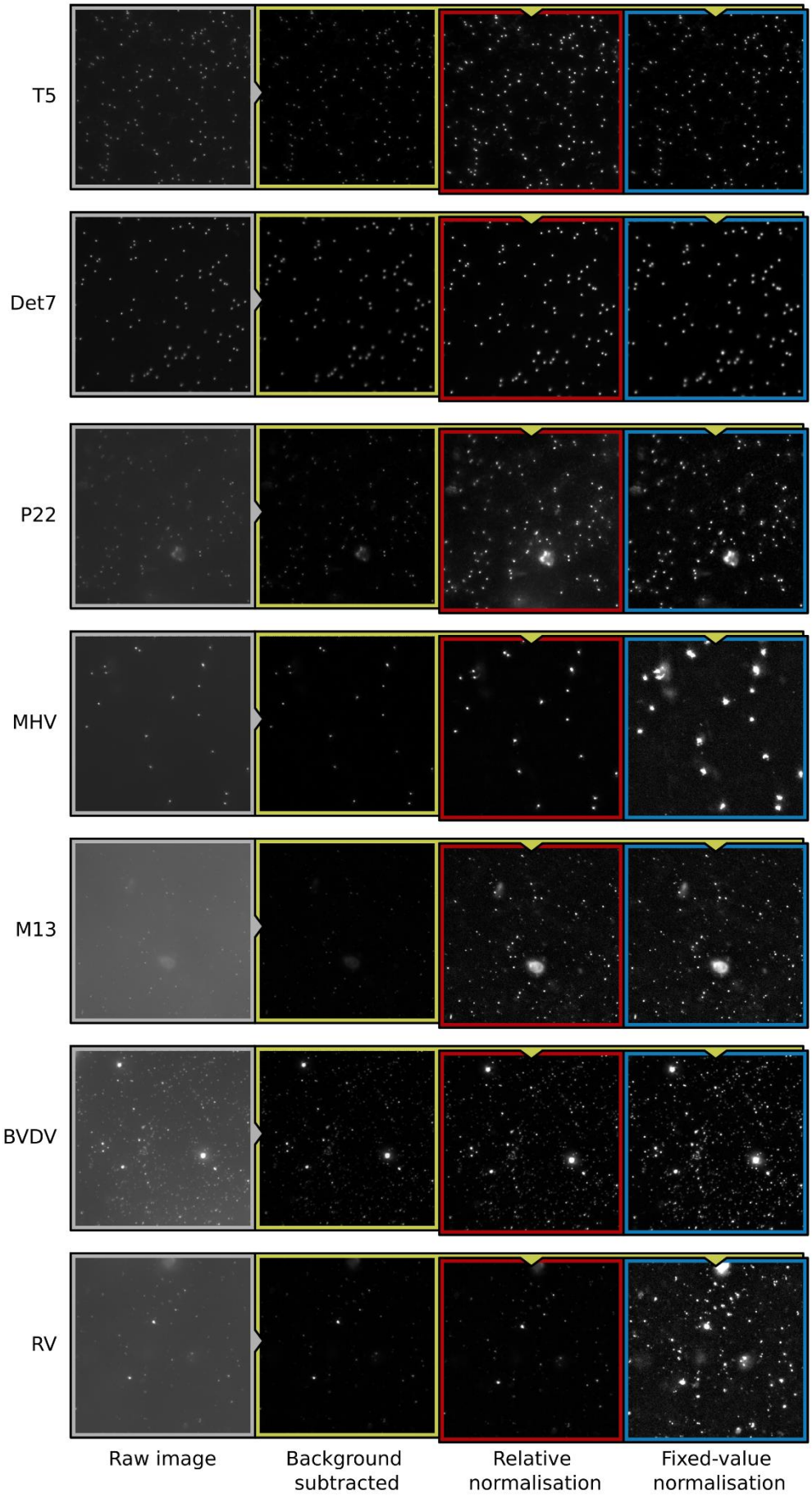
T5
Det7
P22
MHV
M13
BVDV
RV
Raw image
Background subtracted
Relative normalisation
Fixed-value normalisation

Figure 2.8: Relative and fixed-value contrast enhancement of mock community viruses. After subtraction of the background (yellow-outlined images) from the raw images (grey-outlined images), relative normalisation is applied to the images using the "Enhance contrast" function in ImageJ. Using this function, images are normalised to saturation of 0.3% of pixels in each individual image (red-outlined images). Maxima are then detected in the image to locate virus particles, and the brightness value of the Background-subtracted image is taken at each position. The 75-percentile brightness is then taken and used to apply a fixed-value normalisation on all images of the same set, by normalising the Background-subtracted images with the maximum set to that value (blue-outlined images).

brightness distribution resembled a symmetrical bell curve (Fig. 2.7A). For P22, multiple peaks were visible in several replicates. The intensity distribution of MHV-68 peaked close to the maximum brightness for two of the replicates, whereas the third peaked at the low end of the dynamic range of the camera, suggesting issues with the staining process for this replicate. For M13, BVDV-1, RV-A longer exposure times of >500 ms were required. The brightness distribution of these samples resembled a skewed bell curve at the lower end of the dynamic range of the camera suggesting that a fraction of the particles fell below the limit of detection, and the method likely underestimates the actual particle count. However, increasing exposure times increased the brightness of the background and so to prevent bleaching exposure time did not exceed 2,000 ms. Phage Qb was omitted due to difficulties in quantifying by EFM caused by a high level of background noise in the control sample.

2.3.6 Particle detection of non-dsDNA by fixed value normalisation

Manual counting of particles is time consuming and prone to inconsistencies. To speed up the counting process and increase reproducibility, image analysis functions of the image analysis software ImageJ were incorporated into a Python script. To automatically count virus particles in microscopy images, a Jupyter Notebook (Kluyver *et al.*, 2016) was developed in Python that uses the PyImageJ module (Rueden *et al.*, 2021) to interact with ImageJ (Schneider, Rasband and Eliceiri, 2012). Integration of ImageJ with Python facilitates the use of existing ImageJ macros in batch mode to reduce image background, enhance contrast, and detect particles (Fig. 2.8), and allows the automatic display of plots and statistics to evaluate sample quality, greatly streamlining image processing.

For all viruses, a blank was included, consisting of deionised water or PBS, and imaged at the same exposure settings as the virus. For the phages, a negative control consisting of a host culture subjected to the same extraction procedure as the respective phage cultures was also included. For the dsDNA viruses T5, Det7, P22 and MHV-68, blanks and negative controls did not contain any detectable virus-like particles. For the non-dsDNA viruses the automated

contrast enhancement function led to inaccuracies. Due to their lower intrinsic particle brightness, particle detection was more sensitive to differences between the overall intensities of some of the images leading to differences in normalisation values and an inconsistent particle intensity threshold. In addition, longer exposure times increased the signal and noise in the blank and negative control samples leading to false positives. To overcome this, a method was devised to normalise images of a sample to the same value. First, the automated contrast enhancement method was used to detect the particles and their intensities in the virus samples. The 75% percentile intensity value was taken as the maximum value for normalisation, which led to improved image contrast in viruses of low fluorescent intensity (Fig. 2.8).

The fixed-value normalisation method helped distinguish between virus particle and background noise, as it resulted in a clear difference between particle counts in the plateaus of controls and virus samples (Fig. 2.7B) and helped define a threshold for BVDV-1 and RV-A by producing a more pronounced plateau in the particle detection curve. In addition, while this approach did not change the particle count for the dsDNA viruses (T5, Det7, P22, MHV-68), the counts and consistency across replicates was increased for the ssDNA (M13), ssRNA (BVDV-1) and dsRNA (RV-A) viruses (Fig. 2.7C), although the standard error of the mean was increased when applying fixed-value normalisation for some replicates for some viruses.

The final titre estimate for each of the virus stocks was not affected by the normalisation method for phage T5 and Det7while there was a reduction for P22 and MHV-68. Estimates for

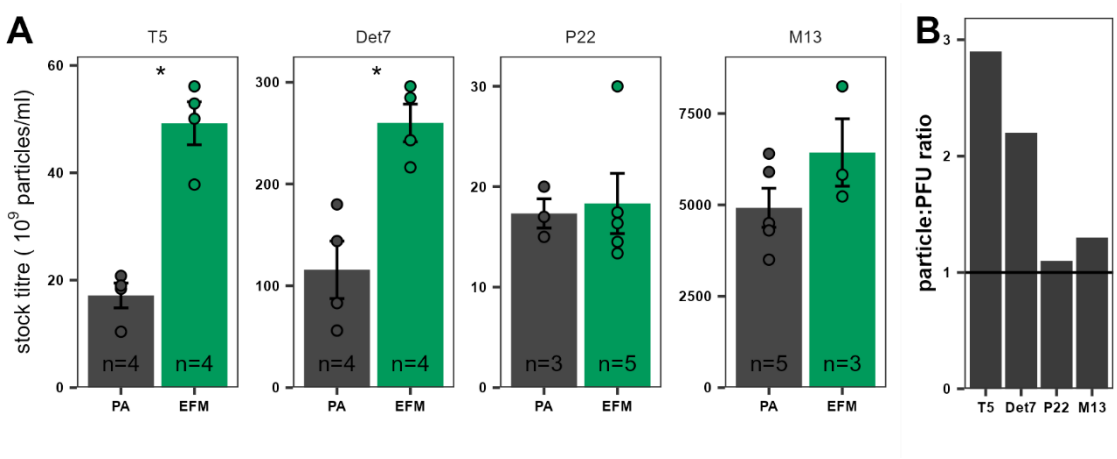

Figure 2.9: Comparison of stock titre estimation of phages by plaque assay and EFM. Stock titres of phages T5, Det7, P22 and M13 were measured by PA and EFM. **A:** estimated stock titre estimate based on PA and EFM. Points show values from individual replicates, error bars show the standard deviation, and sample size is printed at the bottom of each bar. **B:** Ratio of PFU to EFM-based particle count for each of the phages. The horizontal line shows a PPR of 1.

the non-dsDNA virus titres was higher by fixed-value normalisation than for automated normalisation, and the standard deviation and coefficient of variation were reduced, suggesting increased precision of the EFM-based estimation of the stock titre for the non-dsDNA viruses when applying fixed-value normalisation (Fig. 2.7D). The final stock titres for the viruses as measured using EFM with fixed-value normalisation are listed in Table 2.2.

2.3.7 EFM reveals phage particle-to-PFU ratios

For the phages, the stock titre determined through PA was compared to the EFM-based titre estimate to obtain a particle-to-PFU ratio (PPR). Coefficient of variation for EFM was lower than for PA in the case of T5 (PA: 0.27, EFM: 0.16) and Det7 (PA:0.49, EFM: 0.14), while it was comparable for both methods in the case of M13 (PA: 0.24, EFM: 0.25) and higher for EFM for P22 (PA: 0.14, EFM: 0.37) (Fig. 2.9A). For T5 and Det7, the PPR was 2.9 and 2.2, respectively (Fig. 2.9B). For M13, the PPR was lower (1.3), and for P22 the titre estimates by PA and EFM were comparable, with a PPR of 1.1. However, for P22 the four lowest EFM replicates were on average below the average PA measurement, with the highest EFM replicate almost twice as high. Ignoring this data point, the average titre as measured by EFM was 1.54×10^{10} p/ml, while the PFU count determined by PA was 1.73×10^{9} PFU/ml. While the EFM estimate for P22 was imprecise, collectively the data show the value of using EFM over PA, as PA leads to almost a 3-fold underestimation of the number of nucleic acid-containing particles while rarely underestimating particle number.

Table 2.2: Final virus stock titres.

Virus	Stock titre (particles/ml)		SD
T5	4.92	$\times 10^{10}$	$\pm 8.0 \times 10^{9}$
Det7	2.60	$\times 10^{11}$	$\pm 3.7 \times 10^{10}$
P22	1.83	$\times 10^{10}$	$\pm 3.7 \times 10^{9}$
MHV-68	5.74	$\times 10^{8}$	$\pm 1.5 \times 10^{8}$
M13	6.43	$\times 10^{12}$	$\pm 1.6 \times 10^{12}$
BVDV-1	5.91	$\times 10^{10}$	$\pm 2.1 \times 10^{10}$
RV-A	3.40	$\times 10^{9}$	$\pm 7.4 \times 10^{8}$

SD: Standard deviation.

2.4 Discussion

While EFM is routinely used to enumerate virus particles in environmental samples, only dsDNA viruses are typically counted, while ssDNA and RNA viruses require improved methods. In this chapter, T5 and Qb are used to optimise virus staining for EFM. We have found that for EFM, deionised water is the preferred diluent, although PBS should be used for eukaryotic viruses. All viruses were visible by EFM when stained with 25X SYBR Gold. Of the viruses used here, BVDV-1 is the only virus for which visibility is enhanced by fixation with glutaraldehyde.

Using a combination of Python scripts and ImageJ image analysis software, particle counting is streamlined, and virus particles in a variety of prokaryotic and eukaryotic, dsDNA, ssDNA, dsRNA and ssRNA virus stocks are enumerated to establish a virus MC.

2.4.1 Optimal staining parameters

The NTA experiments show that particle brightness is not affected by incubation time and is enhanced at increased dilutions of the virus stock. While the dynamic range of the ZetaView was unable to determine a change particle intensity, we took advantage of the ease and speed of the device to measure and compared fractions of fluorescent particles. As the particle intensity is approximately normally distributed (see Fig. 2.5B, panel H_2O T5), when the particle intensity increases a higher fraction of particles is above the limit of detection of the device. This is supported by the simultaneous increase in particle intensity by EFM and fraction of fluorescent particles by NTA found with increasing dilution of the stock. An increased stock dilution effectively increases the number of dye molecules available for a single virion.

Using plaque assay, staining with 25X SYBR Gold is optimal for virus concentrations of approximately 5 x 10^7 PFU/ml. Depending on the PPR of the virus, the stock might need to be diluted further, meaning that initial titration experiments are required for individual eukaryotic viruses to determine the optimal dilution. One limitation of the EFM technique is that it requires virus stock of ≥ 5 × 10^7 particles/ml, which can be difficult to achieve for some eukaryotic viruses. At least ~10^7 particles are needed to achieve sufficient particle counts. Lower titre stocks requiring a higher sample volume potentially increases non-specific, background, signals originating from the media. In the present protocol, the area of a single microscope image equated to 1/17,000 of the effective filter disc area, meaning a sample with 10^7 particles yields on average ~600 particles in each image. A recommendation therefore is that a 200 µl sample volume containing ~10^7 particles is optimal for staining with 25X SYBR Gold, although larger volumes containing a minimum of 10^6 particles can also be used.

For phages, the optimal diluent is deionised water, with PBS, TE buffer and OptiMEM negatively affecting particle fluorescence for T5 and Qb. The reasons for this are unclear, but our findings suggest that these media interfere with the staining process. Additionally, OptiMEM is not

colourless, even when phenol red-free variant is used, which could attenuate particle fluorescence. For eukaryotic viruses PBS was preferred due to possible negative osmotic effects of using water as a diluent. In our results, fixation of virus stocks was beneficial for observation of BVDV-1, but not other viruses (Fig. 2.5B, 6). It is therefore suggested that fixation is tested for new viruses.

2.4.2 Imaging and image processing script

While the dsDNA viruses required exposure times between 30 and 500 ms and produced images with good contrast, the ssDNA and RNA viruses required longer exposure times. The maximum exposure time was limited by a trade-off between increase particle intensity and bleaching of the sample, and by the increased brightness of the background which limited the achievable contrast. Ideally, the exposure time is such that the virus particle intensity falls in the middle of the dynamic range of the camera, but this was not achieved for these viruses and the EFM method might have underestimated the virus stock titres for M13, BVDV-1 and RV-A, since their brightness distribution curves are skewed and close to the lower limit of the camera's dynamic range (Fig. 2.7A). For the eukaryotic viruses, imaging might be improved using deionised water, as PBS reduced T5 particle brightness. However, the effects of deionised water on the virion integrity in the eukaryotic viruses should first be assessed.

Particle counting is improved by implementing Python scripts that run ImageJ macros and plot the resulting data. This accelerated the counting process and increased consistency by reducing the possibility of user error. User input is however still required to set appropriate threshold for each sample and further improvements can be made by further automating the process. For instance, the slope at of the particle count curve depending on the detection threshold can be determined empirically and used to set the detection threshold. In addition, the Jupyter Notebooks can be converted into a command line script to further streamline the process.

2.4.3 Advantages and disadvantages

EFM provides an accurate and easy to use method for quantification of virus stocks. While optimising particle staining requires some effort, it can be used for a wide variety of viruses with minimal optimisation required for each virus. It thereby offers a clear advantage over qPCR, which can be more technically challenging to setup up, particularly for a large group of viruses for which individual assays and amplification parameters need to be optimised. While it offers a direct measure of nucleic acid-containing particles in a stock, as opposed to only the infective virus particles, or the total nucleic acid in the sample, both can deviate from the actual number of particles.

EFM is more sensitive than NTA and allows measurements of smaller viruses than is possible using NTA, as shown by the unsuccessful use of NTA to quantify Qb virus particles. EFM measures nucleic acid-containing particles, and it cannot measure differences in genomic material present within the particle although compared to qPCR it is less sensitive to genomic mutations common in RNA viruses that lead to defective interfering particles (Vignuzzi and López, 2019).

While nucleic acid-containing particles are visualised in EFM, this can also include non-viral particles such as extracellular vesicles of bacterial, eukaryotic origin or media origin with associated nucleic acid and other auto fluorescent material. Inclusion of controls is therefore important. Negative controls from cell cultures of MHV-68, BVDV-1, and RV-A were lacking as pre-existing stocks were used, leaving open the possibility that non-viral material was included in the analysis. For phages, negative controls can be produced, although it is unclear if production of EVs from infected versus uninfected bacterial cultures are different. Based on the particle intensity profile, high-intensity outlier can be excluded from the analysis to reduce false positives. Further purification of virus stocks, for example through density gradient centrifugation or polyethylene glycol (PEG) precipitation (Villafane, 2009; Hernandez *et al.*, 2019), and removal of EVs by chloroform treatment of non-enveloped viruses (Biller *et al.*, 2017) may limit non-viral particle contamination and increase accuracy.

In our EFM analysis, single pinpricks of light source are assumed to be single viruses, ignoring the possibility of virus aggregates. Clumping of particles will increase the intensity of the light source with distinct peaks only seen for P22. For this phage, particle intensity varied across replicates, and it is not clear whether this reflects background noise or clumping of particles. TEM could help in addressing this issue, although this is a costly technique.

2.4.4 EFM and phage PPR

A 2.9-fold difference is seen between EFM particle counts and PFU estimate for phages. The same reagents and protocol were used for plaque assays although they are not optimised for each phage, which may reduce plating efficiency of different phages (Kutter, 2009). For eukaryotic viruses with a large disparity between number of virus particles and infectious particles the PRR can range from 1 to 10,000 (McCormick and Mermel, 2021). While phages are assumed to have a PPR close to 1, several environmental phages are reported to have PRR values of 1.7 to 10 (Ghanem *et al.*, 2018). In comparing EFM-based particle estimates for the phages to the PFU estimates produces a PPR of approximately 1 for P22 and 2.5 for Det7, consistent with the literature. This shows that accurate estimates of virus particles are required even for phages, since PAs can lead to an underestimating the number of particles.

2.4.5 Conclusion

Fluorescent staining of virus particles has been optimised to measure the titre of virus stocks by EFM. Additionally, a set of Python scripts was developed to count virus particles quickly and reproducibly in EFM images. Measurements on phages and eukaryotic viruses comprising enveloped and non-enveloped viruses, and dsDNA, ssDNA, ssRNA and dsRNA viruses show that a wide range of viruses can be titrated using EFM. The advantage of this technique is its simplicity, and the fact that it offers a direct measurement of nucleic acid containing particles, regardless of infectivity. While optimisation of the staining process and the development of the analysis scripts was time consuming measuring virus stock concentrations itself was straightforward and this technique can be used to measure stock concentrations of a variety of viruses.

CHAPTER 3: ASSESSING BIAS AND REPRODUCIBILITY OF VIRAL METAGENOMICS METHODS

3.1 Introduction

Metagenomics methods enable the collective analysis of microbial genomes (i.e., including bacterial and viral genomes: the microbiome) and were first used to analyse faecal viruses over a decade ago (Breitbart *et al.*, 2003). To date, most virome studies have focussed on dsDNA viruses, as they were assumed to be mostly prokaryotic viruses (phages). However, the abundance of prokaryotic RNA viruses has been underestimated (Callanan *et al.*, 2021), as is reflected by the substantial increase in known RNA viruses following recent metatranscriptomic studies (Callanan *et al.*, 2020; Neri *et al.*, 2022). Moreover, since a large portion of human viruses have RNA genomes, comprehensive metagenomic analysis of gut virus genomes (the virome) require a method that can capture both types of viruses. Assessment of biases and reproducibility of these methods is vital, in particular for longitudinal studies since viromes of individuals have low variation over time (Flores *et al.*, 2014; Shkoporov *et al.*, 2019).

Every stage of the virome metagenomic sequencing pipeline can affect recovery of viruses, and several studies have provided optimisations. Short-term storage at 4 °C and ambient temperature, and long-term storage at -80 °C have negligible effects on phage composition, while operator bias has been shown to have a small but significant contribution to between-sample variation (Shkoporov *et al.*, 2018). Some widely used VLP extraction and purification methods, like ultracentrifugation, CsCl density gradient centrifugation, ultrafiltration, and tangential flow filtration, lead to loss of specific viruses (Conceição-Neto *et al.*, 2015; Kleiner, Hooper and Duerkop, 2015; L. Li *et al.*, 2015; Parras-Moltó *et al.*, 2018; d'Humières *et al.*, 2019). Instead, minimal treatment of the sample is recommended, consisting of homogenisation without bead beating, brief centrifugation, filtration using 0.80 µm or 0.45 µm filters, and digestion of free nucleic acid using DNase and RNase to remove extracellular and host genetic material before extraction of viral genomes (Conceição-Neto *et al.*, 2015; Kleiner, Hooper and Duerkop, 2015; L. Li *et al.*, 2015). Direct ligation of sequencing adapters without subsequent amplification produces the least bias (Karlsson, Belák and Granberg, 2013; d'Humières *et al.*, 2019; Hsieh *et al.*, 2021). This requires relatively large samples (3 – 5 g) and concentration using PEG precipitation (Shkoporov *et al.*, 2018; d'Humières *et al.*, 2019). However, the bias associated with PEG precipitation remains unclear (Conceição-Neto *et al.*, 2015). Also, PEG

precipitation is time consuming, and direct adapter ligation excludes RNA viruses. Thus, amplification-based methods that include RT of RNA are preferred for comprehensive large-scale viral metagenomic research.

Several approaches for metagenome amplification exists. One method is multiple displacement amplification (MDA), in which Φ29 DNA polymerase is used for isothermal amplification of DNA (Blanco *et al.*, 1989; Gonzalez, Portillo and Saiz-Jimenez, 2005), and can be adapted to include an RT step for amplification of RNA (Shkoporov *et al.*, 2018). However, Φ29 DNA polymerase strand engages in rolling circle amplification, causing amplification bias for circular ssDNA viruses (Kim and Bae, 2011; Roux *et al.*, 2016; Parras-Moltó *et al.*, 2018). MDA also introduces other amplification biases (Abulencia *et al.*, 2006; Yilmaz, Allgaier and Hugenholtz, 2010) that cannot be overcome by pooling (Marine *et al.*, 2014) as proposed by some studies (Shkoporov *et al.*, 2018). Sequence-independent single primer amplification (SISPA) (Reyes and Kim, 1991) randomly amplifies DNA and RNA virus genomes using a reverse transcriptase and random primers with a 5' universal sequence (Froussard, 1992; Djikeng *et al.*, 2008). While SISPA does not have a bias towards ssDNA viruses and produces better representation of community relative abundances, both MDA and SISPA libraries have higher GC bias and non-uniform genome coverage compared to non-amplified libraries (Parras-Moltó *et al.*, 2018). On the other hand, while amplification introduces bias and reduces the number of unique reads, it also produces longer contigs in assemblies (L. Li *et al.*, 2015; Kallies *et al.*, 2019). Linker-amplified shotgun library (LASL)-based approaches (Breitbart *et al.*, 2002) work by ligation of a linker sequence to dsDNA (Duhaime *et al.*, 2012) and ssDNA (Roux *et al.*, 2016), which is then amplified using PCR. This method introduces less bias than MDA (Kim and Bae, 2011; Roux *et al.*, 2016). However, accuracy of the method depended on a correction factor for nucleic acid extraction efficiency, which was determined for one phage, but it is unclear if extraction efficiencies of other viruses can be extrapolated from this. Also, while LASL bias is lower than SISPA and MDA, it negatively affects low-abundant viruses (Hsieh *et al.*, 2021) and moreover, it is unable to amplify RNA. The NetoVIR protocol, which was optimised by assessing bias of individual steps in the virome sequencing pipeline on the recovery of viruses in an MC, used a Whole Transcriptome Amplification kit (WTA2), which had high reproducibility and good correlation of read numbers with qPCR-based virus abundance (Conceição-Neto *et al.*, 2015). The WTA2 approach is less time consuming than SISPA, although the kit is costly and, in our experience, susceptible to supply chain issues. Previous studies have evaluated various aspects of the virome sequencing pipeline using virus MCs (Conceição-Neto *et al.*, 2015; Kleiner, Hooper and Duerkop, 2015; L. Li *et al.*, 2015; Roux *et al.*, 2016; Parras-Moltó *et al.*, 2018) faecal samples

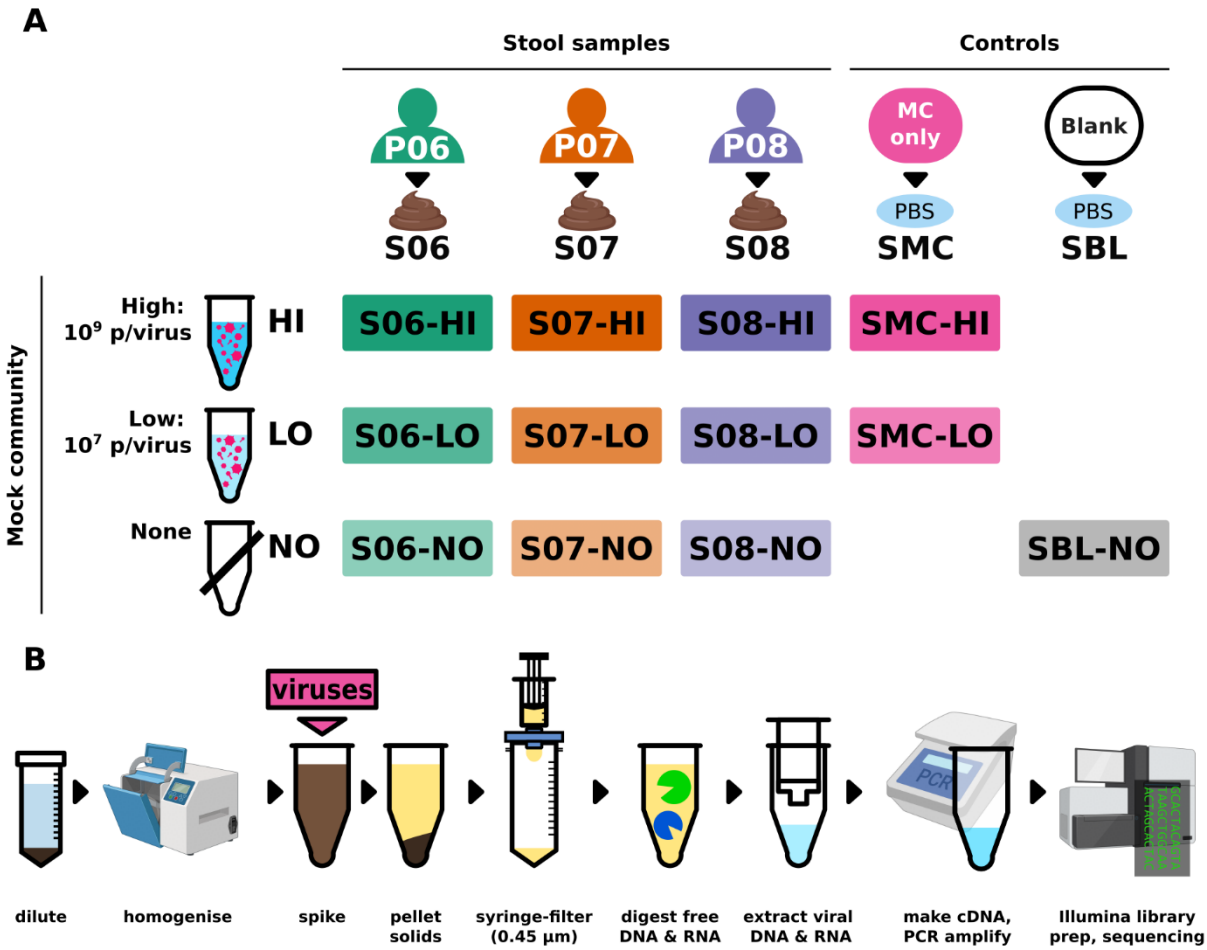

Figure 3.1: Experimental design. A: Stool samples (S06, S07 and S08) were obtained from three ME/CFS patients, P06**,** P07, and P08. After homogenisation, three aliquots are taken from each stool sample. To one aliquot of each sample, high-concentration MC is added (HI), to another aliquot a low-concentration MC is added (LO), and the last remaining aliquot is processed untreated (NO). Additionally, an MC-only sample (SMC) each of the HI and LO MC and a blank sample (SBL) were processed. **B:** Extraction of virus-like particles (VLPs) and viral nucleic acid to generate a dsDNA library and Illumina library for sequencing.

(Shkoporov *et al.*, 2018; d'Humières *et al.*, 2019; Hsieh *et al.*, 2021; Wang *et al.*, 2023; Soria-Villalba *et al.*, 2024), and faecal samples spiked with viruses or MCs (Kleiner, Hooper and Duerkop, 2015; Shkoporov *et al.*, 2018). However, most studies have omitted RNA viruses (Kleiner, Hooper and Duerkop, 2015; Roux *et al.*, 2016; d'Humières *et al.*, 2019; Hsieh *et al.*, 2021) and use different combinations of VLP extraction and amplification methods. Each study has focussed on distinct aspects of bias and reproducibility, including recovery bias of MC and stool viruses, sequencing biases like GC-bias, over- or underrepresentation of different virus sequences, genome coverage and genome coverage depth, diversity of stool viruses, and taxonomic variation. WTA2 was shown to produce good correlation of read numbers to qPCR-

Table 3.1: Characteristics of the MC viruses.

Host	Virus		Genome			Virion		Mock community	
	Abbr.	Species	Type	Top.	Size (kb)	Env.	Size (nm)	HI (p)	LO (p)
Prokaryotic	Det7	*Kuttervirus Det7*	dsDNA	linear	157.5		90[c], 110[t]	1.0×10^9	1.0×10^7
	M13	*Inovirus M13*	ssDNA	circular	6.4		6.5[d], 860[l]	1.0×10^9	1.0×10^7
	P22	*Lederbergvirus P22*	dsDNA	linear	41.7		60	1.0×10^9	1.0×10^7
	T5	*Tequintavirus T5*	dsDNA	linear	121.7		90[c], 160[t]	1.0×10^9	1.0×10^7
Eukaryotic	BVDV-1	*Pestivirus bovis*	ssRNA	linear	12.5	Yes	50	1.0×10^9	1.0×10^7
	MHV-68	*Rhadinovirus muridgamma4*	dsDNA	linear	119.5	Yes	220	2.5×10^8	2.5×10^6
	RV-A	*Rotavirus A*	dsRNA	linear (11 sg.)	18.6		80	1.0×10^9	1.0×10^7

Abbr: abbreviation used in this text, env: enveloped, HI: high-concentration MC, LO: low-concentration MC, c: capsid size, t: tail length, d: virion diameter, l: virion length, p: particles, sg: genome segments, top: genome topology

based abundance of MC viruses (Conceição-Neto *et al.*, 2015), but a comprehensive evaluation of other bias and reproducibility characteristics and its performance with stool samples has not been published. A comparison of WTA2 to other amplification methods is also lacking. This would be especially valuable given the increased cost and availability of WTA2. Thus, a comprehensive analysis of bias and reproducibility of a virome sequencing pipeline based on WTA2 and a comparison of WTA2 and SISPA is warranted.

3.1.2 Specific aims

The main aim of this chapter is to investigate and compare the bias and reproducibility of viral metagenomics methods based on WTA2 and SISPA. The specific aims are:

- To construct two MCs consisting of diverse viruses (Table 3.1) with different concentrations using the virus titre measurements in Chapter 2.
- To perform VLP purification, and viral nucleic acid extraction using faecal samples spiked with the MCs (Fig. 3.1), perform reverse-transcription and amplification using WTA2 and SISPA (Fig. 3.2), and perform sequencing of the WTA2 and SISPA libraries.
- To determine the recovery bias for the MC viruses and compare WTA2 and SISPA.
- To determine the uniformity and reproducibility MC virus sequencing depth of WTA2 and SISPA.
- To compare assembly quality between WTA2 and SISPA
- To determine the reproducibility of faecal virus recovery of WTA2 and SISPA
- To determine the consistency of virome taxonomic analysis and compare WTA2 and SISPA.

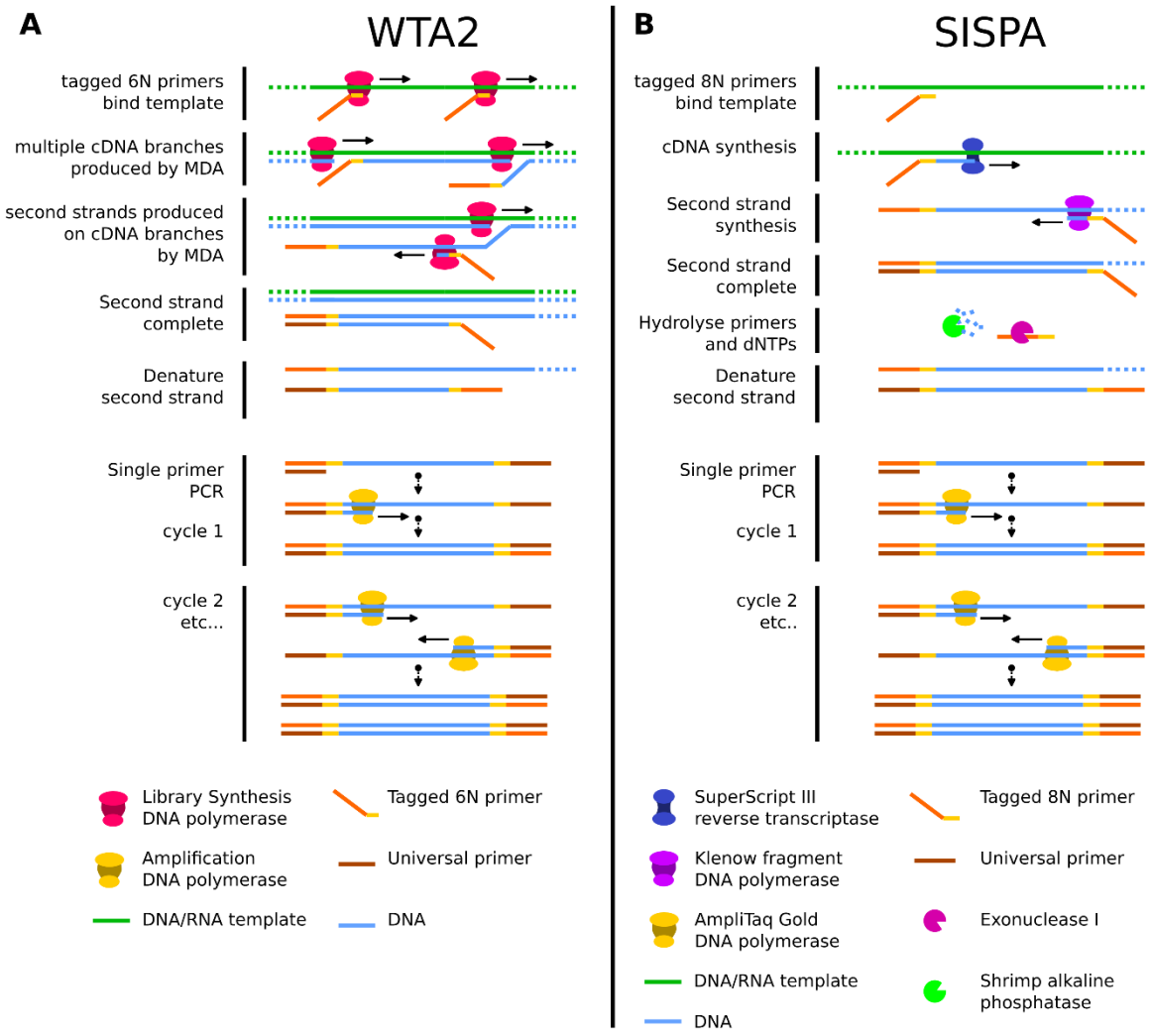

Figure 3.2: Generation of dsDNA libraries using WTA2 and SISPA. A: The whole-transcriptome amplification (WTA2) kit method consists of two steps: 1) cDNA library synthesis followed by 2) library amplification. In the first step, Library Synthesis DNA polymerase (LSDP) with both RNA- and DNA-directed activity and strand displacement capacity, performs MDA. After the binding of quasi-random hexamer primers with a universal sequence tag to the template, the LSDP syntesises a complementary DNA (cDNA) strand. Following extension of primers bound upstream, the cDNA strand is displaced by LSDP, allowing it to serve as a template for primers to bind and for LSDP generate a second strand that is flanked on both ends by a universal sequence tag. In the second step, amplification DNA polymerase extends primers bound to the universal sequence tags flanking the second strands and amplifies the second strand fragments through standard PCR. **B:** In sequence-independent single primer amplification (SISPA), random octamer primers with a universal tag sequence bind DNA and RNA templates. SuperScript III reverse transcriptase with RNA- and DNA-directed DNA polymerase activity (Georgiadis *et al.*, 1995) extends the primer, producing

cDNA. Random octamer primers bind the cDNA strand and a second strand with flanking universal sequences is synthesised by DNA polymerase I large (Klenow) fragment. Free nucleotides are then dephosphorylated by shrimp alkaline phosphatase and primers are degraded by exonuclease I. Using AmpliTaq Gold polymerase, the DNA library fragments are amplified in a standard PCR reaction.

3.2 Materials and methods

3.2.1 Ethics

Samples were originally obtained for the study "Autoimmunity in ME/CFS" (AI-ME/CFS), ClinicalTrials.gov ID NCT03254823 (Seton, 2022). Ethical approval for that study was obtained from the National Research Ethics Service (NRES) Committee London Hampstead (17/LO/1102). Participants provided informed consent for the use of samples in subsequent research. All research was performed in accordance with the Declaration of Helsinki (World Medical Association, 2013), and the International Council for Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH)-Good Clinical Practice (ICH-GCP) guidelines. Data was handled following the European Union General Data Protection Regulation (GDPR) and United Kingdom Data Protection Act 2018.

3.2.2 Stool samples

Stool samples were donated by two same household healthy controls (P06, P08) and one severe myalgic encephalomyelitis/chronic fatigue syndrome (P07). Stool samples were collected in July 2018 (S06) and December 2019 (S07, S08) and processed as described previously (Seton, 2022). Briefly, stool samples were homogenised by mixing, divided into 100 mg aliquots, and stored at -80 °C. Sample processing was performed by Dr Katharine Seton, Dr Shen-Yuan Hsieh, and Dr Fiona Newberry.

3.2.3 Mock virus community

An MC was constructed using virus stocks and consisted of the three tailed phages *Tequintavirus T5* (T5), *Kuttervirus Det7* (Det7), and *Lederbergvirus P22* (P22), the filamentous phage *Inovirus M13* (M13), and three eukaryotic viruses: the *Rhadinovirus muridgamma4* strain Murid gammaherpesvirus-68 (MHV-68), the *Pestivirus bovis* strain Bovine viral diarrhoea virus Ky1203nc (BVDV-1), and the *Rotavirus A* stain Simian rotavirus SA/11 (RV-A). Virus titres were determined through epifluorescence microscopy as described in Chapter 2. Each virus stock was diluted in PBS to obtain the same final concentration, except for MHV-68, for which the concentration was 25% of the others. A high concentration mock community (MC HI) was therefore constructed with virus concentrations such that 35 µl of mock community contained 0.25×10^7 particles of MHV-68 and 1×10^7 of each of the other viruses, for a total of 6.25×10^7 virus particles. As the estimated number of virus particles in human stool is up to

10^9 particles/g (Hoyles *et al.*, 2014), the MC HI concentration corresponded to the estimated 5 × 10^7 particles in 50 mg stool. Subsequently, a low concentration MC (MC LO) was produced by 100-fold dilution of MC HI in PBS. MCs were stored at 4 °C.

3.2.4 Spiking experiment

The following procedure was performed by Dr Oliver Charity. Of each stool sample, 1.9 – 3.5 g was taken and diluted to 10% (w/v) in PBS. The sample was then homogenised using a Stomacher 400 Circulator Lab Blender (Seward, Worthing, UK) set to 260 RPM for 3 minutes. For stool samples S06, S07 and S08, one 1 ml aliquot was taken and 35 µl of MC HI was added, producing samples S06-HI, S07-HI, and S08-HI, respectively (Fig. 3.1). A second 1 ml aliquot of each sample was spiked with MC LO, producing samples S06-LO, S07-LO and S08-LO, and a third aliquot was taken without spiking, designated S06-NO, S07-NO and S08-NO. Additionally, 35 µl MC HI and 35 µl MC LO were each added to 1 ml PBS, to produce blank samples SMC-HI and SMC-LO, respectively, and a blank PBS sample was included, designated SBL-NO. Samples were centrifuged at 16,000 × *g* for 3 minutes at 21 °C, and the supernatant was syringe filtered using a 0.45 µm syringe filter unit (Starlab, Milton Keynes, UK, ref. E4780-1456) to extract VLPs. The VLP extract was stored at -80 °C overnight and 700 µl extract was then treated with a nuclease cocktail consisting of 1 µl Benzonase (Millipore, Gillingham, UK, ref. E1014-5KU), 4 µl RNAse I (Thermo Fisher Scientific, Loughborough, UK, ref. AM2294), 16 µl DNAse I (Thermo Scientific, Loughborough, UK, ref. EN0521), 40 µl TURBO DNase (Thermo Fisher Scientific, Loughborough, UK, ref. AM2238), and 125 µl of the respective 10X buffers, and incubated at 37 °C for 2 hours to digest unprotected nucleic acid. The samples were then incubated at 75 °C for 1 hour and 10 µl 0.5 mM EDTA (pH=8.0) (Thermo Fisher Scientific, Loughborough, UK, ref. AM9260G) was added, to deactivate nucleases. Viral nucleic acid was then extracted using the QIAmp Viral RNA Mini Kit (QIAGEN, Manchester, UK, ref. 52904) according to the manufacturer's instructions, without addition of carrier RNA. Nucleic acid extracts were stored at -20 °C.

3.2.5 Reverse-transcription and amplification

To produce a library of dsDNA, the RNA was reverse transcribed, and the DNA replicated, and the resulting dsDNA fragments were amplified. Two approaches were tested, one using the Complete Whole Transcriptome Amplification (WTA2) kit (Sigma-Aldrich, Gillingham, UK, ref. WTA2-10RXN) and another using SISPA (Kramná and Cinek, 2018).

3.2.5.1 WTA2

The WTA2 approach was performed by Dr Oliver Charity the day following nucleic acid extraction. The dsDNA library was produced according to the instructions in the NetoVir protocol (Conceição-Neto *et al.*, 2015). Briefly, 2.82 µl sample was added to 0.5 µl Library Synthesis Solution containing universal sequence-tagged quasi-random hexamer primers. The

sample was denatured at 95 °C for 2 minutes and then cooled to 18 °C to prime RNA and DNA. Then, 1.68 µl library synthesis master mix was added, containing 0.5 µl Library Synthesis Buffer, 0.78 µl RNAse-free water, and 0.4 µl Library Synthesis Enzyme was added and incubated in a thermocycler set to 18 °C for 10 minutes, 25 °C for 10 minutes, 37 °C for 30 minutes, 42 °C for 10 minutes, and 70 °C for 20 minutes, to produce a dsDNA library (Fig 2A). Of the dsDNA library, 5 µl was added to 69.95 µl amplification master mix was added, containing 7.5 µl Amplification Mix with universal sequence primers, 60.2 µl nuclease-free water, 1.5 µl WTA2 dNTP mix, and 0.75 µl Amplification Enzyme. The sample was incubated in a thermocycler set to 94 °C for 2 minutes, followed by 17 cycles of 94 °C for 2 minutes and 70 °C for 5 minutes, to amplify the dsDNA fragments. DNA was then extracted using the QIAquick PCR Purification Kit (QIAGEN, Manchester, UK) according to manufacturer's instructions and stored at -20 °C.

3.2.5.2 SISPA

Ten months after viral nucleic acid was extracted, SISPA was used to produce a dsDNA library for sequencing from the same nucleic acid extracts that had been processed using the WTA2 kit. Samples were processed as described previously (Kramná and Cinek, 2018) (Fig 2B). Briefly, 4 µl sample was added to 9 µl master mix containing 1 µl RNasin (Promega, Chilworth Southampton, UK, ref. N261A), 1 µl 10 mM dNTPs (New England Biolabs, Hitchin, UK, ref. N0447S), 1 µl 20 µM primer D2_8N (5'-AAGCTAAGACGGCGGTTCGGNNNNNNNN-3') and 6.5 µl nuclease-free water (Thermo Fisher Scientific, Loughborough, UK, ref. AM9937), incubated at 65 °C and then cooled to 4 °C. Then, 6 µl master mix containing 4 µl 5X first strand buffer, 1 µl 0.1 mM dithiothreitol, and 1.0 µl SuperScript III reverse-transcriptase (200 U/µl) (Thermo Fisher Scientific, Loughborough, UK, ref. 18080044) were added and incubated in a thermocycler at 50 °C for 1 hour for first strand synthesis. Then, 1.5 µl master mix containing 0.85 µl nuclease-free water, 0.15 µl 10X Klenow buffer and 0.5 µl DNA polymerase I large (Klenow) fragment (5 U/µl) (New England Biolabs, Hitchin, UK, ref. M0212L) was added, and the sample was incubated at 37 °C for 1 hour for second strand synthesis, followed by incubation at 75 °C for 10 minutes to inactivate the enzyme. Free primers and nucleotides were then digested and dephosphorylated by incubation with 20 µl master mix containing 17 µl nuclease-free water, 1.0 µl 10X Shrimp Alkaline Phosphatase (SAP) buffer, 1.0 µl Exonuclease I (20 U/µl) (New England Biolabs, Hitchin, UK, ref. M0293S) and SAP (1 U/µl) (New England Biolabs, Hitchin, UK, ref. M0371S), respectively, at 37 °C for 1 hour, followed by 15 minutes at 75 °C to inactivate the enzymes. The sample was then frozen at -20 °C for 16 hours. The following day, the dsDNA library was generated and amplified from 8 µl of the reverse-transcribed sample, by adding it to 42 µl master mix containing 26 µl nuclease-free water, 5 µl 10X PCR buffer, 6 µl 25 mM MgCl2, 1.5 µl 10mM dNTP, 3 µl 20 µM primer D2

('5-AAGCTAAGACGGCGGTTCGG-3'), and 0.5 µl AmpliTaq Gold (5 U/µl) (Thermo Fisher Scientific, Loughborough, UK, ref. 10685095) DNA polymerase by incubation on a thermocycler. The thermocycler program consisted of: 1) denaturation for 5 minutes at 95 °C, 2) 5 cycles of denaturation at 95 °C for 1 minute, annealing at 55 °C for 1 minute and extension at 72 °C for 1:30 minutes, 3) 25 cycles of denaturation at 95 °C for 30 seconds, annealing at 55 °C for 30 seconds, and extension at 72 °C for 1:30 minutes, adding 2 seconds to the extension time every cycle, 4) final extension at 72 °C for 10 minutes. Samples were kept on ice between incubation steps throughout. PCR product was loaded onto a 2% agarose (Melford, Ipswich, UK, ref. 3913900099) gel in 0.5X TBE buffer (Thermo Fisher Scientific, Loughborough, UK, ref. J62788.K2), and inspected for a smear between 200 – 500 bp. DNA was then extracted using the Cleanup & Concentrator-10 kit (Zymo Research, Freiburg im Breisgau, Germany, ref. D4010) following manufacturer's instructions. Purified DNA of the WTA2 and SISPA methods was quantified using a Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, Loughborough, UK, ref. Q32851) on a Qubit 3.0 Fluorometer (Thermo Fisher Scientific, Loughborough, UK, ref. Q33216) and samples were normalised to 5 ng/µl prior to library preparation.

3.2.6 Fresh MC sample

Virus stocks used to produce the first MC HI were stored at 4 °C for 10 months. A fresh mock community was assembled from these stocks, containing viruses at the same ratios as described above for MC HI. The MC was filtered using a 0.45 µm PVDF membrane centrifugal filter columns (Millipore, Gillingham, UK, ref. UFC40HV00) at 5,000 x g for 4 min at 4 °C and 780 µl of the filtrate was processed further. The sample was then incubated with 42 µl nuclease buffer (1M Tris, 100 mM CaCl2 and 30 mM MgCL2, pH=8), 12 µl Benzonase (Millipore, Gillingham, UK, ref. E1014-5KU) and 6 µl Micrococcal nuclease (New England Biolabs, Hitchin, UK, ref. M0247S) at 37 °C for 2 hours, after which 42 µl 0.5 M EDTA (pH=8.0) (Thermo Fisher Scientific, Loughborough, UK, ref. AM9260G) was added to chelate ions and inhibit nuclease enzymes. Viral nucleic acid was then extracted using the QIAmp Viral RNA Mini Kit and stored at -20 °C overnight. The viral nucleic acid was then processed using SISPA as described above.

3.2.7 Illumina library preparation and sequencing

Library preparation and sequencing was performed by David Baker (QIB Sequencing Facility). Briefly, 0.5 µl of Tagmentation Buffer 1 was mixed with 0.5 µl Bead Linked Transposomes (Illumina, Cambridge, UK, ref. 20018704) and 4 µl nuclease-free water in a master mix, and 5 µl added to a 96 well plate. 2 µl of DNA normalised to 5 ng/µl was pipette-mixed with 5 µl of the Tagmentation mix and heated to 55 °C for 15 minutes. A PCR master mix was made using 10 µl KAPA 2G Fast Hot Start Ready Mix (Merck, Gillingham, UK, ref. KK5601) and 2 µl PCR grade water per sample. Of the PCR master mix, 12 µl was added to each well to be used in a 96-well plate, and 1 µl of 10µM primer mix containing both P7 and P5 Illumina barcodes

(Perez-Sepulveda *et al.*, 2021) were added to each well. For the WTA2 samples, custom 9 bp dual barcodes were used, while for the SISPA samples custom 10 bp unique dual index barcodes were used. Finally, 7 µl tagmentation mix was added and mixed. The PCR was run at 72 °C for 3 minutes, 95 °C for 1 minute, then 14 cycles of 95 °C for 10 seconds, 55 °C for 20 seconds and 72 °C for 3 minutes. The libraries were quantified using the Promega QuantiFluor dsDNA System (Promega, Chilworth Southampton, UK, ref. E2670) and measured on a GloMax Discover Microplate Reader (Promega, Chilworth Southampton, UK, ref. GM3000). Libraries were pooled following quantification in equal quantities. The final pool was size selected using solid phase reversible immobilisation (SPRI) beads at 0.5X concentration, followed by SPRI beads size selection at 0.7X concentration, using Illumina DNA Prep, (M) Tagmentation sample purification beads (Illumina, Cambridge, UK, ref. 20060059). The final pool was quantified on a Qubit 3.0 Fluorometer and run on a D5000 ScreenTape (Agilent, Stockport, UK, ref. 5067-5579) using the Agilent TapeStation 4200 (Agilent, Stockport, UK) to calculate the final library pool molarity. The WTA2 pool was run at a final concentration of 1.8 pM on an Illumina NextSeq 500 instrument with a high output 300-cycle flow cell (Illumina, Cambridge, UK, ref. 20024908). The SISPA pool was run at a final concentration of 750 pM on an Illumina NextSeq 20000 instrument using a P3 300-cycle flow cell (Illumina, Cambridge, UK, ref. 20040561). Each were run following the Illumina recommended denaturation and loading recommendations and included a 1% PhiX Control v3 spike-in (Illumina, Cambridge, UK, ref. FC-110-3001).

3.2.8 Bioinformatics analysis

An overview of the bioinformatics pipeline used for this analysis is depicted in Fig. 3.3.

3.2.8.1 Quality control

Raw reads were mapped against a pre-built bowtie2 index of the human genome version GRCh38 (Langmead, 2020) using Bowtie 2 v.2.4.5 (Langmead and Salzberg, 2012; Langmead *et al.*, 2019).[1] Read quality was analysed using FastQC v.0.11.9 (*Braham Informatics*, 2019). Adapter sequences were trimmed and reads were filtered based on quality using fastp v.0.23.2 (Chen *et al.*, 2018; Chen, 2023) with the default settings, except for the options automatic adapter detection and base correction for pair-ended data, cutting of the tail sequence if the mean phred quality of a sliding window of 4 bases is <20, and trimming the first 7 bases from the front, and poly-G and poly-X sequences from the tail. Reads were analysed again in FastQC to ensure high-quality reads remained.

[1] **Note:** After removal of human reads, subsequent analysis was inadvertently performed using the unfiltered read files instead of the filtered read files.

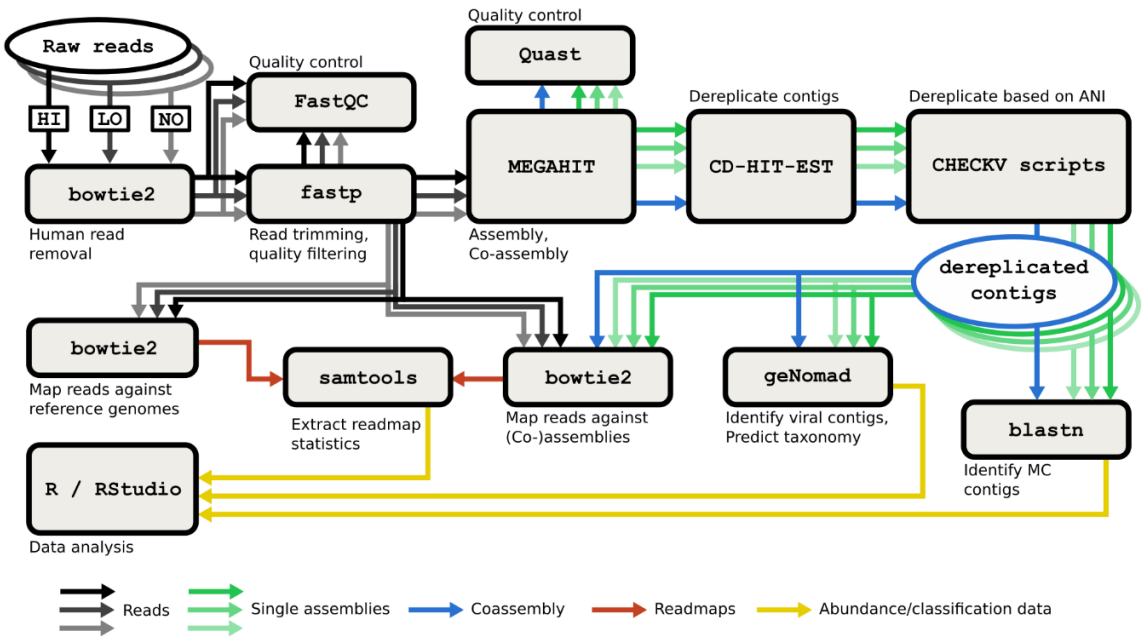

Figure 3.3: Bioinformatics workflow. Black and grey arrows depict flow of read data from raw reads to assembler and read mapping. Green and blue arrows indicate flow of assemblies of single samples and co-assembly of pooled samples from the same stool sample, respectively, from assembler to contig classification and read mapping. Red arrows indicate the flow of read mapping data, and yellow arrows indicate the flow of abundance and classification data to the data analysis stage.

3.2.8.2 Reference-based relative abundance of MC viruses

To determine the relative abundance of MC viruses in the samples, reads were mapped against an index of mock community reference genomes (Table 3.2). First, reference genomes were downloaded from the National Center for Biotechnology Information (NCBI) nucleotide database using the reference genome accession numbers with NCBI Entrez Direct program efetch v.16.6, and an index was built using Bowtie 2 command bowtie2-build with default settings. Reads were then mapped to the reference index using Bowtie 2. The number of reads mapping to respective viruses was calculated using the SAMtools v.1.17 (Danecek *et al.*, 2021) idxstats command. MC abundances were calculated as the number of reads mapped to the respective genomes, per 1000 bases of genome, per 1 million reads in the samples (RPKM). The MC relative abundance was calculated as the percentage of all MC reads that map to the respective virus. The expected abundance of each virus was calculated by multiplying the genome length by the number of particles in the MC and the number of strands in the genome. The expected relative abundance was then calculated as a fraction of the cumulative expected abundance of all viruses. For the WTA2 libraries of SMC-HI and SMC-LO samples, sample labels were inadvertently switched, which was corrected in all analysis.

Table 3.2: Overview of MC reference sequences used.

Virus	NCBI Accession		Genome	
Abbreviation	**Accession nr.**	**Sequence name**	**Bases**	**GC%**
BVDV-1	NC_001461.1	Bovine viral diarrhoea virus 1, complete genome	12,573	45.79
Det7	NC_027119.1	Salmonella phage Det7, complete genome	157,498	44.6
M13	NC_003287.2	Enterobacteria phage M13, complete genome	6407	40.74
MHV-68	NC_001826.2	Murine herpesvirus 68 strain WUMS, complete genome	119,451	47.23
P22	NC_002371.2	Salmonella phage P22, complete genome	41,724	47.09
RV-A	NC_011507.2	Rotavirus A segment 1, complete genome	3,302	33.74
RV-A	NC_011506.2	Rotavirus A segment 2, complete genome	2,693	32.97
RV-A	NC_011508.2	Rotavirus A segment 3, complete genome	2,591	28.91
RV-A	NC_011510.2	Rotavirus A segment 4, complete genome	2,362	34.67
RV-A	NC_011500.2	Rotavirus A segment 5, complete genome	1,614	31.16
RV-A	NC_011509.2	Rotavirus A segment 6, complete genome	1,356	38.57
RV-A	NC_011501.2	Rotavirus A segment 7, complete genome	1,105	33.81
RV-A	NC_011502.2	Rotavirus A segment 8, complete genome	1,059	33.3
RV-A	NC_011503.2	Rotavirus A segment 9, complete genome	1,062	35.78
RV-A	NC_011504.2	Rotavirus A segment 10, complete genome	751	40.21
RV-A	NC_011505.2	Rotavirus A segment 11, complete genome	667	38.53
T5	NC_005859.1	Enterobacteria phage T5, complete genome	121,750	39.27

3.2.8.3 Mock community coverage depth

Coverage depth and breadth of mock community viruses was calculated using the SAMtools depth command. Coverage breadth was calculated as the fraction of bases with a depth >0. To

compare sequencing depth across samples, the normalised sequencing depth was calculated as a fraction of the total number of bases sequenced for each virus in each sample.

3.2.8.4 GC content of MC reads and genomes

The GC content of MC reads was calculated using the SAMtools view command to extract read sequences and a custom script to calculate the GC fraction by dividing the number of GCs in all reads mapping to a virus genome by the total number of bases mapping to the genome. Genome GC content was calculated using the SeqKit v.2.2.0 (Shen *et al.*, 2016) fx2tab command.

3.2.8.5 Assembly of reads

High-quality reads corresponding to the same stool sample were pooled. Pooled reads and reads from individual samples were co-assembled and assembled, respectively, using MEGAHIT v.1.2.9 (D. Li *et al.*, 2015; Li *et al.*, 2016) with default settings. Contigs were dereplicated in two steps. First, CD-HIT-EST v.4.8.1 (Li and Godzik, 2006; Fu *et al.*, 2012) was used cluster contigs based on ≥95% sequence identity, and ≥85% alignment coverage of the shorter sequence. Using BLAST v.2.12.0, a database of contigs from CD-HIT-EST was then built, and an all-versus-all alignment was performed using blastn. Using the accessory scripts "anicalc.py" and "aniclust.py" from CheckV v.1.0.1 (Nayfach, Camargo, *et al.*, 2021), contigs were again clustered at 95% average nucleotide identity and 85% coverage of the shortest sequence to remove circularly permutated redundant sequences. Contigs were filtered from the original assembly using the list of contig IDs with SeqKit using the grep command. (Co-)assemblies before and after dereplication were analysed for quality control using Quast v.5.2.0 (Gurevich *et al.*, 2013).

3.2.8.6 Identification and classification of viral and MC sequences

For identification and taxonomic annotation of viral contigs, geNomad v.1.5.0 (Camargo *et al.*, 2023) was used with geNomad database v.1.2. Dereplicated contigs were analysed using the end-to-end pipeline in geNomad with default settings. For downstream analysis, only non-provirus contigs with a virus score >0.8 were used. Contigs originating from MC viruses were identified using BLAST v.2.12.0. A BLAST database was built of MC virus reference genomes. Dereplicated contigs were then aligned against the MC database. Contigs that aligned with ≥98% nucleotide identity and ≥95% coverage of the contig were kept. Contig genome coverage was calculated from the sum of contig lengths.

3.2.8.7 Virus abundance and relative abundance

To determine viral contig abundances, reads were mapped against the dereplicated contigs using Bowtie 2 with default settings. Using SAMtools command idxstats, the number of reads mapping to each contig were extracted. Contig abundance was calculated on the sample and virome levels, by normalising the number of reads mapping to each contig by the contig length in kb and millions of reads (RPKM) in the total sample and the virome, respectively. The virome

was defined as those contigs identified as viral by geNomad, excluding mock community sequences to maintain consistency between replicates with different spike-in treatments. Relative abundance was then calculated as a fraction of the cumulative total and virome abundance, respectively.

3.2.8.8 Variation in relative abundance and rank

To calculate the variation in abundance, relative abundance, and abundance ranking, reads from each of the individual replicates were mapped to the respective co-assemblies. Abundance and relative abundance were calculated on the sample and virome levels as described above. Contigs with at least one read in each of the replicates were included in the analysis. The range in abundance was calculated as the log2-transformed ratio of the highest and lowest abundance of each contig in three replicates. Abundance rank was calculated by sorting contigs by their abundance from highest to lowest in each replicate, and the range was determined as the difference between the highest and lowest ranking of a contig among the three replicates. The coefficient of variation in relative abundance was calculated as the coefficient of variation of the relative abundance of the contig in the three replicates.

3.2.8.9 Virome taxonomic and diversity analysis

Alpha diversity based on assemblies and co-assemblies was calculated using the R package phyloseq v.1.46 (McMurdie and Holmes, 2013). Beta diversity was calculated using the R package vegan v.2.6-4, by calculating the Bray-Curtis dissimilarity between pairs of the HI, LO, and NO replicates of each stool sample based on the co-assembly data. The top taxonomic ranks were calculated and displayed using the R package fantaxtic v.0.2.0 (Teunisse, 2022).

3.2.8.10 Data analysis and visualization

Data was analysed in R v.1.4.3 with RStudio v.2023.06.1. Data was handled using tidyverse v.2.0.0 (Wickham *et al.*, 2019) packages and displayed using the packages ggplot2 v.3.4.4 (Wickham, 2016), ggExtra v.0.10.1 (Attali and Baker, 2023), ggpubr v.0.6.0 (Kassambara, 2023), ggsci v.3.0.0 (Xiao, 2023) and ggh4x v0.2.6 (van den Brand, 2023).

3.3 Results

Using the virus stock EFM measurements obtained in Chapter 2, a high-concentration MC and a low-concentration MC were composed, with an equivalent of 10^7 and 10^5 particles per sample, respectively, for each of the component viruses. To assess differences in recovery efficiency between viruses, compare recovery of individual viruses between samples, determine an appropriate spike-in concentration for further viral metagenomics research, and assess sequence bias and stool virome variation, the MC was added to aliquots at high concentration (HI) or low concentration (LO), while a third aliquot was left unaltered (NO) (Fig. 3.1A). Viral nucleic acid extracts of each sample were process using WTA2 and SISPA

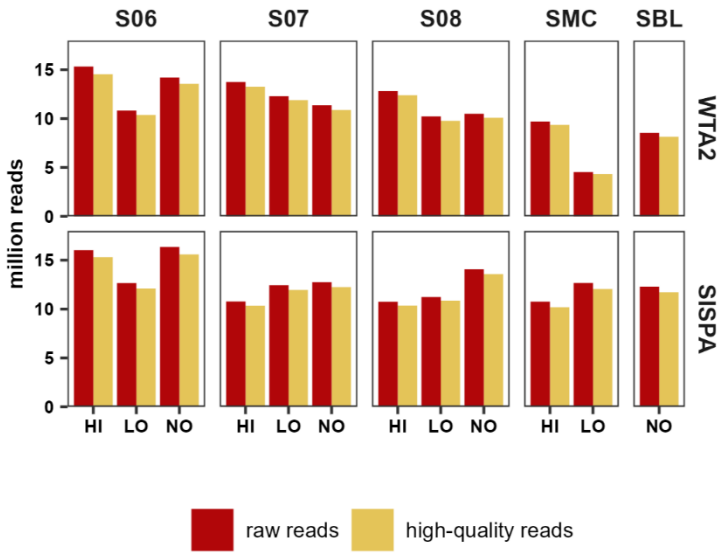

Figure 3.4: WTA2 and SISPA yield equivalent numbers of high-quality reads. The graphs depict million reads for stool sample aliquots and from mock-community-only and blank samples, processed using WTA2 and SISPA

methods for RT and PCR amplification of nucleic acids (Fig. 3.2) and sequenced. Sequencing of WTA2 and SISPA samples returned an average of 12 million reads per sample. A custom bioinformatics pipeline was employed to analyse the sequencing data, including removal of host sequences, quality filtering of reads, assembly of reads into contigs, and classification of contigs (Fig. 3.3). After quality control, an average 4.1% of reads was removed (Fig. 3.4). WTA2 and SISPA returned equivalent numbers of reads. However, mapping high-quality reads to MC reference genomes in spiked stool samples showed a 2- to 45-fold higher fraction of MC virus reads in WTA2 samples compared to SISPA (Table 3.3). In stool samples spiked with MC HI, an average of 3.2% of reads in the WTA2 samples and an average of 0.77% reads in the SISPA samples mapped to MC viruses. In the SMC-HI and -LO samples, the percentage of MC reads was similar between WTA2 and SISPA libraries, as was the percentage of MC reads in the blank sample.

3.3.2 Recovery bias and consistency of mock virus community

Reads were mapped to the MC virus reference genomes to determine variation in virus recovery between samples and replicates, and differences in overall recovery between viruses. There was considerable variation in virus levels between samples for all viruses, although for most viruses, the same pattern was observed for both WTA2 and SISPA libraries (Fig. 3.5A). The exceptions were M13, which was increased, and RV-A, which was virtually absent in SISPA libraries.

Table 3.3: WTA2 yields more MC virus reads than SISPA.

		Reads			Reads per 10,000		
Sample	**Method**	**HI**	**LO**	**NO**	**HI**	**LO**	**NO**
S06	WTA2	316,842	3,308	302	218	3.19	0.22
	SISPA	67,649	426	6	44	0.35	0.00
S07	WTA2	464,892	4,231	64	350	3.56	0.06
	SISPA	181,895	654	7	175	0.55	0.01
S08	WTA2	481,952	2,304	107	389	2.36	0.11
	SISPA	11,254	57	6	11	0.05	0.00
SMC	WTA2	4,311,857	68,580	-	4,643	158.23	-
	SISPA	5,228,808	180,368	-	5,152	149.64	-
SBL	WTA2	-	-	39303	-	-	48.26
	SISPA	-	-	60115	-	-	51.35

HI: sample spiked with high MC HI, LO: sample spiked with MC LO, NO: non-spiked sample, SISPA: sequence-independent single primer amplification, WTA2: whole transcriptome amplification kit.

Based on the number of particles in the MC and single- or double-stranded genome and the genome length of each virus, an expected percentage of reads of total MC reads was calculated. Virus levels were closest to the theoretical level in the MC-only samples, particularly in the low concentration MC. However, there were large differences between the observed and expected levels of all viruses in all samples. Phages M13 and P22 were detected up to 22 and 7 times higher, respectively, than expected, while all other viruses were reduced. Levels of the enveloped eukaryotic viruses MHV-68 and BVDV-1 were lowest, particularly BVDV-1 with at most a few reads present in only some samples.

To determine whether observed differences in virus abundance were accurate, the abundance of MC viruses in the stool samples spiked with MC HI and MC LO was compared. Since MC HI was 100-fold more concentrated than MC LO, around a 100-fold increase would be expected in the HI samples. While the ratio was close to 100 for some viruses in some samples, large differences were observed, particularly in the SISPA libraries (Fig. 3.5B). In the WTA2 libraries, the ratios were between 41 and 225 for phages, while the ratios were lower for the eukaryotic viruses, except for RV-A which was close to 100 in two samples. In the SISPA libraries, the ratios were much greater, up to 1089 for phage M13, likely reflecting the lower MC virus read counts in the SISPA libraries. Ratios for RV-A and BVDV-1 could not be calculated in SISPA due to absence of reads in at least one of the sample pairs. The increased ratio and variation in ratio in the SISPA libraries was likely to be due to the much lower read numbers mapping to MC viruses.

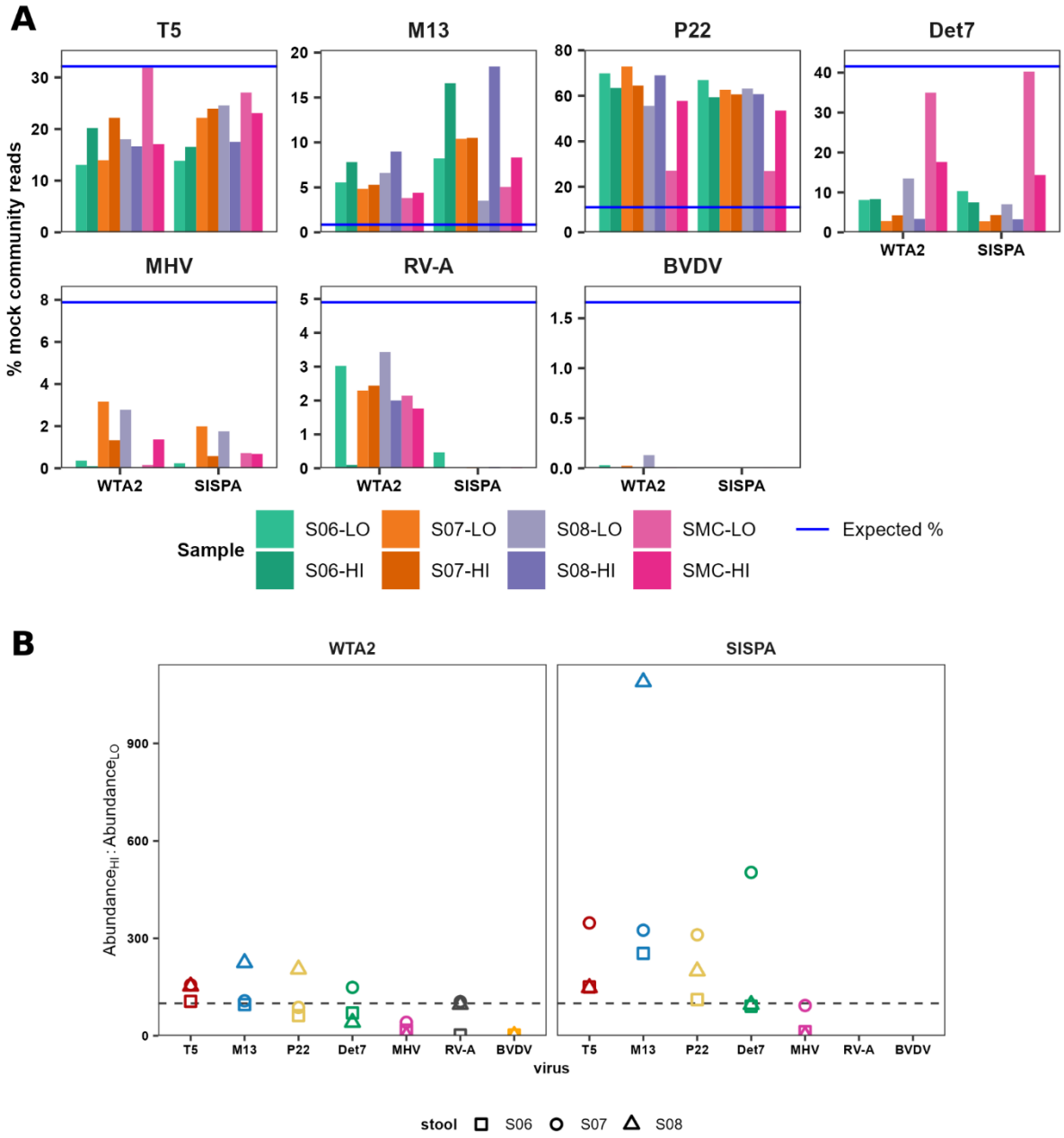

Figure 3.5: Recovery bias and consistenty of MC viruses. A: Percentage of reads mapping to each MC virus, comparing samples processed using WTA2 to SISPA. The blue horizontal line shows the expected fraction of reads mapping to each virus based on the number of particles added, the total genome length and number of strands. **B:** Ratio of virus abundance (read counts normalised for genome length in kb and millions of reads (RPKM)) for each virus in the samples spiked with mock communities HI ($Abundance_{HI}$) and LO ($Abundance_{LO}$).

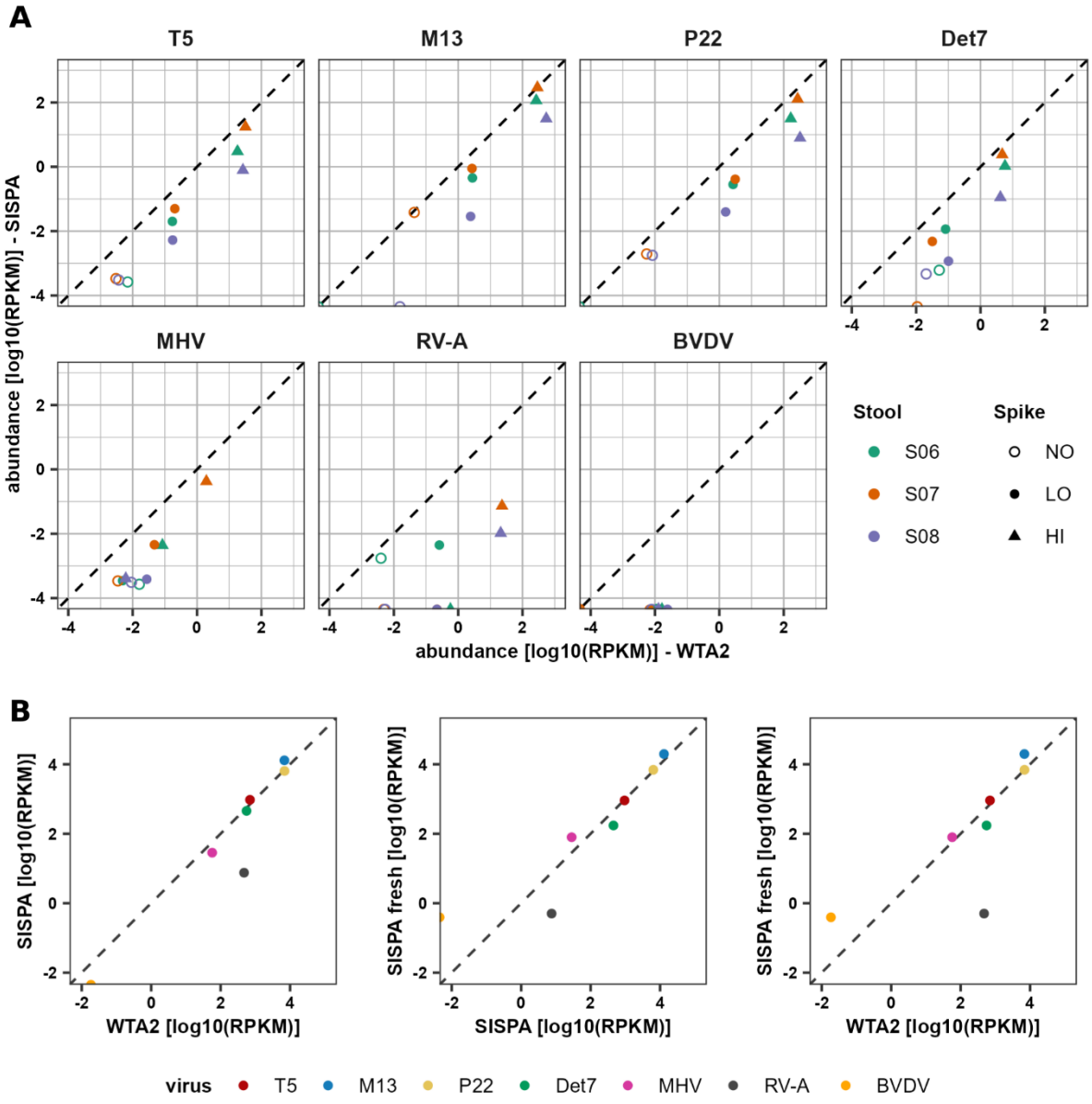

Figure 3.6: Abundance of MC viruses in WTA2 and SISPA samples. A: Comparison of the abundance of each virus in the two spiked and one non-spiked aliquots of the three stool samples, generated using the WTA2 and SISPA methods. The diagonal dashed line depicts a perfect correlation. **B:** Comparison of MC virus abundance in the MC HI control sample processed using WTA2 and SISPA. After extraction of VLPs from (spiked) stool samples, VLP extracts were frozen overnight at -20°C. Additionally, after nucleic acid extracts using WTA2, the extracts were kept at -20°C for 10 months before processing using SISPA. Therefore, a new HI MC sample was processed using SISPA and sequenced. The first plot compares the normalised read counts of the original sample processed using WTA2 and SISPA.

Next, virus recovery in the WTA2 and SISPA libraries was compared (Fig. 3.6A). Overall, abundance in WTA2 libraries was higher than in SISPA libraries, while variation between

samples was lower. Comparing the abundance of the HI and LO replicate of each sample indicated a linear relationship and around a 100-fold difference between the two. Virus reads were also detected in the non-spiked aliquots, possibly indicating erroneous mapping, mapping of reads from conserved regions of related stool viruses, or cross-contamination from spiked samples. The abundance of phages T5 and P22 in those samples was around two orders of magnitude below those spiked with MC LO in the WTA2 libraries and around one order of magnitude in the SISPA samples, possibly due to the lower number of MC reads in the SISPA libraries.

Comparing the abundances of the MC viruses in the MC HI-only WTA2 and SISPA samples showed that abundances of all viruses, except RV-A and BVDV-1, were in close agreement (Fig. 3.6B). In the SISPA library, RV-A abundance was lower, and BVDV-1 was not detectable. As VLP extracts were frozen for 16-18 hours before nuclease treatment and viral nucleic acid extraction, and nucleic acid extracts were stored at -20 °C for over nine months between production of the WTA2 and SISPA libraries, a fresh sample of MC HI was processed to assess the effects of both overnight freezing and long-term nucleic acid storage. For this, a slightly adapted method was used, substituting syringe filters for a centrifugal filter column, and using a different nuclease cocktail. In the fresh MC sample, RV-A levels were lower than the original WTA2 and SISPA libraries, while BVDV-1 levels were higher. Thus, freezing of the VLP extract negatively affected recovery of BVDV-1 RNA. The decrease in RV-A levels in the fresh MC HI suggests degradation of the virus stock.

3.3.3 Sequence bias of WTA2 and SISPA libraries

To determine sequence bias of WTA2 and SISPA, the uniformity, and consistency of coverage depth profiles of the MC virus genomes was investigated. For T5, P22, Det7, and M13 >80% coverages were obtained in the WTA2 and SISPA libraries of at least two samples. Additionally, >80% was obtained for MHV-68 and segment 1 of RV-A in the WTA2 libraries for at least two samples. In both WTA2 (Fig. 3.7) and SISPA (Fig. 3.8) libraries, the coverage depth profile was highly non-uniform, but consistent between replicates. Coverage depth profiles of WTA2 and SISPA showed similarity in the high-depth regions, but higher peaks were observed in the SISPA libraries. (Fig. 3.9A). This was reflected in the coefficient of variation (CV) of the coverage depth of MC virus genomes in the WTA2 and SISPA libraries, with a mean CV for Det7, P22, T5 and M13 in all samples of 87% in the WTA2 libraries and 160% in the SISPA libraries (Fig. 3.9B). The CV was higher for WTA2 than SISPA only for RV-A segments 6 and 10 in sample S06.

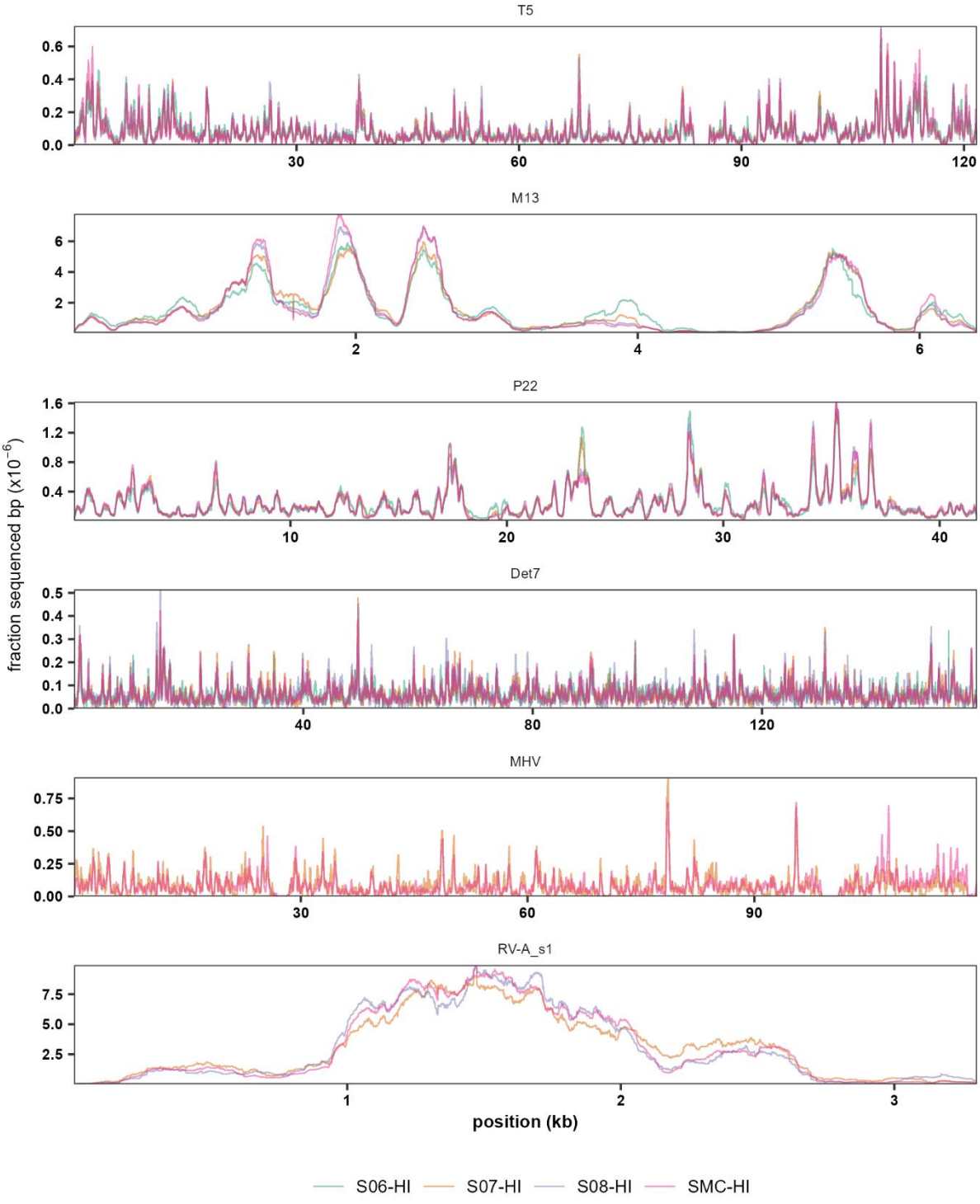

Figure 3.7: Consistency of WTA2 sequencing depth of virus genomes. Sequencing depth normalised by the total number of sequenced based in for each virus in each sample. Only data for samples with >80% genome coverage are displayed. For RV-A, only the depth of the first segment (RV-A_s1) is shown.

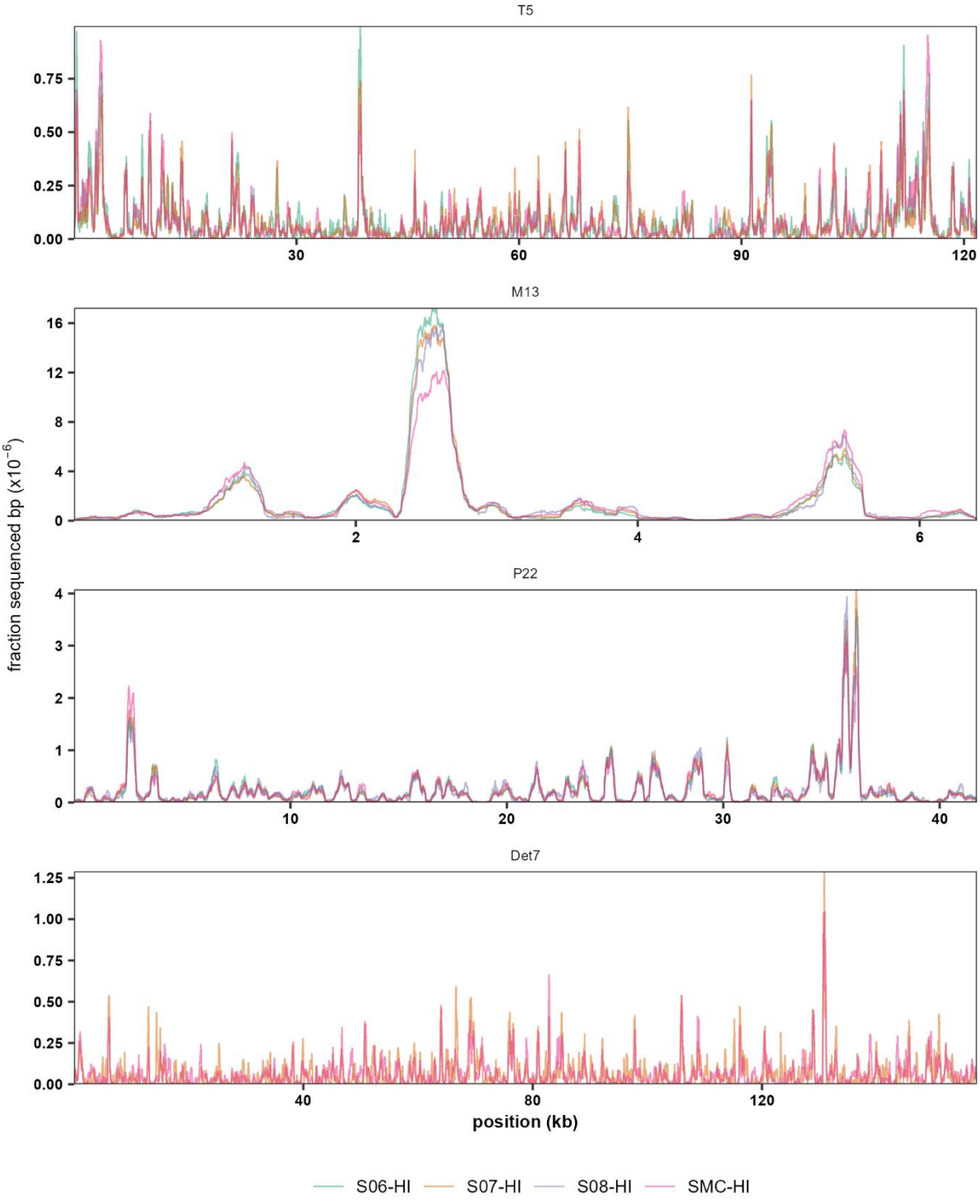

Figure 3.8: Consistency of SISPA sequencing depth of virus genomes. Relative sequencing depth, calculated as the fraction of sequenced bases for each virus in each sample mapping to each position. Only data for samples with >80% genome coverage are displayed.

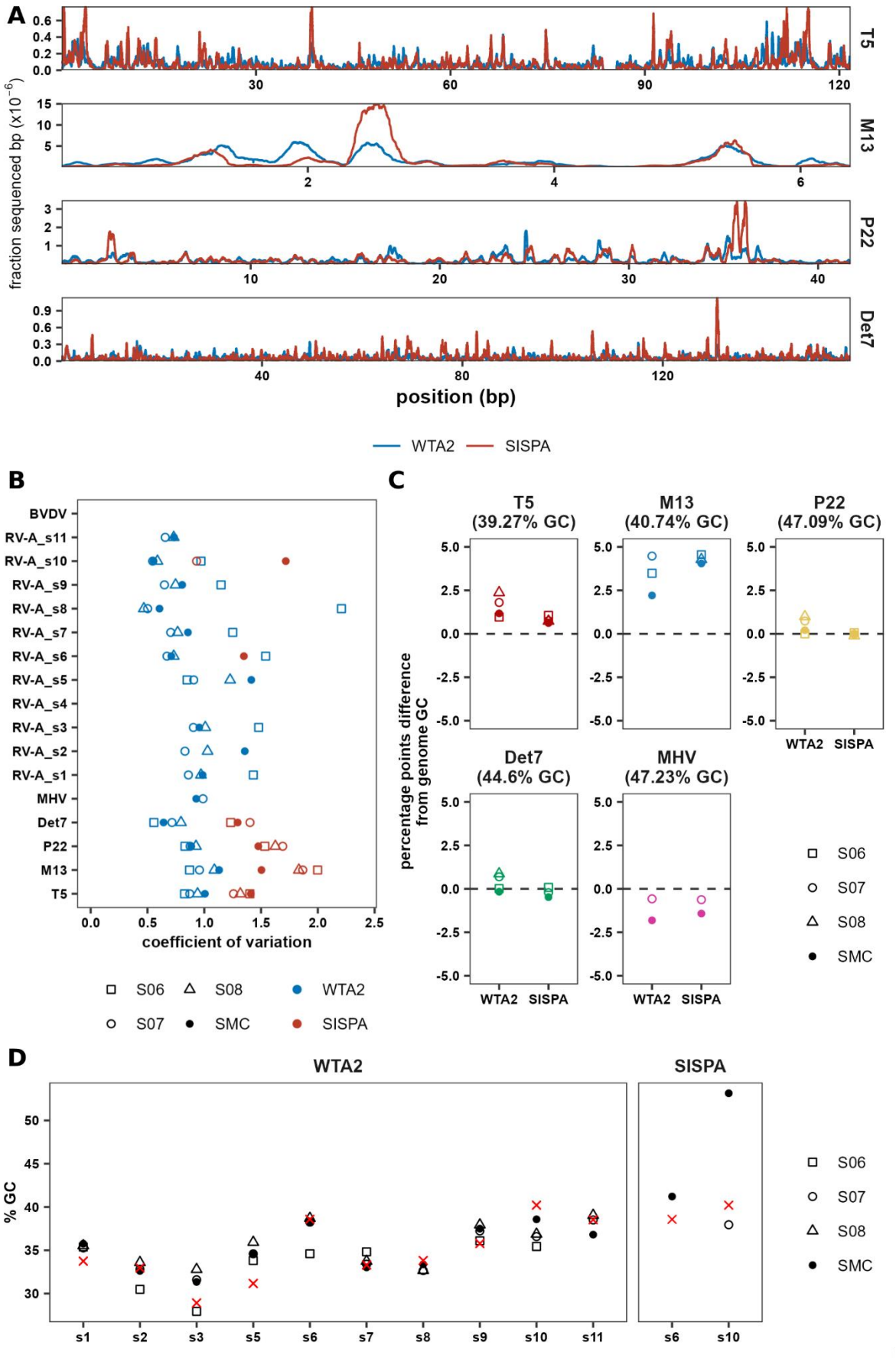
A
T5
M13
P22
Det7
fraction sequenced bp (x10⁻⁶)
position (bp)
WTA2
SISPA
B
BVDV
RV-A_s11
RV-A_s10
RV-A_s9
RV-A_s8
RV-A_s7
RV-A_s6
RV-A_s5
RV-A_s4
RV-A_s3
RV-A_s2
RV-A_s1
MHV
Det7
P22
M13
T5
coefficient of variation
S06
S07
S08
SMC
WTA2
SISPA
C
T5 (39.27% GC)
M13 (40.74% GC)
P22 (47.09% GC)
Det7 (44.6% GC)
MHV (47.23% GC)
percentage points difference from genome GC
D
WTA2
SISPA
% GC
s1
s2
s3
s5
s6
s7
s8
s9
s10
s11

Figure 3.9: Sequence bias of WTA2 and SISPA. A: Comparison of the mean of the normalised sequencing depth of T5, M13, P22 and Det7 in WTA2 and SISPA libraries. **B:** Coefficient of variation of sequencing depth in neat and spiked MC HI. **C:** Difference in the average GC-content of reads mapping to reference genomes of viruses with a non-segmented genome, compared to the actual GC-content of the virus genome, for samples with >20% genome coverage. **D:** Comparison of read GC-content and RV-A segment GC content for samples with >20% coverage of the segment. Red crosses mark the GC content of each RV-A segment. Segment 4 did not achieve >20% coverage in any samples produced by WTA2 and SISPA, while only segments 6 and 10 were sufficiently covered in the SISPA samples.

Sequence bias was further investigated by comparing the for GC-content of virus reads to the genomic GC-content in samples with >20% genome coverage. For DNA viruses with a genomic GC-content of <41%, the read GC-content was higher than the virus genome GC-content, while for phages P22 and Det7 the GC-content was close to the actual genomic GC-content, and for MHV-68, with the highest genomic GC-content, the read GC-content was below the genomic GC-content (Fig. 3.9C). For all DNA viruses, the read GC-content was more consistent between replicates for the SISPA libraries than the WTA2 libraries. For the dsRNA virus RV-A, the genomic GC-content of genome segments varied between 28.9% and 40.2%. In the case of the WTA2 libraries, the segments with a GC-content <36%, read GC-content in most replicates was equal to or higher than the genomic GC-content. On the other hand, for the segments with a GC-content >36%, the read GC-content was equal to or lower than the genomic GC-content in most samples. In the case of the SISPA libraries, only two segments were sequenced to sufficient coverage and no consistent pattern could be determined.

3.3.4 Assembly quality of WTA2 and SISPA libraries.

To determine differences in assembly quality between WTA2 and SISPA, both single sample assemblies, as well as co-assemblies of the replicates of each stool samples were generated from WTA2 and SISPA libraries. Assembly of WTA2 libraries consistently yielded more contigs overall (Fig. 3.10A) and more contigs larger than 10 kb (Fig. 3.10B). Similarly, the three co-assemblies of the WTA2 libraries yielded more contigs than the SISPA libraries (Fig 3.10C, 3.10D). The number of contigs in the co-assembly of the SISPA libraries was between 14% and 56% lower than the WTA2 library co-assemblies. The N50, which is the length at which all contigs of that length or longer compose 50% of the total assembly, was slightly higher for the WTA2 than SISPA libraries, indicating a higher overall quality of assemblies (Table 3.4) and co-assemblies (Table 3.5). Additionally, the largest contig in samples S06 and S08 were 202 kb and 119 kb in the WTA2 co-assemblies and 75 kb and 40 kb in the SISPA co-assemblies, respectively. For SISPA library assemblies, the mean depth of the contigs was between 2.4 and

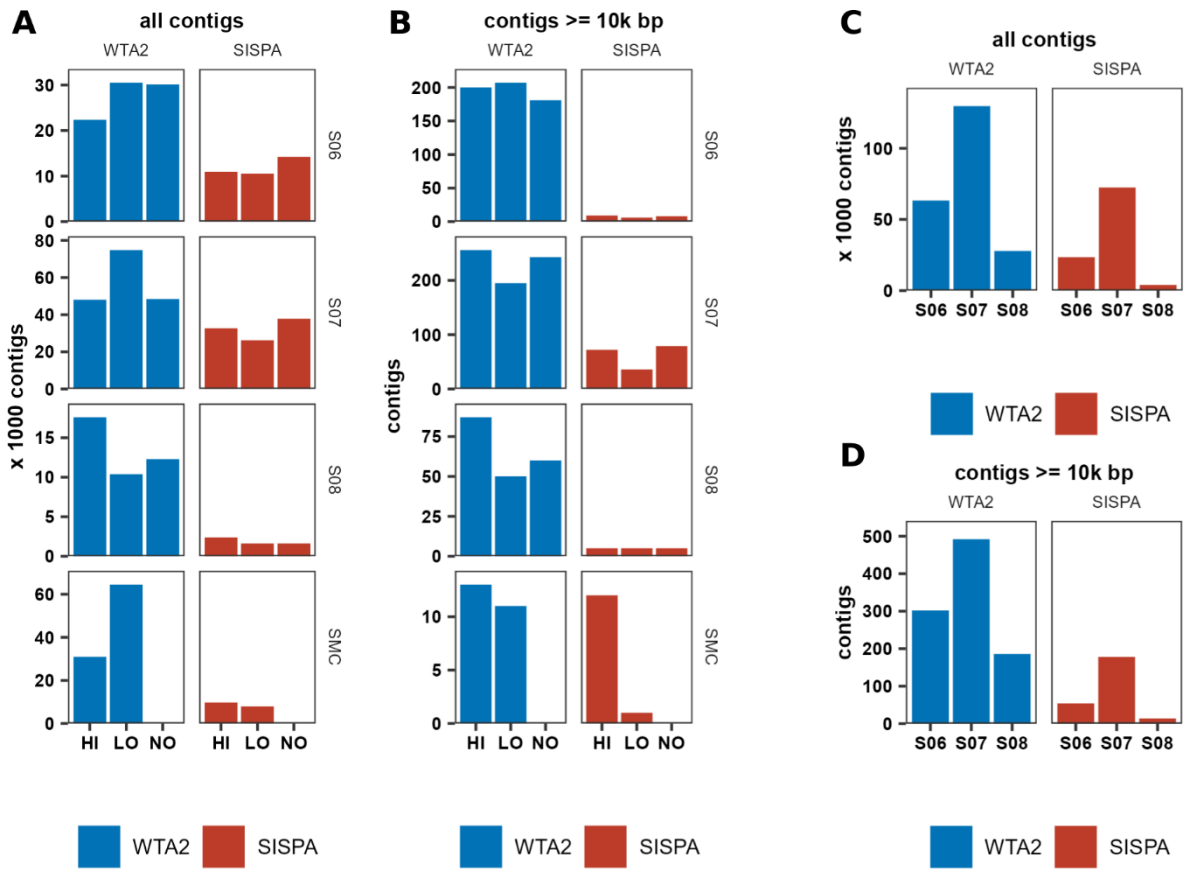

Figure 3.10: WTA2 samples yield more and larger contigs than SISPA. **A:** Total number of contigs in single assemblies of each of the samples. **B:** Total number of contigs ≥ 10 kb in single assemblies. **C:** Total number of contigs in the co-assemblies generated from the three aliquots of each of the stool samples. **D:** Total number of contigs ≥10 kb in length in the co-assemblies generated from the three aliquots of each of the stool samples.

18.4 times higher than WTA2 library assemblies. The 250 longest contigs in the WTA2 library were on average 3, 2, and 6 times longer in WTA2 than SISPA co-assemblies of samples S06, S07, and S08, respectively, suggesting reduced fragmentation and increased diversity in the WTA2 co-assemblies. Overall GC-content of assemblies was more consistent for assembled SISPA libraries than WTA2 libraries (Table 3.4).

Next, geNomad (Camargo *et al.*, 2023) was used to identify viral sequences, and sequences were aligned to reference genomes to identify MC virus sequences. In assemblies and co-assemblies of both WTA2 and SISPA libraries, a large fraction of contigs could not be identified as viral, indicating non-viral sequences and false negatives. The relative abundance of the virus sequences was calculated by mapping reads of individual samples to the contigs of the respective assemblies and co-assemblies. Taken samples overall and ignoring mock community contigs, 28% and 29% of contigs were classified as viral in WTA2 and SISPA co-

Table 3.4: Assembly statistics for the WTA2 and SISPA libraries of spiked stool samples and MC-only samples.

Sample	Spike	# contigs (≥10 kbp)	# contigs (x1000)	Total length (Mbp)	GC (%)	N50 (bp)	L50 (#)
WTA2							
S06	HI	200	22.4	20.5	43.56	1278	2145
	LO	207	30.5	24.6	49.22	924	4383
	NO	181	30.1	23.7	47.36	885	4599
S07	HI	256	48.0	39.4	49.21	975	7208
	LO	195	74.8	53.8	48.69	767	15048
	NO	243	48.5	37.6	51.09	864	7676
S08	HI	87	17.6	15.1	48.11	1129	2382
	LO	50	10.4	8.2	46.36	894	1668
	NO	60	12.3	10.2	47.26	991	1836
SMC	HI	13	30.9	13.8	42.36	420	11126
	LO	11	64.5	28.5	48.12	414	23355
SISPA							
S06	HI	9	10.9	7.8	45.57	823	2290
	LO	6	10.5	6.8	46.21	685	2451
	NO	8	14.2	9.0	46.46	667	3230
S07	HI	72	32.6	21.2	46.89	645	6839
	LO	36	26.2	16.1	46.60	607	6271
	NO	79	37.8	23.6	47.04	600	8291
S08	HI	5	2.4	1.6	46.13	739	435
	LO	5	1.6	1.1	46.09	738	300
	NO	5	1.6	1.2	45.40	778	288
SMC	HI	12	9.8	4.2	49.74	399	3539
	LO	1	7.9	3.3	51.03	395	3042

Table 3.5: Co-assembly statistics for the WTA2 and SISPA libraries of spiked stool samples.

Sample	# contigs (≥10 kbp)	# contigs (x1000)	Total length (Mbp)	GC (%)	N50 (bp)	L50 (#)
WTA2						
S06	302	63.3	51.8	48.08	963	9651
S07	492	129.6	101.4	49.77	875	22994
S08	186	27.8	23.3	47.69	1135	3174
SISPA						
S06	54	23.5	16.7	46.37	783	4027
S07	178	72.4	46.3	47.37	632	16005
S08	14	4.0	2.7	46.22	684	793

assemblies, respectively. Stool viruses contributed on average 21.8% and 10.6% of the relative abundance in the WTA2 and SISPA libraries, respectively. While SISPA (co-)assemblies had a

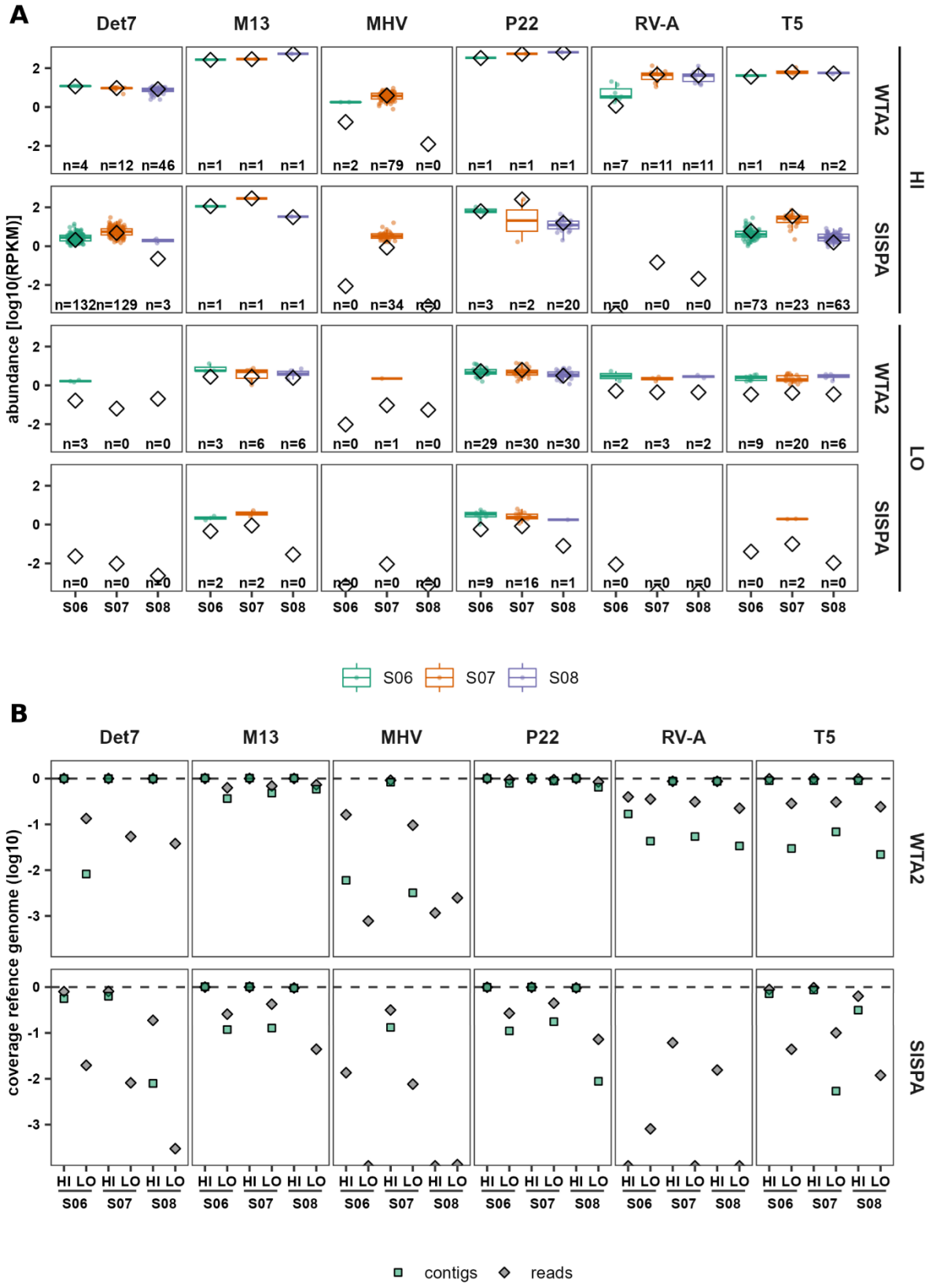

Figure 3.11: Assembly-based abundance is overestimating actual abundance when contig coverage of the genome is low. **A:** Abundance (RPKM) of MC virus contigs from samples spiked with MC HI (HI) and MC LO (LO), compared to the virus abundance based on reads mapping directly to the virus reference genome (diamonds). Numbers along the x-axis represent the number of contigs. **B:** Genome coverage by contigs in the assembly (green squares) and total mapped reads (grey diamonds) in each sample. Dashed horizontal lines represent 100% coverage.

lower proportion of MC virus contigs, the collective abundance of MC contigs relative to all viral contigs was similar for both WTA2 and SISPA libraries and between 0.8% and 5.43%. On the other hand, the relative abundance of contigs not identified as viral was higher in the SISPA than the WTA2 library co-assemblies.

3.3.5 Accuracy of assembly-based virus abundance

The relative abundance of MC contigs was compared to the reference genome-based abundances of the MC viruses to determine the accuracy of assembly-based virus abundances. In the assemblies of WTA2 libraries, samples spiked with MC HI yielded between one and 46 MC contigs for phages T5, Det7, P22 and M13 (Fig. 3.11A), all covering ≤90% of the reference genome (Fig. 3.11B). For phages T5, Det7 and P22, the SISPA libraries yielded more contigs, while genome coverage of contigs was lower for T5 and Det7. Coverage for Det7 varied between 0.79% and 63%, and coverage for T5 between 31% and 87%, indicating reduced fragmentation in the WTA2 library assemblies. For the eukaryotic viruses, the coverage was typically much lower in both libraries, in line with the reduced abundance of these viruses as determined by direct read mapping. The exception was a single WTA2 library which yielded 79 MHV-68 contigs, covering 83% of the genome.

In the samples spiked with MC LO, more contigs were found for each virus with lower total coverage, suggesting higher fragmentation of the genome in these assemblies. Additionally, several samples did not yield any contigs for Det7, M13, MHV-68 RV-A and T5 in the WTA2 and SISPA libraries. This suggests the lower limit for reliable assembly-based detection of viruses in this experiment is for viruses with an abundance $>10^5$ particles in 50 mg of stool. For both libraries, samples spiked with MC LO resulted in lower genome coverage.

In both WTA2 and SISPA libraries, abundance of contigs in fragmented genomes varied up to 10-fold between the highest and lowest contig abundance. When multiple contigs aligned to the genome, there was up to a 32-fold difference between the highest and lowest abundance. Nonetheless, when contigs collectively covered a ≥90% of the reference genome, the mean abundance was close to the reference abundance. However, abundance was overestimated

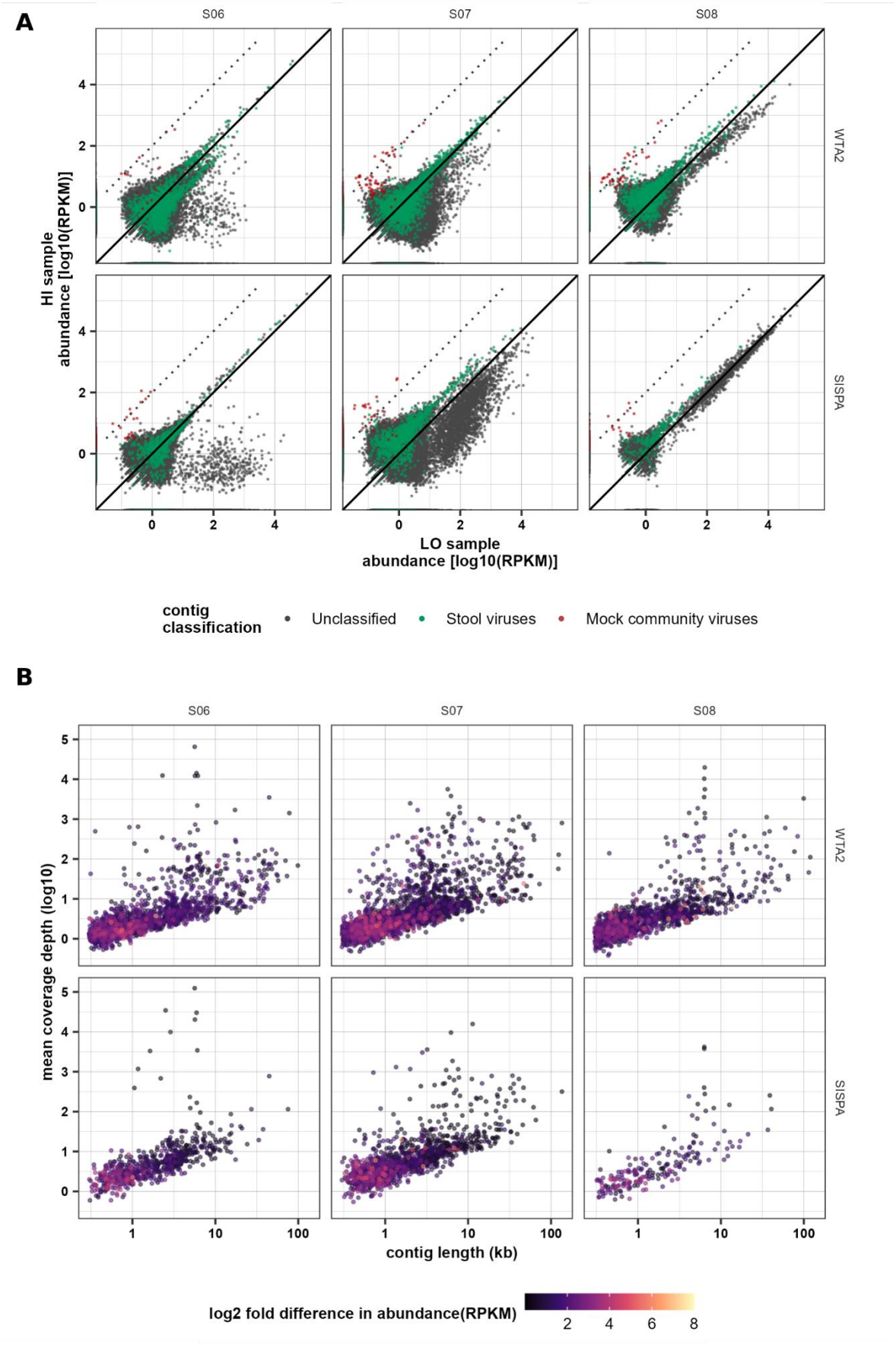
A
S06
S07
S08
WTA2
SISPA
HI sample abundance [log10(RPKM)]
LO sample abundance [log10(RPKM)]
contig classification
Unclassified
Stool viruses
Mock community viruses
B
mean coverage depth (log10)
contig length (kb)
log2 fold difference in abundance(RPKM)

Figure 3.12: Variation in virus abundance between replicates is lowest for high abundance contigs. **A:** Comparison of normalised read counts (RPKM) of stool samples spikes with MC HI and LO. The black solid line depicts perfect agreement. The black dotted line indicates a 100-fold increase in the HI sample over the LO sample. **B:** Mean coverage depth of virus contigs, calculated from the coverage depth of the three replicates, compared to the length of the contig. The colour scale shows the log2-transformed ratio between the highest and lowest abundance in RPKM.

when contigs covered a smaller portion of the reference genome. At a maximum overestimation of 100%, 50%, 20% and 10%, contigs covered collectively at least 31%, 56%, 63%, and 72%, respectively, of the genome. This happened despite genome coverage of reads always being higher than of contigs, with up to 30-fold higher read coverage than contig coverage. This is because the read counts are normalized by the full genome length rather than the contig length, the reference-based abundance is lower.

3.3.6 Consistency of assembly-based virus abundance

Variation in relative abundance of stool virus contigs across replicates was determined by mapping reads of each of the three replicates of the stool samples to the respective co-assemblies. Comparing the abundance of viral contigs between HI and LO replicate, variation in abundance decreased with increasing abundance (Fig. 3.12A). In samples S06 and S07, a large group of unclassified contigs was present at 10- to 100-fold higher concentration in the LO replicate. This difference was particularly pronounced in the SISPA libraries of sample S07. MC virus relative abundance in the HI replicate was around 100-fold higher than in the LO replicate, consistent with the difference in concentration between spike-ins.

To obtain a parameter to distinguish sequences with high and low abundance variation, the ratio between the highest and lowest abundance of the three replicates was compared against the contig length and its coverage (Fig. 3.12B). Only viral contigs, but not MC contigs, were included in this analysis. While there was a linear relationship between the contig length and coverage depth, the abundance range was limited more by the coverage depth than the length. Contigs with low variation were present in the entire range of contig length and coverage depth. However, nearly all contigs with high variation (≥4-fold difference between highest and lowest abundance) had a mean coverage depth of <10X. For contigs with ≥10X coverage, the median ratio between the highest and lowest abundance was 1.7 and 1.4 for the WTA2 and SISPA co-assemblies, compared to 2.5 and 2.4 for contigs with <10X coverage, respectively.

In addition to abundance, abundance-based ranking was considered by ordering contigs from highest to lowest abundance in each replicate. The difference between highest and lowest

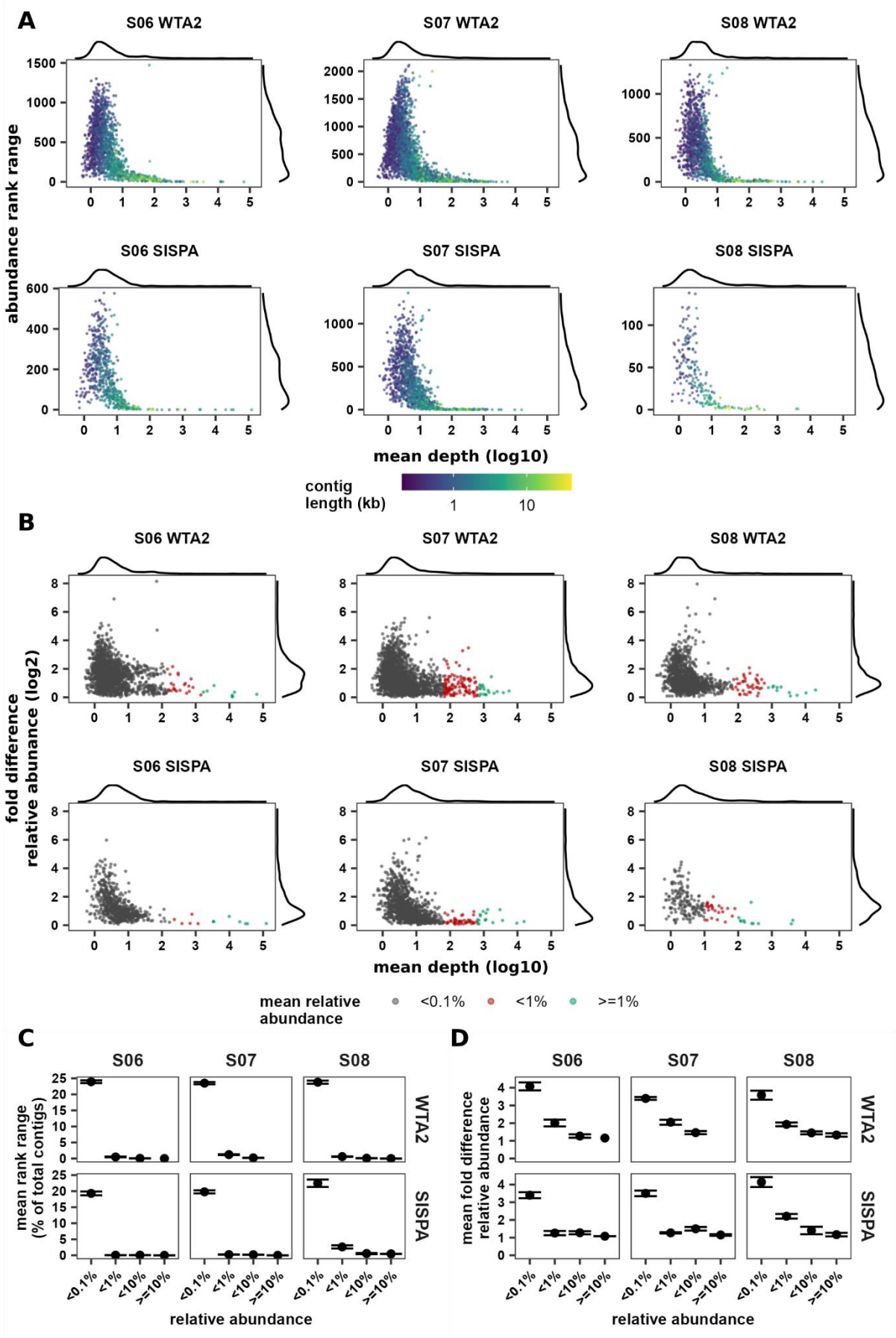
A
S06 WTA2
S07 WTA2
S08 WTA2
S06 SISPA
S07 SISPA
S08 SISPA
abundance rank range
mean depth (log10)
contig length (kb)
1
10
B
fold difference relative abunance (log2)
mean depth (log10)
mean relative abundance
<0.1%
<1%
>=1%
C
S06
S07
S08
WTA2
SISPA
mean rank range (% of total contigs)
<0.1%
<1%
<10%
>=10%
relative abundance
D
mean fold difference relative abundance
relative abundance

Figure 3.13: Abundance-based ranking and relative abundance variation of contigs. The abundance, abundance rank, and relative abundance are calculated for viral contigs in the stool sample co-assemblies. Only contigs detected in all three samples are included. **A:** The difference between the highest and the lowest rank of the three replicates of each stool sample is plotted against the mean coverage depth of the contig in the three replicates. Since the maximum rank range depends on the number of contigs, the y-axis is scaled differently for each sample. Graphs along the x- and y-axis show the density along the respective axis. **B:** Log-transformed ratio between the highest and lowest relative abundance of the three replicates, compared to mean coverage depth. Graphs along the x- and y-axis show the density along the respective axis. **C, D:** Variation in abundance-based ranking (**C**) and relative abundance (**D**), measured as the mean range and fold difference, respectively, of contigs binned by relative abundance <0.1%, ≥0.1% and <1% (<1%), ≥1% and <10% (<10%), and ≥10% (≥10%).

ranking was calculated, showing an increase in the range for contigs below 10X coverage depth (Fig. 3.13A). This leads to an average range of 78 and 70 places on the abundance ranking for high-coverage contigs and 569 and 315 places in the low-abundance contigs in the WTA2 and SISPA co-assemblies, respectively. As expected, given the linear relationship between coverage depth and contig length (Fig. 3.12B), the longer contigs tended towards higher coverage depth, while there was no clear relationship between abundance ranking and contig length (Fig. 3.13A).

The same pattern was observed for the relative abundance of contigs. To estimate variation in abundance, the ratio between the highest and lowest relative abundance in the three replicates was calculated. The majority of contigs had low coverage (<10X), with between 1- and 4-fold variation between highest and lowest relative abundance (Fig. 3.13B). The lowest variation occurred for contigs with a mean coverage depth ≥10X. Overall, variation in high coverage depth contigs was higher between WTA2 libraries than SISPA libraries.

When grouping contigs into bins by the log10-transformed coverage depth in bins of 0.5, the highest median ratio was found in contigs with ≤1X mean coverage (3.2 and 3.8 in WTA2 and SISPA co-assemblies, respectively). The first bin with a median ratio below 2 was the 10X bin for both libraries, while the lowest median variation was in the ≥1000X coverage bin, with a 1.3- and 1.2-fold difference between the highest and lowest relative abundance in the WTA2 and SISPA co-assemblies, respectively.

Although most contigs have a coverage depth <10X, they only contributed a small percentage of viral reads, between 0.42% and 2.1% of reads across all libraries. Across replicates, contigs

with <10X coverage depth represented 2.9%, 9.6% and 6.2% viral relative abundance in WTA2 libraries and 1.0%, 7.1% and 4.8% in SISPA for samples S06, S07 and S08, respectively. Additionally, nearly all contigs with a relative abundance ≥0.1% had ≥100X coverage, except for SISPA sample S08 (Fig. 3.13B). As a result, the range in abundance-based ranking was at most three places in both the WTA2 and SISPA libraries for contigs with an average relative abundance ≥1%, and zero contigs with an average relative abundance ≥10%, except SISPA sample S08 (Fig. 3.13C). Additionally, there was an average 1.4- and 1.3-fold difference between the highest and lowest relative abundance in the WTA2 and SISPA libraries, respectively for contigs with ≥1% average relative abundance (Fig. 3.13D). This means that for the higher abundance contigs, the abundance ranking of contigs with a relative abundance ≥1% is consistent across replicates, even though there is up to 40% difference between the highest and lowest relative abundance.

3.3.7 Taxonomic analysis and sample diversity

The taxonomic content of the individual sample assemblies (Fig. 3.14A) of WTA2 and SISPA libraries was determined using the taxonomic classifications generated by geNomad, excluding MC sequences. Virtually all contigs that were identified as viral could be classified down to the class level, although the classification rate at lower clades was considerably lower, with on average only 8% and 4% of viral sequences classified at the order and family level, respectively. Overall, the phyla of *Phixviricota*, *Uroviricota*, *Pisuviricota* and *Cressdnaviricota, Kitrinoviricota*, were the five most abundant among all samples in both libraries. The phyla *Phixviricota* contain phages with small protein capsids and small circular ssDNA genomes, while *Uroviricota* are tailed phages with large dsDNA genomes. While phages of the phylum *Phixviricota* were the most abundant viruses in samples S06 and S08 for both WTA2 and SISPA libraries, abundance of *Phixviricota* was greater in the SISPA libraries than WTA2 libraries. For sample S07, the difference between the libraries was much greater, with a mean relative abundance 22.6% and 56.8% in the WTA2 and SISPA libraries, respectively. The phyla *Pisuviricota* (ssRNA(+) genomes), *Cressdnaviricota* (circular ssDNA genomes), and *Kitrinoviricota* (ssRNA) infect eukaryotic cells. In sample S07, *Kitrinoviricota* and *Cressdnaviricota* were more abundant in the SISPA library than the WTA2 library. *Kitrinoviricota* and *Cressdnaviricota* had 4.0% and 4.5% relative abundance, respectively, in the SISPA library, while the WTA2 libraries of sample S07 contained 0.26% and 0.54% of these phyla, respectively. For *Cressdnaviricota*, the pattern was reversed in samples S06 and S08, with 0.064% and 0.15% relative abundance in the WTA2 library, and 0.013% and 0.033% in the SISPA libraries, respectively. Such large differences were not found for *Kitrinoviricota* in the other two stool samples. Lastly, the phylum *Pisuviricota* had a consistently higher relative abundance in the WTA2 libraries than the SISPA libraries of all three stool samples.

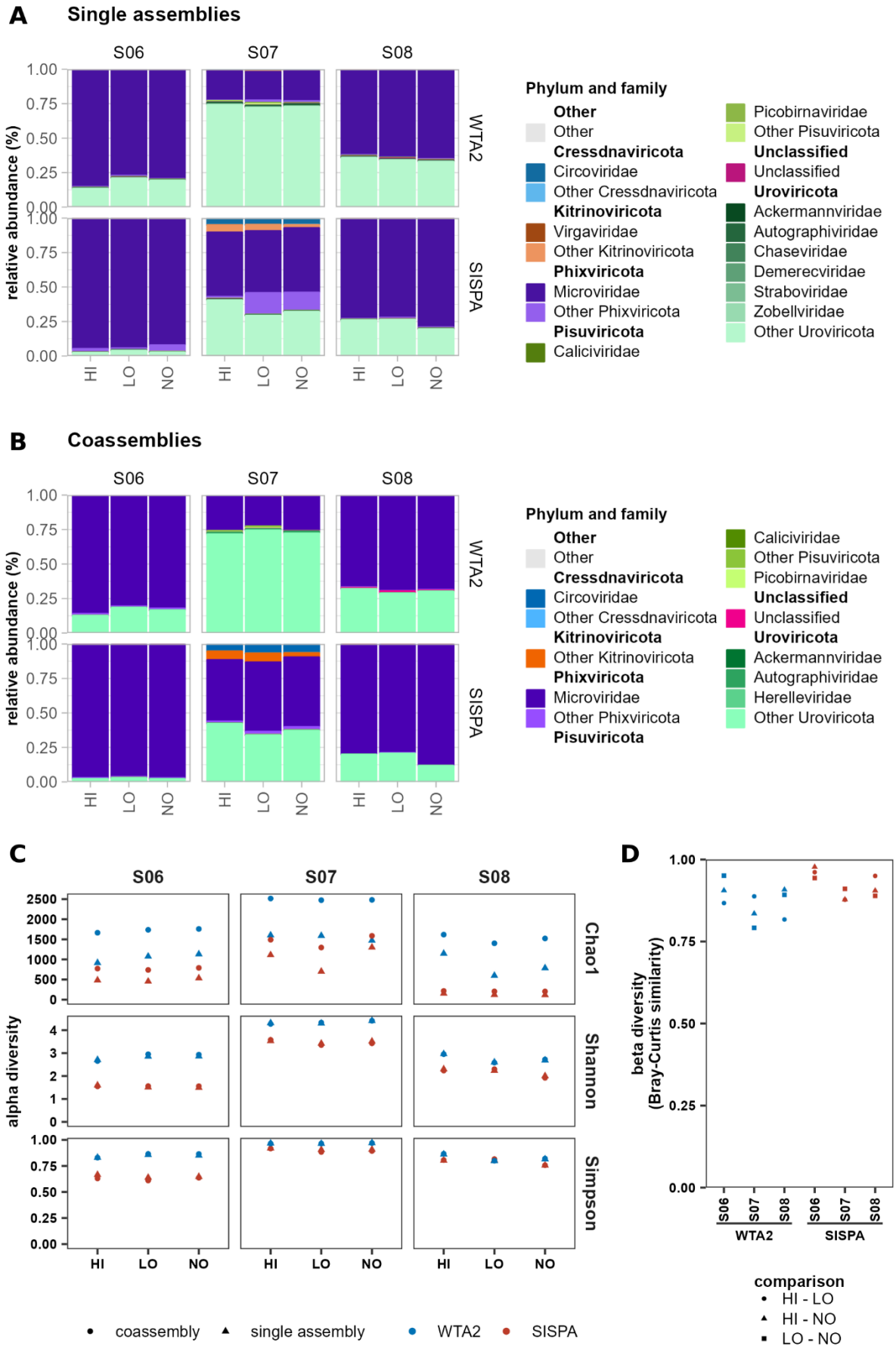
A
Single assemblies
S06
S07
S08
relative abundance (%)
1.00
0.75
0.50
0.25
0.00
WTA2
SISPA
HI
LO
NO
Phylum and family
Other
Other
Cressdnaviricota
Circoviridae
Other Cressdnaviricota
Kitrinoviricota
Virgaviridae
Other Kitrinoviricota
Phixviricota
Microviridae
Other Phixviricota
Pisuviricota
Caliciviridae
Picobirnaviridae
Other Pisuviricota
Unclassified
Unclassified
Uroviricota
Ackermannviridae
Autographiviridae
Chaseviridae
Demerecviridae
Straboviridae
Zobellviridae
Other Uroviricota
B
Coassemblies
Herelleviridae
C
alpha diversity
2500
2000
1500
1000
500
0
4
3
2
1
Chao1
Shannon
Simpson
coassembly
single assembly
WTA2
SISPA
D
beta diversity
(Bray-Curtis similarity)
comparison
HI - LO
HI - NO
LO - NO

Figure 3.14: Taxonomic composition and diversity of stool viromes. A,B: The top six phyla and the top three families in each phylum based on the assemblies of the individual samples (**A**) and the co-assemblies of the stool samples (**B**). **C:** Sample alpha diversity, as calculated by the Chao1 index for species richness and the Simpson and Shannon indices for species evenness. **D:** Beta diversity, calculated through the Bray-Curtis similarity index, of two of the replicates of each stool sample.

The ratio between the highest and lowest relative abundance was lowest for the phage phyla, with 1.07 – 1.10 and 1.02 – 1.26 for *Phixviricota* and 1.03 – 1.57 and 1.36 – 1.66 for *Uroviricota* in the WTA2 and SISPA libraries, respectively. For the eukaryotic viruses *Pisuviricota*, *Cressdnaviricota,* and *Kitrinoviricota*, the ratio was higher, and the ratio could not always be determined due to their absence in one of the replicates. However, the relative abundance of these phyla was <0.1% in most samples, and the variation was between 0.007 and 1.589 percentage points. To compare the effect of increased sequencing depth, taxonomic classification of co-assemblies was also performed (Fig. 3.14B). While the overall pattern did not change, the relative abundance of *Phixviricota* without family assignment was reduced. On

Table 3.6: Variation in relative abundance at the phylum level between replicates of stool samples, based on individual sample assemblies.

		S06		S07		S08	
Phylum	**Method**	**Mean**	**Range[1]**	**Mean**	**Range[1]**	**Mean**	**Range[1]**
Phixviricota	WTA2	83.546%	5.725	23.787%	3.901	68.375%	2.737
	SISPA	97.665%	0.908	50.364%	7.104	82.621%	8.929
Uroviricota	WTA2	15.718%	5.842	73.814%	2.580	30.202%	3.367
	SISPA	2.246%	0.899	37.832%	8.517	17.253%	8.923
Pisuviricota	WTA2	0.065%	0.106	1.010%	1.589	0.013%	0.019
	SISPA	0.025%	0.050	0.049%	0.059	0.000%	0.000
Cressdnaviricota	WTA2	0.033%	0.019	0.585%	0.242	0.194%	0.073
	SISPA	0.011%	0.007	5.772%	1.310	0.062%	0.007
Kitrinoviricota	WTA2	0.575%	0.245	0.299%	0.397	0.094%	0.025
	SISPA	0.023%	0.031	5.298%	3.415	0.064%	0.008

[1]Range is the difference in percentage points between the highest and lowest relative abundance of three replicates.

Table 3.7: Variation in relative abundance at the phylum level between replicates of stool samples, based on stool sample co-assemblies.

		S06		S07		S08	
Phylum	**Method**	**Mean**	**Range[1]**	**Mean**	**Range[1]**	**Mean**	**Range[1]**
Phixviricota	WTA2	81.92%	7.842	22.64%	1.528	64.61%	2.991
	SISPA	97.16%	1.544	56.81%	12.578	75.90%	6.826
Uroviricota	WTA2	17.86%	7.614	75.34%	2.326	34.69%	2.872
	SISPA	2.75%	1.453	34.22%	11.560	23.95%	6.976
Pisuviricota	WTA2	0.07%	0.109	0.84%	1.353	0.01%	0.020
	SISPA	0.03%	0.052	0.07%	0.020	0.00%	0.000
Cressdnaviricota	WTA2	0.06%	0.053	0.54%	0.247	0.15%	0.058
	SISPA	0.01%	0.004	4.47%	0.118	0.03%	0.050
Kitrinoviricota	WTA2	0.03%	0.035	0.26%	0.365	0.08%	0.084
	SISPA	0.03%	0.041	3.96%	3.069	0.05%	0.098

[1]Range is the difference in percentage points between the highest and lowest relative abundance of three replicates.

the other hand, no sequences were classified at the species level for *Kitrinoviricota*. Variation in relative abundance at the Phylum level in assemblies (Table 3.6) and co-assemblies (Table 3.7) did not differ systematically between WTA2 and SISPA libraries, nor between assemblies and co-assemblies.

Species richness was calculated for each sample using the Chao1 index, showing a greater richness in co-assemblies than in single assemblies, and a greater richness in the WTA2 libraries than the SISPA libraries (Fig. 3.14C). The mean CV of species richness in the WTA2 and SISPA libraries was 16% and 18%, respectively, in the case of single assemblies. Species evenness was measured by the Shannon and Simpson indices and did not differ between single and co-assemblies, although WTA2 libraries had greater evenness than the SISPA libraries. The Bray-Curtis similarity index was measured between each of the replicates of each stool sample, to determine the variation in overall sample composition between replicates of each stool sample (Fig. 3.14D). For all three stool samples, the similarity was higher in the SISPA libraries than in the WTA2 libraries, while the samples with the highest and lowest similarity were the same in the WTA2 and SISPA co-assemblies. Interestingly, the level of similarity corresponded

to the percentage of contigs with <10X coverage depth in each co-assembly, although including only contigs with ≥10X coverage depth did not qualitatively improve the similarity.

3.4 Discussion

In this chapter, the reproducibility and bias of virus metagenomics analysis was characterised through sequencing of stool samples spiked with an MC. The MC shows variability between spiked stool samples in the recovery of individual viruses and increased recovery of phages M13 and P22 compared to other viruses. Virus genome coverage depth profiles are highly consistent between samples of both libraries and profiles are more uniform in WTA2 libraries. WTA2 libraries produce assemblies and co-assemblies with greater overall quality and higher virus richness. On the other hand, SISPA assemblies have greater overall coverage depth and lower variability in overall composition. Abundance of stool viruses is most consistent for sequences with ≥10X coverage depth, which represent over 90% of relative abundance in samples. While variability in the relative abundance at the taxonomic level was low down to the class level for both methods, only a small percentage of contigs could be assigned taxonomy at the order or family level and was more variable. In SISPA libraries, the relative abundance of ssDNA viruses is higher, while some RNA viruses are reduced, compared to WTA2 libraries.

The level of recovery of individual MC viruses varies between samples, although variation is larger for some viruses (e.g., Det7 and MHV-68) than for others (e.g., T5 and P22). Since the pattern is comparable for both SISPA and WTA2 libraries this suggests that the variation arises before generation of the dsDNA library. Additionally, while richness was lower for SISPA libraries, the replicate with the highest and lowest richness for each stool sample was the same for SISPA and WTA2 libraries (Fig. 3.14C), which suggests that most variation is introduced during the VLP nucleic acid extraction process, while there is an overall reduction in SISPA libraries. This is also supported by the similar levels in species evenness of WTA2 and SISPA libraries, particularly the Simpson index, which is least affected by species richness. Differential recovery across stool samples of MC viruses could be due to differences in the stool sample composition. For example, some phage capsids contain Ig-like domains that help attach to the intestinal mucus (Barr *et al.*, 2013), including phages T5 (Fraser *et al.*, 2006) and P22 (Parent *et al.*, 2010). Differences in, for example, mucus content, could produce different interactions between viruses and faecal material that affect downstream recovery of viral nucleic acid.

For the phages, the differences in abundance between samples spiked with MC HI and MC LO corresponded to the difference in virus concentration between the two MCs in WTA2 samples, in line with previous results (Conceição-Neto *et al.*, 2015). This was not the case for eukaryotic

viruses in the WTA2 libraries and all viruses in the SISPA libraries, possibly due increased uncertainty associated with low read counts. Future characterisation experiments would benefit from the use of at least three MC concentrations, with a minimum of 10^6 particles added for each virus, to determine the correlation virus concentration and read counts more accurately.

In addition to sample-to-sample variation in virus recovery, the efficiency of recovery also differs between viruses. Phages M13 and P22 exceed expected abundance, with reduced abundance of other viruses, in WTA2 and SISPA libraries. Kleiner *et al*. used P22 in their MC and found a higher PPR than our study, but even when correcting for the PPR, P22 recovery was higher than other phages (Kleiner, Hooper and Duerkop, 2015). One explanation for this is a higher extraction efficiency of P22 genomes. Eukaryotic viruses were underrepresented, particularly the enveloped viruses. MHV-68 levels were much lower than expected and BVDV-1 was almost completely absent from any samples. Loss of herpesvirus was also observed in the NetoVIR publication, even when 0.8 µm filters were used, and coronavirus, another enveloped virus, was sensitive to centrifugation (Conceição-Neto *et al.*, 2015). Possibly, MHV-68 and BVDV-1 virions form aggregates or adsorb to materials in the stool that are filtered out (L. Li *et al.*, 2015).

While each of the MC viruses could have interacted differently with stool content, leading to differential recovery rates, this is unlikely to have played a major role, as similar recovery rates were found in MC controls. The filter size and centrifugation step used here had small effects on virus recovery as noted previously (Conceição-Neto *et al.*, 2015). Freezing of the VLP extracts for 16-18 hours may also play a role, as some viruses are sensitive to freeze-thawing. Omission of freezing increased BVDV-1 concentration 100-fold, although other viruses were unaffected (Fig. 3.6B). Additionally, 0.45 µm PVDF centrifugal filters were used instead of 0.45 µm PES syringe filters, and a different nuclease cocktail was used. One or more of these factors may have contributed to the increase in BVDV-1.

Nucleic acid extraction efficiency is another possible cause for the observed differences in recovery between MC viruses. As extraction efficiency of some dsDNA phages is only 24%-30% (Roux *et al.*, 2016), this could have also contributed to differential recovery of viruses. Experimentally determined recovery efficiencies can be used to correct virus abundance in sequencing data (Roux *et al.*, 2016), although to what degree recovery efficiencies of one virus can be applied to another is unclear.

Lastly, WTA2 and SISPA differentially amplify some viruses. The viruses of the phylum *Phixviricota*, as well as MC virus M13, are phages with circular ssDNA genomes and both M13 and stool *Phixviricota* viruses had higher relative abundance in the SISPA than the WTA2

libraries. Another striking difference is the nearly complete absence of RV-A in the SISPA libraries. While long-term storage of the nucleic acid extracts should be considered, this suggests that amplification of RV-A dsRNA by SISPA is less efficient than WTA2.

3.4.1 Effects of sequence bias

Sequence bias is a known phenomenon in many sequencing methods (Sabina and Leamon, 2015). Both WTA2 and SISPA produce non-uniform coverage depths. Coverage depth profiles of MC viruses are highly consistent in WTA2 and SISPA libraries (Fig. 3.7, 3.8). While there is overlap in the depth profiles, WTA2 and SISPA produce different profiles, with higher peaks in some regions leading reduced uniformity in SISPA (Fig 3.9A, B). Overall, the assembly quality of the WTA2 libraries was higher than the SISPA libraries, with equivalent numbers of reads in the SISPA and WTA2 libraries yielding more and longer contigs in the WTA2 library (co-) assemblies. Together with the fact that MC phages had more contigs in the MC HI SISPA libraries than the WTA2 libraries, with equal or lower coverage, this is indicative of higher fragmentation, which has been reported for SISPA (L. Li *et al.*, 2015). Decreased uniformity of the sequence depth profiles means that certain regions of the genome are more likely to produce reads and contigs, and higher sequencing depth is required to fully recover the genome (Smits *et al.*, 2014). Additionally, preferential amplification of some genomes over others increases their relative abundance and decrease the proportion of reads produced from other genomes. Indeed, while the fraction of MC virus reads in the MC control samples is equivalent in WTA2 and SISPA libraries, SISPA libraries of stool samples spiked with MC HI contained a highly reduced fraction of MC virus reads. The same fraction of contigs could not be identified as viral by geNomad in the WTA2 and SISPA (co-)assemblies, which can be caused by non-viral sequences and false negatives. The median coverage depth and relative abundance of these unclassified contigs was higher in the SISPA libraries, suggesting an increase in non-viral sequences. PCR amplification in SISPA uses the DNA polymerase AmpliTaq Gold, which preferentially amplifies microbial sequences with >50% GC contaminating human DNA extracts (Dabney and Meyer, 2012). However, SISPA libraries have a lower GC bias (Fig. 3.9C), and assemblies and co-assemblies generally have lower GC-content (Table 3.4, 3.5) than WTA2. Increased numbers of non-identifiable reads in SISPA have been reported before (Karlsson, Belák and Granberg, 2013), although the origin of these reads is not clear. Identifying the origin of the unknown sequences will help determine the cause of the increased fragmentation observed in the SISPA libraries and provide guidance for improvements to the protocol.

Several factors introduce sequence bias due to preferential amplification of certain genomic regions. For example, different DNA polymerases are affected by GC-content to different degrees (Dabney and Meyer, 2012), amplification efficiency in tagged random primers is

affected by the tag sequence, random nucleotide sequence and length, and bases downstream of the binding site (Rosseel *et al.*, 2013; Pan *et al.*, 2014), DNA polymerases have different sequence preferences (Hansen, Brenner and Dudoit, 2010), and the transposase used in the Illumina DNA Prep kit has a preferred sequence motif (Gunasekera *et al.*, 2021). Optimisation of the WTA2 protocol is not straightforward due to the proprietary nature of kit components. Use of longer random nucleotide sequences in the tagged primers, and using multiple primers with different tag sequences has been suggested to increase uniformity of the coverage depth (Rosseel *et al.*, 2013). Incorporation of multiple primers with twelve random nucleotides did not improve coverage depth uniformity in an oral virome study, however (Parras-Moltó *et al.*, 2018). And in the present study, WTA2 produces more uniform profiles despite using random hexamers while SISPA uses random octamers. Nevertheless, optimisation of SISPA primers should be explored. WTA2 performs 17 cycles of PCR amplification, while SISPA performs 30 cycles, which will amplify any bias inherent to the method. For WTA2, increasing PCR cycles past 17 did not improve yield of viruses (Conceição-Neto *et al.*, 2015) and a recent benchmarking study showed reduced total assembly length when using 30 amplification cycles, compared to 15 or less cycles (Wang *et al.*, 2023). Thus, while this may require higher input volumes, a reduced number of cycles should be tried for SISPA.

3.4.2 Accuracy and precision of abundance measures

Comparing the abundance and abundance ranking of contigs between replicates, shows a large difference between the highest and lowest abundance and abundance-based ranking for most contigs. However, most of these high variation contigs have a relative abundance of <0.1%, and collectively have a relative abundance between 1% and 10%, suggesting that they play a minor role in the overall variation. Overall, a cut-off of 10X coverage depth of the contig removes most of the highly variable contigs. This corresponds to 66 reads mapping to a 1kb contig, assuming 150 bp reads. Nevertheless, this still includes contigs with more than 4-fold difference between the highest and lowest relative abundance in three replicates, although the median ratio is lower than 2. Even for contigs with a mean depth of >100X, the median fold difference in the WTA2 samples is on average 1.6, although for SISPA it is lower, at 1.2, suggesting that precision of abundance is higher for SISPA. The overall dissimilarity of replicates of WTA2 and SISPA libraries corresponds to the percentage of contigs with <10X coverage depth. However, removal of low coverage contigs does not increase similarity, showing that variability in overall virome composition is not only due to a higher percentage of low-coverage sequences.

While nearly all viral contigs were classified to the class level, only a small percentage of sequences could be assigned at the order or family level. This prevents assessment of variability of taxonomic analysis and differences between WTA2 and SISPA at lower taxa. At the class level, there was only minor variation in relative abundance of the top taxa at the

Phylum and Family level. SISPA does favour ssDNA viruses more than WTA2, considering the higher relative abundance of *Phixviricota*. Although WTA2 (co-)assemblies have reduced fragmentation and increased diversity, the rate of classification at lower levels is similar for both SISPA and WTA2, as well as the variation in species richness. Comparing co-assemblies to assemblies shows that a 3-fold increase in sequencing depth results in an increase of viral sequences. However, as taxonomic classification below class level is similar in assemblies and co-assemblies, a higher increase in sequencing depth is required to increase classification sensitivity.

MC virus genomes are represented in the assemblies by several contigs that can have up to 30-fold difference in abundance. Theoretically contigs in an assembly have a coverage of at least 1X, whereas the actual genome coverage can be much lower, leading to an overestimation of the abundance of the virus. Provided that contigs together sufficiently cover the virus genome (>70%), the abundance is overestimated by up to only 10%, whereas a coverage <30% leads to a doubling of the calculated abundance. The accuracy of the calculated virus abundance therefore benefits from grouping of contigs into genome bins that belong to the same virus, as well as estimation of the genome completeness. The former can be accomplished by virus genome binning tools like PHAMB (Johansen *et al.*, 2022), vRhyme (Kieft *et al.*, 2022), MetaBAT 2 (Kang *et al.*, 2015, 2019), and CoCoNet (Arisdakessian *et al.*, 2021), and the assembly contiguity improvement tool COBRA (Chen and Banfield, 2024), whereas the latter can be performed by CheckV (Nayfach, Camargo, *et al.*, 2021).

In future characterisation experiments, adapting primers should be considered. The use of a virus MC has provided insights into the consistency of virus recovery across samples, biases in virus recovery, sequence bias, and accuracy and precision of calculated abundance. The spike concentration of MC HI was sufficient to recover full genomes for most dsDNA viruses in the assemblies of WTA2 libraries. However, SISPA libraries contained fewer reads, particularly for the smaller RNA viruses. Therefore, a higher spiking concentration is recommended, in the order of 1×10^8 particles of each virus per 50 mg stool. Alternatively, a MC in which virus concentrations are adjusted for the nucleotide content of the genome should in theory produce the same coverage of each virus.

3.4.3 Limitations

The present study has several limitations. While WTA2 and SISPA libraries were generated from the same viral nucleic acid extracts, the library preparation methods were not the only difference between the two data sets. Storage time for the SISPA libraries was increased, Illumina libraries were prepared on separate occasions, and were sequenced on different sequencing platforms. An effect of each of these differences on the final data cannot be

excluded. Inaccuracies in the epifluorescence microscopy measurements used to calibrate virus stocks may contribute to deviations in the mock community virus abundance from the expected levels. M13 as one of the smaller viruses was more difficult to visualise than the dsDNA viruses, which could have led to an underestimation of the stock titre and an excess of M13 in the MC. However, RV-A and BVDV-1 were similarly difficult to visualise and were underrepresented in the sequencing data. Nonetheless, inaccuracies in the determined stock titres could have contributed to the observed differences, particularly since each virus stock is unique and different effects might have played a role for different viruses. A sizeable portion of the sequencing data was attributable to sequences of unknown origin. These were especially pronounced in the SISPA libraries. It is not clear whether these are non-viral sequences or viral sequences that were not recognised by geNomad. Additional virus detection tools, improved assembly contiguity and genome binning of contigs, and comparing unrecognised sequences to microbial and other sequence databases can help identify the source of these sequences. This would show whether SISPA libraries contain more contaminations, or whether the increased number of unclassified sequences is due to increased fragmentation hampering identification of virus sequences.

3.4.4 Conclusion

Using a set of stool samples analysed in triplicate and spiked with a mock viral community, the reproducibility and bias of a virome sequencing method has been assessed, including the comparison of two methods for the reverse-transcription and amplification of viral nucleic acid. The results show that individual viruses have different recovery efficiencies, and that recovery of individual viruses varies between replicates. In assemblies, minimum variability of abundance is achieved for contigs with ≥10X coverage. Nonetheless, contigs with lower coverage are a small fraction of the total virome at higher taxonomic levels and compositional variability is low even when low coverage depth contigs are included. Taxonomically, the greatest difference between WTA2 and SISPA libraries was the increased relative abundance of ssDNA viruses in SISPA libraries. Assembly-based abundance calculation of viruses is most accurate when genome completeness is >70%. Differences between WTA2 and SISPA include increased assembly contiguity and virus richness in WTA2 libraries, and reduced variability in virus abundance, GC-content, and overall virome composition in SISPA. Based on these results, the WTA2 kit provides higher sensitivity, at the cost of higher variability. SISPA on the other hand is less sensitive and accurate, but more consistent. While WTA2 is a proprietary kit, SISPA uses commonly used reagents and is thus less costly and less sensitive to supply chain issues, since reagents can be substituted more easily. Future improvements to the SISPA protocol should be investigated to increase sensitivity, particularly to RNA viruses, as this would make SISPA a competitive alternative to WTA2. Incorporation of a mock viral community was key in

this analysis, and is a valuable control to be included in metagenomic studies (Knight *et al.*, 2018; Boers, Jansen and Hays, 2019).

CHAPTER 4: INVESTIGATING THE EFFECTS OF FMT ON THE GI VIROME IN ME/CFS

4.1 Introduction

A large subset of people with ME/CFS experience GI disturbances including nausea, diarrhoea, constipation, abdominal pain, and bloating (Maes *et al.*, 2014; Tschopp *et al.*, 2023). While estimates vary between studies, a recent analysis found 65% of ME/CFS patients reported GI symptoms (Tschopp *et al.*, 2023) and 28% of patients described a GI-related infectious trigger for their illness with 38% reporting co-morbid IBS (Johnston, Staines and Marshall-Gradisnik, 2016). Across studies, IBS is highly associated with ME/CFS, with comorbidity estimates ranging from 17% to 92%, versus 10% to 20% in the general population (Chu *et al.*, 2019). Inflammation as a result of an increase in gut permeability, also called "leaky gut", has been hypothesized to contribute to ME/CFS aetiology (Navaneetharaja *et al.*, 2016; Morris *et al.*, 2019), supported by evidence of increased bacterial translocation (Maes, Mihaylova and Leunis, 2007; Maes and Leunis, 2008; Maes *et al.*, 2012, 2014; Navaneetharaja *et al.*, 2016; F. Martín *et al.*, 2023). Recent analysis of the antibody repertoire in the blood of ME/CFS patients has shown an increase in antibodies against bacterial flagellins, particularly from the genus *Lachnospiracae* (Vogl *et al.*, 2022). Interestingly, increased antibodies against bacterial flagellins was also found in patients with CD (Alexander *et al.*, 2021; Bourgonje *et al.*, 2023), and IBD, particularly CD, which increases the risk for ME/CFS (Tsai *et al.*, 2019).

Several studies have shown changes in the gut microbiome in ME/CFS (Frémont *et al.*, 2013; Giloteaux *et al.*, 2016; Giloteaux, Hanson and Keller, 2016; Nagy-Szakal *et al.*, 2017; Guo *et al.*, 2023; Xiong *et al.*, 2023). However, the changes described are inconsistent across studies, likely due to differences in methodology, sample size and patient cohorts (Du Preez *et al.*, 2018; König *et al.*, 2022). Still, several studies found that gut microbiome diversity in ME/CFS patients is marked by lower species diversity, with reduced species evenness and richness, and heterogeneity between people with ME/CFS, compared to heterogeneity amongst healthy controls (Giloteaux *et al.*, 2016; Nagy-Szakal *et al.*, 2017; Guo *et al.*, 2023; Xiong *et al.*, 2023). Additionally, IBS is an determining factor in ME/CFS microbiome composition, with ME/CFS patients with and without IBS having distinct microbiomes (Nagy-Szakal *et al.*, 2017; Guo *et al.*, 2023). Two large-scale multi-omics studies identified a reduction in short chain fatty

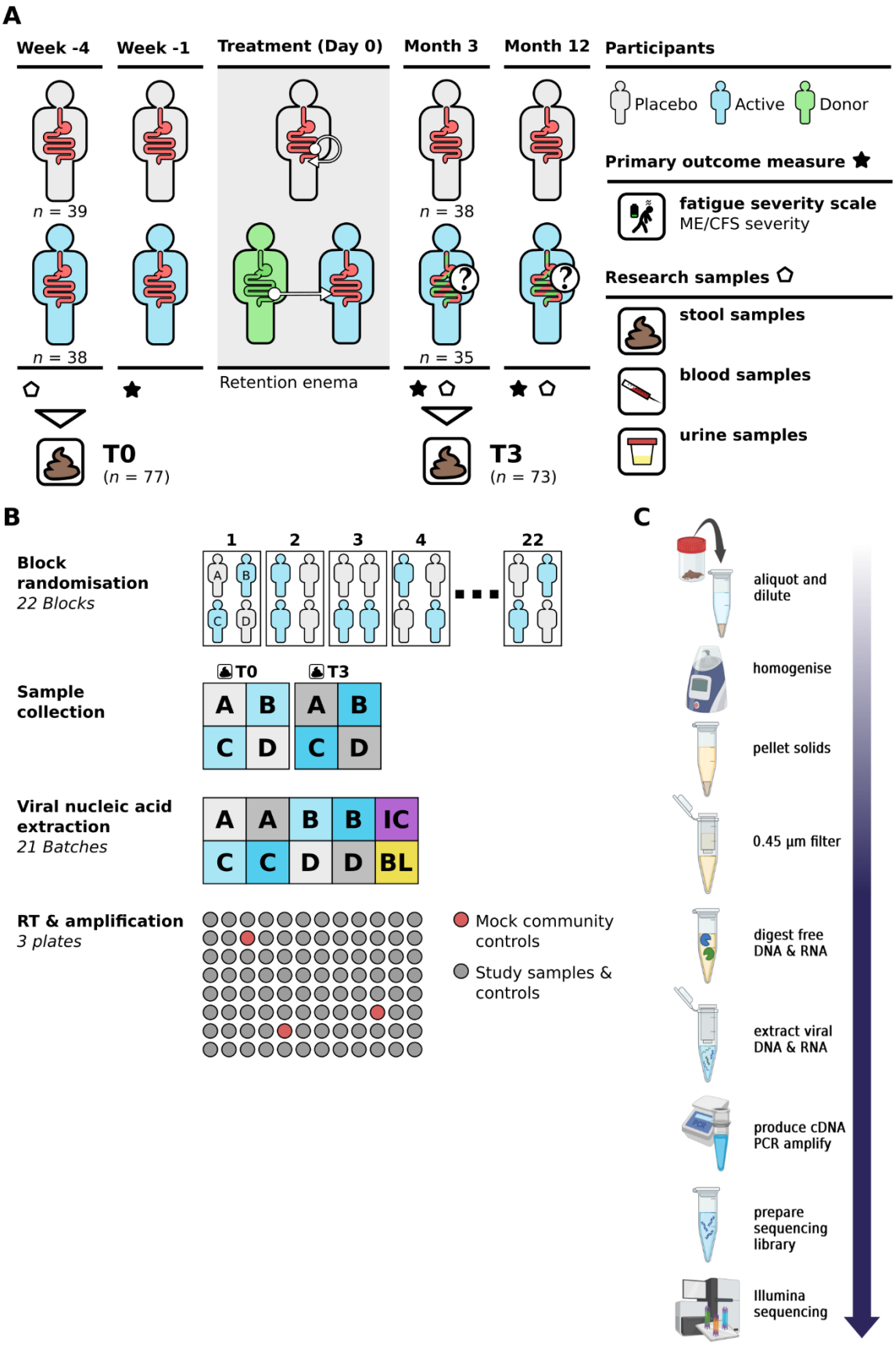

Figure 4.1: Study overview. A: The Comeback Study is a randomised, double-blinded, placebo-controlled phase IIb clinical trial, investigating the efficacy and safety of FMT in

patients with ME/CFS. A total of 80 patients were recruited, randomised, and allocated to the placebo or active FMT transplantation. Patients received an FMT *via* retention enema, with the placebo group receiving an autologous transplant, while the active group received a transplant from a healthy donor. The change in fatigue, as measured by the Fatigue Severity Scale, after three months is the primary outcome measure of the study. Stool samples from the baseline (T0) and three-month follow-up (T3) are used for virome analysis. The numbers indicate the number of samples from T0 and T3 used for virme analysis. **B:** Patients are randomised in blocks of four, assigning two patients to placebo transplant and two to active transplant. To maintain these treatment blocks and to control for batch variation, T0 and T3 samples of all patients in each block are processed together in a single batch for viral nucleic acid extraction. In each batch an internal control (IC), consisting of a homogenised stool sample, and a blank (BL) are included. Donor samples were processed in an additional batch. Due to drop-outs and unavailable samples, some blocks were combined and thus some batches consisted of multiple incomplete blocks, leading to a difference between the number of blocks and batches. During RT and amplification of the viral nucleic acid, batches are allocated in their entirety to a 96-well plate. On each plate, three mock virus community nucleic acid controls are included. **C:** Viral nucleic acid extraction started with aliquoting of the stool sample, dilution in PBS, homogenisation of the sample, pelleting of solids and filtering the supernatant to remove cells and other materials. Free nucleic acid was then degraded and protected nucleic acid was extracted. Then, double-stranded DNA fragments were generated through RT and amplified, and an Illumina library was prepared for pair-ended short read sequencing and sequenced on the Illumina platform. Created with BioRender.com

---

acid-producing bacteria associated with reduced plasma butyrate (Guo *et al.,* 2023; Xiong *et al.,* 2023). Particularly, the studies corroborated earlier findings of reduced levels of the SCFA-producing *Faecalibacterium prausnitzii* (Giloteaux *et al.,* 2016; Nagy-Szakal *et al.,* 2017; Kitami *et al.,* 2020), which has anti-inflammatory properties and contributes to intestinal homeostasis (R. Martín *et al.,* 2023). Interestingly, *F. prausnitzii* was also found to be reduced in post-acute sequelae of COVID (Liu, Mak, *et al.,* 2022). One study found an inverse correlation between fatigue and levels of SCFA in IBS patients (El-Salhy *et al.,* 2021), which suggests a role for SCFA or SCFA-producing bacteria in fatigue (König *et al.,* 2022).

A major contributor to shaping the gut microbiome are phages which infect bacteria with consequences for not only microbiome structure but also human health (Mirzaei and Maurice, 2017). Phages might therefore contribute to alterations in the GI prokaryome of ME/CFS patients and play a role in ME/CFS pathogenesis (Newberry *et al.,* 2018). In turn, infectivity of human-infecting (eukaryotic) viruses in the GI tract is affected by GI microbiota (Roth, Grau

and Karst, 2019) with eukaryotic viruses activating immune responses that affect microbiota structure (Li, Handley and Baldridge, 2021). Several viruses have been associated with the onset of ME/CFS (Rasa *et al.*, 2018; Hwang *et al.*, 2023), including herpesviruses (Ariza, 2021) and enteroviruses (O'Neal and Hanson, 2021; Hanson, 2023). The GI tract is a potential reservoir for eukaryotic viruses (Neurath, Uberla and Ng, 2021). However, few studies have investigated the GI virome in ME/CFS, although recent studies found no association of human viruses (Briese *et al.*, 2023) or overall dsDNA virome composition (Hsieh *et al.*, 2023) in ME/CFS. Large-scale and comprehensive analyses of the GI virome in ME/CFS, encompassing both RNA and DNA phages and eukaryotic viruses, are lacking.

Given the association of GI dysbiosis with ME/CFS, modulation of the GI microbiome might provide an avenue for treatment (Navaneetharaja *et al.*, 2016; Stallmach *et al.*, 2024). Antibiotics, probiotics and FMT have been investigated in ME/CFS, with mixed results (Seton *et al.*, 2024). Of these, FMT offers the most comprehensive approach to replace the microbiota in the patient with that of a healthy donor. FMT was first approved for the treatment of rCDI (Cammarota *et al.*, 2017), where it achieves around 90% treatment success (Allegretti *et al.*, 2019). FMT is now under investigation for the treatment of many diseases associated with an altered gut microbiota composition, for example for liver disease, IBS, IBD, obesity and metabolic disorder, arthritis, cancer, and autism, all with mixed results so far (Biazzo and Deidda, 2022). For instance, treatment of IBD (Cheng *et al.*, 2021; Feng *et al.*, 2023; Imdad *et al.*, 2023) and IBS (Rodrigues *et al.*, 2023) have produced promising results. Interest in treatment of ME/CFS using FMT increased after a retrospective study reported successful treatment of 60 ME/CFS patients receiving a transfusion of a mixture of GI bacteria (Borody, Nowak and Finlayson, 2012). Yet, studies on FMT in ME/CFS are limited, with a small RCT reporting no effect in 5 patients who received active transplant (Salonen *et al.*, 2023), while a retrospective study found improvements in 21 patients who received FMT (Kenyon, Coe and Izadi, 2019). Larger RCTs are required to provide evidence of the safety and efficacy of FMT in ME/CFS.

4.1.2 Specific aims

The main aim of this chapter is to determine whether FMT leads to changes in the GI virome in ME/CFS patients participating in the Comeback Study (Skjevling *et al.*, 2024) (Fig. 4.1A). The specific aims are:

- To perform quality control, including removal of contaminating sequences, of the sequencing data using an MC, an internal control, and blank samples.
- To determine changes in taxonomic composition, alpha diversity, and beta diversity of the virome following FMT.

- To assess evidence of engraftment of donor viruses following FMT and determine which types of viruses are engrafted.
- To enrich eukaryotic viral sequencing data to enhance sensitivity for eukaryotic viruses and determine changes in the abundance of eukaryotic viruses.

4.2 Materials and methods

4.2.1 Ethics considerations

Ethical approval for The Comeback Study was given by the Regional Committee for Medical Research Ethics Northern Norway (REK Nord), application number 2018/180. All research was performed in accordance with the Declaration of Helsinki (2013) and ICH-Good Clinical Practice and monitored by Department of Clinical Research (The Centre of Quality and Development, KVALUT) at the UNN. All data was handled in accordance with the EU GDPR and the UK Data protection Act 2018.

4.2.2 Patient recruitment

Patients were recruited through local general practitioners, posters at doctor's offices, and a Facebook group. Eligibility was assessed by telephone according to the ME-ICC and severity rating for ME/CFS (Carruthers *et al.*, 2011). Eligible patients were then referred to the UNN for reassessment. Prior to the reassessment appointment, patients were provided informed written consent information, filled out a modified version of the DePaul Questionnaire (DSQ) (Jason *et al.*, 2010), Hospital Anxiety Depression Scale (HADS) (Zigmond and Snaith, 1983), Fatigue Severity Scale (FSS) (Krupp *et al.*, 1989), MOS 36-Item Short Form Survey (SF-36) (Stewart, Hays and Ware, 1988), and an antibiotics and food supplement questionnaire (AFQ). During reassessment, patients underwent a physical exam, and blood, faecal, and urine samples taken to exclude differential diagnosis in accordance with the Norwegian National Guidelines for Assessment of CFS/ME (Helsedirektoratet, 2014). Patients also underwent cognitive screening through the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) test under the guidance of a neuropsychologist (Randolph *et al.*, 1998; Dickson, Toft and O'Carroll, 2009).

Patients were included in the study if they met the ME-ICC, were between 18 and 65 years of age, had mild to severe ME/CFS, scored 5 – 7 on the FSS, and had a symptom duration of 2 – 15 years. Exclusion criteria were: kidney failure, congestive heart failure, immunodeficiency or use of immunosuppressive drugs, other disease that may explain ME/CFS symptoms discovered during diagnostic work up, use of antibiotics in the last three months, use of low-dose naltrexone or isoprinosin, pregnancy or breastfeeding, serious endogenous depression, chronic infectious disease, introduction of new food supplements, change in diet or introduction of new medications in the last three months, assessed not to be able to follow the

instructions for data and sample collection, very severe ME/CFS (WHO class IV), symptom duration of less than 24 months or more than 15 years, history of abdominal surgery, with the exception of appendectomy, cholecystectomy, caesarean section and hysterectomy, and previous treatment with FMT.

4.2.3 Donor recruitment and screening

FMT donors were recruited locally and informally. Donors were healthy and between 16 and 30 years of age. Donors were excluded in case of peroral antibiotics during past 3 months, use of topical antibiotics in the past 2 months, a new tattoo or piercing in the past 6 months, former imprisonment, a history of chronic diarrhoea, constipation, IBD, IBS, colorectal polyps, colorectal cancer, immunosuppression, obesity, metabolic syndrome, ME/CFS, psychiatric disorders, or serious autoimmune disease, close relatives with serious autoimmune disease, high risk sexual behaviour, bowel movements not corresponding to Bristol Stool Scale type 3 or 4, journeys abroad within six months to countries high in antibiotic resistance, use of pre-, pro- or symbiotics food supplements within 1 month, or dysbiosis grade 3 or more by the GA-map Dysbiosis Test (Genetic Analysis SA, Oslo, Norway). Additionally, donors underwent full screening through physical examination and routine blood tests, psychiatric evaluation, pathogen test of blood and faeces, including bacterial pathogens and norovirus, rotavirus, sapovirus, and adenovirus, and COVID-19 nasopharyngeal swab and serology tests at UNNH. This screening was performed before the first donation and every four weeks thereafter.

4.2.4 Study design

The Comeback Study is a double-blinded, placebo-controlled, parallel-group, single centre, phase II RCT (ClinicalTrials.gov: NCT03691987). The study was conducted at the UNN, Harstad, Norway. During the assessment period five weeks before treatment, patients completed the modified DSQ, HADS, FSS, SF-36 and AFQ at home. Four weeks before treatment, blood, urine, and stool samples were taken, cognitive assessment was performed using the RBANS, a Food Frequency Questionnaire (FFQ) was filled out, and IBS symptoms were recorded according to the Rome IV criteria at the UNN Harstad. Patients were randomly allocated to placebo or active treatment in a 1:1 ratio. On the day of the treatment, patients completed the FSS, SF-36, and antibiotics and food supplement questionnaire. After treatment, there were follow-ups at 1, 3, 6, 9, and 12 months post-treatment. The one-month follow-up consisted of completion of the FSS, patient-reported adverse event questionnaire (PRAEQ), and answering the disease worsening question (DWQ). At the three-month follow-up, cognitive assessment was performed using the RBANS, and the HADS, FSS, FFQ, AFQ, PRAEQ and DWQ were completed. At six and nine months, the FSS and AFQ were again completed, and GI-related items of the DSQ were completed at six months. At the final follow-up 12 months post-FMT, the HADS, FSS, FFQ, SF-36, AFQ, PRAEQ, DWQ, and GI-related items of the DSQ were completed,

and blood, urine, and stool samples were taken. Data were recorded in REDCap (Harris *et al.*, 2009, 2019), except for the FFQ and RBANS.

The primary endpoint of the study was the proportion of patients in the placebo and treatment groups with treatment success, defined as an increase in score of at least 1.2 points on the FSS. Additionally, secondary end points were the difference in change in fatigue as measured by FSS between placebo and treatment groups at 12-month follow-up, quality of life as measured by SF-36 up to 12 months after treatment, neurocognitive function as measured by the RBANS score at baseline to 3 months after treatment, anxiety and depression as measured by the HADS score, and GI-related complaints as measured by the GI-related items on the DSQ at baseline and up to 12 months after treatment.

The Comback Study was funded by the Norwegian Research Council through the BEHOV-ME research program, as well as the Norwegian ME/CFS patient organisation Norges ME-forening. The full study protocol has been published elsewhere (Skjevling *et al.*, 2024).

4.2.5 Donor and placebo sample preparation

Donor and patient stool were used to produce active and placebo transplants, respectively. Patient stool taken during the inclusion process four weeks before treatment was used to produce placebo. Donor stool was processed between one week and two years before treatment. For both treatment and placebo, between 50 and 80 grams of stool was mixed with 120 ml isotonic saline and 25 ml 85% glycerol and homogenised in a blender for 30 seconds. The mix was poured through a 0.5 mm steel mesh strained and stored in 60 ml Luerlock syringes at -80 °C. Donor transplants were stored for up to two years.

4.2.6 Randomisation and blinding

Patients were randomised in blocks of four, with two patients allocated to placebo, one to one donor and the last to another donor. The randomisation sequence was generated using REDCap. Randomisation of transplants was performed in a separate, blinded room with a freezer containing patient and donor transplants. Transplants were tagged with donor or patient IDs. Placebo transplants were placed on a table by a member of the research team. After the researcher left the room, an independent allocator then entered. The allocator used a computer to access the randomisation sequence, which was only accessible to the allocator. The tags from the placebo transplants of the patients allocated to active treatment, were removed, and placed onto the respective allocated donor transplants. The untagged placebo transplants were then discarded, and the remaining tagged transplants were placed in a freezer. Investigators, assessors, and patients remained blinded to the allocation.

4.2.7 FMT procedure

FMT was administered at the gastroenterology department of the UNN. Patients received a bowel preparation containing sodium picosulfate, magnesium oxide and citric acid (Picoprep, Ferring Pharmaceuticals, Saint Prex, Switzerland). FMT was administered by enema. A positioning procedure was performed to enhance perfusion of the transplant (Skjevling *et al.*, 2023). Briefly, patients first lay on their left side, tilted 15° with their heads down, on an examination bench. The transplant was then administered using an enema kit. The patient was then left on their left side for 2 minutes, then turned onto their abdomen for 2 minutes, and turned slowly onto their right side for another 2 minutes. The bench was then tilted in the opposite direction, and the patient remained at 15° head-up for an additional 2 minutes. The enema kit was removed, and the patient remained on their right side for 5 to 10 minutes before slowly standing up. Patients allocated to treatment received a donor transplant and those allocated to placebo received a autologous transplant produced from the faecal sample donated during the baseline work-up. Patients were asked to maintain their current diet without introduction of food supplements and probiotics during the follow-up period.

4.2.8 Stool sample collection and storage

Patient stool samples were collected at home using a home collection kit. Samples were stored at home in a freezer for up to eight weeks. Samples were collected and transported on dry ice to the UNNH for storage at -80 °C.

4.2.9 Internal control sample preparation

Stool of a study drop-out was used to produce internal control samples. Using a #10 surgical blade (Swann-Morton, Sheffield, UK, ref. 0201), 12.25 g of stool was transferred into a falcon tube (VWR, Oslo, Norway, ref. 525-1109) with 4.8 ml PBS and vortexed to homogenise the sample. Using plastic inoculation loops (Thermo Fisher Scientific, Oslo, Norway, ref. 254437), 0.4 – 0.7 g was transferred into 2.0 ml Eppendorf tubes and stored at -80 °C.

4.2.10 Sample virome processing

Patients were treated in blocks of four, with two each receiving a treatment and two others receiving a placebo (Fig. 4.1B). VLP purification and nucleic acid extraction was performed at the Arctic University of Norway, Tromsø, Norway. Reverse transcription, amplification, and library preparation was performed at the Quadram Institute. Samples were organised into batches based on these blocks by including both timepoints of each block in the same batch, to ensure both balanced processing of placebo and treatment, and to ensure both patient samples are processed under the same conditions. In some cases, multiple participants from the same block had dropped out, in which case the block was merged with another block to optimise sample processing speed. In addition, viral nucleic acid extraction of both samples of participants P027, P065, P032 and P067 were repeated and therefore not processed in their

original blocks. Of participants P073, P013, P061 and P075, only the first timepoint was available and these were also processed in a different batch from the other samples in their blocks. Faecal samples corresponding to the transplants of each donor were all processed together in one batch. Each batch included one blank sample (PBS) and one internal control sample.

4.2.11 Nuclease buffer

Using 100 ml volumetric flasks, 2M Tris (Millipore, 648311-1KG), 1M $CaCl_2$ (Sigma, C1016-100G) and 1M $MgCl_2$ (Sigma, M8266-100G) solutions were prepared in deionised water. Then, 50 ml 2M Tris, 10 ml 1M $CaCl_2$ and 3 ml $MgCl_2$ were combined in a beaker with 15 ml deionised water. The pH was adjusted to 8.0 by dropwise addition of 8M HCl. Deionised water was added to a total volume of 100 ml and the pH was checked and corrected with 8M HCl until pH=8.0. The buffer was autoclaved and stored at 21 °C.

4.2.12 Total viral nucleic acid extraction

The viral nucleic acid extraction method was adapted from the NetoVIR protocol (Fig. 4.1C) (Conceição-Neto *et al.*, 2015). Stool samples were kept on dry ice until ready for processing. Samples were taken out of the dry ice 5 minutes before processing. Using a surgical scalpel (Swann-Morton, Sheffield, UK, ref. 0201), 50 – 100 mg of was sampled from the frozen stool and placed into a MagNA Lyser Green Beads tube (Roche, Oslo, Norway, ref. 03358941001) with the beads discarded beforehand. A fresh blade was used for each sample and the scalpel handle was wiped and submerged in 4% sodium hypochloride (Klorin, Lilleborg, Oslo, Norway, ref. 4124) for 2 minutes to disinfect and decontaminate.

After aliquoting all samples, 500 – 1,000 µl PBS (Merck, Oslo, Norway, ref. D8537-500ML) was added for a 10% (m/v) suspension. Samples were homogenised in a MagNA Lyser (Roche, Oslo, Norway, ref. 3358976001) at 3000 rpm for 60 seconds at 4 °C. Solids were settled in a centrifuge at 17,000 x *g* for 3 min. Then, 500 µl of the supernatant was filtered through a 0.8 µm PES filter column (Merck, Oslo, Norway, ref. UFC40HV00) in a centrifuge at 5,000 x *g* for 4 minutes at 4 °C. To digest free nucleic acids, 130 µl of the filtrate was incubated with 7 µl nuclease buffer (1M Tris, 100 mM CaCl2 and 30 mM MgCL2, pH=8), 2 µl Benzonase nuclease (Merck, Oslo, Norway, 70746-3) and 1 µl Micrococcal nuclease (New England Biolabs, Hitchin, UK, ref. M0247SVIAL) at 37 °C for 2 hours. The reaction was stopped by adding 7 µl of 0.5 mM EDTA (Thermo Fisher Scientific, Oslo, Norway, AM9260G).

Viral nucleic was extracted using the QIAamp Viral RNA Mini kit (Qiagen, Oslo, Norway, ref. 52906) following manufacturer's instructions. Briefly, to a RNAse-free 2.0 ml Eppendorf tube, 560 µl AVL buffer and 140 µl sample were added and mixed by pulse vortexing. After a 10-minute incubation at 21 °C, 560 µl ethanol was added and mixed by vortexing. To a spin

column, 630 µl sample was added and the column was centrifuged at 6,000 x g for 1 minute, repeatedly until all sample was added to the column. Then, 500 µl AW1 buffer was added to the column and spun in a centrifuge at 6,000 x g for 1 minute. Then, 500 µl AW2 buffer was added and spun in a centrifuge at 16,000 x *g* for 1.5 minutes twice. The column was transferred to a 1.5 ml Eppendorf tube and 60 µl AVE buffer was added to the column. After 1 minute at 21 °C, the tube was spun at 6,000 x *g* for 1 minute and the eluate was split across three 0.5 ml PCR tubes on dry ice and then stored at -80 °C.

4.2.13 Mock virus community

A mock virus community of prokaryotic and eukaryotic viruses was produced from virus stocks of *Tequintavirus T5* (T5), *Kuttervirus Det7* (Det7), *Lederbergvirus P22* (P22), *Inovirus M13* (M13), *Qubevirus durum* (Qbeta), *Rhadinovirus muridgamma4* strain Murid gammaherpesvirus-68 (MHV-68), *Pestivirus bovis* strain Bovine viral diarrhea virus Ky1203nc (BVDV-1), *Sindbis virus* laboratory strain AR339 (SINV) and *Rotavirus A* strain Simian rotavirus SA/11 (RV-A). SINV was produced by Charlotte Dixon in the lab of Dr. Penny Powell, University of East Anlgia, Norwich, UK, using baby hamster kindey (BHK) cells in Dulbecco's Minimum Essential Medium with GlutaMAX supplement (Thermo Fisher Scientific, Loughborough, UK, ref. 10566016), with 10% foetal bovine serum (Thermo Fisher Scientific, ref. 16000044) and 1% penicillin-streptomycin (Thermo Fisher Scientific, ref. 15070063) added. Stocks of RV-A, BVDV-1 and MHV-68 were kind gifts of, respectively, Dr. Edward Mee (National Institute for Biological Standards and Control, Potters Bar, UK), Dr. Joe Brownlie (Royal Veterinary College, London, UK) (Clarke, Brownlie and Howard, 1987), and Dr. James Stewart (University of Liverpool, Liverpool, UK). Phage stocks were produced as described in Chapter 3. SINV, MHV-68 and BVDV-1 were resuspended using a syringe and needle to dissociate clumped particles. Of each virus, 3.2×10^8 particles were added. For the phages, the volume was based on PFU counts. For BVDV-1, MHV-68 and RV-A, stocks were added based on EFM-based titre estimates obtained in Chapter 2. SINV was added based on titres obtained from plaque assays performed by Charlotte Dixon. To the mock community, PBS was added to a total volume of 1,800 µl and then filtered using a 0.8 µm PES filter column (Merck, Oslo, Norway, ref. UFC40HV00) in a centrifuge at 5,000 x *g* for 4 minutes at 4 °C. The filtrate was incubated in 1× nuclease buffer, 0.36 U/µl Benzonase nuclease (Millipore, Oslo, Norway, ref. 70746-3) and 29 U/µl Micrococcal nuclease at 37 °C for 2 hours. The reaction was stopped by adding 0.5 mM EDTA (Thermo Fisher Scientific, ref. AM9260G) to a final concentration of 83 mM. Total nucleic acid was extracted using a single spin column of the QIAamp Viral RNA Mini kit (Qiagen, Oslo, Norway, ref. 52906). The complete mock community volume of 1.9 ml was aliquoted into 2.0 ml Eppendorf tubes containing lysis buffer. After 10 minutes incubation at 21 °C, 630 µl of the lysate was added to the spin column and centrifuged at 6,000 × *g* for 1 minute, until all aliquots

were processed. The spin column was then further processed according to manufacturer's instructions.

4.2.14 Reverse transcription and amplification

Nucleic acid extract was converted to double-stranded DNA and amplified using a protocol for SISPA (Kramná and Cinek, 2018, pp. 63–68). Samples were processed in 96 well format, maintaining the batches from the nucleic acid extraction step. The batches were distributed across three plates. On each plate three mock community samples were added, equally distributed between batches. All equipment and materials were cleaned using disinfectant wipes containing chlorhexidine (Starlab, Milton Keynes, UK, ref. XTM353-C) to remove contaminations. The reverse-transcription step was performed in a class II microbiological safety cabinet. Separate thermocyclers and equipment were dedicated to the reverse-transcription and amplification steps and kept in separate areas of the laboratory. SISPA was performed as described in Chapter 3. Successful PCR amplification was confirmed using agarose gel electrophoresis. A 2% agarose gel in 1× Tris-acetate-EDTA buffer and SYBR Safe (Thermo Fisher Scientific, Loughborough, UK, ref. S33102) was used to run 2 µl of sample in 1 µl loading dye and 2 µl nuclease-free water (Thermo Fisher Scientific, Loughborough, UK, ref. AM9937).

Amplified DNA was purified using the Genomic DNA Clean & Concentrator-10 kit (Zymo Research, Freiburg im Breisgau, Germany, ref. D4011). To each sample, 240 µl sample binding buffer was added. The sample was then added to the spin column and centrifuged at 15,000 × *g* for 1 minute. The spin column was transferred to a new collection tube. Then 200 µl wash buffer was added and spun down at 15,000 × *g* for 1 minute, twice. Then the spin column was transferred to a 1.5 ml Eppendorf tube and 20 µl elution buffer heated to 65 °C was added according to manufacturer's recommendations. Samples were stored at -20 °C.

4.2.15 Library preparation and sequencing

Library preparation and sequencing was performed by David Baker, Molly Millar and Tristan Seecharran. Purified DNA was normalised to 5ng/µl with 10mM Tris-HCl. Library preparation and sequencing was performed by David Baker. Briefly, 0.5 µl of Tagmentation Buffer 1 was mixed with 0.5 µl Bead Linked Transposomes (Illumina, Cambridge, UK, ref. 20018704) and 4 µl nuclease-free water in a master mix, and 5 µl added to a 96 well plate. 2 µl of DNA normalised to 5 ng/µl was pipette-mixed with 5 µl of the Tagmentation mix and heated to 55 °C for 15 minutes. A PCR master mix was made using 10 µl KAPA 2G Fast Hot Start Ready Mix (Merck, Gillingham, UK, ref. KK5601) and 2 µl PCR grade water per sample. Of the PCR master mix, 12 µl was added to each well to be used in a 96-well plate, and 1 µl of 10µM primer mix containing both P7 and P5 Illumina barcodes (Perez-Sepulveda *et al.*, 2021) were added to

each well with 10 bp unique dual index barcodes used. Finally, 7 µl tagmentation mix was added and mixed. The PCR was run at 72 °C for 3 minutes, 95 °C for 1 minute, then 14 cycles of 95 °C for 10 seconds, 55 °C for 20 seconds and 72 °C for 3 minutes. The libraries were quantified using the Promega QuantiFluor® dsDNA System (Promega, Chilworth Southampton, UK, ref. E2670) and measured on a GloMax® Discover Microplate Reader (Promega, Chilworth Southampton, UK, ref. GM3000). Libraries were pooled following quantification in equal quantities. The final pool was size-selected using SPRI beads at 0.5X concentration, followed by SPRI beads size selection at 0.7X concentration, using Illumina DNA Prep, (M) Tagmentation sample purification beads (Illumina, Cambridge, UK, ref. 20060059). The final pool was quantified on a Qubit 3.0 Fluorometer and run on a D5000 ScreenTape (Agilent, Stockport, UK, ref. 5067-5579) using the Agilent TapeStation 4200 (Agilent, Stockport, UK) to calculate the final library pool molarity. The pool was sent to Novogene for 150 base-pair pair-ended sequencing on five lanes of the Illumina NovaSeq X Plus platform, including a 1% PhiX Control. Samples P007-T3, P008-T0, P009-T0, P045-T0 and Donor-B-3 produced no reads during the sequencing run at Novogene and were sequenced at the Quadram Institute. The libraries were pooled and run on an Illumina NextSeq 2000 at a final concentration of 750 pM on an Illumina NextSeq 20000 instrument using a P3 300-cycle flow cell (Illumina, Cambridge, UK, ref. 20040561). Each were run following the Illumina recommended denaturation and loading recommendations and included a 1% PhiX Control v3 spike-in (Illumina, Cambridge, UK, ref. FC-110-3001)

4.2.16 Bioinformatics analysis

4.2.16.1 Sequencing quality control

Host reads were removed, adapter sequences were trimmed, and reads were quality-filtered by Andrea Telatin using the cleanup v1.4-11-gf422ea8 Nextflow pipeline (Telatin, 2022), run using Nextflow v23.04.2 (Di Tommaso *et al.*, 2017) (Fig. 4.2). Briefly, reads were mapped against a custom database containing a modified version of human reference genome GRCh37.p13 (hg19) with masked viral sequences (Handley, 2020), the phage *Sinsheimervirus phiX174* genome (NCBI ref. AF176027.1), and the *Severe acute respiratory syndrome coronavirus-2* genome (NCBI ref. NC_045512.2). Reads were then processed using fastp v0.23.2 (Chen *et al.*, 2018; Chen, 2023), to trim adapter sequences, poly-G sequences and filter reads below 50 bp. Read quality was then assessed through a MultiQC v1.13.dev0 (Ewels *et al.*, 2016) report generated using the fastp output.

Cleaned reads of the mock community controls were then mapped to the reference genomes of the respective viruses (Table 4.1) using Bowtie 2 v2.5.1 (Langmead and Salzberg, 2012). Read mapping statistics were extracted using SAMtools v1.17 (Danecek *et al.*, 2021) idxstats command and reformatted, for further analysis in R 4.3.1 in RStudio v2023.06.1 using the

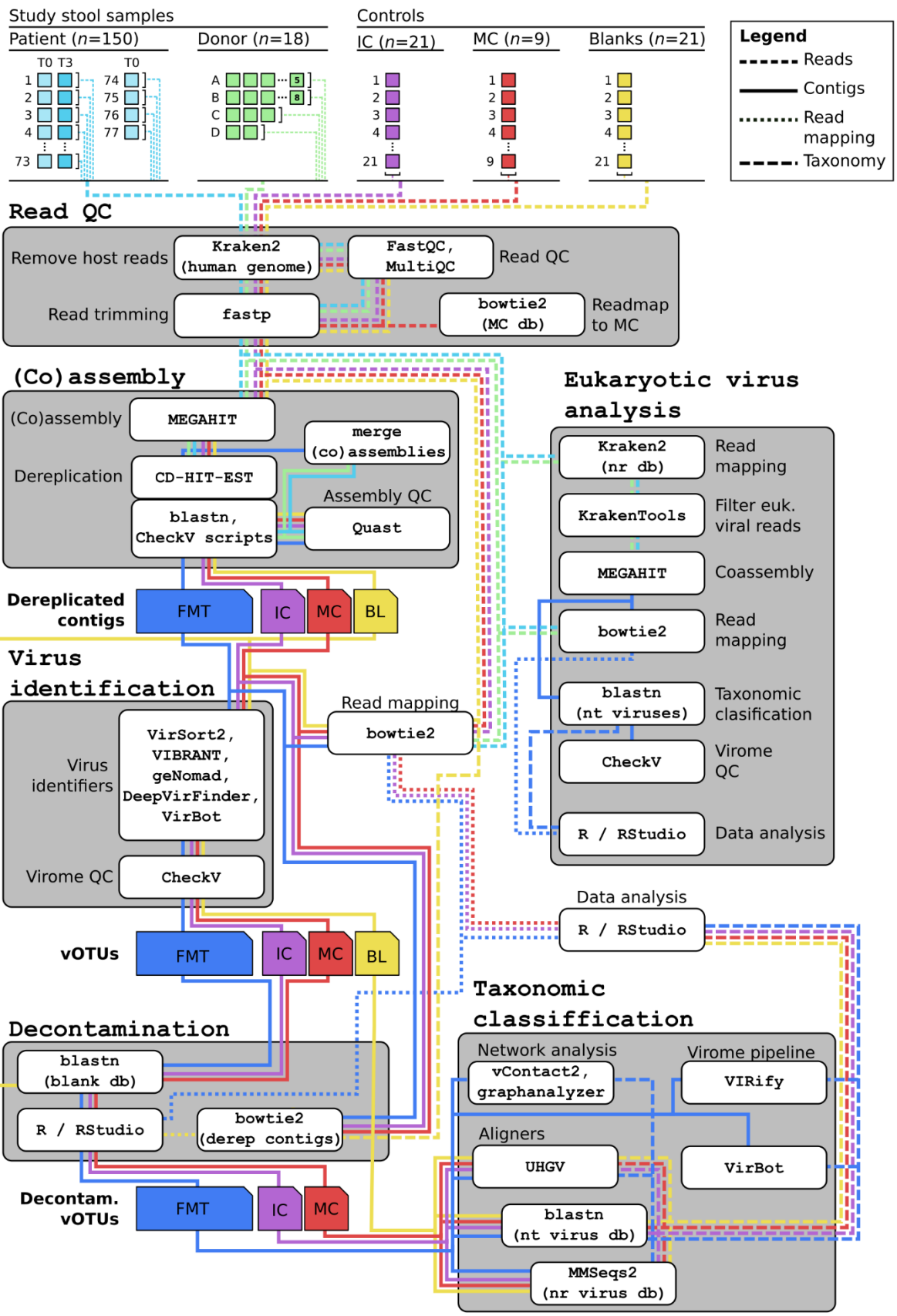
Study stool samples
Patient (n=150)
Donor (n=18)
Controls
IC (n=21)
MC (n=9)
Blanks (n=21)
T0 T3
T0
Legend
Reads
Contigs
Read mapping
Taxonomy
Read QC
Remove host reads
Kraken2 (human genome)
FastQC, MultiQC
Read QC
Read trimming
fastp
bowtie2 (MC db)
Readmap to MC
(Co)assembly
(Co)assembly
MEGAHIT
merge (co)assemblies
Dereplication
CD-HIT-EST
Assembly QC
blastn, CheckV scripts
Quast
Dereplicated contigs
FMT
IC
MC
BL
Virus identification
Read mapping
bowtie2
Virus identifiers
VirSort2, VIBRANT, geNomad, DeepVirFinder, VirBot
Virome QC
CheckV
vOTUs
FMT
IC
MC
BL
Decontamination
blastn (blank db)
R / RStudio
bowtie2 (derep contigs)
Decontam. vOTUs
FMT
IC
MC
Eukaryotic virus analysis
Kraken2 (nr db)
Read mapping
KrakenTools
Filter euk. viral reads
MEGAHIT
Coassembly
bowtie2
Read mapping
blastn (nt viruses)
Taxonomic clasification
CheckV
Virome QC
R / RStudio
Data analysis
Data analysis
R / RStudio
Taxonomic classiffication
Network analysis
vContact2, graphanalyzer
Virome pipeline
VIRify
Aligners
UHGV
VirBot
blastn (nt virus db)
MMSeqs2 (nr virus db)

Figure 4.2: Bioinformatics pipeline. Data flow through the bioinformatics pipeline. Lines are coloured by the sample origin, with light blue and green corresponding to patient and donor samples, respectively, and dark blue corresponding to the FMT data set, consisting of the merged data of all patient and donor samples. Purple, red, and yellow lines indicate data relating to IC, MC, and blank samples, respectively. In the sample overview at the top of the figure, small brackets indicate groups of samples that were co-assembled in the main analysis pipeline. There are four patients for which only T0 samples were available, and these were assembled individually.

tidyverse 2.0.0 (Wickham *et al.*, 2019), ggpubr 0.6.0 (Kassambara, 2023), ggh4x (van den Brand, 2023) and ggsci 3.0.0 (Xiao, 2023) packages. MC virus abundance in each control sample was calculated by normalising the read count by the genome length in kilobases and millions of reads of each sample (RPKM).

Table 4.1: Mock community virus strains and reference genomes.

Abbr.	Strain name	Species	Nucleic acid	Segm.	Length (kb)	%GC	NCBI Accession
BVDV-1	Bovine viral diarrhoea virus 1 Ky1203nc	*Pestivirus bovis*	ssRNA	-	12.3	45	MW250798.1
Det7	Salmonella virus Det7	*Kuttervirus Det7*	dsDNA	-	157.5	44	NC_027119.1
M13	Escherichia virus M13	*Inovirus M13*	ssDNA	-	6.4	40	NC_003287.2
MHV-68	Murine gammaherpesvirus-68	*Rhadinovirus muridgamma4*	dsDNA	-	119.5	47	NC_001826.2
P22	Salmonella virus P22	*Lederbergvirus P22*	dsDNA	-	41.7	47	NC_002371.2
Qbeta	Escherichia virus Qbeta	*Qubevirus durum*	ssRNA	-	4.2	48	NC_001890.1
RV-A	Simian rotavirus SA-11	*Rotavirus A*	dsRNA	1	3.3	33	NC_011507.2
				2	2.7	32	NC_011506.2
				3	2.6	28	NC_011508.2
				4	2.4	34	NC_011510.2
				5	1.6	31	NC_011500.2
				6	1.4	38	NC_011509.2
				7	1.1	33	NC_011502.2
				8	1.1	33	NC_011501.2
				9	1.1	35	NC_011503.2
				10	0.8	40	NC_011504.2
				11	0.7	38	NC_011505.2
SINV	Sindbis virus AR339	*Sindbis virus*	ssRNA	-	11.7	50	NC_001547.1
T5	Escherichia virus T5	*Tequintavirus T5*	dsDNA	-	121.8	39	NC_005859.1

Abbr., abbreviation; Segm., segment number.

4.2.16.2 Co-assembly

Reads were assembled using a custom co-assembly pipeline built in Nextflow. For each individual patient, read files of pre- and post-FMT samples were co-assembled using MEGAHIT v.1.2.9 (D. Li *et al.*, 2015; Li *et al.*, 2016) with default settings. Similarly, the reads from all samples from each donor were co-assembled, and all samples corresponding to each of the controls, i.e., IC, MC and BL, were co-assembled. Contigs were then dereplicated in two steps. First, using CD-HIT-EST v.4.8.1 (Li and Godzik, 2006; Fu *et al.*, 2012) contigs were clustered based on ≥95% sequence identity, and ≥85% alignment coverage of the shorter sequence, keeping only the longest sequence in each cluster. To additionally remove circularly permutated redundant sequences, BLAST v.2.14.0 was used to create a database of contigs from CD-HIT-EST, and an all-versus-all alignment was performed using blastn. Using the accessory scripts "anicalc.py" and "aniclust.py" from CheckV v.1.0.1 (Nayfach, Camargo, *et al.*, 2021), contigs were again clustered at 95% average nucleotide identity and 85% coverage of the shortest sequence. Contigs were filtered from the original assembly using the list of contig IDs with SeqKit v.2.2.0 (Shen *et al.*, 2016). Reads from patient samples were mapped collectively to their respective co-assemblies for quality control using Bowtie 2, and read mapping statistics were calculated using the SAMtools coverage command. The dereplicated patient and donor co-assemblies were then merged, and a second round of dereplication was performed to remove redundant sequences among individuals, producing the initial FMT data set. Quast v.5.2.0 (Gurevich *et al.*, 2013) was used to assess the quality of the coassemblies.

4.2.16.3 Virus identification

For the identification of viral sequences in the FMT and IC, MC and BL control data sets, an ensemble of virus identification tools was employed, consisting of VirSorter2 (J. Guo *et al.*, 2021), VIBRANT (Kieft, Zhou and Anantharaman, 2020), geNomad (Camargo *et al.*, 2023), DeepVirFinder (Ren *et al.*, 2020), and VirBot (Chen *et al.*, 2023) to identify a large variety of viruses. For DeepVirFinder, data sets were split into blocks of 50,000 sequences, and the program was run on each block in parallel. DeepVirFinder output files were then concatenated, and only sequences with a score >0.9 and a p-value <0.01 were considered a positive identification. Sequences identified as viral by any of the tools were selected and analysed using the CheckV v.1.0.1 (Nayfach, Camargo, *et al.*, 2021) end-to-end command and CheckV database version 1.5. Contigs meeting the following criteria were selected as viral sequences: 1) at least "Low quality", 2) at least 1 kb in length, 3) at least one viral gene detected, 4) no warnings. A list of contig IDs was saved and used to extract sequences from the dereplicated data sets using the SeqKit grep command to produce virus operational taxonomic unit (vOTU) data sets.

4.2.16.4 Decontamination

Data from the BL samples was used to remove likely contaminants from the FMT, IC and MC data sets. For the FMT data set, three criteria were used to identify possible contaminants. First, the number of positive samples was determined for each contig. A contig was marked as present in a sample if there was at least 25% coverage at 0.25X coverage depth. Contigs present in ≥20 samples exceeded 1.5 × IQR and were marked as overrepresented sequences. Second, FMT contigs were aligned to BL contigs using blastn. Contigs aligning with ≥95% sequence identity and ≥85% coverage of the shortest sequence were marked as a blast hit against the BL data set. Third, reads of the blank samples were mapped to the FMT data set. Contigs with ≥20% coverage exceeded 1.5 × IQR and were marked as read hits. Any contig meeting at least two out of three criteria 1) overrepresented sequence, 2) blast hit, or 3) read hit, was marked as a contaminant. For the IC and MC, blast and read hits were determined in the same way. IC and MC contigs were removed if a contig was a blast and/or a read hit. Using the SeqKit grep command, contaminated contigs were removed from the data sets. Contigs that failed the decontamination check were extracted and aligned to NCBI non-redundant nucleotide database (NT) viral sequences for taxonomic annotation. KronaTools v2.8.1 (Ondov, Bergman and Phillippy, 2011) command ktImportBLAST was used to produce a Krona plot and visualise taxonomy of contaminants.

4.2.16.5 Abundance calculation

Reads of individual patient and donor sample were mapped to the dereplicated FMT data set and control sample reads were mapped to their respective dereplicated data sets using Bowtie 2. Read mapping statistics were extracted using the SAMtools coverage and idxstats commands. For the FMT data set samples, custom Python scripts were used to filter read mapping statistics of the viral contigs using the previously generated list of vOTU IDs to reduce the file size. vOTU abundance was calculated from the read count, normalised by the contig length in kilobases and the number of viral reads in the sample in millions of reads (RPKM). Relative abundance was calculated as the percentage of the total RPKM in each sample.

4.2.16.6 Taxonomic annotation

Decontaminated FMT vOTUs were annotated using an ensemble of taxonomic annotation approaches. First, phage annotation was performed on the FMT data set by Dr Shen-Yuan Hsieh through gene sharing network analysis using vContact2 v0.11.3 (Jang *et al.*, 2019). Briefly, genome annotations were produced using Pharokka v1.0.1 and Pharokka database version 1.4. The protein sequences and protein-to-genome mappings were generated using a perl script (Cook, 2022). Then, the 11 November 2023 curated phage protein database generated by INPHARED (Cook *et al.*, 2021) was downloaded and merged with the FMT protein sequences and protein-to-genome mappings. Using vContact2, a gene sharing network was generated, and the graphanalyzer script v1.6.0 from the MetaPhage pipeline (Pandolfo *et al.*, 2022) was

used to assign taxonomies to the contigs based on the gene sharing network. In parallel, three alignment-based annotations of FMT, IC, and MC were performed by Rik Haagmans. First, vOTUs were classified against the unified human gut virome (UHGV) database v0.4 using the accompanying analysis tool uhgv-tools (Nayfach, 2023). The database was downloaded using the uhgv-tools download command, and contigs were aligned to the database using the uhgv-tools classify command. Second, the FMT, IC, and MC datasets were aligned against NT database virus sequences by blastn using a custom script to enable parallel alignment of sequences to improve alignment speed. After alignment, the KronaTools ktClassifyBLAST command was used to determine the taxonomic classification from the blastn output, and the TaxonKit v.0.15.0 (Shen and Ren, 2021) lineage command was used to determine the taxonomic lineage. Third, contigs were classified against the NCBI non-redundant protein (NR) database virus sequence using MMSeqs2 (Steinegger and Söding, 2017, 2018; Mirdita *et al.*, 2021). The NR database was downloaded on the 24 November 2023 and built using the MMSeqs2 databases command, and the MMSeqs2 filtertaxseqdb was used with the option taxid-list set to 10239, the NCBI taxon ID for viruses, to reduce the database to virus protein sequences only. The MMSeqs2 easy-taxonomy command with the options report-mode and tax-lineage set to 1 was then used to classify contigs based on protein sequences. Additionally, FMT sequences were analysed by Dr Shen-Yuan Hsieh using VIRify (Rangel-Pineros *et al.*, 2023) and VirBot (Chen *et al.*, 2023). For the FMT sequences, taxonomic annotation was performed as follows: network analysis-based taxonomy was used where available. Remaining contigs were assigned UHGV-based annotation where available unless NT- and NR-based alignment agreed on a different taxonomy. When UHGV annotation was absent, NT-based annotation was used, then NR-based annotation. Any remaining unannotated contigs were annotated using VIRify, then VirBot. For the IC and MC data sets, contigs were annotated only based on UHGV, NT and NR using the same approach as for the FMT data set.

4.2.16.7 Alpha and beta diversity analysis

Alpha diversity of individual samples was calculated using the R package phyloseq v.1.46 (McMurdie and Holmes, 2013). Beta diversity of all patient and donor sample pairs was calculated using the R package vegan v.2.6-4, by calculating the Bray-Curtis dissimilarity. Additionally, the Aitchison distance was calculated between each sample pair (Gloor *et al.*, 2017). Briefly, first read numbers are statistically imputed to remove zeroes from the count table using the R package zCompositions v.1.5.0 (Palarea-Albaladejo, 2024) cmultRepl function, and the centred log ratio (CLR) was calculated for the vOTUs in each sample. The Aitchison distance was then calculated on the CLR-transformed read counts using the base R distance function.

4.2.16.8 Virus transfer analysis

Transferred vOTUs were determined by the coverage and average coverage depth of each. vOTUs were marked as present in a sample when at least 50% of the contig was covered by reads at an average depth of at least 1X over the entire length of the contig.

4.2.16.9 Targeted analysis of eukaryotic viruses

To enrich for eukaryotic virus reads, a Kraken2 database was built of the NR sequences database using Kraken2 v2.1.3 (Wood, Lu and Langmead, 2019). First, the NR database was downloaded on 18 November 2023 using the kraken2-build command with the download-taxonomy and protein options, followed by the kraken2-build --download-library command with the nr and protein options. Then, the NR database including GI-less proteins was downloaded and merged with the NR database. Then, the merged NR database was used to build a NR Kraken2 database with the kraken2-build command and the build and protein options. Reads of patient and donor samples were mapped to the database using Kraken 2. Then, the KrakenTools (Lu *et al.*, 2022) "combine_kreports.py" script was used to create a reads table, and TaxonKit was used in combination with a custom script to generate taxonomy tables of hits. Next, the KrakenTools script extract_kraken_reads.py was used to extract reads mapping to eukaryotic viruses, using the options exclude and include_children, and providing the following NCBI taxon IDs: 131567 (cellular organisms), 3044425 (Ainoaviricetes), 2731619 (Caudoviricetes), 2732411 (Faserviricetes),2732449 (Huolimaviricetes), 2732010 (Laserviricetes), 2842243 (Leviviricetes), 2732413 (Malgrandaviricetes), 2841637 (Tokiviricetes), and 2732460 (Vidaverviricetes). This effectively excluded all cellular organisms, leaving all viruses except for the described virus classes. Classes were selected based on the International Committee on Taxonomy of Viruses (ICTV) Virus Metadata Resource – Master Species List 38 v2, selecting all classes of bacterial or archaeal origin. The list of classes was then cross-referenced with the virus host overview on Expasy ViralZone (Hulo *et al.*, 2011), and any class for which a eukaryotic host was noted was removed from the list, leaving only the classed described above. The read extraction script was used in a custom Nextflow pipeline to enable extraction of reads from multiple samples in parallel. Then, the eukaryotic virus-enriched reads of all donor and patient samples were co-assembled using MEGAHIT and the resulting co-assembly was aligned to the NT virus sequences database using blastn. Taxonomic annotation was produced from the alignments using the KronaTools ktClassifyBLAST command, and taxonomic lineages were added using the TaxonKit lineage command. Reads of each donor and patient sample were then mapped to the co-assembly, and the abundance of contigs was calculated by normalising the read count by the length of the contig in kbp and the total number of reads in each sample in Mbp (RPKM). Contig ≥1 kbp were extracted and their quality was assessed using the CheckV end-to-end command.

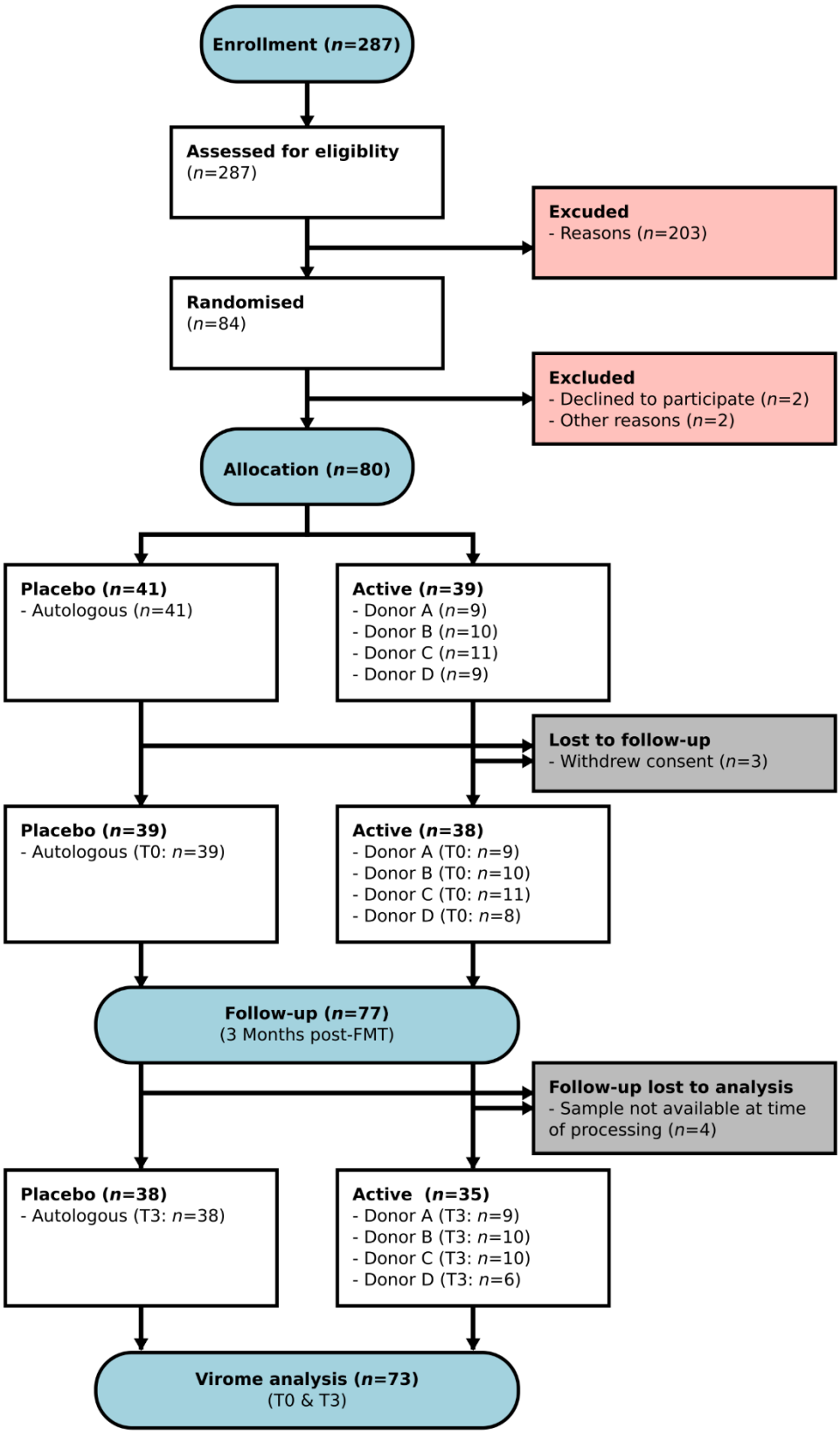

Figure 4.3: CONSORT flow chart. Flow chart of the number of patients enrolled and allocated in the study, and the number of baseline (T0) and 3-month follow-up (T3) samples used in the virome analysis for each group.

Table 4.2: Patient demographics

Characteristic	**Placebo**, N = 41[1]	**Donor**, N = 39[1]
Sex		
Female	35 (85%)	33 (85%)
Male	6 (15%)	6 (15%)
Assigned treatment		
Donor-A	-	9 (23%)
Donor-B	-	10 (26%)
Donor-C	-	11 (28%)
Donor-D	-	9 (23%)
Placebo	41 (100%)	-
Samples used in virome study		
None (drop-out)	2 (4.9%)	1 (2.6%)
T0 only	1 (2.4%)	3 (7.7%)
T0 & T3	38 (93%)	35 (90%)

[1] *n* (%)

4.2.16.10 Data processing and visualization

All final data analysis was performed using R markdown files with R v.4.3.1 with RStudio v.2023.06.1. Data was handled using the tidyverse 2.0.0 (Wickham *et al.*, 2019) package collection, and ggpubr 0.6.0 (Kassambara, 2023), ggh4x (van den Brand, 2023) and ggsci 3.0.0 (Xiao, 2023) were used for the plotting of graphs.

4.3 Results

Patients were recruited between 23 May 2019 and 26 January 2022. In total, 287 patients were assessed for eligibility, 203 were excluded and 84 patients were randomised (Fig. 4.3). Of the 84 randomised patients, two declined to participate and two did not participate for other reasons, leaving 39 patients randomised to active FMT and 41 patients to placebo FMT. Patients randomised to active transplant were allocated at equal ratios to four donors (Table 4.2). A different number of transplants was produced for each donor: Donor A (*n*=5), B (*n*=8), C (*n*=3), D (*n*=2). Three patients withdrew from the study after allocation, two who received placebo and one who received active treatment. Of the remaining patients, 86% were female (*n*=66) and 14% were male (*n*=11). A baseline (T0) faecal sample was collected during the initial assessment of each patient four to six weeks before FMT (Fig. 4.1A). Post-

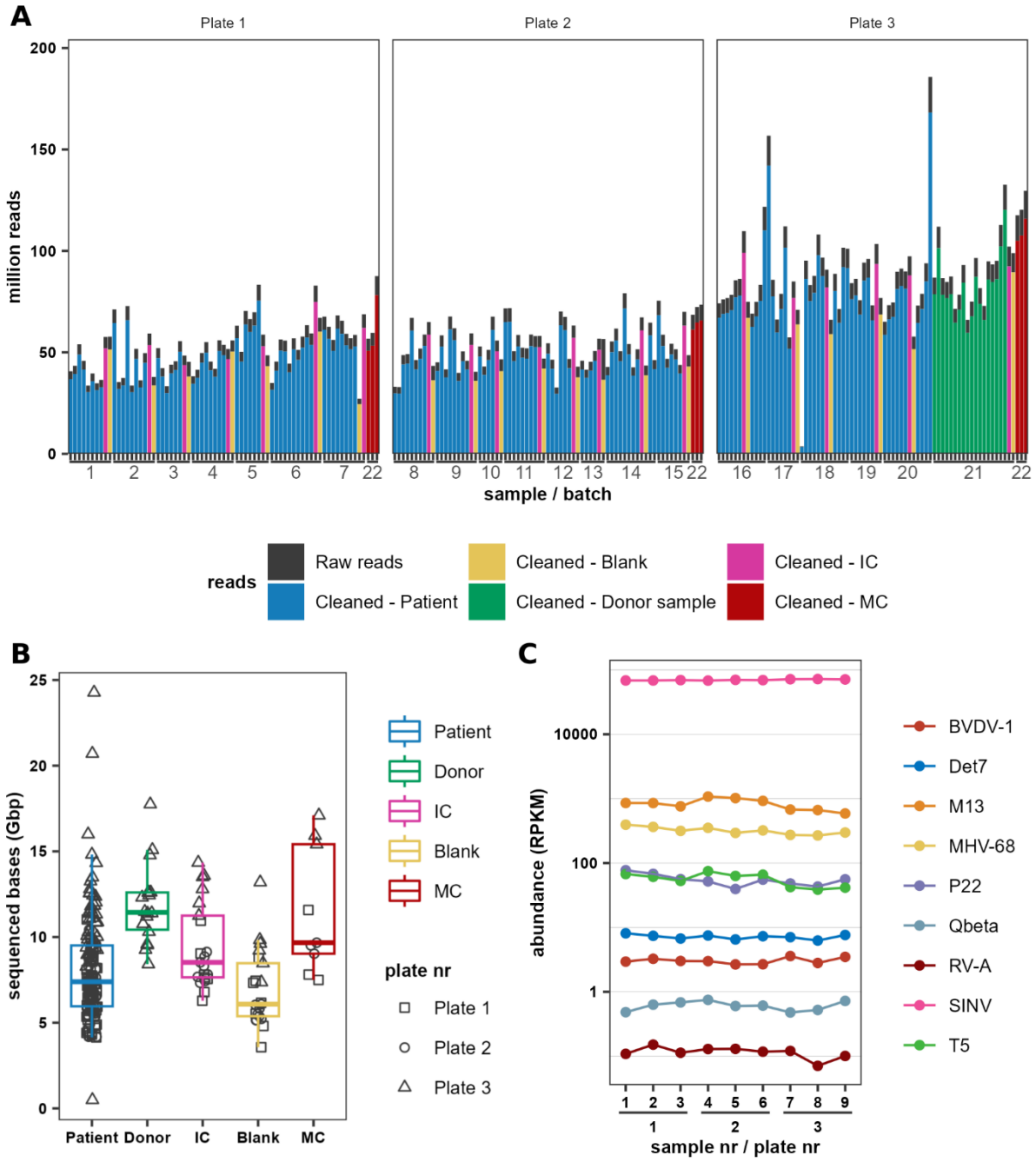

Figure 4.4: Read quality control. **A:** Number of reads acquired for each sample, ordered by VLP extraction batch and library preparation plate. Gray bars indicate reads that did not pass quality control. IC: internal control, MC: mock community. **B:** Number of sequenced base pairs by sample type. Shapes of data points indicate the plate number. **C:** Abundance of mock community viruses, as measured by the number of reads normalised for virus genome length in kb and millions of reads in each sample (RPKM).

FMT, three patients withdrew from the study. During the follow-up three months post-FMT (T3), a second faecal sample was collected. At the time of sample processing, T3 samples of four patients were not available, one in the placebo group and one in the active group (Fig 4.2).

A total of 150 patient samples, consisting of 73 pairs of T0 and T3 samples and four single T0 samples, and 18 healthy donor samples were processed in batches, each including an IC and BL control sample. Reverse-transcription and amplification of viral nucleic acid and subsequent generation of Illumina sequencing libraries was performed on three 96-well plates on which three MC samples were included (Fig. 4.1B). In total, 73 pairs of T0 and T3 patient samples, four patient T0 samples without matching T3 sample, 18 donor samples, 21 IC samples, 21 BL samples, and 9 MC samples were processed and sequenced (Fig. 4.1C), yielding a total of 14.38 billion raw reads and a total of 1.96 Tbp. Filtering of host-derived and low-quality reads removed 9.9% of reads on average, leaving on average 56.13 million patient sample reads (min: 3.394×10^6, max: 168.2×10^6 reads) and 81.55 million donor sample reads (min: 59.66×10^6, max: 120.3×10^6) (Fig. 4.4A). Samples amplified on plate 3 produced on average 66.6% more reads than plates 1 and 2, increasing the sequencing depth of all donor samples and the patient samples processed on that plate (Fig. 4.4B). There was no systematic difference in MC virus abundance between samples from different plates, suggesting the increased sequencing depth on plate 3 samples did not differentially affect virus abundances (Fig. 4.4C).

Co-assemblies were produced from the cleaned reads of the samples corresponding to individuals and different control sample types, producing 77 patient co-assemblies, 4 donor co-assemblies, and an IC, an MC and a BL co-assembly (Fig. 4.3). Co-assemblies were dereplicated at the species level (Roux *et al.*, 2019) and contained on average 102,285 (min: 11,062; max: 510,504) contigs for patients and 320,119 (min:145,114; max: 537,354) contigs for donors. For the controls, the IC co-assembly contained 117,698 contigs, the MC co-assembly contained 36,089 contigs, and the BL co-assembly contained 66,529 contigs. There were large differences between patients and donors in the number of contigs and the total length of the co-assembly (Fig. 4.5A). Taken across all co-assemblies, the median number of contigs was 25,177 contigs ≥500 bp, 6,520 contigs ≥1 kbp, 218 contigs ≥5 kbp, 48.5 contigs ≥10 kbp, and 3 contigs ≥50 kbp (Fig. 4.5B). Co-assemblies of two patients had no contigs ≥10 kbp, and 12 patient co-assemblies did not have contigs >50 kbp. To assess assembly quality, patient reads were mapped collectively to their respective co-assemblies, showing that on average 57.7% of reads map to contigs ≥500 bp (Fig. 4.5C). While there was large variation in sequencing depth and number of contigs between samples, there was no statistically significant correlation between the sequencing depth (Fig. 4.5D). Additionally, the average coverage depth of contigs was higher for co-assemblies with fewer contigs, suggesting reduced diversity in samples with lower contig counts and not insufficient sequencing depth.

The patient and donor co-assemblies were merged, producing an FMT data set of 6,857,145 contigs. The contigs were dereplicated at the species level (Roux *et al.*, 2019), removing 66% of contigs (Table 4.3). An ensemble of virus identification tools was used to identify viral

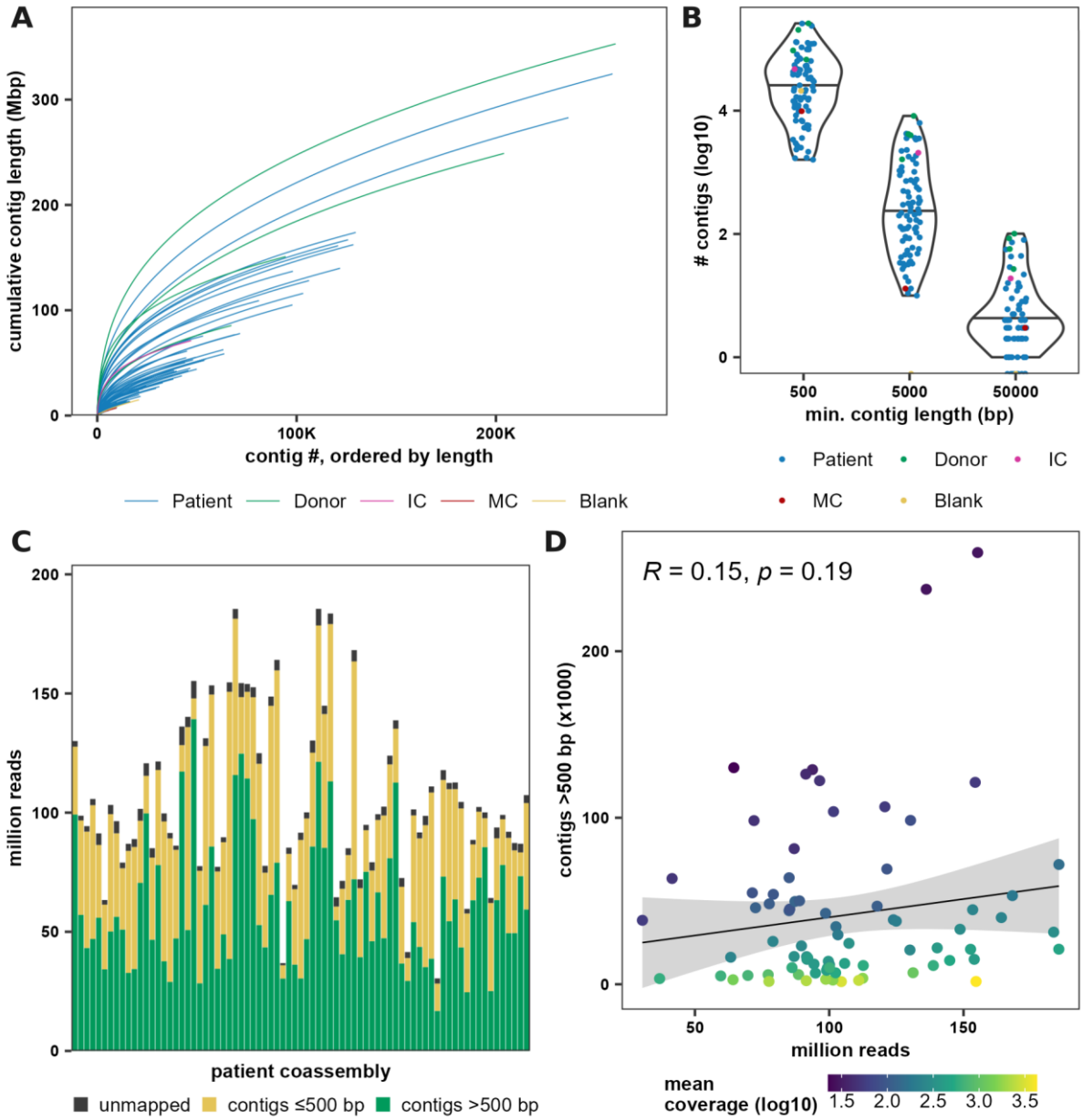

Figure 4.5: Co-assembly quality control. **A:** Cumulative contig length of contigs ordered by length from largest to smallest. **B:** Number of contigs larger than 500, 5000 and 50,000 bp. **C:** Total number of reads of all samples from each patient that are unmapped, or map to contigs shorter and longer than 500 bp in the respective patient co-assembly. **D:** Relationship between number of contigs in each patient co-assembly, and number the total number of reads, with each patient co-assembly coloured by the mean coverage of contigs in each sample.

sequences in the FMT (*n*=202,612), IC (*n*=7,917), MC (*n*=4,002) and BL (*n*=6704) data sets. The FMT data set contained a sub-population of sequences corresponding to around 5 kbp total genome length, likely representing phages in the *Microviridae* family, and another population of genomes corresponding to around 50 kbp genome length, likely corresponding to

Table 4.3: Number of contigs in the participant and control data sets at various stages of refinement.

	Data set			
Refinement stage	**FMT**[1]	**IC**[1]	**MC**[1]	**BL**[1]
Raw contigs	6,857,145 (100%)	117,698 (100%)	36,089 (100%)	66,529 (100%)
Dereplicated contigs	2,335,952 (34%)	103,207 (88%)	31,487 (87%)	62,064 (93%)
Classified viral	202,612 (3.0%)	7,917 (6.7%)	4,002 (11%)	6,704 (10%)
vOTUs (passed QC)	22,841 (0.3%)	533 (0.5%)	34 (0.1%)	124 (0.2%)

Abbreviations: FMT: faecal microbiota transplantation, IC: internal control, MC: mock community, BL: blank, QC: quality control, vOTU: virus operational taxonomic unit. [1] *n* (% of total)

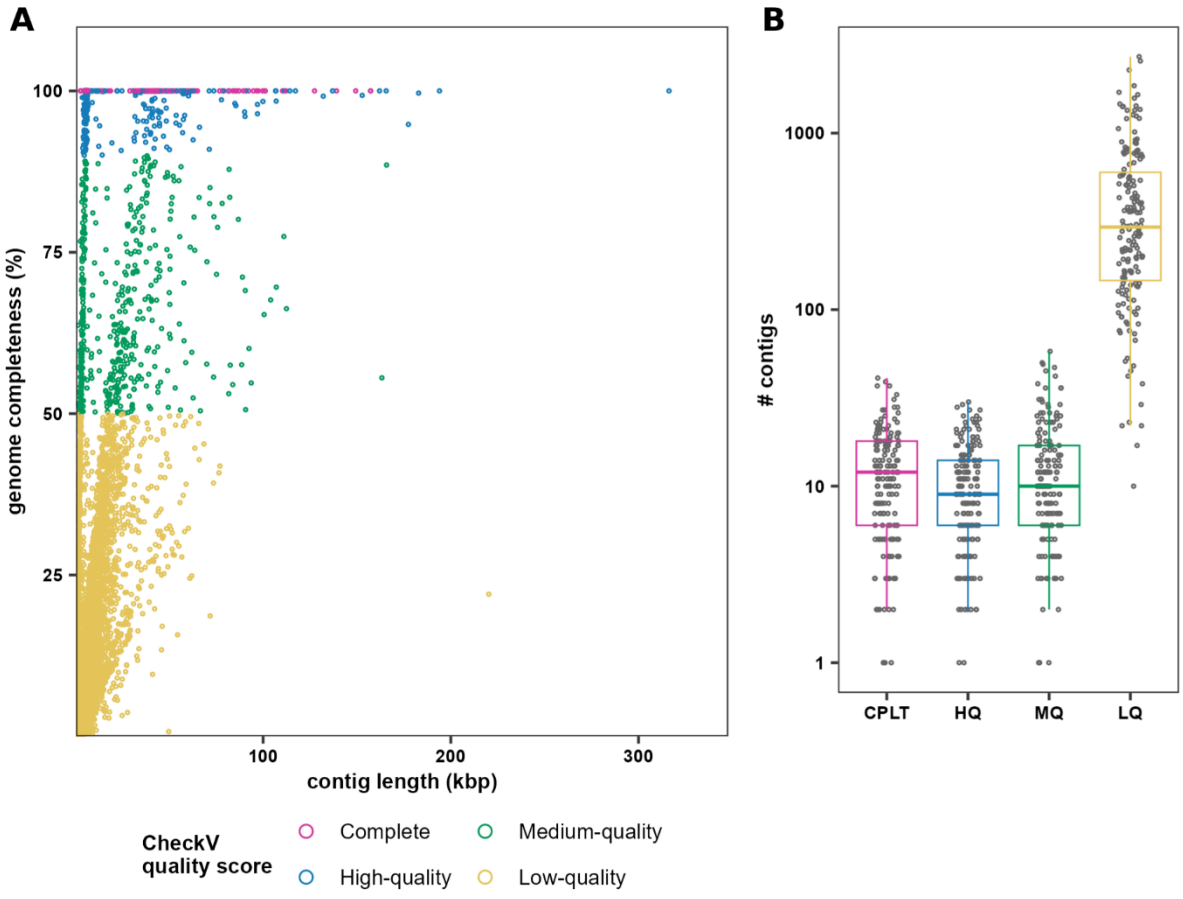

Figure 4.6: Virus sequence quality control. A: Sequence length and genome completeness, as estimated by CheckV, of vOTUs in the FMT data set. **B:** Number of contigs from the FMT data set that are present in patient and donor samples, separated by genome quality as estimated by CheckV. CPLT: complete, HQ: high-quality, MQ: medium-quality, LQ: low-quality.

Caudoviricetes phages (Fig. 4.6A). Quality assessment using CheckV (Nayfach, Camargo, *et al.*, 2021) indicated 517 complete genomes, 332 high-quality genomes (≥90% complete), 522 medium-quality genomes (≥50% complete), and 88,442 low-quality genomes (<50% complete) in the FMT data set, while the quality of 112,799 contigs could not be determined (Fig. 4.6B). After quality filtering, a total of 22,841 vOTUs with sequences ≥ 1 kb and evidence of viral origin remained in the FMT data set, and 533, 34, and 124 vOTUs in the IC, MC, and BL data sets, respectively (Table 4.3).

4.3.1 Mitigation of sample contamination

BL samples produced similar amounts of reads to faecal and control samples (Fig. 4.4A, B) and the BL co-assembly contained several low-quality viral contigs. Thus, taxonomic annotation of the BL viral sequences was performed, and BL reads and contigs were compared to the FMT, IC and MC data sets to remove possible contaminants. While the BL co-assembly did not contain any contigs ≥5 kb, there were over a hundred sequences detected as viral. Of the viral sequences that passed the quality control criteria, those with the highest completeness (>10%) included *Circoviridae* and other viruses in the *Cressdnaviricota* phylum, the families *Parvovirirdae*, *Ackermannviridae*, and *Microviridae*, and the genus *Gammaretrovirus*. Most sequences with lower completeness matched *Caudoviricetes* genomes, and notably six sequences matched the *Orthoherpesviridae* family. The *Gammaretrovirus* sequence had a close match to various murine leukaemia virus sequences. Murine leukaemia viral genomes have been found to contaminate the SuperScript reverse-transcriptase used in this study (Sato, Furuta and Miyazawa, 2010), as well as other lab reagents and clinical samples in general (Smith, 2010; Erlwein *et al.*, 2011; Asplund *et al.*, 2019). *Cressdnaviricota*, *Parvoviridae*, *Herpesviridae* and *Caudoviricetes* viruses have been identified as contaminants in various reagents used for viral metagenomics, including the QIAamp Viral RNA Mini kit used in this study (Smuts *et al.*, 2014; Asplund *et al.*, 2019; Porter *et al.*, 2021). Alignment of unfiltered contigs yielded matches against a wide variety of viruses in the NT virus sequences database. This included matches against phages in the order *Caudoviricetes*, and family *Microviridae*, and taxa present in the MC. Several sequences aligned to the RV-A genus *Rotavirus*, M13 family *Inoviridae*, BVDV-1 genus *Pestivirus*, SINV genus *Alphavirus*, and Qbeta family *Fiersviridae,* indicating that some level of cross contamination between samples may have occurred.

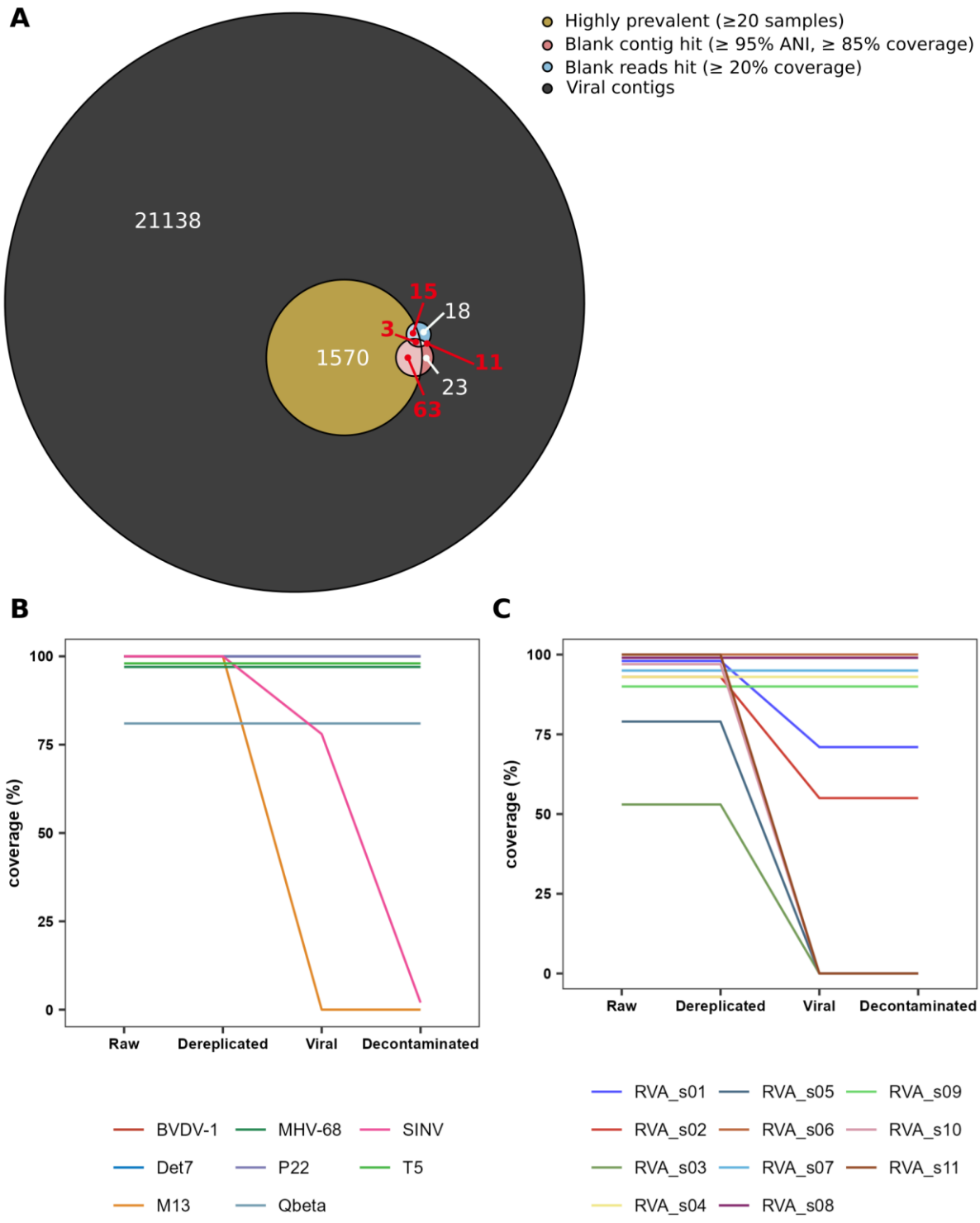

Figure 4.7: Contaminant removal. A: Euler diagram of vOTUs meeting any of three contaminant criteria: 1) highly prevalent sequences, present in ≥20 samples, 2) vOTUs that align to contigs in the blank data set at the species level, and 3) vOTUs to which reads from blank samples map with ≥ 20% coverage of the vOTU. Any vOTU meeting at least two criteria are marked as a contaminant, as shown by the red numbers and markers in the figure. Numbers indicate the number of vOTUs in the respective section. **B, C:** Percentage of the virus

reference genome (**B**) and RV-A segments (**C**) covered by contigs from the MC data set, as determined by the BLAST function in Bandage, during subsequent stages of refinement.

To detect possible contaminations in the FMT data set, FMT vOTUs, were scored based on three criteria, 1) significant alignment of the vOTU sequence to a BL contig, 2) BL reads mapping to the vOTU sequence, and 3) high prevalence of the vOTU among patient and donor samples. Alignment against the BL data set produced 47 vOTUs with at least one hit, and while most vOTUs had only one hit, five vOTUs had more than 150 hits. Nearly all significant hits were against BL contigs <1 kb, and covered <20% of the FMT vOTU, suggesting increased fragmentation in the BL co-assembly and low levels of these virus sequences in the blank samples. Next, BL reads were mapped against the FMT dereplicated data set. While a distinct portion of contigs had close to 100% coverage, only a 1045 bp contig corresponding to a vOTU had 100% coverage at 60,000X coverage depth. Additionally, 17% of vOTUs had <10% coverage, but > 1X average coverage depth, suggesting blank reads mapped to conserved sequences. Therefore, a minimum coverage threshold of 20% was determined to count a vOTU as read hit, based on 1.5 × IQR of contig coverage, to reduce false positives. This marked 100 vOTUs as read hits. Lastly it was noted that for some vOTUs, at least one read mapped to the sequence in >90% of all samples, indicating very high prevalence. To reduce false positives, first a cut-off was empirically determined. To count every vOTU as positive in at least one sample, a minimum coverage of 50% was needed. The number of sequences and the fraction of positive samples at a minimum of 0%, 25% and 75% coverage were compared to the 50% coverage cut-off and a 25% cut-off was decided for the detection of overrepresented vOTUs as this was more sensitive than the 50% cut-off, but less sensitive than counting all contigs with at least one read. At a cut-off of 25% coverage at 1X average coverage depth, 20% of contigs were present in ≤2 samples, 50% of contigs were present in ≤5 samples, and 90% of contigs were present ≤16 samples. The 1.5 × IQR rule was used to determine a cut-off of 20 samples for overrepresentation. Since there were 8 transplants from Donor B and 10 patients received a transplant from Donor B, up to 18 patient and donor samples could reasonably be expected to have viruses in common, even if all individuals had completely unique viromes, which falls below this limit. This marked 1651 vOTUs as overrepresented. To further reduce the chance of false positives, vOTUs that met at least two out of the following three criteria 1) contig hit, 2) read hit, and 3) overrepresentation, were counted as a contaminant (Fig. 4.7A). This removed 92 sequences from the FMT data set. Removed vOTUs mostly matched *Caudoviricetes* and *Microviridae* phages, as well as a *Cytomegalovirus* sequence and *Tomato brown rugose fruit virus* sequences.

BL read and contig hits were also assessed for the IC and MC data sets. vOTUs with a read hit and/or a contig hit were removed. There were 17 vOTUs with read hits and 2 vOTUs with contig hits in the IC data set, with one vOTU meeting both criteria, leading to the removal of 18 vOTUs from the IC data set. Removed vOTUs matched *Microviridae*, *Tobamovirus*, including *Tomato brown rugose fruit virus*, *Caudoviricetes* and *Cytomegalovirus* sequences. Of the 24 vOTUs in the MC data set, 9 had read hits and 6 contig hits, with 5 meeting both criteria. A total of 10 vOTUs were removed from the MC data set, with hits against *Microviridae* and *Sindbis*

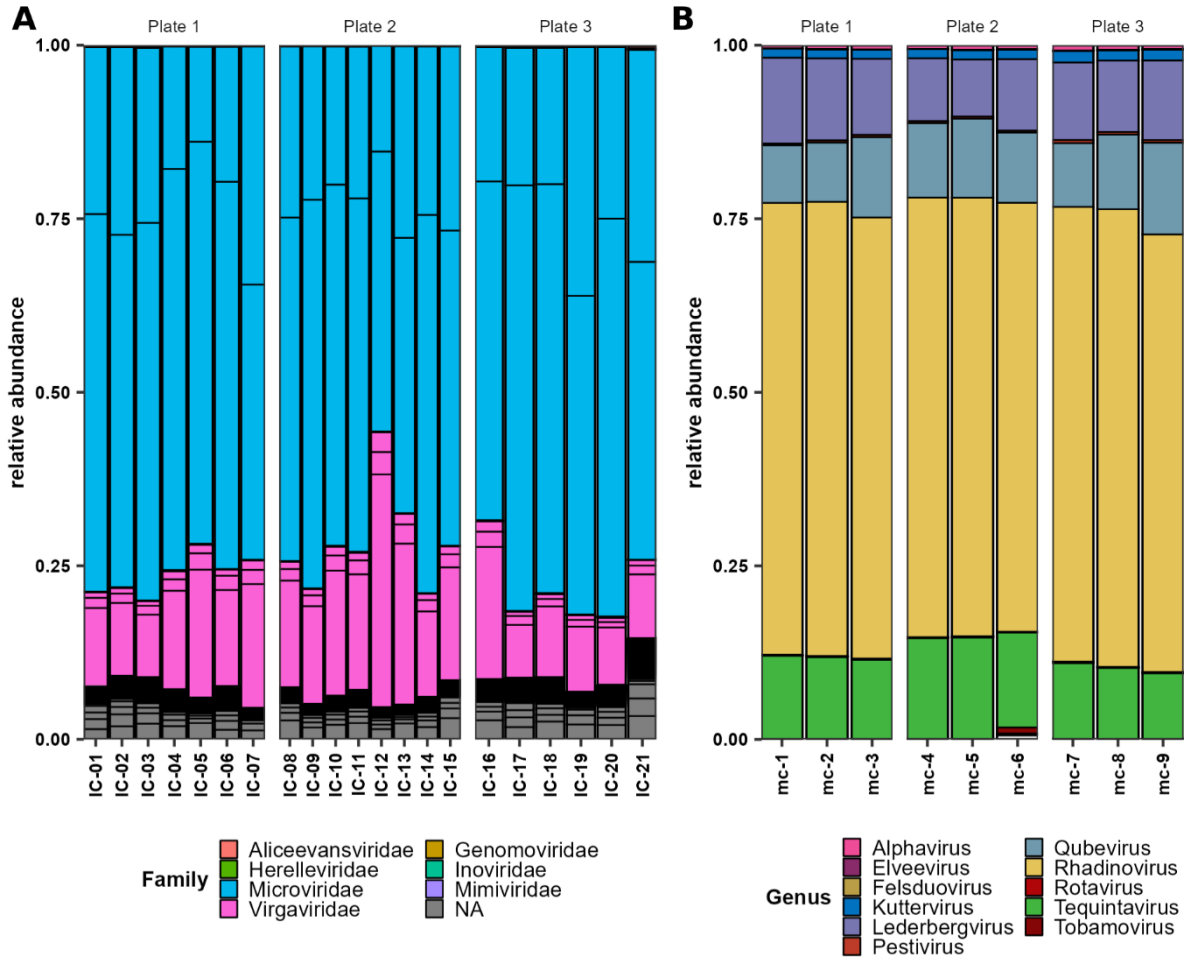

Figure 4.8: Taxonomic analysis of IC and MC samples. Taxonomic annotation of the IC (**A**) and MC (**B**) data set after viral sequence detection and decontamination was performed using the UHGV data set classification tool, alignment of vOTU sequences to the NCBI non-redundant virus nucleotide database using BLASTn, and the NCBI non-redundant virus protein database using MMSeqs2. vOTU relative abundances were determined by read mapping, controlling for contig length and total number of reads in the sample mapping to vOTUs.

virus. To assess the effects of the virus classification and decontamination process, the MC co-assembly, dereplicated co-assembly, vOTU data set and decontaminated vOTU data set were analysed in Bandage (Wick *et al.*, 2015) to determine genome coverage of the assembly through BLAST alignment against the reference genomes (Table 4.1). Before identification of viral sequences, 100% of Det7, 100% of M13, 100% of P22, 99.7% of BVDV, 99.6% of SINV, 98.3% of T5, 97.5% of MHV-68, and 81.0% of Qbeta genomes were covered by the co-assembly (Fig. 4.7B). For RV-A, the co-assembly achieved 52.8% to 100% coverage of genome segments (Fig. 4.7C). Viral identification altered the genome coverage of SINV to 77.8% due to removal of 384 sequences and all 16 M13 sequences were removed, while other viruses remained unchanged. Closer inspection of the removed M13 sequences showed that one sequence was a full M13 genome for which CheckV detected genes, but was unable to mark genes as viral, leading to removal of the sequence. Among the sequences that were removed from the FMT data set due to lack of viral genes, 25 sequences matched other *Inoviridae* genomes, indicating a difficulty in detecting viral genes in this family of viruses by CheckV. For RV-A, segments 3, 5, 10, and 11 had between 55.4% and 100% coverage after dereplication. These segments were 667 - 2,591 bp and contigs did not meet the size requirement of 1 kbp. Decontamination only affected SINV, for which only 1.89% coverage of SINV remained, indicating the presence of SINV reads in the BL samples.

4.3.2 Taxonomic annotation of control samples

Next, alignment-based taxonomic annotation of MC and IC vOTUs was performed using the UHGV and NT virus nucleotide sequences databases, and the NR virus protein sequences database. On average, the five most abundant vOTUs were two *Microviridae* phages (51.2% and 23.5%, respectively), a *Virgaviridae* plant virus (14.7%), a *Caudoviricetes* phage (2.06%) and a *Tobamovirus* plant virus (1.73%). While there was some variation in the relative abundance of individual viruses between samples, there was no systematic difference between plates (Fig. 4.8A). For the MC samples, the genus of each virus was recovered, except M13: *Alphavirus* (SINV), *Kuttervirus* (Det7), *Lederbergvirus* (P22), *Pestivirus* (BVDV-1), *Qubevirus* (Qbeta), *Rhadinovirus* (MHV-68), *Rotavirus* (RV-A), and *Tequintavirus* (T5) (Fig. 4.8B). One *Tobamovirus* sequence was detected, at an average relative abundance of >0.1%, with 98% estimated completeness. Additionally, the phage genera *Elveevirus* and *Felsduovirus* were detected, although their average relative abundance was <0.01% and sequences had low estimated completeness (<4%), together suggesting some contaminants remain in the MC data set after decontamination. While removal of low completeness sequences would remove two of the contaminants, this would have also removed RV-A sequences, as the completeness of RV-A genomic sequences was underestimated by CheckV. Sequences with low completeness (<50%) were therefore kept, to preserve sensitivity in subsequent analysis.

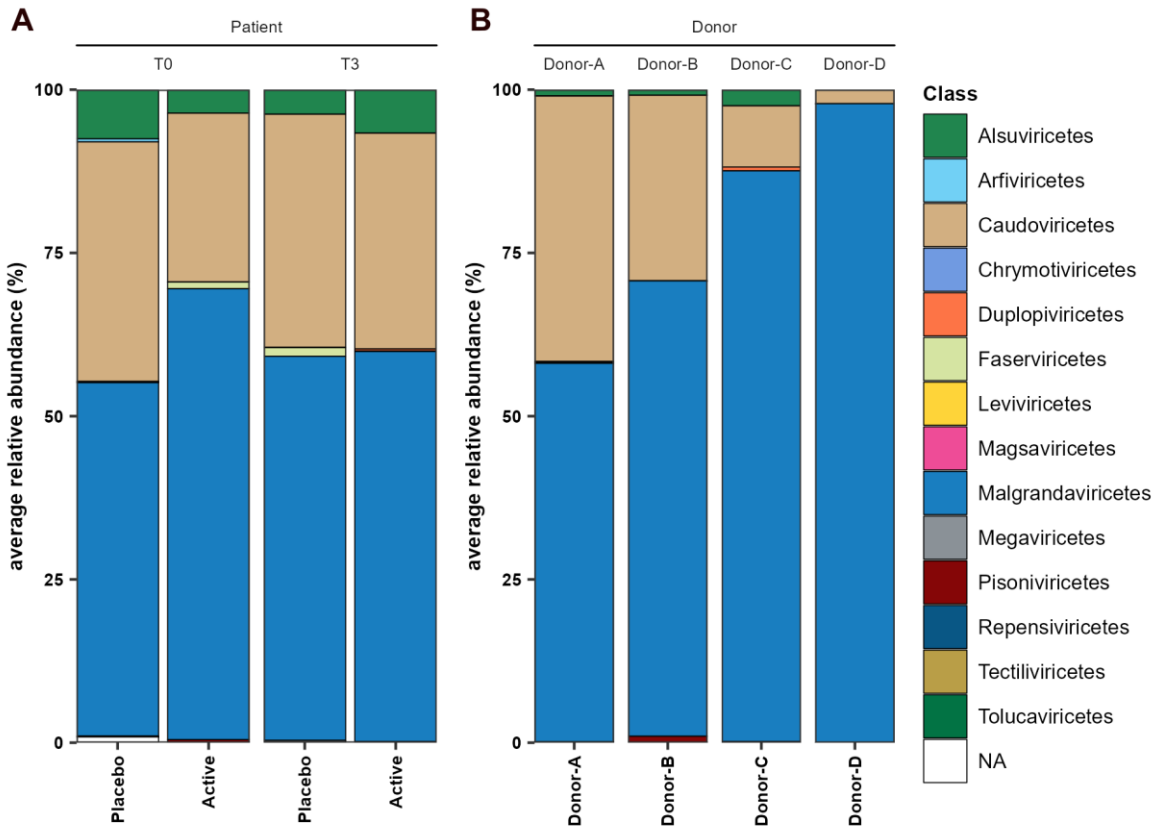

Figure 4.9: Total relative abundance of virus classes in patients and donors. Bars represent the fraction of reads of all samples of the patients in the indicated treatment groups at T0 and T3 (**A**) and of all donor samples of individual donors (**B**).

4.3.3 Taxonomic annotation of patient and donor viromes

A combination of gene sharing network- and reference database alignment-based approaches were employed for taxonomic annotation of vOTUs in the FMT dataset, supplemented by results from the VIRify and VirBot pipelines. Up to 99.6% of vOTUs could be assigned a class, and 9.31% of sequences were assigned a family. In both patients and donors, the phage class *Malgrandaviricetes* which includes the *Microviridae* family, is the predominant class, followed by tailed phages of the class *Caudoviricetes*. At baseline, patients in the placebo group have on average more *Caudoviricetes*, while the active group has more *Malgrandaviricetes*, while both groups have similar levels of each class post-FMT (Fig. 4.9A). For the donors, each donor on average has distinct levels of *Malgrandaviricetes* and *Caudoviricetes* (Fig. 4.9B). The average relative abundance of *Caudoviricetes* varies, with 40%, 28%, 2% and 9% in samples from Donor A, B, C, and D, respectively.

Individually, there was a large variation in the taxonomic composition of the virome of patients. In 66% of patients, more than 50% of the virome is composed of *Malgrandaviricetes* at baseline

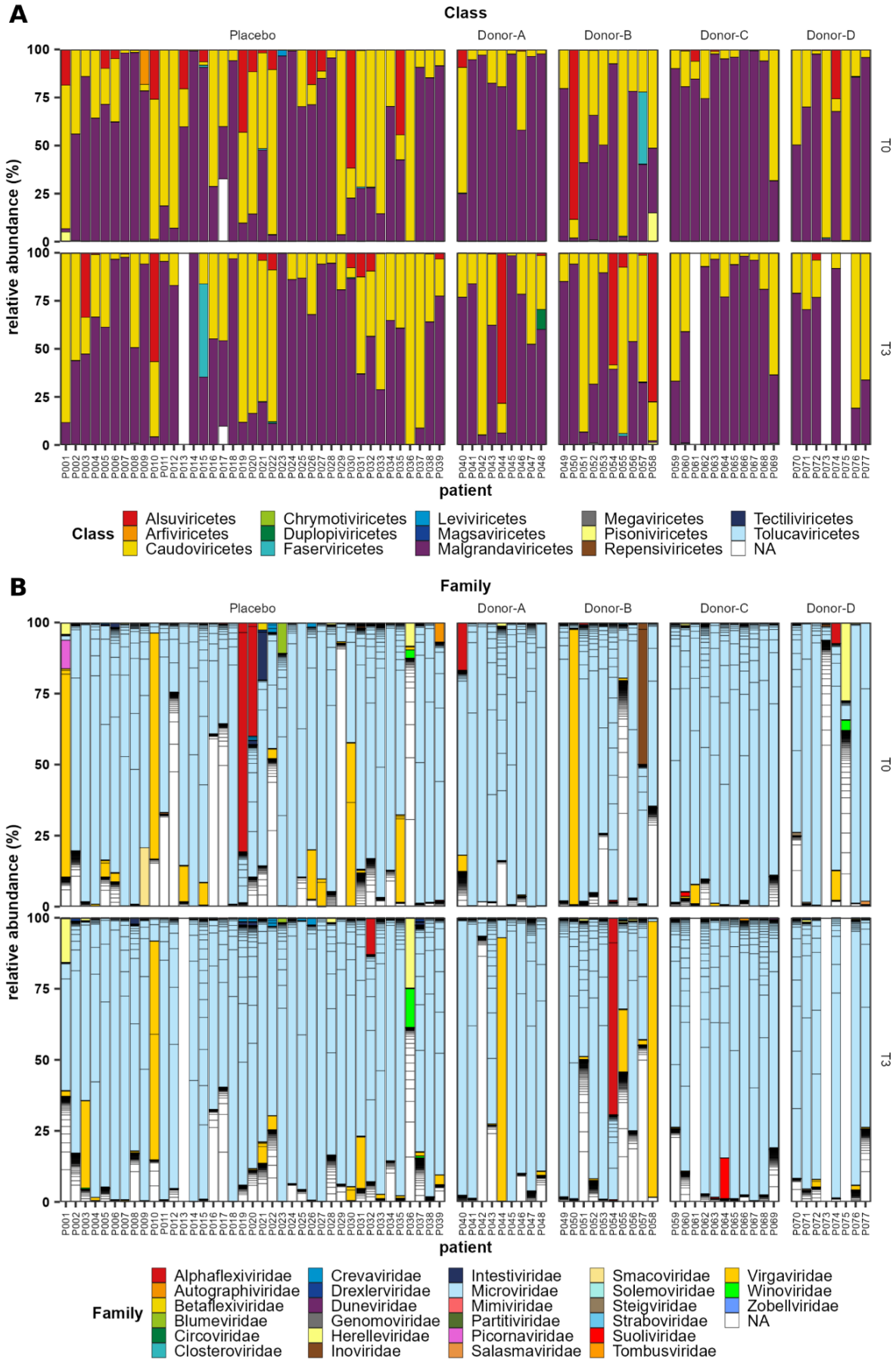
A
Class
Placebo
Donor-A
Donor-B
Donor-C
Donor-D
relative abundance (%)
T0
T3
patient
Class
Alsuviricetes
Arfiviricetes
Caudoviricetes
Chrymotiviricetes
Duplopiviricetes
Faserviricetes
Leviviricetes
Magsaviricetes
Malgrandaviricetes
Megaviricetes
Pisoniviricetes
Repensiviricetes
Tectiliviricetes
Tolucaviricetes
NA
B
Family
Placebo
Donor-A
Donor-B
Donor-C
Donor-D
relative abundance (%)
T0
T3
patient
Family
Alphaflexiviridae
Autographiviridae
Betaflexiviridae
Blumeviridae
Circoviridae
Closteroviridae
Crevaviridae
Drexlerviridae
Duneviridae
Genomoviridae
Herelleviridae
Inoviridae
Intestiviridae
Microviridae
Mimiviridae
Partitiviridae
Picornaviridae
Salasmaviridae
Smacoviridae
Solemoviridae
Steigviridae
Straboviridae
Suloviridae
Tombusviridae
Virgaviridae
Winoviridae
Zobellviridae
NA

Figure 4.10: Consistent taxonomic compsition of the GI virome at the class and family level in both placebo and active transplant FMT. Taxonomic composition of patient stool samples at the class (**A**) and family (**B**) level, at baseline (T0) and at the three-month follow up after FMT (T3). Patients are grouped by the received FMT sample. For patients P013, P061, P073, and P075 the three-month follow-up samples were not available. **A:** Taxonomic composition at the class level. **B:** Taxonomic composition at the family level. Only vOTUs with ≥50% completeness are included as the family could not be assigned to the majority of vOTUs with lower genome completeness. NA: No taxonomy assigned at the respective taxonomic level.

(Fig. 4.10A). Of the 26 patients with less than 50% *Malgrandaviricetes*, 22 had on average 75% *Caudoviricetes*, while three were dominated by *Alsuviricetes*, a class consisting of mainly plant and fungal viruses. Among the vOTUs with >50% genome completeness, virome composition was more consistent between patients at the family level. The phage family *Microviridae* formed most viruses in almost all patients, and the filamentous phage family *Inoviridae* had 37% relative abundance in one patient (Fig. 4.10B). While *Caudoviricetes* were the second most dominant phage class, individual families from this clade only formed a small fraction of any single virome in most patients, although the family *Intestiviridae* had a relative abundance of 21% in one patient. While most patient viromes were dominated by phages, some patients were instead dominated by plant viruses from the *Alphaflexiviridae* and *Betaflexiviridae* familie. Additionally, eukaryotic viruses of the *Picornaviridae* and *Smacoviridae* were present at 18% and 4.9% relative abundance, respectively in two patients.

Post-FMT, there were no obvious systematic differences between the virome of active and placebo transplant receivers. In the placebo group, 60% of patients had >50% *Malgrandaviricetes*, compared to 65% in the active treatment group. Tailed phages of the *Caudoviricetes* class increased in 60% of patients in the active group and in 55% of patients in the placebo group, with on average 28 and 18 percentage points, respectively. However, baseline levels were different in the placebo and active treatment groups, with on average 37% and 25% *Caudoviricetes*, respectively. The difference between placebo and active treatment at T3 was small at 36% and 33% average *Caudoviricetes* relative abundance, respectively. Differences in the virome composition at baseline between the donor groups complicated comparing the effects of different donors. Nonetheless, the largest differences in *Caudoviricetes* are found in the donor groups with the lowest baseline *Caudoviricetes* levels: an increase of 14 and 10 percentage points was seen in recipients of Donor A and C, respectively, who had on average relative abundance of 18% and 13% at baseline, respectively, which is much lower than the baseline overall average of 31%. However, the average relative abundance of

Caudoviricetes in recipients of Donor A and C was 32% and 24%, respectively, which is still lower than the 36% average relative abundance in the placebo group.

Comparing the taxonomic composition of donor transplants, showed variation not only between donors, but also between samples of the same donor. At the class level, *Malgrandaviricetes* was most abundant in most samples, with >50% relative abundance in 16 out of 18 donor samples (Fig. 4.11A). Only one sample of Donor A, and one of Donor B, had less

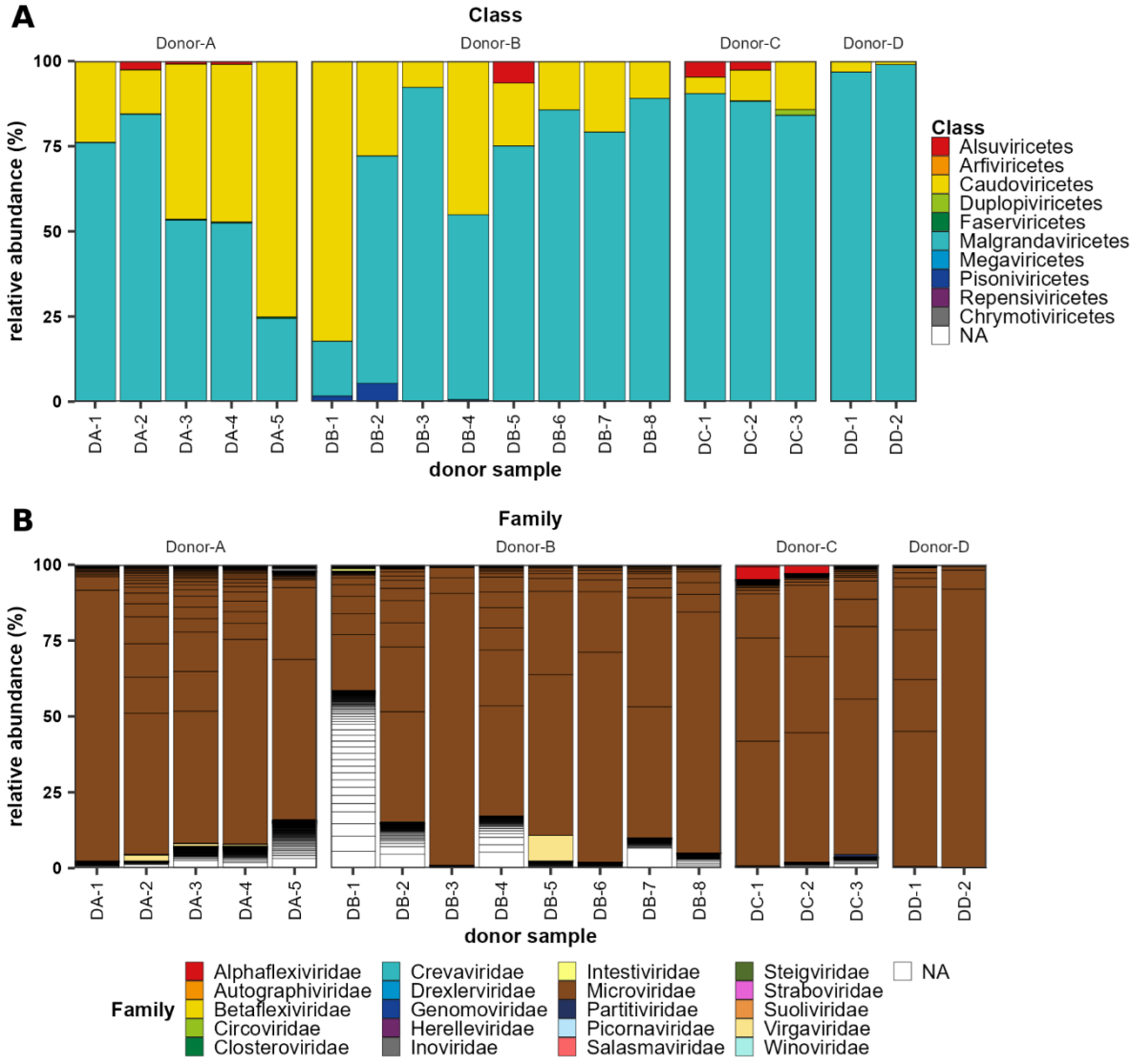

Figure 4.11: Inter- and intra-donor variability in taxonomic composition of the GI virome at the class and family level. Taxonomic composition of donor stool samples at the class (**A**) and family (**B**) level. **A:** Taxonomic composition at the class level. **B:** Taxonomic composition at the family level. Only vOTUs with ≥50% completeness are included, as the family could not be assigned to the majority of vOTUs with lower genome completeness. NA: No taxonomy assigned at the respective taxonomic level.

than 50% *Malgrandaviricetes*, and were instead dominated by *Caudoviricetes*. This class was the second-most abundant virus class in all other samples. Samples 1 to 4 of donor A were donated 3 and 5 days apart, while sample 5 was donated 3,5 weeks after sample 4. Over this period, *Caudoviricetes* increased from 23% to 75% relative abundance (Fig. 4.11A). Donor B donated samples 6 days to 3 weeks between each donation over a two-month period, which saw a decrease in *Caudoviricetes* from 82% to 11%. Samples of donors D and C, however, were donated within a week and had smaller differences in *Caudoviricetes* levels. Several donor samples additionally contained eukaryotic viruses, particularly *Alsuviricetes* and *Arfiviricetes* plant viruses, and *Pisoniviricetes*. At the family level, only vOTUs with ≥50% genome completeness were analysed, and *Microviridae* was the most dominant family in all samples (Fig. 4.11B). Additionally eukaryotic viruses were detected and annotated at the family level, with *Alphaflexiviridae* and *Virgaviridae* plant viruses present in some donor samples, as well as *Picornaviridae* at low abundance (<0.001%) in Donor B transplants 1 – 4 and 8.

To further analyse whether any differences were apparent at the class or family level between active and placebo transplant post-FMT, the abundance of vOTUs with at least 50% genome completeness was compared pre- and post-FMT in both groups. Among the virus classes, an increase in *Caudoviricetes* was visible in the active transplant group, but not in the placebo group, although again the relative abundance was higher in the placebo group at baseline (Fig. 4.12). Although the relative abundance of the filamentous phage class *Faserviricetes* was low, between 0.47% and 0.0000010%, the median relative abundance was 0.0053% in the active transplant group and 0.00082% in the placebo group post-FMT. No patterns were observed for the classes *Alsuviricetes*, *Arfiviricetes*, *Duplopiviricetes*, *Leviviricetes*, *Pisoniviricetes*, *Repensiviricetes*, and *Tolucaviricetes*, as only a few samples were positive for most of these. Comparing post-FMT levels of active and placebo transplant recipients, the median relative abundance of *Caudoviricetes* families *Crevaviridae*, *Drexlerviridae*, and *Salasmaviridae* were higher in the active transplant group, while there was a small decrease in *Herelleviridae*, *Intestiviridae*, *Soluoviridae*, and *Winoviridae*. The filamentous phage family *Inoviridae* remained the same, even though its parent class *Faserviricetes* was increased (Fig. 4.13). Comparison of the mean relative abundances of pre- and post-FMT in the active and placebo groups did not find any statistically significant differences using the Wilcoxon singed-rank test between timepoints in either group.

4.3.4 Alpha and beta diversity of ME/CFS and donor viromes

To assess the effect of FMT on the GI virome diversity, viral richness and evenness was analysed. Comparing species richness between pre- and post-FMT in the placebo and active transplant groups, showed a significant (p<0.05) increase in species richness in the active group, but not in the placebo group as measured by the Chao1 index (Fig. 4.14A). There was no

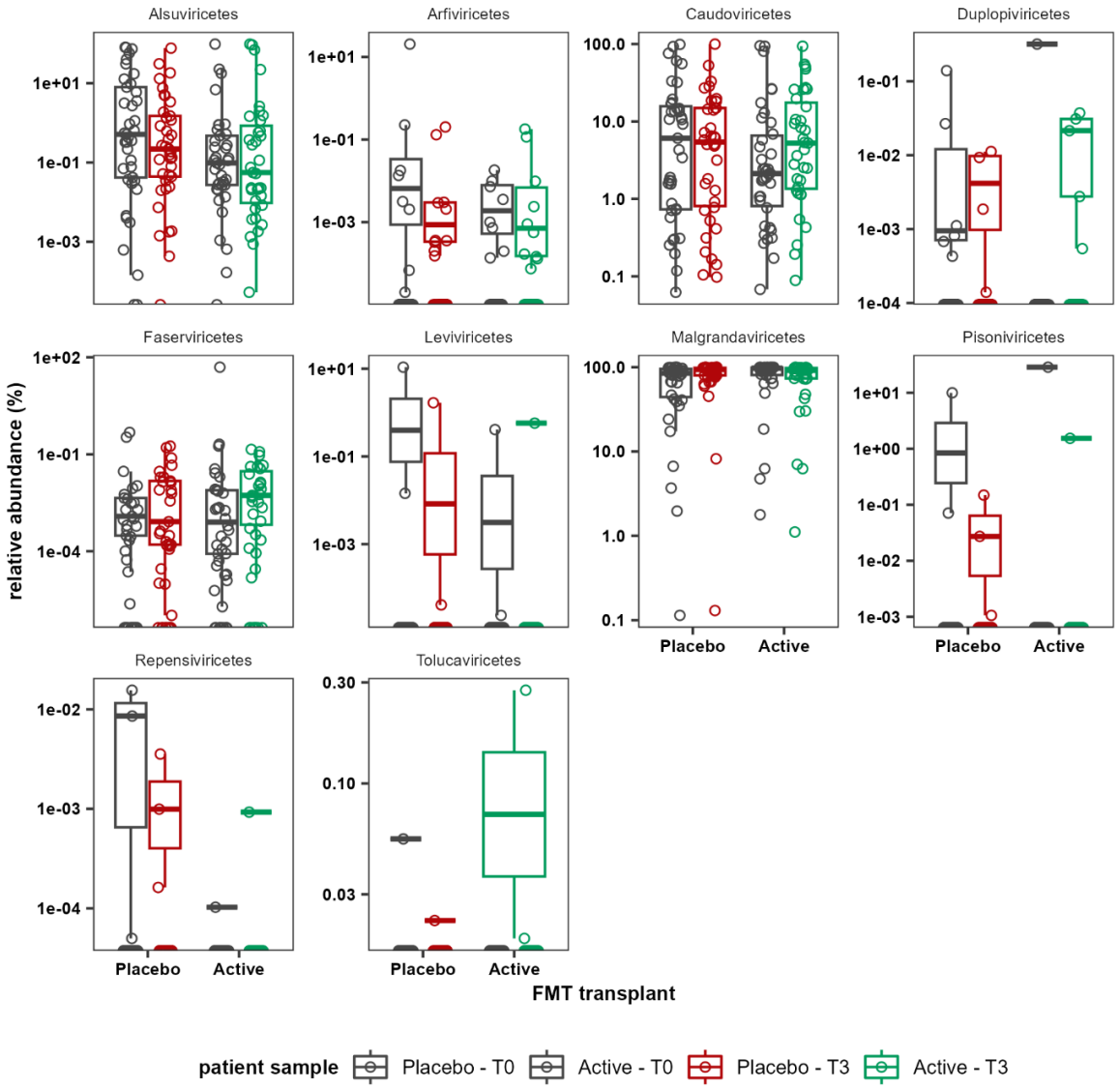

Figure 4.12: No change in relative abundance of virus classes between per- and post-FMT. Relative abundance of virus classes in each patient in the placebo and treatment groups at T0 and T3 for vOTUs with ≥50% complete genomes as estimated by CheckV. Relative abundance before and after treatment was compared between placebo and active FMT by Wilcoxon signed-rank test, with correction for multiple comparisons using the Holmes method. A significant difference between timepoints was not detected for any class.

change in species richness-weighted (Shannon index) and species abundance-weighted (Simpson index) alpha diversity. Separating patients by transplant donor group, differences between pre- and post-FMT were not as clear, although the richness in the groups of Donor B, C, and D were higher than placebo (Fig. 4.14B).

Next, the beta diversity of the pre- and post-FMT samples of each patient were analysed by

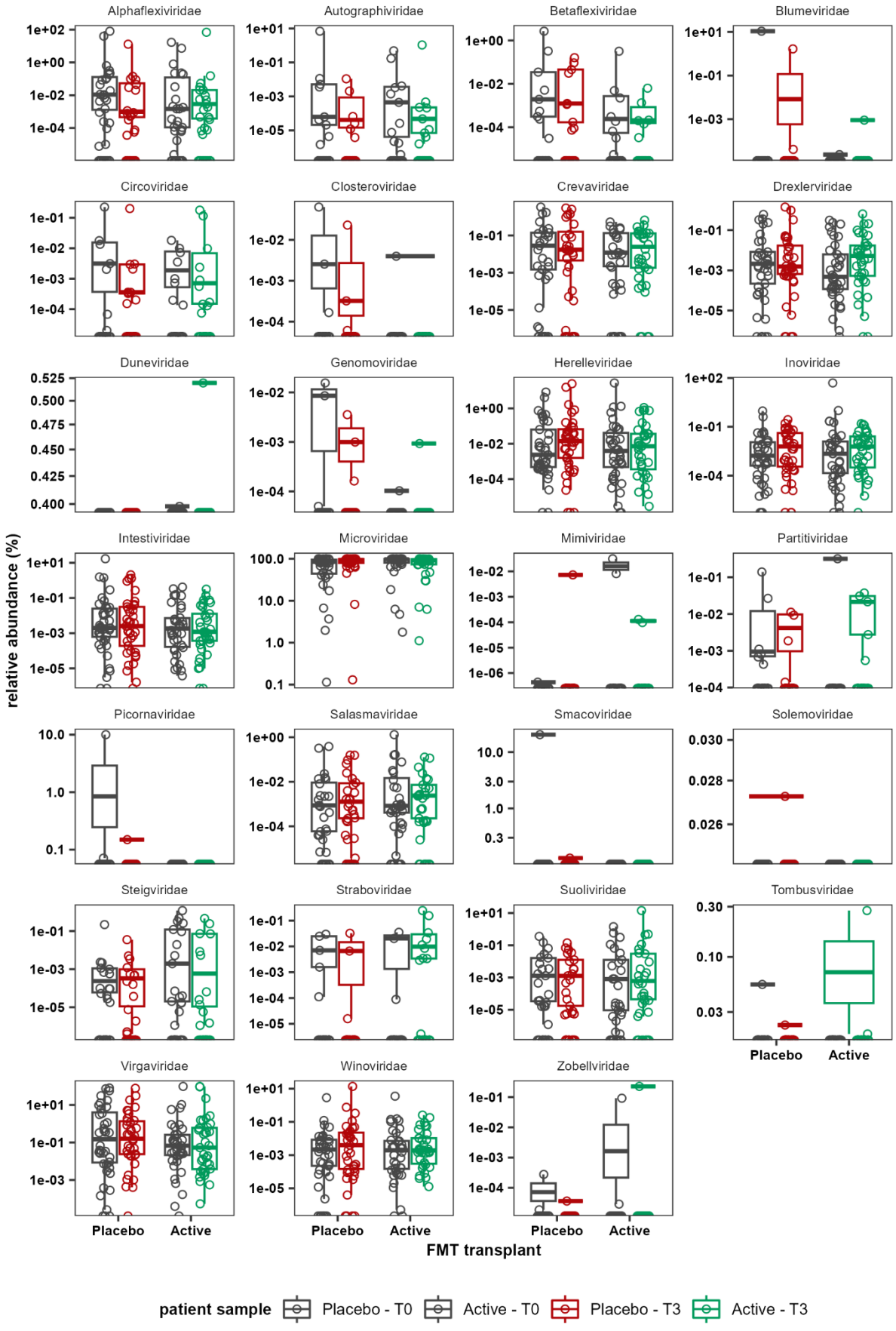
Alphaflexiviridae
Autographiviridae
Betaflexiviridae
Blumeviridae
Circoviridae
Closteroviridae
Crevaviridae
Drexlerviridae
Duneviridae
Genomoviridae
Herelleviridae
Inoviridae
Intestiviridae
Microviridae
Mimiviridae
Partitiviridae
Picornaviridae
Salasmaviridae
Smacoviridae
Solemoviridae
Steigviridae
Straboviridae
Suoliviridae
Tombusviridae
Virgaviridae
Winoviridae
Zobellviridae
relative abundance (%)
Placebo
Active
FMT transplant
patient sample
Placebo - T0
Active - T0
Placebo - T3
Active - T3

Figure 4.13: No significant changes in relative abundance of virus families between pre- and post-FMT. Relative abundance of virus families for each of the patients in the placebo and treatment groups, at T0 and T3. Relative abundance after before and after treatment in the placebo and active FMT groups were compared by Wilcoxon signed-rank test, with correction for multiple comparisons using the Holmes method. N significant difference between timepoints was detected for any family.

calculating Bray-Curtis (BC) dissimilarity and the Aitchison distance (Fig. 4.15A). The beta diversity of pre- and post-FMT samples was significantly higher in the active treatment group for both measures, indicating that FMT of a healthy donor transplant induces larger changes in the GI virome than autologous FMT. When comparing donor groups, there was a similar difference, with statistically significant differences in the BC dissimilarity between each donor group and the placebo group (Fig. 4.15B). Beta diversity measured by Aitchison distance was

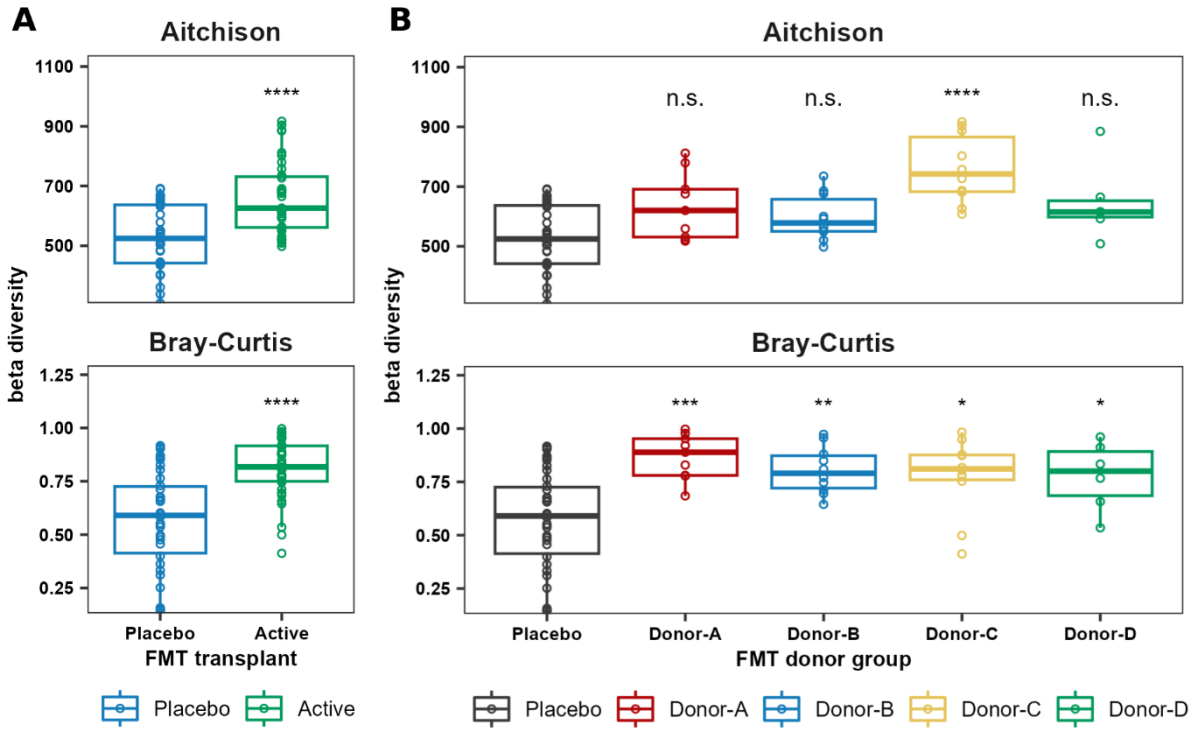

Figure 4.14: Beta diveristy of the faecal virome pre- and post-FMT. Beta-diversity, as determined by the Aitchison distance and Bray-Curtis dissimilarity, of the virome of T0 and T3 samples of each patient. The active transplant group as a whole (**A**) and donor groups groups (**B**) are compared to the placebo group using a Wilcoxon signed-rank test to determine significant differences in beta diversity. n.s. not significant ($p \geq 0.05$), * $p<0.5$, ** $p<0.01$, *** $p<0.001$, **** $p<0.0001$.

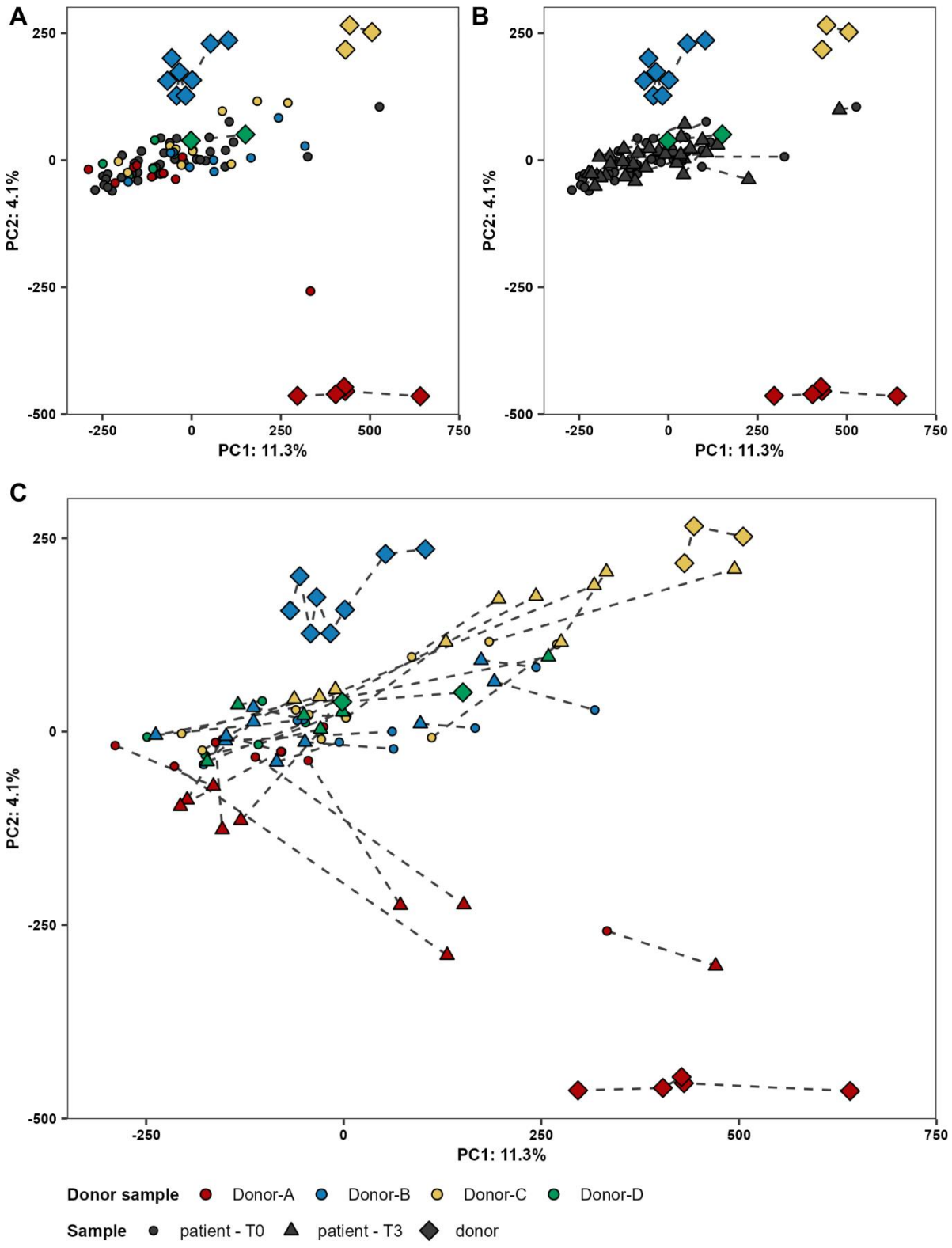

Figure 4.15: Principal component analysis (PCA) of the Aitchison distance of patient and donor samples shows a shift in the active FMT group virome composition towards the donor. Diamonds depict donor samples, circles depict patient T0 samples, and triangles depict patient T3 samples. Patients in the treatment group and donors are coloured correspondingly, while patients in the placebo group are coloured grey. Samples

corresponding to the same patient or donor are connected by a dashed line. A single PCA was performed on all patient and donor samples. For clarity, sets of samples are displayed in separate subfigures. **A:** Baseline samples of both treatment groups and donor samples. **B:** Baseline and post-FMT samples of the placebo transplant group and donor samples. **C:** Baseline and post-FMT samples of the active transplant group and donor samples

---

higher than placebo for each donor group, although statistical significance was only achieved in the recipients of donor C.

Using principal component analysis of the Aitchison distance of all patient and donor samples, the overall similarity of patient and donor samples was investigated. Samples of individual donors are more like each other than to other donors, and donor A being more distinct from the other three (Fig. 4.16A). Patient baseline samples formed a large cluster between donors A, B, and C (Fig 4.16A), while Donor D was placed within the cluster. Donors A and C were furthest removed from patient baseline samples. Post-FMT, changes in virome composition were apparent in the placebo and active transplant groups. However, post-FMT viromes in the placebo group remained within the baseline sample cluster (Fig. 4.16B), while viromes of the active group shifted towards donors, particularly for Donor A and C, and to a lesser extent Donor B. Recipients of Donor D did not clearly shift towards donor D, likely due to the high similarity of Donor D to patients at baseline.

As the initial analysis was performed blinded, the beta diversity of the post-FMT sample and each of the donor transplant samples was calculated, to examine which patients were most like which donor post-FMT. After unblinding, the similarity of the post-FMT samples and donor sample of each donor group was compared. Compared to samples of Donor A, post-FMT similarity was between 0.00 and 0.12 for the placebo group, while post-FMT similarity was between 0.01 and 0.61 for Donor A patients (Fig. 4.17). Interestingly, some of the Donor A patients were more like a sample other than that from Donor A that was used for their FMT.

For the placebo group, the range in similarity to Donor B samples was slightly lower than for Donor B patients. Post-FMT samples of the placebo group had low similarity to samples of Donor C, while nearly all Donor C patients had higher similarity to Donor C samples. Placebo samples had the lowest similarity to Donor D and three out of six patients who received a transplant from Donor D were more like Donor D than placebo. These results show that Donors A and C induce a larger shift of the recipient virome towards their own virome than Donor B and D. Interestingly, the virome of some patients shifted more towards a donor sample other than their donor. Nonetheless, the median difference between patient and the received transplant was reduced after FMT in all donor groups and by both BC dissimilarity and

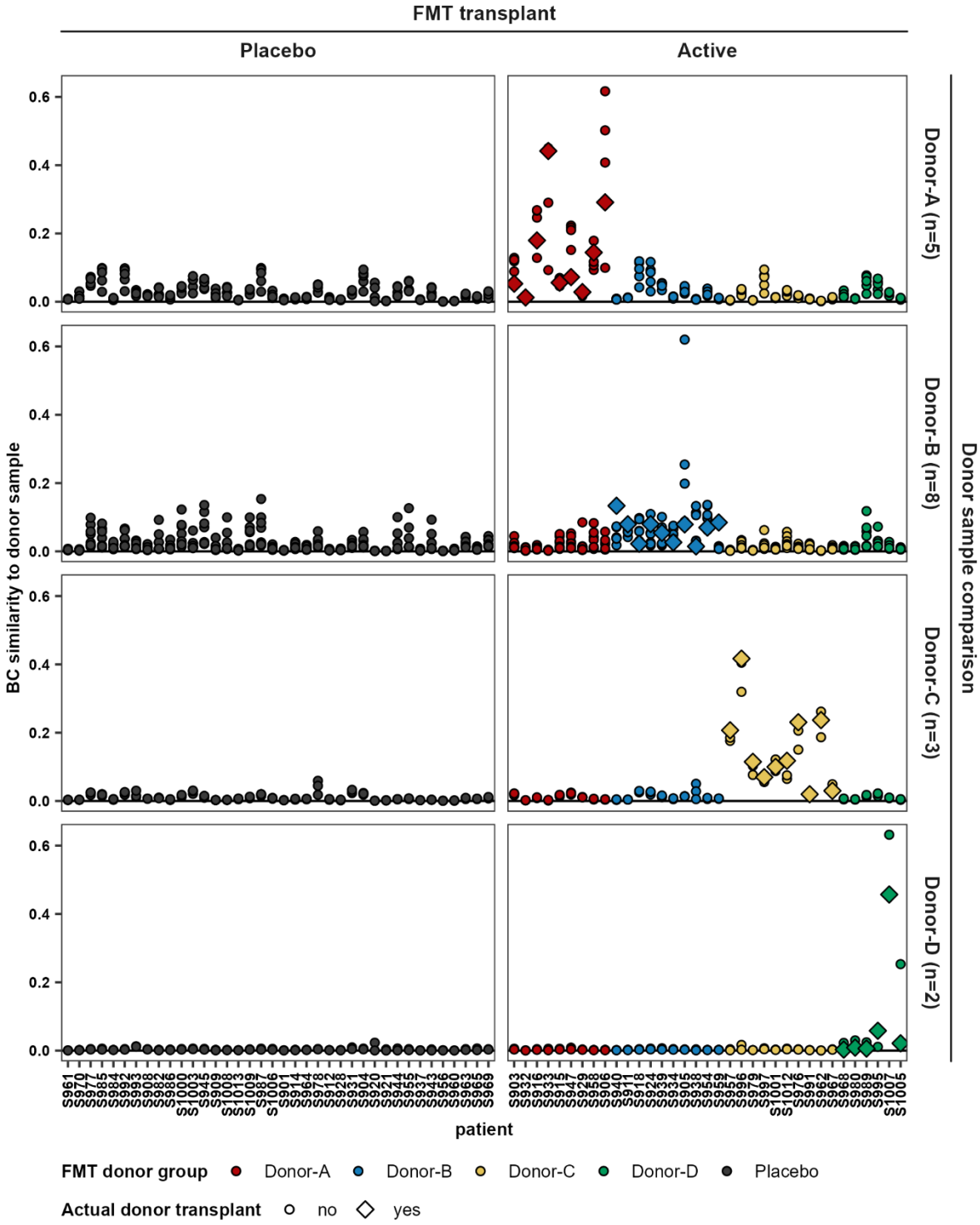

Figure 4.16: Change in similarity to donor samples. Change in similarity to donor samples, as calculated by the difference between the BC dissimilarity to donor samples at T0 and T3. Patients are compared to all FMT samples of Donor-A, Donor-B, Donor-D and Donor-E (top to bottom panels). Patients are coloured by the received FMT sample, and the diamonds show change in similarity to the received donor sample.

Aitchison distance, except for the Aitchison distance of Donor A patients (Fig. 4.18). The difference between the beta diversity of the pre-FMT and transplant sample and the post-FMT and transplant sample of each was statistically significant for all groups when measured by the Bray-Curtis dissimilarity.

4.3.5 FMT-mediated transfer and engraftment of viruses

As there was evidence of a shift in the composition of the virome of ME/CFS patients following FMT towards the donor virome, the presence of vOTUs in patients pre- and post-FMT was compared to their presence in the received FMT transplant. Each donor donated multiple samples from which transplants were produced with each patient receiving a transplant produced from only one of those samples. Therefore, overlap in the presence of vOTUs with the donor sample corresponding to the received transplant, as well as overlap with other samples from the same donor was analysed. Additionally, vOTUs shared between patient and

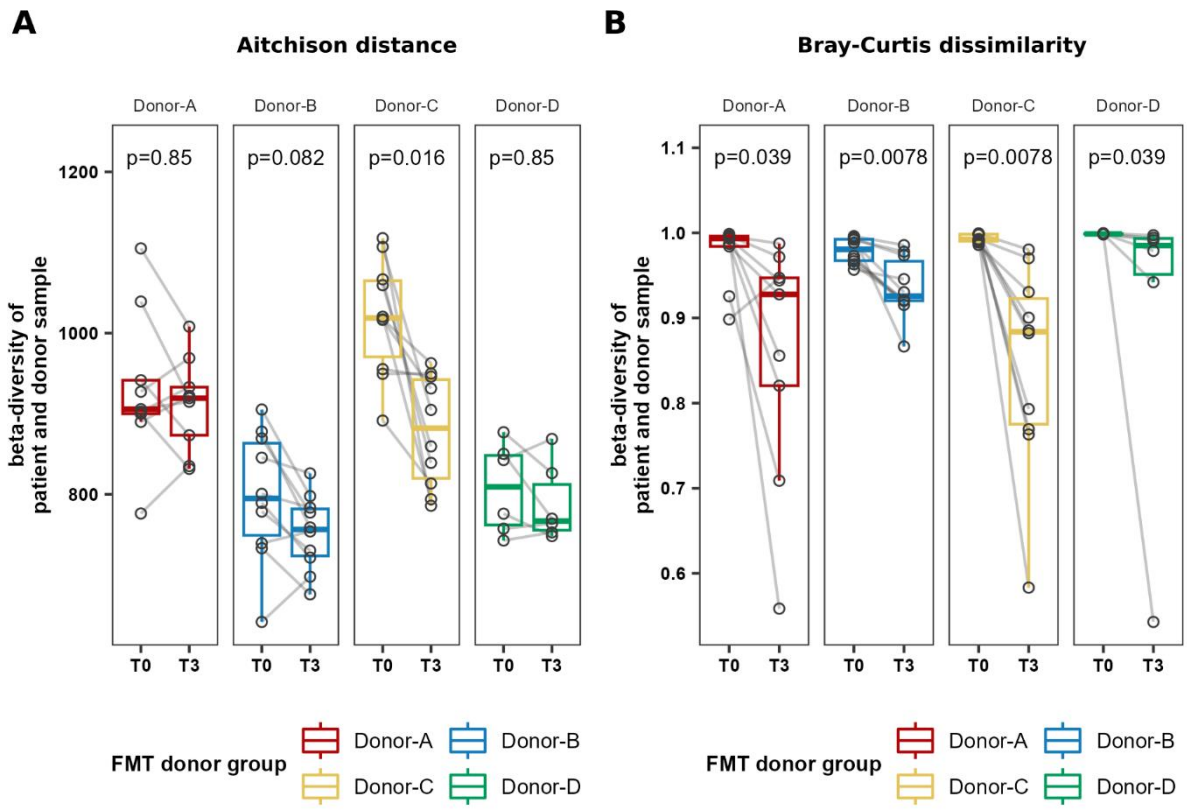

Figure 4.17: Change in dissimilarity of patient faecal virome to the virome of the received donor sample. Beta diversity of patients at T0 and T3 and the donor faecal sample correspondin to the transplant received by the patient, as determined by the Aitchison distance (**A**) and Bray-Curtis dissimilarity (**B**). A pair-wise Wilcoxon signed-rank test was performed to detect significant changes in the distance and dissimilarity between T0 and T3. Significance values, corrected for multiple comparisons using the Holmes method, are displayed at the top of each plot.

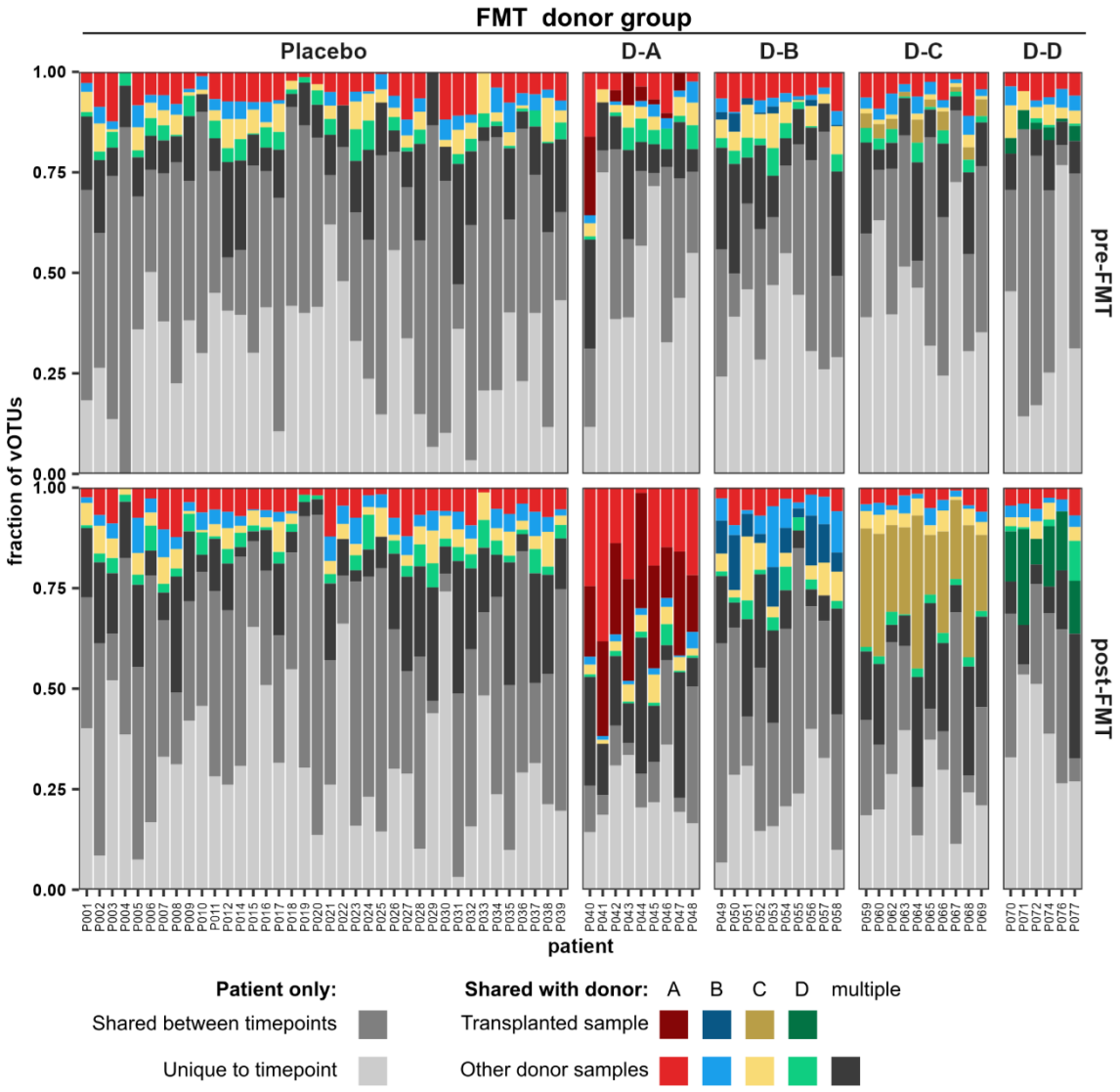

Figure 4.18: An increase in donor viruses shared with the transplant donor in the active transplant group post-FMT shows transfer of viruses between donor and patient. The vOTUs present in patient samples at T0 (top panels) and T3 (bottom panels) in the placebo and active transplant groups are coloured depending on their presence in both patient samples, in one (lighter shaded colours) or more (dark grey) donor samples, and presence in the donor sample corresponding to the received transplant (darker shaded colours).

other donor samples was analysed to determine the shared vOTUs with all donors at baseline.

At baseline, patients shared between 8.8% and 69% and on average 28% of vOTUs with at least one donor (Fig. 4.19). After FMT, patients in the placebo group had on average 30% of vOTUs in common with at least one donor, while the average was 50% for the active treatment group. Moreover, patients in the active transplant group had on average 27% of vOTUs in common

with that of their FMT donor sample. The largest difference was observed for Donor C patients, with on average 40% of vOTUs in common with that of the donor sample. Donor A patients shared on average 33% of vOTUs with their donor sample, and on average 57% of vOTUs shared with any Donor A sample. For Donor B and D, a smaller fraction of vOTUs was shared. with their transplant sample, on average 11% and 21%, respectively. In addition to viruses transferred from donors, the number of vOTUs in patient samples that were only found in the patient baseline sample was higher in the active group than the placebo group. On average 71% and 74% of vOTUs in the active and placebo group, respectively, were not shared with a donor at baseline. However, the percentage of vOTUs unique to the baseline sample was 41% in the active group and 30% in the placebo group. This suggests that not only did FMT from healthy donors lead to transfer of viruses from donor to patient, but also that donor derived viruses were lost.

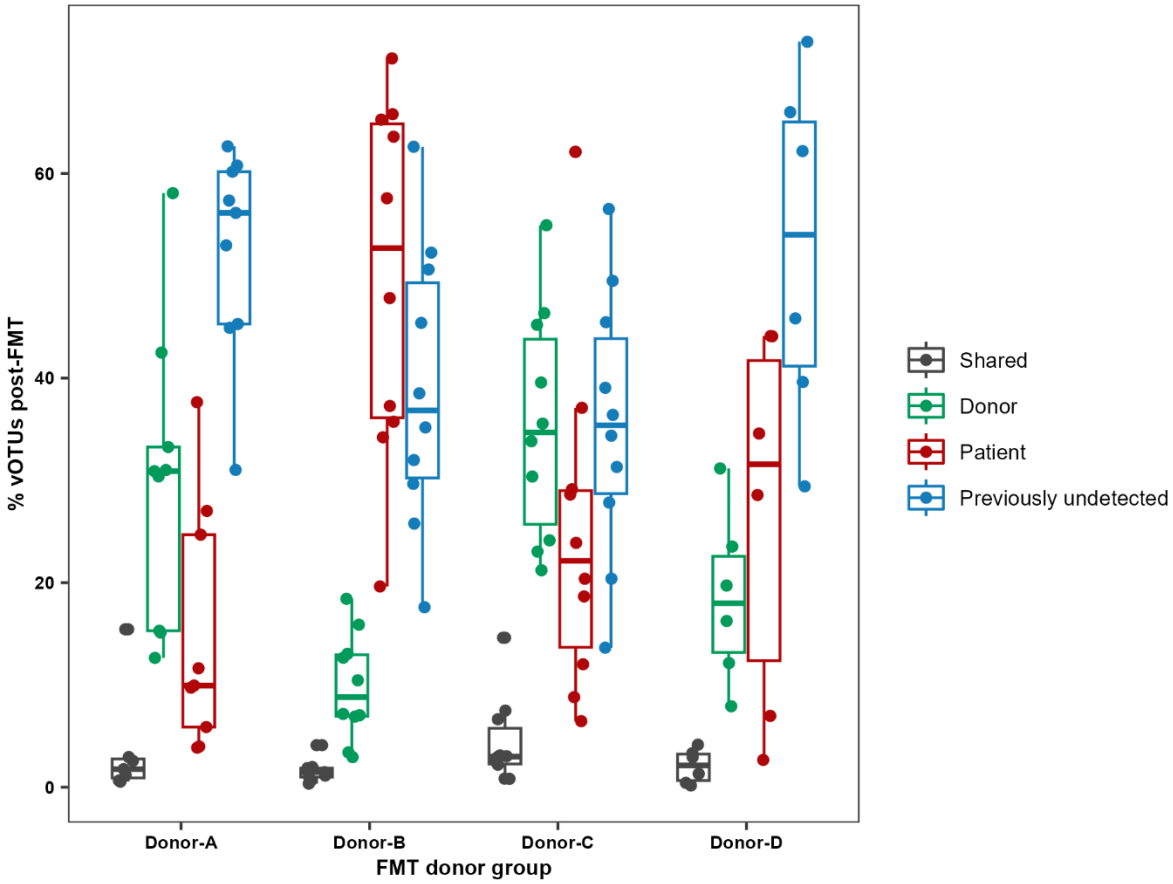

Figure 4.19: The fraction of donor-derived vOTUs post-FMT varies by donor group. The percentage of vOTUs in the T3 sample that are present in both T0 and donor sample (Shared), only the T0 sample (Patient), only the sample corresponding the received transplant (Donor), or not detected in the T0 or donor sample (Previously undetected). Patients are grouped by FMT donor.

Overall, patients in the active treatment group shared between 2.9% and 58% of vOTUs with their donor post-FMT. Donor C patients had on average the highest content of donor-derived vOTUs (35%), followed by donor A (29%), D (18%) and B (9.8%) (Fig. 4.20). Post-MFT, there was a high fraction of vOTUs not detected at baseline or in the transplant sample, with 45% on average and up to 73% of vOTUs previously undetected. On average, 30% of vOTUs were present at baseline and post-FMT, but not shared with the donor, while only 3% of vOTUs were shared between donor and patient at baseline.

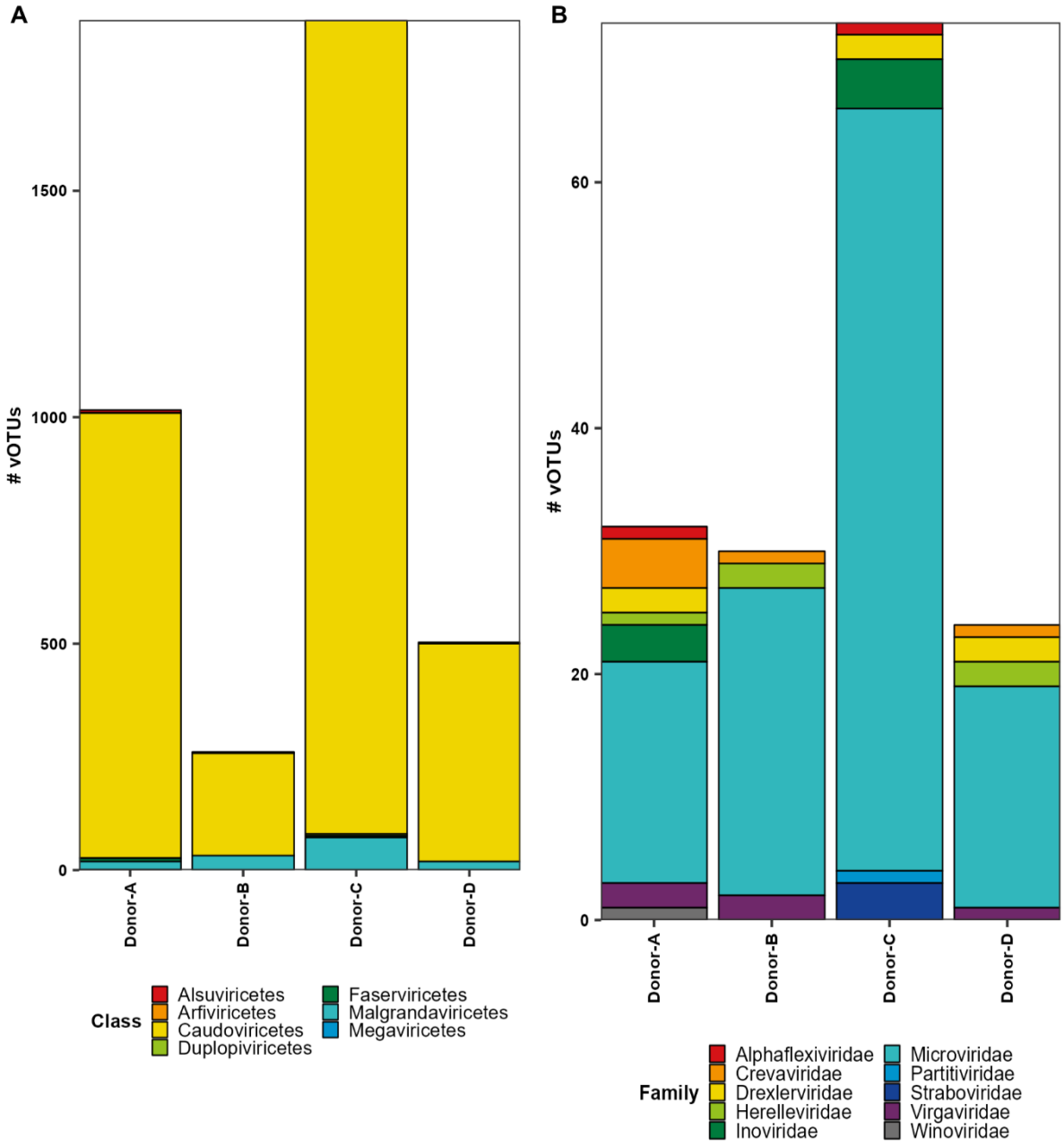

Figure 4.20: Taxonomy of transferred vOTUs. Taxonomic composition at the class (**A**) and family (**B**) level, of vOTUs of each donor that transferred to a patient at least once. Only vOTUs with >50% genome completeness are included in the family-level plot.

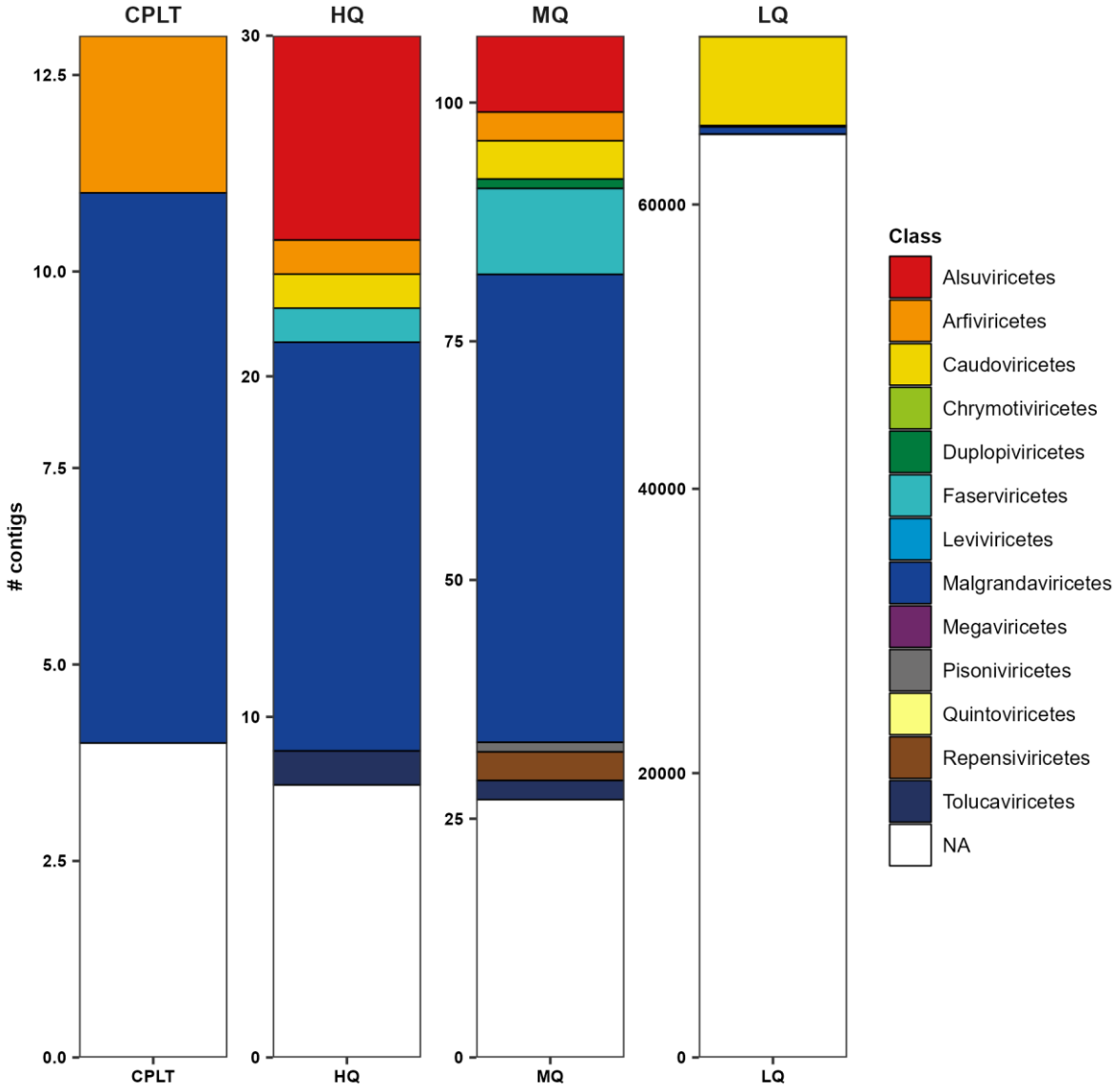

Figure 4.21: Taxonomic composition of vOTUs recovered from eukaryotic virus-enriched reads. vOTUs recovered after co-assembly of all patient and donor reads after enriching eukaryotic virus reads. vOTUs are ordered by the estimated genome completeness as determined by CheckV. CPLT: Complete genome; HQ: High quality genome; LQ: Low quality genome; MQ: Medium quality genome; NA: No class could be assigned.

Next, the taxonomic annotation of transferred vOTUs was examined to determine which viruses were transferred. At the class level, the majority of vOTUs belonged to *Caudoviricetes*, while most phages in donor samples were *Microviridae*, suggesting that tailed phages are more effectively transferred (Fig. 4.21A). The fraction of tailed phages was lower at the family level, as most *Caudoviricetes* vOTUs had a genome completeness <50% and family-level annotation was not possible. Transferred viruses also included filamentous phages of the *Inoviridae* family, while most viruses with family level annotation were *Microviridae* in all donors (Fig. 4.21B).

Notably, several vOTUs transferred from donors A, B, and D belonged to the family of crAss-like phages *Crevaviridae*. The only eukaryotic virus clades that were annotated as transferred between donor and patient were the classes *Duplopiviricietes*, *Alsuviricetes*, *Arfiviricetes*, and *Megaviricetes* and the families *Partitiviridae* and *Virgaviridae*. While human viruses like *Enterovirus* belong to the class *Duplopiviricetes*, these vOTUs likely correspond to plant viruses, since vOTUs of the plant virus family *Partitiviridae* were the only vOTUs annotated at the family level in the *Duplopiviricetes* class. The *Megaviricetes* class contains a variety of eukaryotic viruses, but no vOTUs within this class were annotated at the family level. Therefore, most or all detected eukaryotic viruses were plant viruses. Since these viruses are most likely food-derived, this reflects the acquisition of these viruses from food rather than from the FMT.

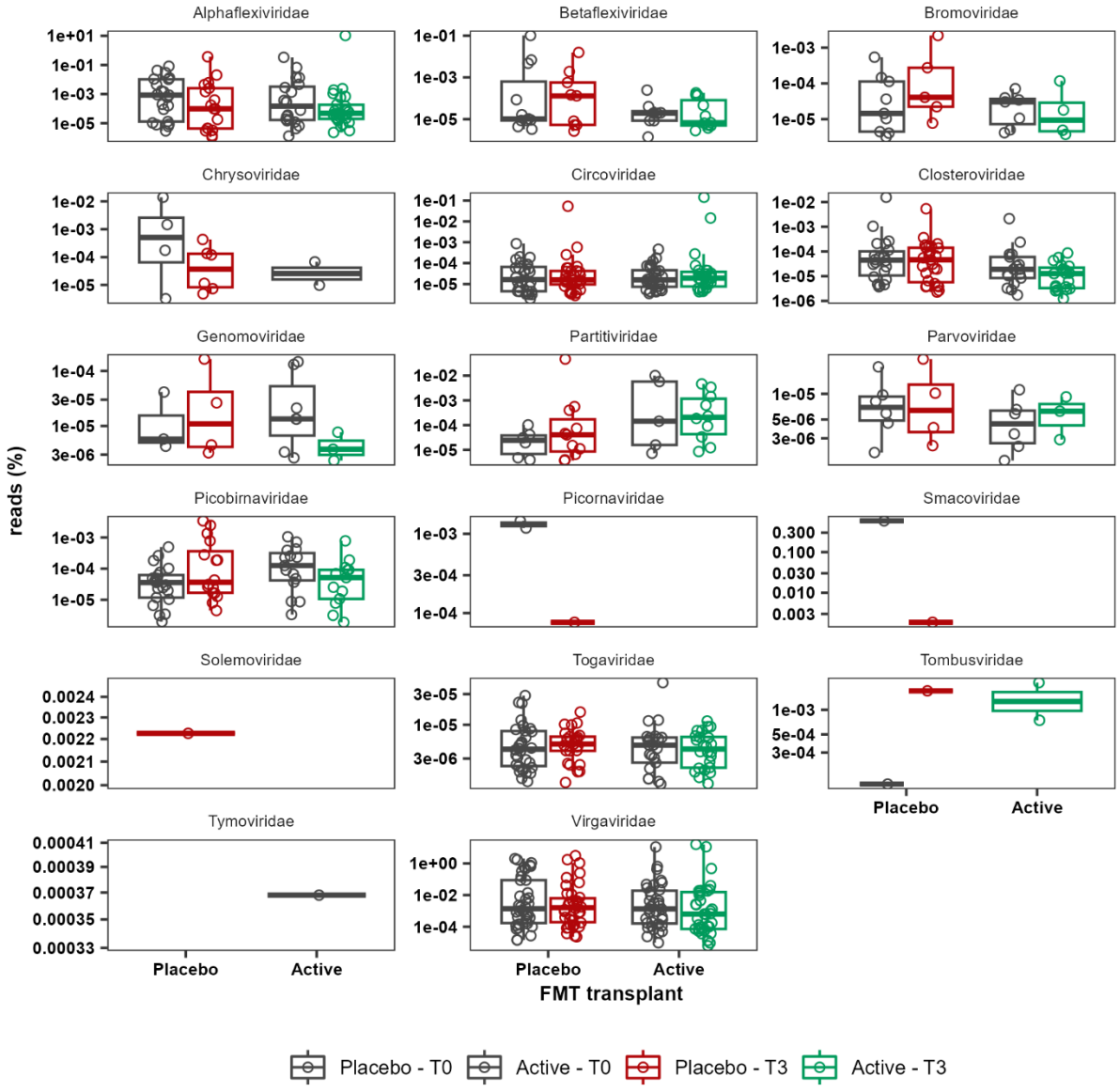

Figure 4.22: Eukaryotic virus relative abundance at T0 and T3. Relative abundance of eukaryotic virus families in patients in the placebo and active FMT groups at T0 and T3.

4.3.6 Changes in eukaryotic virus abundance

To increase the sensitivity of the virome analysis for eukaryotic viruses, reads were mapped against the NR database using Kraken2 (Wood, Lu and Langmead, 2019), and reads from patient and donor samples mapping to eukaryotic viruses were extracted and co-assembled. The co-assembly was aligned against the NT database and quality of sequences was assessed using CheckV (Fig. 4.22). While most sequences were still prokaryotic viruses, several eukaryotic virus sequences were detected. According to CheckV assessment, one complete, one high-quality and three medium-quality *Circoviridae* genomes were found, one full *Smacoviridae* genome was present in the assembly, as well as several high-quality plant virus genomes. Additionally, several low completeness (<50%) genomes were annotated within human virus families, including one *Parviviridae*, several *Picobirnaviridae*, one *Picornaviridae*, and two *Togaviridae*. Reads of individual patient samples were mapped to the co-assembly to determine changes in the relative abundance of eukaryotic virus sequences (Fig. 4.23). No clear differences between the active and placebo groups were apparent for any of the viruses.

4.4 Discussion

The results presented in this chapter provide the first comprehensive description of the GI virome in ME/CFS and the effects of FMT on its composition. Most ME/CFS patient GI viromes in the Comeback Study cohort are dominated by *Microviridae*, while few dominated by *Caudoviricetes* phages pre-FMT. At the class or family level, a systematic difference in GI viruses between the placebo and active transplant post-FMT could not be detected. Nonetheless, GI virome richness increased in the active transplant group, with GI virome composition reflecting that of the donor virome. Moreover, an increase in donor virus sequences was detected in active transplant patients post-FMT, providing evidence of engraftment of donor viruses, with enrichment of *Caudoviricetes* sequences among the transplanted virus sequences. Differences were observed between donors in the degree of virome similarity and the percentage of engrafted virus sequences post-FMT amongst recipients, consistent with not all donors achieving equivalent levels of virus engraftment in FMT recipients.

Around two-thirds of faecal viromes in the Comeback Study cohort are dominated by microviruses of the phylum *Malgrandaviricetes*. A smaller previous virome GI analysis in ME/CFS patients found a majority of *Caudoviricetes* in patient faecal samples (Hsieh *et al.*, 2023). In addition to the smaller sample size (n=9) compared to this study (n=77), the previous study focussed on dsDNA viruses and combined bulk faecal DNA samples with VLP enriched samples, which produces a different virome profile and includes prophages. Additionally, results discussed in Chapter 2 suggest the SISPA method used in this study might preferentially amplify ssDNA viruses, which will overestimate GI ssDNA virus abundance. On the other hand,

single-stranded virus genomes will generate half the number of reads of a double-stranded genome of the same length. Since the relative abundance calculation did not take this into account, the relative abundance of single-stranded virus sequences in the final sequencing libraries is underestimated by a factor of 2. Lastly, geographical differences between patient cohorts might contribute to differences in virome composition (Gregory *et al.*, 2020). A study of the eukaryotic faecal virome in ME/CFS patients (*n*=106) only found sporadic presence of non-dietary eukaryotic viruses, with no difference between patients and healthy controls (Briese *et al.*, 2023). This is in line with our findings where most eukaryotic viruses are associated with plants with *Picornaviridae* and *Smacoviridae* viruses being present in only two patients. A previous study found various eukaryotic viruses in IBS patients, including viruses in the families *Herpesviridae* and *Poxviridae*, that were not detected in this study, possibly due to differences in bioinformatics analysis (Ansari *et al.*, 2020). The few herpesvirus sequences found in the present study were regarded as contaminants. Giant viruses of the class *Megaviricetes* are present in 8 patients, and have previously been found in studies investigating the virome in IBS (Ansari *et al.*, 2020; Coughlan *et al.*, 2021) and IBD (Zuo *et al.*, 2019). However, confirmation of giant viruses will require more rigorous analysis, given that the 0.45 μm filters used to extract VLPs should have removed most giant viruses (Conceição-Neto *et al.*, 2015) and conserved regions in viral sequences can lead to spurious annotation of sequences as giant virus (Sutton, Clooney and Hill, 2020; Coughlan *et al.*, 2021). Some patients in both treatment groups had a high fraction of plant viruses of the phylum *Alsuviricetes* before or after FMT, with no discernible systematic difference between placebo and active transplants. No patient was dominated by plant viruses at both timepoints, suggesting these are temporary expansions of plant virus in the faecal virome. Potential explanations include dietary changes and an overall reduction in phage abundance.

Caudoviricetes abundance and richness has previously been associated with disease states, with abundance increased in *C. difficile* infection (CDI) and decreased in *Norovirus* gastroenteritis, while richness decreased in both (Zuo *et al.*, 2018). While *Caudoviricetes* increased in the active treatment group, there is no significant difference between the placebo and treatment group post-FMT. The relative abundance of viruses in the *Microviridae* and all other virus families did not differ significantly between placebo and active transplant post-FMT, suggesting that changes in the virome in the active transplant group are inconsistent between patients, or that changes take place below the family level. In CDI, successful FMT was associated with reduced abundance of GI mucosal *Caudoviricetes* at baseline (Zuo *et al.*, 2019). Recipients of Donors A and C had the highest degree of engraftment, as measured by similarity to the donor virome and the fraction of donor-derived viral sequences post-FMT. Patients in these donor groups on average had lower abundance of *Caudoviricetes* than patients in donor

groups B and D. On the other hand, FMT induced an increase in *Microviridae* in CDI (Zuo *et al.*, 2018), while a systematic change in *Microviridae* or *Caudoviricetes* was not detected in this study. Colitis induces stochastic restructuring of the GI prokaryotic virome in mice (Duerkop *et al.*, 2018), which could explain the high inter- and intra-personal variability in IBD virome composition between samples taken a month apart compared to healthy controls (Stockdale *et al.*, 2023). GI inflammation has often been hypothesized to play a role in ME/CFS and could increase GI virome variability through a similar mechanism and could explain a lack of observed systematic changes in the GI virome post-FMT. On the other hand, a recent virome analysis found a large fraction of viruses unique to a single timepoint in a two-month time series, on average 66% healthy controls and 73% - 78% in IBD (Stockdale *et al.*, 2023). In the present ME/CFS patient cohort, 30% of sequences are found at only one timepoint in the placebo group. It is not clear whether this is due to lower temporal variability of the virome in ME/CFS, or methodological differences.

In line with previous findings of FMT in CDI and autism spectrum disorder (ASD) (Kang *et al.*, 2017; Draper *et al.*, 2018), FMT in ME/CFS increases the similarity of virome composition to the donor and increases virus richness. On average between 10% and 35% of sequences in post-FMT samples are donor derived. Transfer of viruses from donor to patient by FMT is associated with treatment success in CDI, ASD, ulcerative colitis, and carbapenem-resistant *Enterobacteriaceae* infection (CREI) (Kang *et al.*, 2017; Conceição-Neto *et al.*, 2018; Draper *et al.*, 2018; Zuo *et al.*, 2018; Liu, Zuo, *et al.*, 2022). Temperate phages of the *Caudoviricetes* phylum are transferred during FMT in CDI, IBD and CREI (Chehoud *et al.*, 2016; Draper *et al.*, 2018; Liu, Zuo, *et al.*, 2022), which suggests a subset of phages transfer *via* prophages integrated in the host bacterial genome of transferred bacteria. This mechanism is particularly striking in the treatment of CRE infection, as several *Escherichia* and *Klebsiella* phages are only detected as prophages in donor samples but are present in recipients at higher abundances correlating with the elimination of CRE (Liu, Zuo, *et al.*, 2022). Most transferred virus sequences correspond to *Caudoviricetes*, consistent with but not indicative of a predominant transfer of temperate phages. Notably, the actual number of *Caudoviricetes* phages transferred is likely overestimated, as most *Caudovirivetes* vOTUs consist of sequences with <50% completeness and thus multiple sequences can correspond to the same virus. It is not clear to what degree temperate phages are present in the ME/CFS virome, and which fraction of transferred phages are temperate. Gene function analysis could shed more light on the presence of temperate phages in the future, and incorporation of microbiome analysis data will facilitate host prediction and enable analysis of transfer of phages in the lytic and lysogenic phase of their replication cycle and determine whether any transferred phages leads to loss of a resident host

bacterium. In this regard, it will be interesting to determine whether any *Anaerotruncus* phages are present in the ME/CFS virome (Hsieh *et al.*, 2023).

The change in virus richness, similarity of the virome composition to the donor, and the fraction of transferred viruses by different donors suggests that some donors elicit a larger change in the virome than others, although it is not clear which factors contribute to this. In CDI, treatment success was associated with a lower proportion of *Caudoviricetes* in the recipient than the donor (Zuo *et al.*, 2018). However, in this study Donor A and C had the largest effect, with only Donor A having a larger proportion of *Caudoviricetes* than the recipients. On the other hand, recipients of Donor A and C had a lower proportion of *Caudoviricetes* than those of Donor B and D, suggesting that the baseline level of tailed phages, rather than the difference with the donor affects the capacity for engraftment. For Donor A and B, respectively 5 and 8 different transplants were used that were each donated a few days apart. Despite these small difference in time, there are significant differences in the ratio of *Caudoviricetes* to *Microviridae* in samples from the same donor. Since at most two patients received a transplant from a single sample of Donor A and B, this further complicates determining donor factors that contribute to engraftment. Additionally, temporal variability of the GI virome in ME/CFS should be considered, since patient T0 samples are taken four to eight weeks before FMT. While differences in faecal virome composition at the class and family level between pre- and post-FMT are small for most patients in the placebo group, some patients show significant changes in the relative abundance of *Caudoviricetes*. As temporal variability of the GI virome is increased in IBS (Stockdale *et al.*, 2023), and GI virome composition shows large variation in donor samples taken less than a week apart, it is not clear how well the T0 virome represents the virome at time of treatment, and whether deviations between T0 and time of treatment affect engraftment.

4.4.1 Limitations

Several methodological limitations are associated with this study. While the longitudinal and interventional design of the study enabled investigation of changes in the virome following FMT, healthy controls, ideally age-, sex-, and geography-matching are needed to determine deviations of the baseline ME/CFS virome from healthy viromes. FMT was administered via an enema. While enema delivery is associated with reduced costs, invasiveness and adverse events compared to colonoscopy, enema achieves perfusion of the transverse colon in 83% of people with the cecum being reached in only 50% of people. By comparison, colonoscopy reaches the whole colon in 100% of people (Skjevling *et al.*, 2023). Anatomical differences affect the distribution of faecal transplants via enema (Skjevling *et al.*, 2023), and it is not clear how this affected enema perfusion and microbiota engraftment of the virome in this study. While the Comeback Study included a follow-up 12 months after FMT, only the sample from 3

months post-FMT was analysed. In ASD patients, changes in the virome of recipients were evident at 10 weeks post-FMT, although changes were most pronounced at 18 weeks post-FMT (Kang *et al.*, 2017). The results in this study might therefore underestimate the maximum amount of change elicited by FMT. On the other hand, in a graft versus host disease patient, the ratio of *Microviridae* to *Caudoviricetes* stabilised after 1 week (Zhang *et al.*, 2021).

The bioinformatics pipeline used for the analysis of the faecal virome is based on co-assembly of patient and donor stool samples. To increase the diversity of the data set, and to prevent loss of eukaryotic viruses, genomes with <50% estimated completeness were included and a sequence length cut-off of ≥1 kb was used. However, due to the broad range of virus genome sizes, the FMT data set included many *Caudoviricetes* sequences of <2500 kbp. Since *Caudoviricetes* phages typically have 30 – 200 kbp genomes, it is likely that the FMT data set contains several sequences belonging to the same virus, while they were regarded as separate vOTUs in this analysis. Therefore, the richness, diversity, and number of transferred viruses is likely to be overestimated in this study. Assembly contiguity could be further improved by various genome binning techniques, like those mentioned in the Discussion of Chapter 3, section 3.4.2. However, care should be taken as these techniques also introduce the risk of creating chimeric sequences. Additionally, patients sporadically exhibited an expansion of plant viruses, which are likely to be related to diet, although a reduction in other GI viruses, particularly phages, could also explain this expansion. Excluding diet-derived eukaryotic viruses from future diversity analyses might reduce temporal variability.

While the virus sequence quality control used here attempted a one-size-fits-all approach, future analysis could be improved by using different criteria for distinct groups of viruses. For instance, *Inoviridae* sequences were removed due to lack of detected viral genes and are therefore likely underrepresented in the FMT data set, while some RV-A segments in the MC control data set were recovered with only an estimated <50% completeness. Taxonomic annotation before the final quality control step could inform the appropriate criteria for individual sequences based on the predicted viral class. Further, the decontamination process potentially led to the removal of highly prevalent viruses, since high prevalence was one criterium for removal, and since blank samples contained SINV reads, a low degree of cross-contamination is likely to have occurred. Thus, reads of highly prevalent viruses could have been present in the blank samples, which will lead to the removal of these viruses from the FMT data set.

The GI virome changes along the GI tract and from the GI mucosa to the lumen (Shkoporov *et al.*, 2022). Additionally, storage and processing of the stool sample affect the faecal virome composition (Conceição-Neto *et al.*, 2015; Shkoporov *et al.*, 2018). Therefore, while steps were

taken to minimise bias, it should be noted that there are possible differences between the faecal virome and the GI virome.

4.4.2 Conclusion and future research

This study is the first large-scale and comprehensive investigation of the GI virome in ME/CFS, which includes an analysis of eukaryotic and prokaryotic RNA and DNA viruses. The present data show that FMT leads to engraftment of prokaryotic viruses from healthy donor transplants in ME/CFS patients, increasing virus richness and similarity of virome composition to that of the donor. The level of engraftment differs between donor groups. Future research should focus on elucidating factors in the donor virome that corelate with engraftment success. Patient factors, such as disease severity, diet, and IBS status, were not available at the time of analysis and future inclusion of these data could shed more light on recipient factors that affect engraftment and virome composition. Most importantly, this will also enable correlation of changes in specific viruses and engraftment of donor viruses with changes in disease severity. This will show whether FMT can improve ME/CFS, and how the virome affects ME/CFS improvement.

CHAPTER 5: INVESTIGATING THE PREVALENCE OF GI PATHOGENS AND THE EFFECT OF FMT IN ME/CFS

5.1 Introduction

The GI tract can serve as reservoir for pathogens (Wilen *et al.*, 2018; Cheung *et al.*, 2020; Parasa *et al.*, 2020; Neurath, Uberla and Ng, 2021), and GI pathogens include *Rotavirus*, *Adenovirus*, and *Norovirus*, bacterial pathogens like *Shigella* spp., *Salmonella* spp., *Campylobacter* spp., *Vibrio cholerae*, enteropathogenic and enterotoxigenic *Escherichia coli*, and *Aeromonas* spp., and eukaryotic parasites like *Entamoeba histolytica* and *Cryptosporidium* spp. (Troeger *et al.*, 2017). However, the association of these GI pathogens with ME/CFS has not been investigated.

While clinical diagnosis of GI pathogens traditionally relies on cell culture, enzyme immune assays, antigen assays, and nucleic acid amplification assays, these methods are time consuming and each pathogen can require a separate assay (Burbelo, Iadarola and Chaturvedi, 2019). An emerging technology is the multiplex molecular test panel, in which multiple pathogens are tested using a single test (Greatorex *et al.*, 2014). This approach enables the development of syndromic test panels, that detect a set of pathogens relevant for a specific set of symptoms, like respiratory, GI, and central nervous system test panels (Heaney *et al.*, 2015). Apart from commercially available panels, TaqMan real-time PCR array cards (TAC) exist, that allow selection and customisation of qPCR assays.

An overview of the TaqMan real-time PCR assay is provided in Figure 5.1. Briefly, it comprises an RT step followed by PCR incorporating probes conjugated to a fluorescent reported dye on the 3' side and a non-fluorescent quencher (NFQ) and major groove binder (MGB) on the 5' end. When the DNA polymerase binds to the primer and starts extension, the probe is degraded, releasing the reporter dye and the NFQ. Due to the increased distance between the reporter dye and the NFQ, the reporter dye becomes fluorescent, and a signal is produced. The advantage of TACs is that they enable the use of multiple assays in parallel on a single sample, with each assay individually optimisable and customisable without affecting the other assays (Liu *et al.*, 2013). Syndromic TACs have been successfully used for epidemiological testing (Agoti *et al.*, 2022), and Thermo Fisher has developed a GI pathogen panel TAC (GI-TAC), containing duplicate assays against 22 common GI pathogens (Fig. 5.1B).

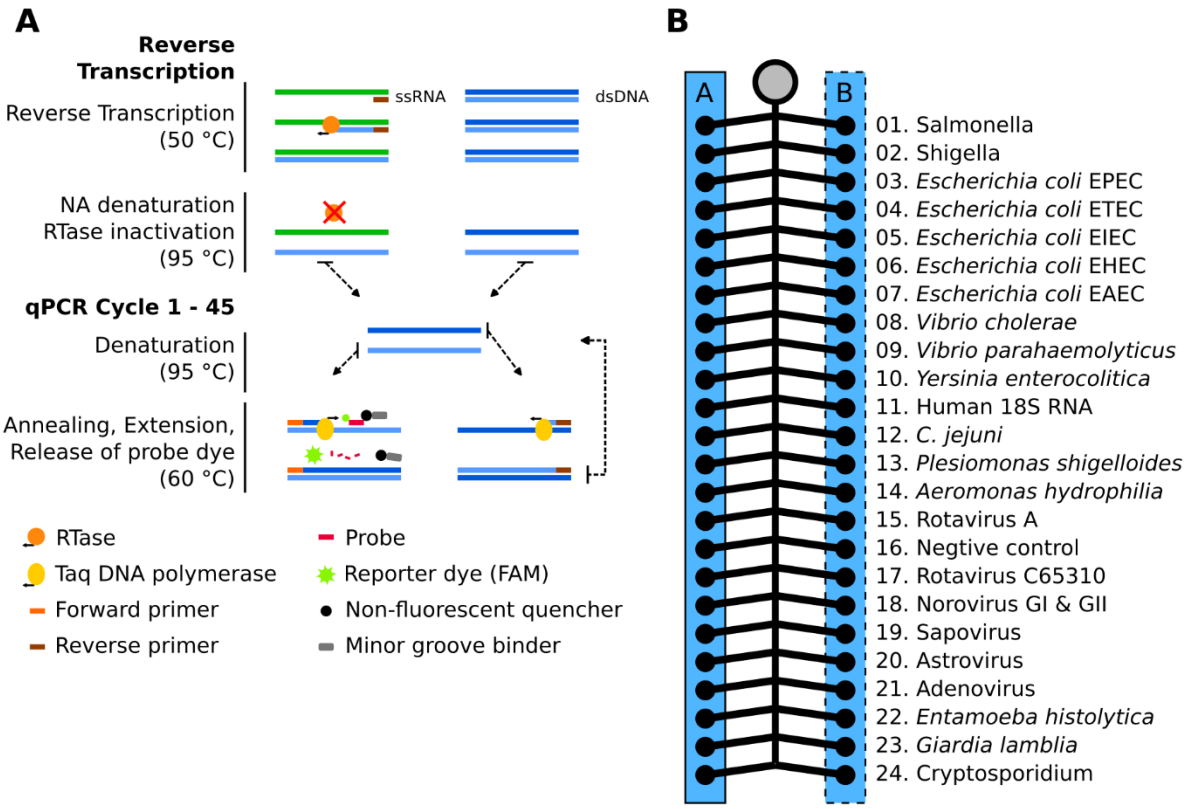

Figure 5.1: GI-TAC chemistry and assay layout. **A:** TaqMan (RT-)qPCR biochemistry overview. Reverse primers bind to ssRNA, which is reverse transcribed. dsDNA remains unaffected during this step. Next, double stranded nucleic acid (NA) is denatured to separate strands, and reverse transcriptase (RTase) is heat inactivated. Forward primers and probes anneal to the complementary DNA. The proximity of the fluorescent dye and the non-fluorescent quencher bound to either ends of the probe inhibits dye fluorescence. When the primer is extended by Taq DNA polymerase, the probe is degraded, releasing the fluorescent dye and quencher and allowing fluorescent light emission from the dye. Each cycle, new primers anneal to template molecules and release fluorescent dye, increasing the fluorescence signal. **B:** Layout of a single lane of the 384-well TaqMan array card. The card contains eight lanes, with each lane consists of two rows of 24 wells. Each column contains identical qPCR assay primers and probes, providing two replicates, A and B, of each assay for each lane. EAEC: enteroaggregative *E. coli*; EHEC: enterohemorragic *E. coli*; EIEC: enteroinvasive *E. coli*; ETEC: enterotoxigenic *E. coli*;

5.1.2 Specific aims

The main aim of this chapter is to assess the prevalence of GI pathogens in ME/CFS patients and determine the effects of FMT, and to evaluate its use as a screening tool. The specific aims are:

- To compare the prevalence of GI pathogens between patients and HCCs from the study "Autoimmunity in ME/CFS" (AI-ME/CFS) (Seton *et al.*, 2023) using the GI-TAC.
- To determine the prevalence of GI pathogens in 73 ME/CFS patients participating in The Comeback Study pre- and post-FMT and determine the effect of FMT.
- To evaluate the use of TACs as a clinical screening tool for large-scale studies on ME/CFS.

5.2 Materials and methods

5.2.1 Ethical considerations

For the pilot study, samples from the study "Autoimmunity in ME/CFS" (AI-ME/CFS), (ClinicalTrials.gov: NCT03254823) (Seton, 2022) were used, as described in Chapter 2. For the Comeback Study samples, the same considerations apply as described in Chapter 3. All research was performed in accordance with the Declaration of Helsinki (World Medical Association, 2013), and the ICH-GCP guidelines. Data was handled following the European Union GDPR and United Kingdom Data Protection Act 2018.

5.2.2 Faecal nucleic acid extraction

For the ME/CFS pilot study, faecal DNA was extracted from 50 – 100 mg of faeces using the Maxwell RSC PureFood GMO and Authentication Kit (Promega, Southamptom, UK, ref. AS1600) on the Promega Maxwell 48 (Promega, Southampton, UK, ref. AS8500) according to manufacturer's instructions. Comeback Study VLP-enriched nucleic acid extracts were obtained as described in Chapter 4.

5.2.3 GI-TAC design

The CCDC GI Panel TaqMan array card was developed by Thermo Fisher Scientific (Thermo Fisher Scientific, Loughborough, ref. 4342253) and wells were loaded with proprietary primer and probe targets of the listed pathogens (Fig. 5.1B). Assay 16 was designed to detect a Phocine herpesvirus spike-in control. This control was omitted in the tested samples and assay 16 is therefore regarded as a negative control in this study.

5.2.4 Validation of the GI-TAC

The GI-TACs were initially validated by Dr Martin Curran (NHS Addenbrookes Hospital, Cambridge, UK) using samples from a previously published study in which GASTRO v4 TACs, developed by Dr Martin Curran, were used to detect pathogens (Agoti *et al.*, 2022). The assay for *Giardia lamblia* was found to be ineffective. As the GI-TACs were stored for more than a year, additional validation was performed by Dr Martin Curran using total nucleic acid extracts from known positive faecal samples, internal controls developed for surveillance during the 2012 London Olympic games, and pooled pathogens from the Zeptometrix NATrol GI Panel (Zeptometrix, Buffalo, NY, USA, ref. NATGIP-BIO) to validate assays for *Adenovirus*, *Astrovirus*, *Norovirus*, *Rotavirus A*, *Sapovirus*, *Campylobacter jejuni*, enteroaggregative *Escherichia coli*

(EAEC), enteropathogenic *E. coli* (EPEC), enterotoxigenic *E. coli* (ETEC), *Plesiomonas shigelloides*, *Shigella*, *Vibrio cholerae*, *Yersinia enterocolitica*, *Cryptosporidium*, and *Entamoeba histolytica*.

5.2.5 GI-TAC quantitative PCR

For each AI-ME/CFS sample, 5 µl of nucleic acid extract was used. For the Comeback Study samples, 1 µl was used. To this, nuclease-free water (Thermo Fisher Scientific, Loughborough, UK, ref. AM9937) was added to a final volume of 75 µl and 25 µl master mix TaqMan Fast Virus 1-step Master Mix (Thermo Fisher Scientific, Loughborough, UK, ref. 4444436) was added and mixed by pipetting. Of this mix, 90 µl was then added to the port of a TaqMan array card. The card was loaded in a centrifuge adaptor (Thermo Fisher Scientific, Loughborough, UK, ref. 75015686) and centrifuged at 300 x *g* for 2 minutes twice, followed by double application of a seal. After removing the loading port strip from the card, the card was loaded into a QuantStudio 7 Flex real-time PCR machine (Thermo Fisher Scientific, Loughborough, UK, ref. 4485690). The reaction was held at 50 °C for 5 minutes for reverse-transcription, then 95 °C for 20 seconds for denaturation and heat inactivation of the reverse-transcriptase, and then for 45 cycles of denaturation at 95 °C for 1 second followed by primer annealing and extension at 60 °C for 20 seconds (Fig. 5.1A). Comeback Study sample qPCR was performed by Dr Martin Curran.

5.2.6 Real-time PCR data analysis

The data was analysed using QuantStudio Real-Time PCR Software v1.3 (Thermo Fisher Scientific, Loughborough, UK). First, the multicomponent curves for the probe dye carboxyfluorescein (FAM) and passive reference dye carboxyrhodamine (ROX) were inspected manually to assess the accuracy of the Ct value estimate for that assay. Then, using the analysis software, the normalised reporter fluorescence (Rn) was calculated by normalising the FAM fluorescence by the ROX fluorescence value. The automatically determined baseline Rn was then subtracted from the Rn at each cycle to obtain the baseline-corrected normalized reporter fluorescence (ΔRn). A threshold of 0.2 was used to calculate the Ct value of assays with exponential growth. An assay was considered positive for a Ct≤40 and a sample was considered positive for a pathogen if both corresponding assay replicates were positive. Dr Martin Curran performed the initial analysis of qPCR results in the QuantStudio software and noted results on result sheets. The result sheets were then compared to the qPCR data by the author and the qPCR data was then exported and loaded into R v.4.1.2. The R packages tidyverse 2.0.0 (Wickham *et al.*, 2019), ggplot2 v3.4.2 (Wickham, 2016), ggpubr 0.6.0 (Kassambara, 2023), ggh4x (van den Brand, 2023) and ggsci 3.0.0 (Xiao, 2023) were used to load and display the data.

5.2.7 *Sapovirus* detection in donor samples

Metagenomic data from patients P056 and P068 and Donor B was acquired as described in Chapter 4. The *Sapovirus* reference genomes were downloaded from the NCBI RefSeq database on 5 May 2024 using the Entrez Direct v.19.9 command line tools esearch with options `-db nuccore -query "txid95341[Organism:exp] AND refseq[filter]"` and efetch options `-format fasta -mode text`. The genomes were indexed using the Bowtie2 (Langmead and Salzberg, 2012) bowtie2-build command with default settings. Then, the reads corresponding to T0 and T3 patient samples and all Donor B samples were mapped against the index using Bowtie2 with the default settings.

5.3 Results

The GI-TAC was previously validated using samples from a study in Kenya (Agoti *et al.*, 2022), that were analysed using a different, validated GI pathogen TAC developed by Dr Martin Curran at the NHS Addenbrooke's hospital in Cambridge, UK, showing all assays were able to detect the target pathogens, with the exception of *Giardia lamblia*.

5.3.1 GI pathogens in severe ME/CFS patients and healthy household controls

For the AI-ME/CFS study, five pairs of severe ME/CFS patients and healthy household controls (HHC) were enrolled. Of four pairs, DNA from faecal samples was analysed using GI-TACs, to detect the presence of GI pathogens. Both the ME/CFS patient and the paired HCC from Pair 2 tested positive for EPEC (Fig. 5.2A). Incidentally, both individuals also self-reported having IBS. Additionally, a single replicate tested was positive for EPEC in the ME/CFS patient of Pair 1. However, as the other replicate was negative, the sample was considered negative for EPEC. The single DNA virus assay on the GI-TAC, *Adenovirus*, was not detected in any of the samples (Fig. 5.2B), nor the parasites (Fig. 5.2C), while the internal positive controls was positive for all samples, indicating the PCR reaction worked effectively (Fig. 5.2D)

5.3.2 Additional validation of GI-TAC

As the GI-TAC was stored at 4 °C for more than two years between the analysis of the AI-ME/CFS samples and the Comeback Study samples, nucleic acid extracts of known positive samples, internal controls developed for surveillance testing during the 2012 London Olympics, and a Zeptometrix GI pathogen control panel were used to re-validate assays on the GI-TAC, with a focus on GI viruses. Collectively, these samples contained *Adenovirus*, *Astrovirus*, *C. jejuni*, *Cryptosporidium*, *E. histolytica*, EAEC, EPEC, ETEC, *Giardia lamblia*, *Norovirus*,

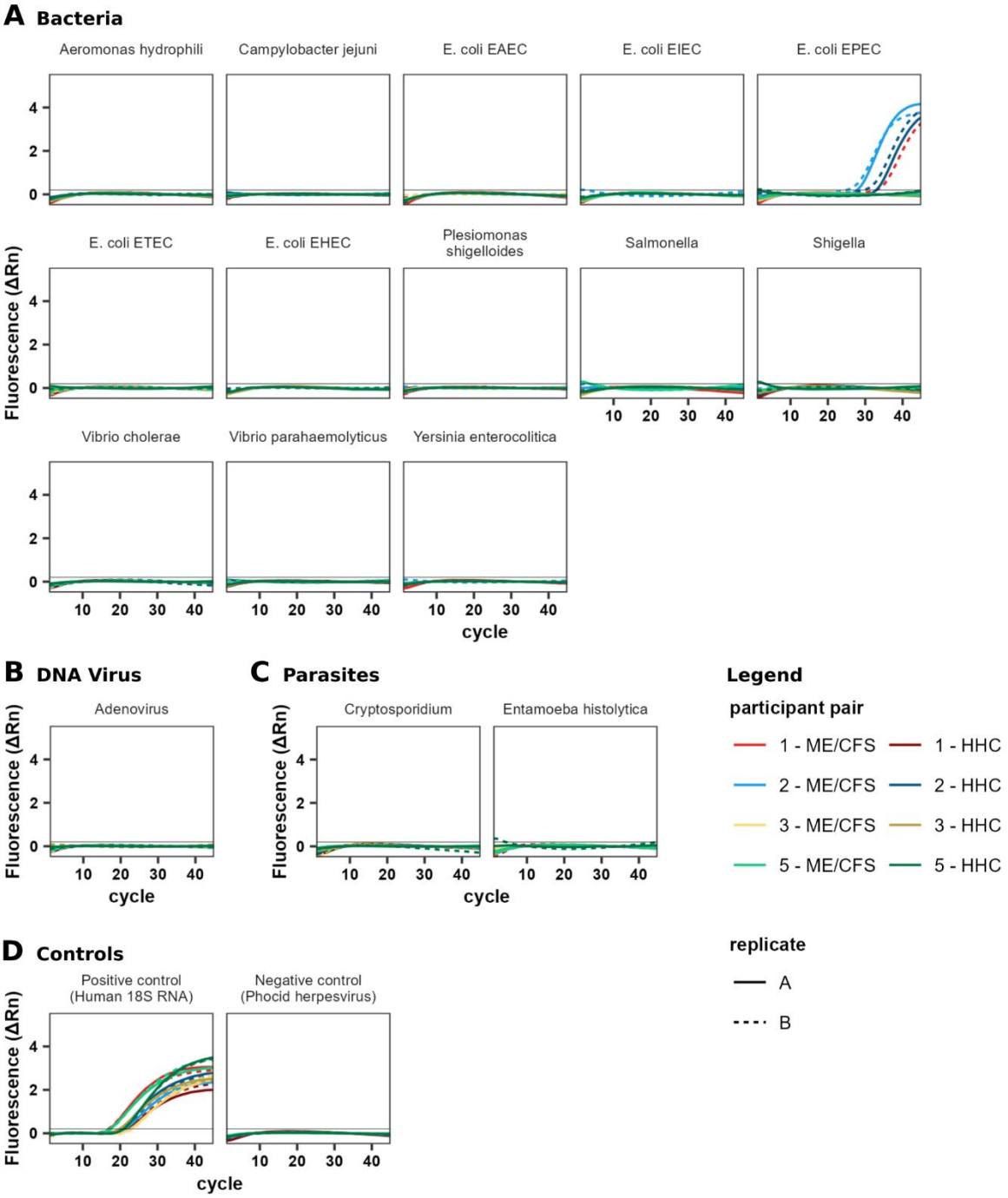

Figure 5.2: Bacterial, DNA viral, and eukaryotic pathogens in severe ME/CFS patients and healthy controls. A, B, C: Faecal DNA extracts were analysed on the GI TaqMan Array Card to detect bacterial (**A**), viral (**B**), and eukaryotic (**C**) pathogens. **D:** Samples were also tested against a positive control, consisting of an assay detecting a human 18S ribosomal RNA gene, and a negative control consisting of an assay against Phocine herpesvirus, which was not spiked into any samples. The legend applies to all subfigures.

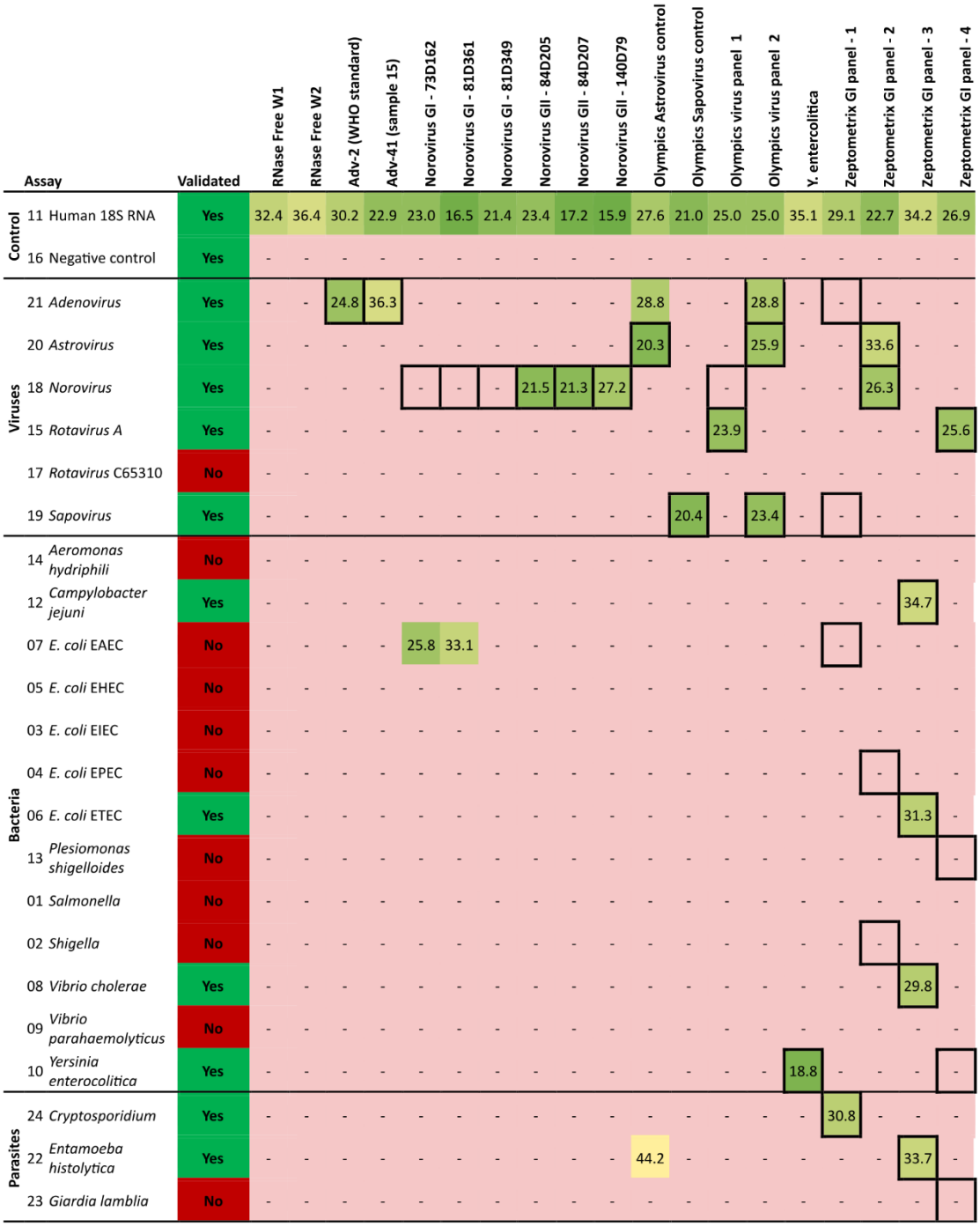

Figure 5.3: GI-TAC assays detect all viruses except norovirus genotype I and rotavirus group C. Due to storage of the array cards past their best by date, all viral and several bacterial and eukaryotic parasite assays were tested using nucleic acid extracts of known positive samples and a Zeptometrix nucleic acid test GI pathogen verification panel. Numbers indicate the Ct value, and thick outlines show the pathogen(s) that the sample is known to contain.

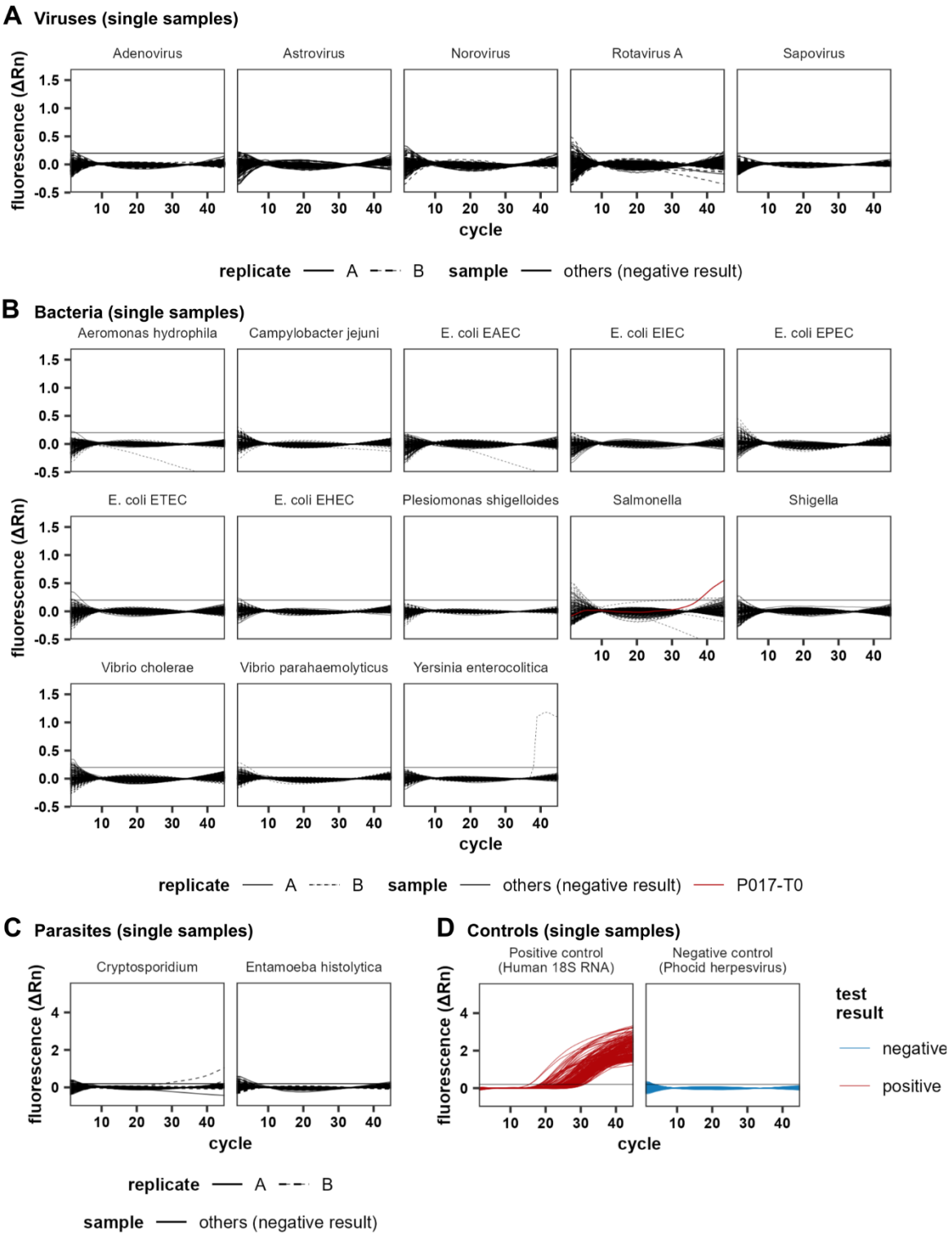

Figure 5.4: No enteric pathogens were detected in Comeback Study samples. Nucleic acid of VLP-enriched faecal extracts were analysed using a GI-TACs. In each graph, only positive sample replicates with a Ct ≤40 are coloured, all other samples are displayed in black. Samples are only considered positive for a pathogen if both replicates are positive. **A, B, C:** Pooled nucleic acid extracts of were analysed on the GI-TAC to detect viral (**A**), bacterial (**B**), and eukaryotic (**C**) pathogens. For each sample, assay replicates A and B are displayed

using a solid and dashed line, respectively. **D:** Samples were also tested against a positive control, consisting of an assay detecting a human 18S ribosomal RNA gene, and a negative control consisting of an assay against Phocine herpesvirus, which was not spiked into any samples.

P. shigelloides, *Rotavirus A*, *Sapovirus*, *Shigella*, *V. cholerae*, and *Y. enterocolitica*. All but one of the viruses contained within the samples were detected (Fig. 5.3). The exception was *Norovirus*, for which only the prevalent genotype II, but not the rare genotype I, was detected. Additionally, the Olympics internal control for *Astrovirus* also tested positive for *Adenovirus*, while *Adenovirus* and *Sapovirus* in Pool 1 of the Zeptometrix test panel were not detected. Together, all virus assays, except for *Rotavirus C65310*, were therefore regarded as validated. Of the 13 bacterial assays only 8 were present among test samples. Of those, *C. jejuni*, ETEC, *V. cholerae* and *Y. enterocolitica* were detected and only these assays were regarded as validated. The EAEC assay gave a positive result for two samples positive for *Norovirus* GI, but since it is unclear whether this is a false positive, the assay was regarded as not validated. The eukaryotic parasites *Cryptosporidium* and *E. histolytica* were both detected in positive samples, while *Giardia lamblia* was not detected, in line with previously conducted validation.

5.3.3 Analysis of Comeback Study patients

To determine the presence of GI pathogens and the effect of FMT in ME/CFS patients, the pre- and post-FMT faecal viral nucleic acid extracts of each patient were analysed using the GI-TAC. While these extracts were produced from isolated faecal VLPs, on average only 14.6% of reads mapped to contigs identified as viral, suggesting a sizeable portion of nucleic acid in the extracts is of non-viral origin. Therefore, non-viral pathogens were also considered. The baseline (T0) and post-FMT (T3) samples of patients for which both samples were available, a total of 146 samples, were randomised with six no-template control samples. The first 13 plates were loaded with single samples, with none of the samples testing positive for any of the viruses (Fig. 5.4A). For bacteria, no samples were positive, although one replicate assay of the baseline sample of patient P017 gave a positive result for *Salmonella* (Fig. 5.4B). Likewise, eukaryotic parasites were not detectable in the samples (Fig. 5.4C), while the positive control assay was positive for all samples (Fig. 5.4D). The remaining samples were pooled in pairs. Of these pairs, sapovirus (SaV) was detected in P068 T0/P056 T3 pair by both assay replicates at an average Ct of 36.8, while no viruses were detected in any of the other pairs (Fig. 5.5A). No bacterial (Fig. 5.5B) or eukaryotic (Fig. 5.5C) pathogens were detected in any of the pairs, while positive controls gave positive results for all (Fig. 5.5D). Lastly, both baseline and post-FMT samples of both patients were analysed separately. SaV was detected by both assay replicates

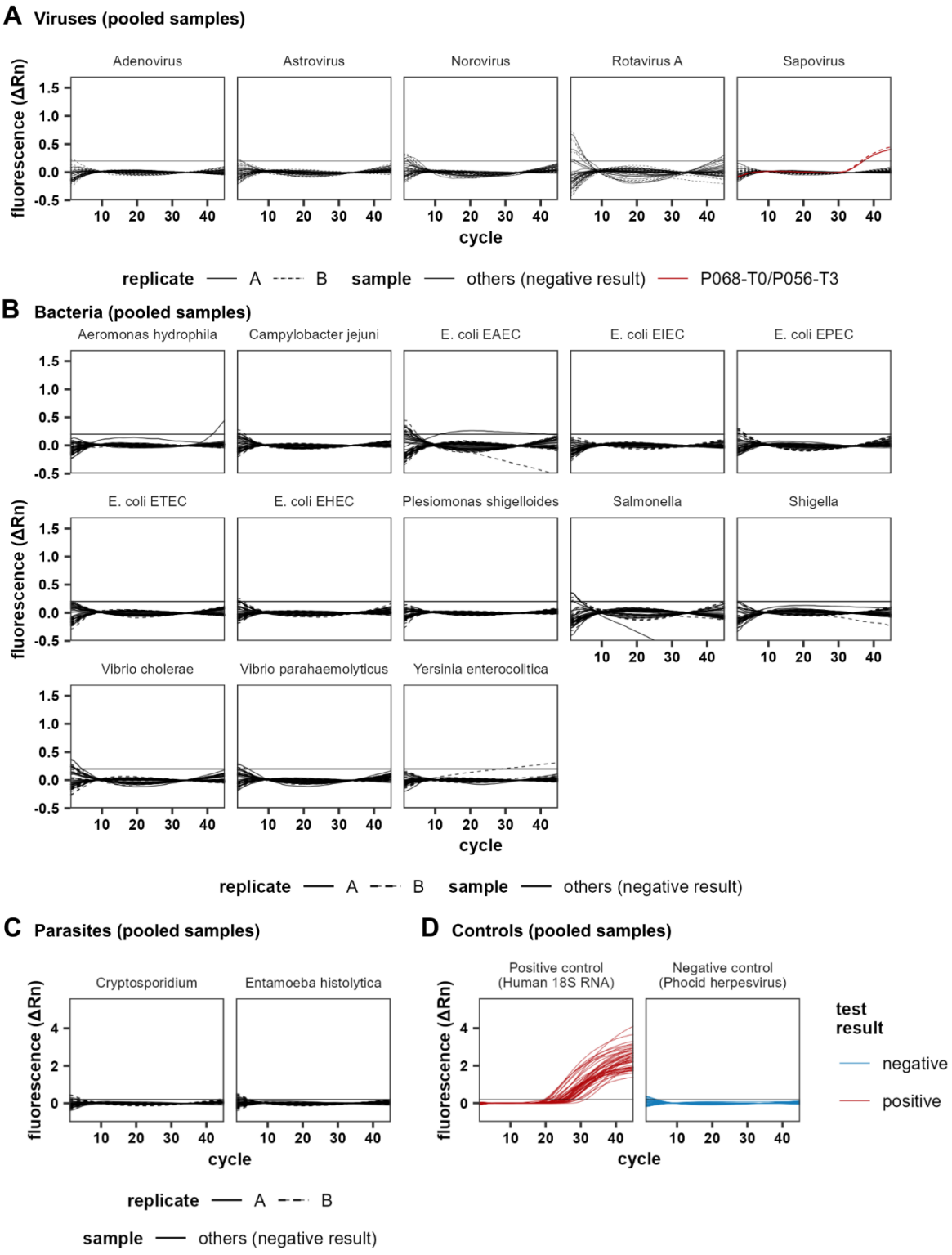

Figure 5.5: A pool of sample P068-T0 and P056-T3 was positive for *Sapovirus*. Two VLP-enriched faecal nucleic acid extracts were pooled randomly and analysed using a GI -TAC. In each graph, only positive sample replicates with a Ct≤40 are coloured, all other samples are displayed in black. Samples are only considered positive for a pathogen if both replicates are positive. **A, B, C:** Pooled nucleic acid extracts of were analysed on the GI-TAC to detect viral (**A**), bacterial (**B**), and eukaryotic (**C**) pathogens. For each sample, assay replicates A and B

are displayed using a solid and dashed line, respectively. **D:** Samples were also tested against a positive control, consisting of an assay detecting a human 18S ribosomal RNA gene, and a negative control consisting of an assay against Phocine herpesvirus, which was not spiked into any samples.

in the post-FTM sample of patient P056 at an average C_T of 38.2, while viral pathogens were not detected in the other samples (Fig. 5.6A, B).

Donors underwent comprehensive screening for GI pathogens, including the GI viral pathogens *Adenovirus*, *Norovirus*, *Rotavirus*, and SaV. To assess whether screening was effective, the possibility of a donor origin of SaV in patient P056 was investigated. However, due to limited availability of GI-TACs, donor samples could not be analysed with this method. Therefore, the metagenomic data acquired in Chapter 4 was analysed, by mapping reads of patients P068 and P056 and Donor B, the donor of patient P056, to 8 SaV reference genomes in the NCBI RefSeq database. No reads mapping to SaV genomes were recovered in any of the samples. As no SaV reads were found in the positive sample in which SaV was detected twice by qPCR this suggests the metagenomics approach cannot be used to confirm the presence or absence of SaV.

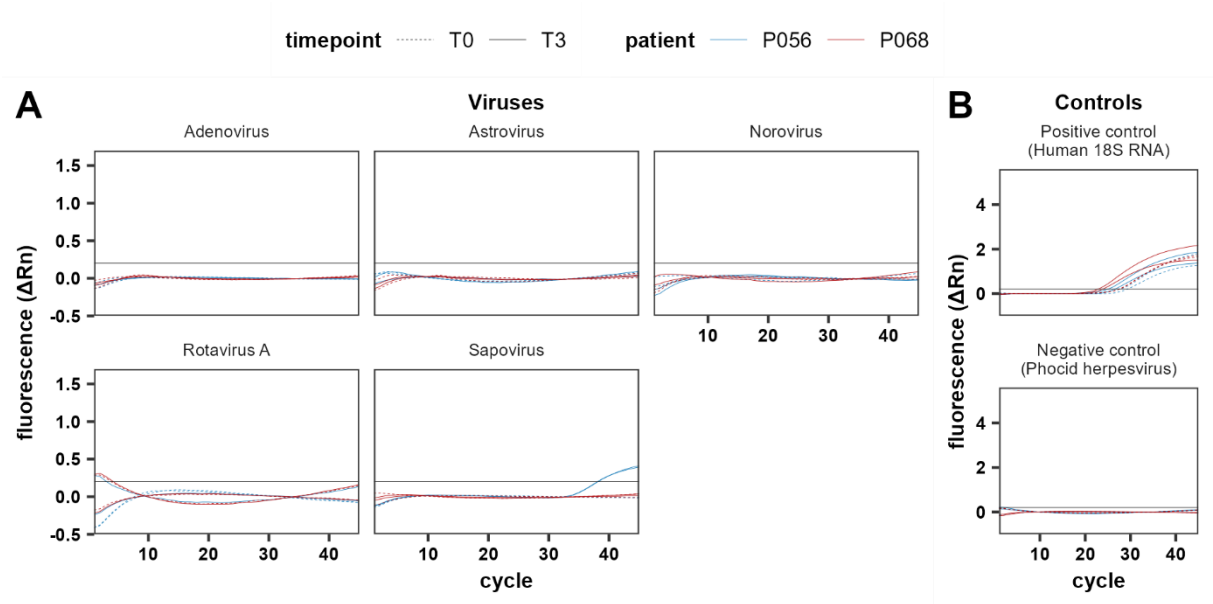

Figure 5.6: Patient P056 tested positive for Sapovirus post-FMT. A: The pre- (T0) and post-FMT (T3) nucleic acid of faecal VLP extracts of patients P056 and P068 were analysed individually on a GI TaqMan array card for the presence of viral enteric pathogens. **B:** Positive (Human 18S RNA) and negative (Phocine herpesvirus) control assays. The legend applies to both subfigures.

5.4 Discussion

To our knowledge, this is the first time multiplex real-time PCR assays have been used to analyse the presence of bacterial, viral, and eukaryotic GI pathogens in ME/CFS patients. GI pathogens were analysed in a small cohort of severe ME/CFS patients, as well as a larger cohort of ME/CFS patients participating in the Comeback Study. A bacterial enteropathogenic *E. coli* strain was detected in one severe ME/CFS patient, while *Sapovirus* was detected in another ME/CFS patient in the Comeback Study. Overall, there was no evidence of high prevalence of GI pathogens in ME/CFS, suggesting the tested pathogens play no role in ME/CFS.

Of the four pairs of severe ME/CFS patient and HHC, one pair tested positive for EPEC. One of the hallmark genes of EPEC is the *eae* gene (Jerse *et al.*, 1990), which encodes for Intimin, an adhesin that enables attachment of the bacterium to the GI epithelium and causes attaching and effacing lesions in the epithelium (Mare *et al.*, 2021). Typical EPEC (tEPEC) contains the *E. coli* adherence factor plasmid (pEAF), while atypical EPEC (aEPEC) does not (Carlino *et al.*, 2020). The *eae* gene is one of the sequences targeted in multiplex qPCR assays (Liu *et al.*, 2013; Sun *et al.*, 2020; Agoti *et al.*, 2022). Differentiating between tEPEC and aEPEC can be done through assays that target genes in the bundle-forming pilus *bfp* operon on the pEAF (Hernandes *et al.*, 2009; Liu *et al.*, 2013; Carlino *et al.*, 2020; Sun *et al.*, 2020; Mare *et al.*, 2021). While the EPEC assay target in the GI-TAC is propriety information, only one assay is present and since *eae* is a marker of both typical and aEPEC (Mare *et al.*, 2021), this is likely to be the target in the GI-TAC. Indeed, the widely used BioFire FilmArray GI Panel only detects the *eae* gene and cannot distinguish between typical and atypical EPEC (Carlino *et al.*, 2020). Of note, EPEC was detected in both individuals of Pair 2, and both individuals reported having IBS. EPEC has been associated with post-infectious IBS (Beatty, Bhargava and Buret, 2014), but the infectious history of the patient and the HHC are unknown. As EPEC was detected in the patient as well as the HHC of Pair 2, this would suggest EPEC is not a contributor to ME/CFS in this patient. Ultimately the lack of information on the individual's medical history prevents further speculation on the relationship between the detected EPEC and IBS and/or ME/CFS symptoms. No other GI pathogens were detected in this cohort using the GI-TAC. It is important to note that RNA viruses were omitted in this analysis, since DNA extracts were used, and the presence of rotavirus, norovirus, sapovirus, and astrovirus could therefore not be determined.

In the patients participating in the Comeback Study, SaV was the only detected pathogen, in the post-FMT faecal sample of a patient in the active FMT group. There is therefore no evidence of increased prevalence of the common GI viral pathogens adenovirus, astrovirus, norovirus, rotavirus, and SaV in ME/CFS. However, this does not exclude the involvement of other viral pathogens, such as enterovirus or herpesvirus, that linger in the GI tract of ME/CFS patients,

that were not tested for. The inability to detect *Sapovirus* sequences in the metagenomic data obtained from the same sample, illustrates the increased sensitivity of qPCR over metagenomic sequencing for detecting rare virus sequences in faecal samples. Thus, targeted qPCR-based assays might be the preferred option for identifying any associations with other eukaryotic viruses, e.g., herpesviruses and enteroviruses in ME/CFS patient samples.

Since detection of SaV occurred in a post-FMT sample of a patient in the active transplant group, this raises the question of whether this virus was transferred from the donor to the patient.While metagenomic analysis of the Donor samples did not return SaV reads, neither did the qPCR-positive patient sample. Thus, based on the current data it remains unclear whether SaV was present in Donor B and transferred through FMT to patient P056, or whether this patient acquired SaV independently. *Sapovirus* infection can be symptomatic and asymptomatic (Usuku *et al.*, 2008; Yoshida *et al.*, 2009; Yamashita *et al.*, 2010) and shedding can continue for up to four weeks (Oka *et al.*, 2015). Chronic infections of *Sapovirus* have only been reported in immunocompromised infants (Rubio-Mora, Carrascoso and Rodríguez, 2024) and adults, with the longest reported duration of shedding of 119 days (Pietsch and Liebert, 2019). SaV is a common cause of acute gastroenteritis (AGE) and has a prevalence of 1% – 10% in Europe in children with AGE (Razizadeh, Khatami and Zarei, 2022). Globally, the prevalence of SaV in asymptomatic people >5 years old is 0.8% (Valcarce *et al.*, 2021). SaV is transmitted via the faecal-oral route, including by asymptomatic food handlers (Usuku *et al.*, 2008; Yoshida *et al.*, 2009). FMT donors are screened for *Sapovirus* (see Chapter 4, Methods section), and transfer from the donor would imply unusually long shedding, given that the post-FMT sample was taken 3 months after transplantation. It is therefore unlikely that the SaV detected in patient P056 was acquired from the donor, and instead was acquired SaV through other means of transmission, including contact with shedding individuals or contaminated food.

For the GI-TAC analysis of the Comeback Study, VLP-enriched nucleic acid extracts were used. While on average 85.4% of reads across samples did not map to viral contigs, it is possible that bacterial pathogens and their genetic material present in the faecal samples were removed, preventing detection by the GI-TAC. The present results could be complemented by future sequencing of total DNA extracts that will allow investigation of pathogenic *E. coli* strains, including EPEC.

5.4.1 Limitations

The study was limited by availability of genetic material, as for the severe ME/CFS pilot study, only total DNA extracts were available, while only virus-enriched total nucleic acid extracts were available for the Comeback Study. Additionally, analysis of the Comeback Study was restricted due to available sample volumes and use of volumes lower than recommended,

possibly reducing the sensitivity of the assays. Future sequencing or qPCR analysis of total DNA acid extracts of Comeback Study faecal samples and RNA or total nucleic acid extracts of the AI-ME/CFS study faecal samples would complement the present analysis. In the future extraction of larger sample volumes will produce more genetic material for analysis, which increases sensitivity. Due to the low number of detected pathogens, the relationship between detection and disease remains unclear, as a positive test result does not necessarily indicate the cause of disease (Levine and Robins-Browne, 2012; Hata, Powell and Starolis, 2023).

5.4.2 Conclusion

Altogether, the data do not support an association of gastroenteritis-associated human viruses in ME/CFS. EPEC was detected in a single severe ME/CFS patient and HHC pair from a cohort of four pairs, but the small sample size of this cohort prevents drawing further conclusions on the association with ME/CFS. While *Sapovirus* was detected in a Comeback Study patient post-FMT, transfer from the donor is unlikely. In the future, results can be complemented by metagenomic sequencing of faecal total nucleic acid extracts to determine the presence of bacterial pathogens in Comeback Study participants. The GI-TAC was amenable for us in the Comeback Study, and the TAC platform could be used in the future to incorporate qPCR assays targeting on other ME/CFS-associated viruses like herpesviruses and enteroviruses in VLP-enriched nucleic acid extracts or bulk nucleic acid extracts.

CHAPTER 6: GENERAL DISCUSSION

6.1 Summary and impact of results

Worldwide, up to 65 million people suffer from ME/CFS and this number is likely to increase significantly following the COVID-19 pandemic (Hanson and Germain, 2020; Wong and Weitzer, 2021). With a significant impact on the quality of life of those with ME/CFS and their carers (Vyas *et al.*, 2022), as well as a significant economic impact on society as a whole (Mirin, Dimmock and Jason, 2022), a treatment is urgently needed. Since a sizeable portion (up to 65%) of ME/CFS patients has co-morbid GI disturbances including IBS (Chu *et al.*, 2019; Tschopp *et al.*, 2023), the GI prokaryome composition deviates from that of healthy controls (König *et al.*, 2022; Wang *et al.*, 2024), and GI microbiota play a vital role in human health and disease (Mirzaei and Maurice, 2017), FMT has been considered as a treatment for ME/CFS (Skjevling *et al.*, 2024). The GI virome is an understudied component of the GI microbiome, and RNA viruses have been understudied in GI virome research. As phages shape GI prokaryome composition and, more importantly, function, it is vital to determine the role of viruses in FMT (Mirzaei and Maurice, 2017). The GI tract also harbours eukaryotic viruses, which includes mostly diet-derived viruses but also disease-associated and non-disease-associated human viruses, and the role of these viruses in FMT is still unknown (Lam *et al.*, 2022). Investigating the role of human viruses is particularly pertinent in the case of ME/CFS, as persistent viral infections and reactivation of latent viral infections, including herpesvirus, enterovirus, and parvovirus, have been implicated in ME/CFS aetiology. The main goal of this thesis was therefore to investigate whether FMT can induce changes in the composition of the GI prokaryotic and eukaryotic virome in ME/CFS patients. To this end, 1) a SYBR Gold-based staining method was optimised and an image analysis pipeline was developed to quantify virus stock titres using EFM, 2) a virus MC was produced, added to faecal samples, and analysed to compare a WTA2-based and a SISPA-based viral metagenomics method, 3) the faecal virome of ME/CFS patients and FMT donors participating in the Comeback Study was analysed to determine the effects of FMT on the GI virome, and 4) the prevalence GI pathogens in ME/CFS patients participating in the AI-ME/CFS study and the Comeback Study was assessed using TACs. The main findings of this thesis are:

- Sufficient dilution of the virus stock, but not incubation time, during staining with SYBR Gold facilitates optimal visualisation of small RNA and DNA viruses through EFM.
- Both WTA2 and SISPA are biased towards ssDNA viruses.

- WTA2 is more sensitive, has higher coverage uniformity, and better assembly quality compared to SISPA.
- FMT from a healthy donor increases virus richness in ME/CFS patients at three months post-FMT.
- Following FMT from a healthy donor, ME/CFS patient virome composition becomes more like the composition of donor virome and leads to engraftment of donor viruses at three months post-FMT.
- Different donors achieve distinct levels of change in composition and of engraftment.
- There is no evidence of an association of gastroenteritis-associated GI pathogens with ME/CFS patients.

6.1.1 Developing a virus reference standard

In support of the virome analysis of the Comeback Study, a virus MC was developed as a reference standard. For this, an EFM protocol has been optimised using 25X SYBR Gold to fluorescently stain virus particles in virus stocks, showing that sufficient dilution of the virus stock at the time of incubation with SYBR Gold is required for optimal detection of virus particles. While EFM has been used before to quantify viruses for use in MCs (Kleiner, Hooper and Duerkop, 2015; Roux *et al.*, 2016; Warwick-Dugdale *et al.*, 2019), this is, to our knowledge, the first time it was used to quantify eukaryotic viruses, enveloped viruses, and RNA viruses in a MC. Additionally, the development of image analysis scripts has increased image processing speed and consistency.

6.1.2 Assessing bias and reproducibility of viral metagenomics methods

The quantification of viruses by EFM enabled spiking of faecal samples with a MC at accurately determined titres, irrespective of particle infectivity. This enabled one of the most comprehensive analyses of viral metagenomics methods to date, assessing methodological bias for individual viruses and showing that ssDNA viruses are preferentially amplified while enveloped virus recovery is reduced, and that accurate estimation of relative abundance requires ≥70% genome completeness. By analysing the faecal virome we determined that the consistency of relative abundance estimates is highest for genome fragments with ≥10X coverage. This level of coverage has previously been shown to be required for successful assembly of viral genomes (Roux *et al.*, 2017), which was also confirmed in the Comeback Study virome analysis, where all vOTUs with ≥90% completeness had >10X coverage. Fragments with >10X coverage typically had >0.1% relative abundance in the experiment. Additionally, taxonomic assignment of faecal viromes is consistent across replicates. Two widely used methods for virome amplification, WTA2 and SISPA, were compared as well, showing that WTA2 has increased sensitivity, reduced bias for ssDNA viruses, increased coverage uniformity, and increased assembly quality, compared to SISPA. To our knowledge, this is the first time these methods are compared comprehensively.

6.1.3 Analysing Comeback Study GI viromes and evaluating the GI-TAC

For the analysis of the Comeback Study virome, an internal control faecal sample and MC sample was used to determine consistency between batches, and blank samples were included to enable removal of contaminating sequences. A bespoke bioinformatics pipeline was developed for the analysis of patient and control samples. Analysis of the GI virome in ME/CFS patients shows that species richness increases following FMT from healthy donors, but not following autologous FMT. The overall composition of the GI virome becomes more like the donor virome post-FMT, although the effect was stronger for some donors than for others. Donor virus sequences are detectable in ME/CFS patient faeces post-FMT but not pre-FMT, indicating engraftment. The majority of engrafted viruses are phages, and engraftment of human viruses was not detected. Eukaryotic virus sequences that were common between donors and post-FMT faecal samples of the recipients belonged to the *Alsuviricetes* and *Arfiviricetes* classes, which predominantly contain plant and fungal viruses, and thus are likely diet derived. Nonetheless, the classes *Alsuviricetes* and *Arfiviricetes* also contain a few human-infecting virus taxa, including the *Orthohepevirus* genus that includes hepatitis E virus, the *Alphavirus* genus that includes Sindbis virus and Chikungunya virus, and the *Redondoviridae* family that includes human viruses that are not associated with disease. Thus, engraftment of human viruses cannot be ruled out completely. Still, overall, both metagenomic analysis and GI-TAC analysis did not identify an association of GI human infecting viruses with ME/CFS. This is the first comprehensive analysis of GI virome in ME/CFS, and one of a few virome analyses on the virome in FMT. Additionally, this is, to our knowledge, the first time a GI-TAC was used to assess the prevalence of GI pathogens in ME/CFS. While the GI-TAC yielded few positive samples, the system is amenable to our large-scale study and fits use cases requiring high throughput.

6.2 Limitations

While the MC experiment in Chapter 3 showed that a minimum of 70% genome completeness was required to determine the relative abundance accurately, few viral sequences (948/22,841) in the vOTU data set had an estimated completeness ≥70%. This cut-off was therefore not applied in the virome analysis to maintain sensitivity in our analysis for lower abundant viruses. As there are likely fragmented viral genomes in the data set and in some cases, multiple sequences originate from the same virus, and genome binning could improve genome completeness and accuracy of the relative abundance.

The most important remaining question is how virus engraftment relates to the primary and secondary end points of The Comeback Study. At the time of writing, clinical outcome data,

including fatigue as measured by the SF-36, GI symptoms, and cognitive performance as measured by RBANS were not available. Future integration of these data with virome and microbiome data will help assess the effect of virome engraftment on the clinical changes in ME/CFS. Another limitation of the Comeback Study virome analysis is the lack of healthy controls, which prevents comparing the viromes of ME/CFS patients and healthy individuals, and thus identification of any functional differences between ME/CFS and healthy individual viromes.

6.3 Future directions

The optimisation of an EFM approach for quantifying diverse viruses and an accompanying image analysis software formed the basis for the evaluation of bias and reproducibility of our viral metagenomics method. It enabled the construction of a diverse virus MC that was crucial to understanding some of the biases of the method, including overrepresentation of ssDNA viruses and underrepresentation of enveloped viruses. The EFM method does not require the development of virus-specific assays and can be adopted by any group with access to a fluorescence microscope. Integration of the EFM analysis pipeline with the MC analysis pipeline could provide a useful platform for other groups wishing to evaluate their viral metagenomics approaches. Use of higher purity virus stocks, for instance as obtained through density gradient centrifugation, more advanced microscopy techniques, and further automation of image analysis could further improve the accuracy of virus quantification.

While experiments indicate the superiority of WTA2 over SISPA, supply chain issues prevented us from using WTA2 in the Comeback Study. Since SISPA is more economical and does not rely on the availability of a single kit and components are more easily exchanged, SISPA is less susceptible to supply chain issues. A recent virome study in IBD modified an adaptase linker amplification method (Roux *et al.*, 2016) to produce sequencing libraries from unamplified VLP nucleic acid extracts (Stockdale *et al.*, 2023). Future experiments should focus on comparing this method to WTA2 and SISPA, as well as further optimisation of SISPA, for instance through the use of different primers and a reduction in amplification cycles.

Phage-bacteria dynamics are a key factor in shaping microbiome structure, but which dynamics are at play during FMT remains to be fully elucidated (Lam *et al.*, 2022). The majority of engrafted viruses were phages, and future integration of our virome data with microbiome data will offer valuable insights on donor and recipient microbiome properties affect engraftment. For instance, it will be interesting to determine which proportion of phages are temperate and lytic phages, whether there is evidence that temperate phages transfer by piggybacking on bacteria through lysogenic replication or as virus particles during a lytic

replication cycle, and how phage transfer correlates with changes in the resident microbiota. Investigating the presence of phage defence systems in donor and resident bacteria, as well as anti-phage defence systems in donor and resident phages, could give valuable new insights. Other factors that could affect engraftment success include method of delivery, e.g., colonoscopy, enema, or oral capsules, and the number of treatments (Biazzo and Deidda, 2022). Future research should focus on comparing these to the methods used in the Comeback Study.

The GI-TAC system could complement future GI virome analysis by including assays for other viruses, like enterovirus and other *Picornaviridae* viruses, herpesvirus, anellovirus, and circovirus assays. Particularly assays for *Picornaviridae* and *Cicroviridae* viruses could be of interest, as sequences of these families were detected in several ME/CFS patient samples. For instance, *Picornaviridae* sequences were detected in several transplant samples of Donor B but none of the recipients. The increased sensitivity of the TAC could give new insights into the prevalence of these viruses in ME/CFS patients, and the effect of FMT.

Another key factor in ME/CFS aetiology is the immune system, which was not addressed in this thesis. Future studies will benefit from analysing correlations of GI virome changes with immunological changes, including antibody repertoire and inflammation markers and correlations of these changes with clinical outcome. GI inflammation has been associated with the induction of prophages, and it will also be interesting to determine whether clinical changes correspond to changes in the number of temperate phages replicating via the lytic replication cycle.

6.4 Conclusion

This thesis has laid the groundwork for a comprehensive analysis pipeline for the evaluation of bias and reproducibility of viral metagenomics methods and has provided the first insight into changes in the GI virome following FMT in ME/CFS. It has shown that after three months, donor GI viruses engraft in the patient microbiota, and that different donors achieve distinct levels of engraftment. This is a first indication that donor microbes can be transferred from healthy donors to ME/CFS patients using FMT, with correlations between engraftment and clinical outcome remaining to be determined.

References

Ablashi, D.V. *et al.* (2000) 'Frequent HHV-6 reactivation in multiple sclerosis (MS) and chronic fatigue syndrome (CFS) patients', *Journal of Clinical Virology*, 16(3), pp. 179–191. Available at: https://doi.org/10.1016/S1386-6532(99)00079-7.

Abulencia, C.B. *et al.* (2006) 'Environmental Whole-Genome Amplification To Access Microbial Populations in Contaminated Sediments', *Applied and Environmental Microbiology*, 72(5), pp. 3291–3301. Available at: https://doi.org/10.1128/AEM.72.5.3291-3301.2006.

Acheson, E.D. (1959) 'The clinical syndrome variously called benign myalgic encephalomyelitis, Iceland disease and epidemic neuromyasthenia', *The American Journal of Medicine*, 26(4), pp. 569–595. Available at: https://doi.org/10.1016/0002-9343(59)90280-3.

Agoti, C.N. *et al.* (2022) 'Differences in epidemiology of enteropathogens in children pre- and post-rotavirus vaccine introduction in Kilifi, coastal Kenya', *Gut Pathogens*, 14(1), p. 32. Available at: https://doi.org/10.1186/s13099-022-00506-z.

Alexander, K.L. *et al.* (2021) 'Human Microbiota Flagellins Drive Adaptive Immune Responses in Crohn's Disease', *Gastroenterology*, 161(2), pp. 522-535.e6. Available at: https://doi.org/10.1053/j.gastro.2021.03.064.

Allegretti, J.R. *et al.* (2019) 'The evolution of the use of faecal microbiota transplantation and emerging therapeutic indications', *The Lancet*, 394(10196), pp. 420–431. Available at: https://doi.org/10.1016/S0140-6736(19)31266-8.

Altschul, S.F. *et al.* (1990) 'Basic local alignment search tool', *Journal of Molecular Biology*, 215(3), pp. 403–410. Available at: https://doi.org/10.1016/S0022-2836(05)80360-2.

Amos, G.C.A. *et al.* (2020) 'Developing standards for the microbiome field', *Microbiome*, 8(1), pp. 1–13. Available at: https://doi.org/10.1186/s40168-020-00856-3.

Anderson, B. *et al.* (2011) 'Enumeration of bacteriophage particles', *Bacteriophage*, 1(2), pp. 86–93. Available at: https://doi.org/10.4161/bact.1.2.15456.

Annesley, S.J. *et al.* (2024) 'Unravelling shared mechanisms: insights from recent ME/CFS research to illuminate long COVID pathologies', *Trends in Molecular Medicine*, 0(0). Available at: https://doi.org/10.1016/j.molmed.2024.02.003.

Ansari, M.H. *et al.* (2020) 'Viral metagenomic analysis of fecal samples reveals an enteric virome signature in irritable bowel syndrome', *BMC Microbiology*, 20(1), pp. 1–12. Available at: https://doi.org/10.1186/s12866-020-01817-4.

Antipov, D. *et al.* (2020) 'Metaviral SPAdes: assembly of viruses from metagenomic data', *Bioinformatics*, 36(14), pp. 4126–4129. Available at: https://doi.org/10.1093/bioinformatics/btaa490.

Aoun Sebaiti, M. *et al.* (2022) 'Systematic review and meta-analysis of cognitive impairment in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)', *Scientific Reports*, 12(1), p. 2157. Available at: https://doi.org/10.1038/s41598-021-04764-w.

Apostolou, E. *et al.* (2022) 'Saliva antibody-fingerprint of reactivated latent viruses after mild/asymptomatic COVID-19 is unique in patients with myalgic-encephalomyelitis/chronic fatigue syndrome', *Frontiers in Immunology*, 13. Available at: https://doi.org/10.3389/fimmu.2022.949787.

Apostolou, E. and Rosén, A. (2024) 'Epigenetic reprograming in myalgic encephalomyelitis/chronic fatigue syndrome: A narrative of latent viruses', *Journal of Internal Medicine*, pp. 1–23. Available at: https://doi.org/10.1111/joim.13792.

Arisdakessian, C.G. *et al.* (2021) 'CoCoNet: an efficient deep learning tool for viral metagenome binning', *Bioinformatics*, 37(18), pp. 2803–2810. Available at: https://doi.org/10.1093/bioinformatics/btab213.

Ariza, M.E. *et al.* (2013) 'Epstein-Barr Virus Encoded dUTPase Containing Exosomes Modulate Innate and Adaptive Immune Responses in Human Dendritic Cells and Peripheral Blood Mononuclear Cells', *PLOS ONE*, 8(7). Available at: https://doi.org/10.1371/journal.pone.0069827.

Ariza, M.E. (2021) 'Myalgic encephalomyelitis/chronic fatigue syndrome: The human herpesviruses are back!', *Biomolecules*, 11(2), pp. 1–17. Available at: https://doi.org/10.3390/biom11020185.

Aroniadis, O.C. *et al.* (2019) 'Faecal microbiota transplantation for diarrhoea-predominant irritable bowel syndrome: a double-blind, randomised, placebo-controlled trial', *The Lancet Gastroenterology & Hepatology*, 4(9), pp. 675–685. Available at: https://doi.org/10.1016/S2468-1253(19)30198-0.

Asplund, M. *et al.* (2019) 'Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries', *Clinical Microbiology and Infection*, 25(10), pp. 1277–1285. Available at: https://doi.org/10.1016/j.cmi.2019.04.028.

Attali, D. and Baker, C. (2023) 'ggExtra'. Available at: https://github.com/daattali/ggExtra (Accessed: 1 September 2023).

Bakken, I.J. *et al.* (2014) 'Two age peaks in the incidence of chronic fatigue syndrome/myalgic encephalomyelitis: a population-based registry study from Norway 2008-2012', *BMC Medicine*, 12(1), p. 167. Available at: https://doi.org/10.1186/s12916-014-0167-5.

Balestrieri, E. *et al.* (2021) 'Evidence of the pathogenic HERV-W envelope expression in T lymphocytes in association with the respiratory outcome of COVID-19 patients', *EBioMedicine*, 66. Available at: https://doi.org/10.1016/j.ebiom.2021.103341.

Bankevich, A. *et al.* (2012) 'SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing', *Journal of Computational Biology*, 19(5), pp. 455–477. Available at: https://doi.org/10.1089/cmb.2012.0021.

Bansal, A.S. *et al.* (2012) 'Chronic fatigue syndrome, the immune system and viral infection', *Brain, Behavior, and Immunity*, 26(1), pp. 24–31. Available at: https://doi.org/10.1016/j.bbi.2011.06.016.

Barbara, G. *et al.* (2019) 'Rome Foundation Working Team Report on Post-Infection Irritable Bowel Syndrome', *Gastroenterology*, 156(1), pp. 46-58.e7. Available at: https://doi.org/10.1053/j.gastro.2018.07.011.

Barr, J.J. *et al.* (2013) 'Bacteriophage adhering to mucus provide a non–host-derived immunity', *Proceedings of the National Academy of Sciences*, 110(26), pp. 10771–10776. Available at: https://doi.org/10.1073/pnas.1305923110.

Bateman, L. *et al.* (2021) 'Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: Essentials of Diagnosis and Management', *Mayo Clinic Proceedings*, 96(11), pp. 2861–2878. Available at: https://doi.org/10.1016/j.mayocp.2021.07.004.

Beatty, J.K., Bhargava, A. and Buret, A.G. (2014) 'Post-infectious irritable bowel syndrome: Mechanistic insights into chronic disturbances following enteric infection', *World Journal of Gastroenterology : WJG*, 20(14), pp. 3976–3985. Available at: https://doi.org/10.3748/wjg.v20.i14.3976.

Beller, L. and Matthijnssens, J. (2019) 'What is (not) known about the dynamics of the human gut virome in health and disease', *Current Opinion in Virology*, 37, pp. 52–57. Available at: https://doi.org/10.1016/j.coviro.2019.05.013.

Berger, A.K. and Mainou, B.A. (2018) 'Interactions between enteric bacteria and eukaryotic viruses impact the outcome of infection', *Viruses*, 10(1), pp. 4–6. Available at: https://doi.org/10.3390/v10010019.

Bhar, S. *et al.* (2022) 'Bacterial extracellular vesicles control murine norovirus infection through modulation of antiviral immune responses', *Frontiers in Immunology*, 13. Available at: https://www.frontiersin.org/articles/10.3389/fimmu.2022.909949.

Biazzo, M. and Deidda, G. (2022) 'Fecal Microbiota Transplantation as New Therapeutic Avenue for Human Diseases', *Journal of Clinical Medicine*, 11(14), p. 4119. Available at: https://doi.org/10.3390/jcm11144119.

Biller, S.J. *et al.* (2017) 'Membrane vesicles in sea water: Heterogeneous DNA content and implications for viral abundance estimates', *ISME Journal*, 11(2), pp. 394–404. Available at: https://doi.org/10.1038/ismej.2016.134.

Blanco, L. *et al.* (1989) 'Highly Efficient DNA Synthesis by the Phage ϕ 29 DNA Polymerase: Symmetrical Mode of DNA Replication', *Journal of Biological Chemistry*, 264(15), pp. 8935–8940. Available at: https://doi.org/10.1016/S0021-9258(18)81883-X.

Blomberg, J. *et al.* (2018) 'Infection Elicited Autoimmunity and Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: An Explanatory Model', *Frontiers in Immunology*, 9(299), pp. 1–20. Available at: https://doi.org/10.3389/fimmu.2018.00229.

Blomberg, J. *et al.* (2019) 'Antibodies to Human Herpesviruses in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome Patients.', *Frontiers in Immunology*, 10(1946), pp. 1–12. Available at: https://doi.org/10.3389/fimmu.2019.01946.

Blundell, S. *et al.* (2015) 'Chronic fatigue syndrome and circulating cytokines: A systematic review', *Brain, Behavior, and Immunity*, 50, pp. 186–195. Available at: https://doi.org/10.1016/j.bbi.2015.07.004.

Boers, S.A., Jansen, R. and Hays, J.P. (2019) 'Understanding and overcoming the pitfalls and biases of next-generation sequencing (NGS) methods for use in the routine clinical microbiological diagnostic laboratory', *European Journal of Clinical Microbiology and Infectious Diseases*, 38(6), pp. 1059–1070. Available at: https://doi.org/10.1007/s10096-019-03520-3.

Bonilla, H. *et al.* (2023) 'Myalgic Encephalomyelitis/Chronic Fatigue Syndrome is common in post-acute sequelae of SARS-CoV-2 infection (PASC): Results from a post-COVID-19 multidisciplinary clinic', *Frontiers in Neurology*, 14. Available at: https://doi.org/10.3389/fneur.2023.1090747.

Borody, T., Nowak, A. and Finlayson, S. (2012) 'The GI microbiome and its role in Chronic Fatigue Syndrome: A summary of bacteriotherapy', *ACNEM Journal*, 31(3), p. 3.

Bourgonje, A.R. *et al.* (2023) 'Phage-display immunoprecipitation sequencing of the antibody epitope repertoire in inflammatory bowel disease reveals distinct antibody signatures', *Immunity*, 56(6), pp. 1393-1409.e6. Available at: https://doi.org/10.1016/j.immuni.2023.04.017.

Braham Informatics (2019) *FastQC: A quality control tool for high throughput sequence data.* Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed: 1 May 2023).

van den Brand, T. (2023) *ggh4x: Hacks for 'ggplot2'*. Available at: https://teunbrand.github.io/ggh4x/ (Accessed: 1 October 2023).

Breitbart, M. *et al.* (2002) 'Genomic analysis of uncultured marine viral communities', *Proceedings of the National Academy of Sciences*, 99(22), pp. 14250–14255. Available at: https://doi.org/10.1073/pnas.202488399.

Breitbart, M. *et al.* (2003) 'Metagenomic Analyses of an Uncultured Viral Community from Human Feces', *Journal of Bacteriology*, 185(20), pp. 6220–6223. Available at: https://doi.org/10.1128/JB.185.20.6220-6223.2003.

Bretherick, A.D. *et al.* (2023) 'Typing myalgic encephalomyelitis by infection at onset: A DecodeME study', *NIHR Open Research*, 3, p. 20. Available at: https://doi.org/10.3310/nihropenres.13421.4.

Briese, T. *et al.* (2023) 'A multicenter virome analysis of blood, feces, and saliva in myalgic encephalomyelitis/chronic fatigue syndrome', *Journal of Medical Virology*, 95(8), p. e28993. Available at: https://doi.org/10.1002/jmv.28993.

Broggi, A. *et al.* (2017) 'IFN-λ suppresses intestinal inflammation by non-translational regulation of neutrophil function', *Nature Immunology*, 18(10), pp. 1084–1093. Available at: https://doi.org/10.1038/ni.3821.

Brum, J.R. and Sullivan, M.B. (2015) 'Rising to the challenge: accelerated pace of discovery transforms marine virology', *Nature Reviews Microbiology*, 13(3), pp. 147–159. Available at: https://doi.org/10.1038/nrmicro3404.

Brurberg, K.G. *et al.* (2014) 'Case definitions for chronic fatigue syndrome/myalgic encephalomyelitis (CFS/ME): A systematic review', *BMJ Open*, 4(2). Available at: https://doi.org/10.1136/bmjopen-2013-003973.

Brussaard, C.P.D. (2004) 'Optimization of Procedures for Counting Viruses by Flow Cytometry', *Applied and Environmental Microbiology*, 70(3), pp. 1506–1513. Available at: https://doi.org/10.1128/AEM.70.3.1506-1513.2004.

Brussaard, C.P.D., Marie, D. and Bratbak, G. (2000) 'Flow cytometric detection of viruses', *Journal of Virological Methods*, 85(1–2), pp. 175–182. Available at: https://doi.org/10.1016/S0166-0934(99)00167-6.

Budinoff, C.R. *et al.* (2011) 'A protocol for enumeration of aquatic viruses by epifluorescence microscopy using Anodisc™ 13 membranes', *BMC Microbiology*, 11(168), pp. 1–5. Available at: https://doi.org/10.1186/1471-2180-11-168.

Burbelo, P.D., Iadarola, M.J. and Chaturvedi, A. (2019) 'Emerging technologies for the detection of viral infections', *Future Virology*, 14(1), pp. 39–49. Available at: https://doi.org/10.2217/fvl-2018-0145.

Call, L., Nayfach, S. and Kyrpides, N.C. (2021) 'Illuminating the Virosphere Through Global Metagenomics', *Annual Review of Biomedical Data Science*, 4(Volume 4, 2021), pp. 369–391. Available at: https://doi.org/10.1146/annurev-biodatasci-012221-095114.

Callanan, J. *et al.* (2020) 'Expansion of known ssRNA phage genomes: From tens to over a thousand', *Science Advances*, 6(eaay5981), pp. 1–8. Available at: https://doi.org/10.1126/sciadv.aay5981.

Callanan, J. *et al.* (2021) 'Biases in Viral Metagenomics-Based Detection, Cataloguing and Quantification of Bacteriophage Genomes in Human Faeces, a Review', *Microorganisms*, 9(3), p. 524. Available at: https://doi.org/10.3390/microorganisms9030524.

Camacho, C. *et al.* (2009) 'BLAST+: architecture and applications', *BMC Bioinformatics*, 10(1), p. 421. Available at: https://doi.org/10.1186/1471-2105-10-421.

Camargo, A.P. *et al.* (2023) 'Identification of mobile genetic elements with geNomad', *Nature Biotechnology*, pp. 1–10. Available at: https://doi.org/10.1038/s41587-023-01953-y.

Camarillo-Guerrero, L.F. *et al.* (2021) 'Massive expansion of human gut bacteriophage diversity', *Cell*, 184(4), pp. 1098-1109.e9. Available at: https://doi.org/10.1016/j.cell.2021.01.029.

Cameron, B. *et al.* (2010) 'Serological and virological investigation of the role of the herpesviruses EBV, CMV and HHV-6 in post-infective fatigue syndrome', *Journal of Medical Virology*, 82(10), pp. 1684–1688. Available at: https://doi.org/10.1002/jmv.21873.

Cammarota, G. *et al.* (2017) 'European consensus conference on faecal microbiota transplantation in clinical practice', *Gut*, 66, pp. 569–580. Available at: https://doi.org/10.1136/gutjnl-2016-313017.

Cao, Z. *et al.* (2022) 'The gut virome: A new microbiome component in health and disease', *eBioMedicine*, 81, p. 104113. Available at: https://doi.org/10.1016/j.ebiom.2022.104113.

Carlino, M. *et al.* (2020) 'Quantitative analysis and virulence phenotypes of atypical enteropathogenic Escherichia coli (EPEC) acquired from diarrheal stool samples from a Midwest US hospital', *Gut Microbes*, 12(1), p. 1824562. Available at: https://doi.org/10.1080/19490976.2020.1824562.

Carpenter, J.E., Henderson, E.P. and Grose, C. (2009) 'Enumeration of an Extremely High Particle-to-PFU Ratio for Varicella-Zoster Virus', *2Journal of Virology*, 83(13), pp. 6917–6921. Available at: https://doi.org/10.1128/JVI.00081-09.

Carruthers, B.M. *et al.* (2003) 'Myalgic Encephalomyelitis/Chronic Fatigue Syndrome', *Journal of Chronic Fatigue Syndrome*, 11(1), pp. 7–115. Available at: https://doi.org/10.1300/J092v11n01_02.

Carruthers, B.M. *et al.* (2011) 'Myalgic encephalomyelitis: International Consensus Criteria', *Journal of Internal Medicine*, 270(4), pp. 327–338. Available at: https://doi.org/10.1111/j.1365-2796.2011.02428.x.

Castro-Marrero, J. *et al.* (2019) 'Unemployment and work disability in individuals with chronic fatigue syndrome/myalgic encephalomyelitis: a community-based cross-sectional study from Spain', *BMC Public Health*, 19(1), p. 840. Available at: https://doi.org/10.1186/s12889-019-7225-z.

Chang, H. *et al.* (2023) 'Increased risk of chronic fatigue syndrome following infection: a 17-year population-based cohort study', *Journal of Translational Medicine*, 21(1), p. 804. Available at: https://doi.org/10.1186/s12967-023-04636-z.

Chapenko, S. *et al.* (2006) 'Activation of human herpesviruses 6 and 7 in patients with chronic fatigue syndrome', *Journal of Clinical Virology*, 37, pp. S47–S51. Available at: https://doi.org/10.1016/S1386-6532(06)70011-7.

Chapenko, S. *et al.* (2012) 'Association of Active Human Herpesvirus-6, -7 and Parvovirus B19 Infection with Clinical Outcomes in Patients with Myalgic Encephalomyelitis/Chronic Fatigue Syndrome', *Advances in Virology*, (205085), pp. 1–7. Available at: https://doi.org/10.1155/2012/205085.

Chehoud, C. *et al.* (2016) 'Transfer of Viral Communities between Human Individuals during Fecal Microbiota Transplantation', *mBio*, 7(2), pp. 1–8. Available at: https://doi.org/10.1128/mBio.00322-16.Editor.

Chen, F. *et al.* (2001) 'Application of Digital Image Analysis and Flow Cytometry To Enumerate Marine Viruses Stained with SYBR Gold', *Applied and Environmental Microbiology*, 67(2), pp. 539–545. Available at: https://doi.org/10.1128/AEM.67.2.539-545.2001.

Chen, G. *et al.* (2023) 'VirBot: an RNA viral contig detector for metagenomic data', *Bioinformatics*, 39(3), p. btad093. Available at: https://doi.org/10.1093/bioinformatics/btad093.

Chen, L. and Banfield, J.F. (2024) 'COBRA improves the completeness and contiguity of viral genomes assembled from metagenomes', *Nature Microbiology*, 9(3), pp. 737–750. Available at: https://doi.org/10.1038/s41564-023-01598-2.

Chen, S. *et al.* (2018) 'fastp: an ultra-fast all-in-one FASTQ preprocessor', *Bioinformatics*, 34(17), pp. i884–i890. Available at: https://doi.org/10.1093/bioinformatics/bty560.

Chen, S. (2023) 'Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp', *iMeta*, 2(2), p. e107. Available at: https://doi.org/10.1002/imt2.107.

Cheng, F. *et al.* (2021) 'Fecal microbiota transplantation for Crohn's disease: a systematic review and meta-analysis', *Techniques in Coloproctology*, 25(5), pp. 495–504. Available at: https://doi.org/10.1007/s10151-020-02395-3.

Cheung, K.S. *et al.* (2020) 'Gastrointestinal Manifestations of SARS-CoV-2 Infection and Virus Load in Fecal Samples From a Hong Kong Cohort: Systematic Review and Meta-analysis', *Gastroenterology*, 159(1), pp. 81–95. Available at: https://doi.org/10.1053/j.gastro.2020.03.065.

Chevallereau, A. *et al.* (2022) 'Interactions between bacterial and phage communities in natural environments', *Nature Reviews Microbiology*, 20(1), pp. 49–62. Available at: https://doi.org/10.1038/s41579-021-00602-y.

Chia, J. *et al.* (2010) 'Acute enterovirus infection followed by myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) and viral persistence', *Journal of Clinical Pathology*, 63(2), pp. 165–168. Available at: https://doi.org/10.1136/jcp.2009.070466.

Chia, J.K.S. and Chia, A. (2003) 'Diverse Etiologies for Chronic Fatigue Syndrome', *Clinical Infectious Diseases*, 36(5), pp. 671–672. Available at: https://doi.org/10.1086/367666.

Chia, J.K.S. and Chia, A.Y. (2008) 'Chronic fatigue syndrome is associated with chronic enterovirus infection of the stomach', *Journal of Clinical Pathology*, 61(1), pp. 43–48. Available at: https://doi.org/10.1136/jcp.2007.050054.

Choutka, J. *et al.* (2022) 'Unexplained post-acute infection syndromes', *Nature Medicine*, 28(5), pp. 911–923. Available at: https://doi.org/10.1038/s41591-022-01810-6.

Chu, L. *et al.* (2017) 'Patients diagnosed with Myalgic encephalomyelitis/chronic fatigue syndrome also fit systemic exertion intolerance disease criteria', *Fatigue: Biomedicine, Health and Behavior*, 5(2), pp. 114–128. Available at: https://doi.org/10.1080/21641846.2017.1299079.

Chu, L. *et al.* (2019) 'Onset Patterns and Course of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome', *Frontiers in Pediatrics*, 7. Available at: https://www.frontiersin.org/articles/10.3389/fped.2019.00012 (Accessed: 12 July 2023).

Cinek, O. *et al.* (2021) 'Eukaryotic viruses in the fecal virome at the onset of type 1 diabetes: A study from four geographically distant African and Asian countries', *Pediatric Diabetes*, 22(4), pp. 558–566. Available at: https://doi.org/10.1111/pedi.13207.

Clarke, M.C., Brownlie, J. and Howard, C.J. (1987) 'Isolation of cytopathic and non-cytopathic bovine viral diarrhoea virus from tissues of infected animals', *Pestivirus infections of ruminants/edited by JW Harkness* [Preprint].

Cliff, J.M. *et al.* (2019) 'Cellular immune function in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)', *Frontiers in Immunology*, 10(MAR), pp. 1–15. Available at: https://doi.org/10.3389/fimmu.2019.00796.

Clinton, N.A. *et al.* (2022) 'Crosstalk between the Intestinal Virome and Other Components of the Microbiota, and Its Effect on Intestinal Mucosal Response and Diseases', *Journal of Immunology Research*, 2022, p. e7883945. Available at: https://doi.org/10.1155/2022/7883945.

Collin, S.M. *et al.* (2016) 'Chronic fatigue syndrome (CFS) symptom-based phenotypes in two clinical cohorts of adult patients in the UK and The Netherlands', *Journal of Psychosomatic Research*, 81, pp. 14–23. Available at: https://doi.org/10.1016/j.jpsychores.2015.12.006.

Conceição-Neto, N. *et al.* (2015) 'Modular approach to customise sample preparation procedures for viral metagenomics: A reproducible protocol for virome analysis', *Scientific Reports*, 5(16532), pp. 1–14. Available at: https://doi.org/10.1038/srep16532.

Conceição-Neto, N. *et al.* (2018) 'Low eukaryotic viral richness is associated with faecal microbiota transplantation success in patients with UC', *Gut*, 67(8), pp. 1558–1559. Available at: https://doi.org/10.1136/gutjnl-2017-315281.

Cook, R. *et al.* (2021) 'INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes', *PHAGE*, 2(4), pp. 214–223. Available at: https://doi.org/10.1089/phage.2021.0007.

Cook, R. (2022) 'single_GenBank_to_vConTACT_inputs.pl'. Available at: https://github.com/RyanCook94/Random-Perl-Scripts/blob/main/single_GenBank_to_vConTACT_inputs.pl (Accessed: 6 September 2023).

Cook, R. *et al.* (2024) 'Nanopore and Illumina sequencing reveal different viral populations from human gut samples', *Microbial Genomics*, 10(4), p. 001236. Available at: https://doi.org/10.1099/mgen.0.001236.

Corbitt, M. *et al.* (2019) 'A systematic review of cytokines in chronic fatigue syndrome/myalgic encephalomyelitis/systemic exertion intolerance disease (CFS/ME/SEID)', *BMC Neurology*, 19(1), pp. 1–11. Available at: https://doi.org/10.1186/s12883-019-1433-0.

Coughlan, S. *et al.* (2021) 'The gut virome in Irritable Bowel Syndrome differs from that of controls', *Gut Microbes*, 13(1), p. 1887719. Available at: https://doi.org/10.1080/19490976.2021.1887719.

Cox, B.S. *et al.* (2022) 'EBV/HHV-6A dUTPases contribute to myalgic encephalomyelitis/chronic fatigue syndrome pathophysiology by enhancing TFH cell differentiation and extrafollicular activities', *JCI Insight*, 7(11), p. e158193. Available at: https://doi.org/10.1172/jci.insight.158193.

Dabney, J. and Meyer, M. (2012) 'Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries', *BioTechniques*, 52(2), pp. 87–94. Available at: https://doi.org/10.2144/000113809.

Dallari, S. *et al.* (2021) 'Enteric viruses evoke broad host immune responses resembling those elicited by the bacterial microbiome', *Cell Host & Microbe*, 29(6), pp. 1014-1029.e8. Available at: https://doi.org/10.1016/j.chom.2021.03.015.

Danecek, P. *et al.* (2021) 'Twelve years of SAMtools and BCFtools', *GigaScience*, 10(2), p. giab008. Available at: https://doi.org/10.1093/gigascience/giab008.

De Meirleir, K.L. *et al.* (2013) 'Plasmacytoid dendritic cells in the duodenum of individuals diagnosed with myalgic encephalomyelitis are uniquely immunoreactive to antibodies to human endogenous retroviral proteins', *In Vivo*, 27(2), pp. 177–187.

Dechaumes, A. *et al.* (2020) 'Coxsackievirus-b4 infection can induce the expression of human endogenous retrovirus w in primary cells', *Microorganisms*, 8(9), pp. 1–16. Available at: https://doi.org/10.3390/microorganisms8091335.

Di Tommaso, P. *et al.* (2017) 'Nextflow enables reproducible computational workflows', *Nature Biotechnology*, 35(4), pp. 316–319. Available at: https://doi.org/10.1038/nbt.3820.

Diard, M. *et al.* (2017) 'Inflammation boosts bacteriophage transfer between Salmonella spp.', *Science*, 355(6330), pp. 1211–1215. Available at: https://doi.org/10.1126/science.aaf8451.

Dickson, A., Toft, A. and O'Carroll, R.E. (2009) 'Neuropsychological functioning, illness perception, mood and quality of life in chronic fatigue syndrome, autoimmune thyroid disease and healthy participants', *Psychological Medicine*, 39(9), pp. 1567–1576. Available at: https://doi.org/10.1017/S0033291708004960.

Dion, M.B., Oechslin, F. and Moineau, S. (2020) 'Phage diversity, genomics and phylogeny', *Nature Reviews Microbiology*, 18(3), pp. 125–138. Available at: https://doi.org/10.1038/s41579-019-0311-5.

Djikeng, A. *et al.* (2008) 'Viral genome sequencing by random priming methods', *BMC Genomics*, 9(1), p. 5. Available at: https://doi.org/10.1186/1471-2164-9-5.

Dlusskaya, E. *et al.* (2021) 'Outer Limits of Flow Cytometry to Quantify Viruses in Water', *ACS ES&T Water*, 1(5), pp. 1127–1135. Available at: https://doi.org/10.1021/acsestwater.0c00113.

Domingues, T.D. *et al.* (2021) 'Herpesviruses Serology Distinguishes Different Subgroups of Patients From the United Kingdom Myalgic Encephalomyelitis/Chronic Fatigue Syndrome Biobank', *Frontiers in Medicine*, 8, p. 686736. Available at: https://doi.org/10.3389/fmed.2021.686736.

Doore, S.M. and Fane, B.A. (2016) 'The microviridae: Diversity, assembly, and experimental evolution', *Virology*, 491, pp. 45–55. Available at: https://doi.org/10.1016/j.virol.2016.01.020.

Dopkins, N. *et al.* (2024) 'Endogenous retroelement expression in the gut microenvironment of people living with HIV-1', *eBioMedicine*, 103, p. 105133. Available at: https://doi.org/10.1016/j.ebiom.2024.105133.

Dowsett, E.G. *et al.* (1990) 'Myalgic encephalomyelitis a persistent enteroviral infection?', *Postgraduate Medical Journal*, 66(777), pp. 526–530. Available at: https://doi.org/10.1136/pgmj.66.777.526.

Draper, L.A. *et al.* (2018) 'Long-term colonisation with donor bacteriophages following successful faecal microbial transplantation', *Microbiome*, 6(220), pp. 1–9. Available at: https://doi.org/10.1186/s40168-018-0598-x.

Du Preez, S. *et al.* (2018) 'A systematic review of enteric dysbiosis in chronic fatigue syndrome/myalgic encephalomyelitis', *Systematic Reviews*, 7(1), pp. 1–12. Available at: https://doi.org/10.1186/s13643-018-0909-0.

Duerkop, B.A. *et al.* (2018) 'Murine colitis reveals a disease-associated bacteriophage community', *Nature Microbiology*, 3(9), pp. 1023–1031. Available at: https://doi.org/10.1038/s41564-018-0210-y.

Duhaime, M.B. *et al.* (2012) 'Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method', *Environmental Microbiology*, 14(9), pp. 2526–2537. Available at: https://doi.org/10.1111/j.1462-2920.2012.02791.x.

Eaton-Fitch, N. *et al.* (2019) 'A systematic review of natural killer cells profile and cytotoxic function in myalgic encephalomyelitis/chronic fatigue syndrome', *Systematic Reviews*, 8(1). Available at: https://doi.org/10.1186/s13643-019-1202-6.

Edgar, R.C. *et al.* (2022) 'Petabase-scale sequence alignment catalyses viral discovery', *Nature*, 602(February), pp. 142–147. Available at: https://doi.org/10.1038/s41586-021-04332-2.

El-Salhy, M. *et al.* (2020) 'Efficacy of faecal microbiota transplantation for patients with irritable bowel syndrome in a randomised, double-blind, placebo-controlled study', *Gut*, 69(5), pp. 859–867. Available at: https://doi.org/10.1136/gutjnl-2019-319630.

El-Salhy, M. *et al.* (2021) 'Changes in fecal short-chain fatty acids following fecal microbiota transplantation in patients with irritable bowel syndrome', *Neurogastroenterology & Motility*, 33(2), p. e13983. Available at: https://doi.org/10.1111/nmo.13983.

Erlwein, O. *et al.* (2011) 'DNA extraction columns contaminated with murine sequences', *PLOS ONE*, 6(8), pp. 8–11. Available at: https://doi.org/10.1371/journal.pone.0023484.

Erttmann, S.F. *et al.* (2022) 'The gut microbiota prime systemic antiviral immunity via the cGAS-STING-IFN-I axis', *Immunity*, 55(5), pp. 847-861.e10. Available at: https://doi.org/10.1016/j.immuni.2022.04.006.

Estévez-López, F. *et al.* (2020) 'Systematic Review of the Epidemiological Burden of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome Across Europe: Current Evidence and EUROMENE Research Recommendations for Epidemiology', *Journal of Clinical Medicine*, 9(5), p. 1557. Available at: https://doi.org/10.3390/jcm9051557.

Ewels, P. *et al.* (2016) 'MultiQC: summarize analysis results for multiple tools and samples in a single report', *Bioinformatics*, 32(19), pp. 3047–3048. Available at: https://doi.org/10.1093/bioinformatics/btw354.

Feng, J. *et al.* (2023) 'Efficacy and safety of fecal microbiota transplantation in the treatment of ulcerative colitis: a systematic review and meta-analysis', *Scientific Reports*, 13(1), p. 14494. Available at: https://doi.org/10.1038/s41598-023-41182-6.

Flores, G.E. *et al.* (2014) 'Temporal variability is a personalized feature of the human microbiome', *Genome biology*, 15(12), p. 531. Available at: https://doi.org/10.1186/s13059-014-0531-y.

Flynn, C.E. *et al.* (2003) 'Viruses as vehicles for growth, organization and assembly of materials11The Golden Jubilee Issue—Selected topics in Materials Science and Engineering: Past, Present and Future, edited by S. Suresh.', *Acta Materialia*, 51(19), pp. 5867–5880. Available at: https://doi.org/10.1016/j.actamat.2003.08.031.

Fonseca, A. *et al.* (2024) 'IgG Antibody Responses to Epstein-Barr Virus in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: Their Effective Potential for Disease Diagnosis and Pathological Antigenic Mimicry', *Medicina*, 60(1), p. 161. Available at: https://doi.org/10.3390/medicina60010161.

Forterre, P. *et al.* (2013) 'Fake virus particles generated by fluorescence microscopy', *Trends in Microbiology*, 21(1), pp. 1–5. Available at: https://doi.org/10.1016/j.tim.2012.10.005.

Fraser, J.S. *et al.* (2006) 'Ig-Like Domains on Bacteriophages: A Tale of Promiscuity and Deceit', *Journal of Molecular Biology*, 359(2), pp. 496–507. Available at: https://doi.org/10.1016/j.jmb.2006.03.043.

Frémont, M. *et al.* (2009) 'Detection of herpesviruses and parvovirus B19 in gastric and intestinal mucosa of chronic fatigue syndrome patients', *In Vivo*, 23(2), pp. 209–214.

Frémont, M. *et al.* (2013) 'High-throughput 16S rRNA gene sequencing reveals alterations of intestinal microbiota in myalgic encephalomyelitis/chronic fatigue syndrome patients', *Anaerobe*, 22, pp. 50–56. Available at: https://doi.org/10.1016/j.anaerobe.2013.06.002.

Froussard, P. (1992) 'A random-PCR method (rPCR) to construct whole cDNA library from low amounts of RNA', *Nucleic Acids Research*, 20(11), p. 2900. Available at: https://doi.org/10.1093/nar/20.11.2900.

Fu, L. *et al.* (2012) 'CD-HIT: accelerated for clustering the next-generation sequencing data', *Bioinformatics*, 28(23), pp. 3150–3152. Available at: https://doi.org/10.1093/bioinformatics/bts565.

Fukuda, K. *et al.* (1994) 'The Chronic Fatigue Syndrome: A Comprehensive Approach to Its Definition and Study', *Annals of Internal Medicine*, 121(12), pp. 953–958. Available at: https://doi.org/10.1016/S0140-6736(63)91642-8.

Galani, I.E. *et al.* (2017) 'Interferon-λ Mediates Non-redundant Front-Line Antiviral Protection against Influenza Virus Infection without Compromising Host Fitness', *Immunity*, 46(5), pp. 875-890.e6. Available at: https://doi.org/10.1016/j.immuni.2017.04.025.

Georgiadis, M.M. *et al.* (1995) 'Mechanistic implications from the structure of a catalytic fragment of Moloney murine leukemia virus reverse transcriptase', *Structure*, 3(9), pp. 879–892. Available at: https://doi.org/10.1016/S0969-2126(01)00223-4.

Ghanem, N. *et al.* (2018) 'Changes of the Specific Infectivity of Tracer Phages during Transport in Porous Media', *Environmental Science & Technology*, 52(6), pp. 3486–3492. Available at: https://doi.org/10.1021/acs.est.7b06271.

Ghosh, T. *et al.* (2024) 'A retroviral link to vertebrate myelination through retrotransposon-RNA-mediated control of myelin gene expression', *Cell*, 187(4), pp. 814-830.e23. Available at: https://doi.org/10.1016/j.cell.2024.01.011.

Giloteaux, L. *et al.* (2016) 'Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome', *Microbiome*, 4(1), p. 30. Available at: https://doi.org/10.1186/s40168-016-0171-4.

Giloteaux, L., Hanson, M.R. and Keller, B.A. (2016) 'A pair of identical twins discordant for myalgic encephalomyelitis/chronic fatigue syndrome differ in physiological parameters and gut microbiome composition', *American Journal of Case Reports*, 17, pp. 720–729. Available at: https://doi.org/10.12659/AJCR.900314.

Gimenez-Orenga, K. *et al.* (2022) 'HERV-W ENV antigenemia and correlation of increased anti-SARS-CoV-2 immunoglobulin levels with post-COVID-19 symptoms', *Frontiers in Immunology*, 13, p. 1020064. Available at: https://doi.org/10.3389/fimmu.2022.1020064.

Glass, K.A. *et al.* (2023) 'Urine Metabolomics Exposes Anomalous Recovery after Maximal Exertion in Female ME/CFS Patients', *International Journal of Molecular Sciences*, 24(4), p. 3685. Available at: https://doi.org/10.3390/ijms24043685.

Gloor, G.B. *et al.* (2017) 'Microbiome Datasets Are Compositional: And This Is Not Optional', *Frontiers in Microbiology*, 8, p. 2224. Available at: https://doi.org/10.3389/fmicb.2017.02224.

Gogokhia, L. *et al.* (2019) 'Expansion of Bacteriophages Is Linked to Aggravated Intestinal Inflammation and Colitis', *Cell Host & Microbe*, 25(2), pp. 285-299.e8. Available at: https://doi.org/10.1016/j.chom.2019.01.008.

Goll, R. *et al.* (2020) 'Effects of fecal microbiota transplantation in subjects with irritable bowel syndrome are mirrored by changes in gut microbiome', *Gut Microbes*, 12(1). Available at: https://doi.org/10.1080/19490976.2020.1794263.

Gonzalez, J.M., Portillo, M.C. and Saiz-Jimenez, C. (2005) 'Multiple displacement amplification as a pre-polymerase chain reaction (pre-PCR) to process difficult to amplify samples and low copy number sequences from natural environments', *Environmental Microbiology*, 7(7), pp. 1024–1028. Available at: https://doi.org/10.1111/j.1462-2920.2005.00779.x.

Gozalbo-Rovira, R. *et al.* (2021) 'Microbiota Depletion Promotes Human Rotavirus Replication in an Adult Mouse Model', *Biomedicines*, 9(7), p. 846. Available at: https://doi.org/10.3390/biomedicines9070846.

Grandi, N. and Tramontano, E. (2018) 'Human endogenous retroviruses are ancient acquired elements still shaping innate immune responses', *Frontiers in Immunology*, 9(SEP), pp. 1–16. Available at: https://doi.org/10.3389/fimmu.2018.02039.

Gravelsina, S. *et al.* (2022) 'Biomarkers in the diagnostic algorithm of myalgic encephalomyelitis/chronic fatigue syndrome', *Frontiers in Immunology*, 13. Available at: https://doi.org/10.3389/fimmu.2022.928945.

Greatorex, J. *et al.* (2014) 'New methods for identifying infectious diseases', *British Medical Bulletin*, 112(1), pp. 27–35. Available at: https://doi.org/10.1093/bmb/ldu027.

Gregory, A.C. *et al.* (2019) 'Marine DNA Viral Macro- and Microdiversity from Pole to Pole', *Cell*, 177(5), pp. 1109-1123.e14. Available at: https://doi.org/10.1016/j.cell.2019.03.040.

Gregory, A.C. *et al.* (2020) 'The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut', *Cell Host & Microbe*, 28(5), pp. 1–17. Available at: https://doi.org/10.1016/j.chom.2020.08.003.

Groeger, D. *et al.* (2013) 'Bifidobacterium infantis 35624 modulates host inflammatory processes beyond the gut', *Gut Microbes*, 4(4), pp. 325–339. Available at: https://doi.org/10.4161/gmic.25487.

Gunasekera, S. *et al.* (2021) 'Evaluating coverage bias in next-generation sequencing of Escherichia coli', *PLOS ONE*, 16(6), p. e0253440. Available at: https://doi.org/10.1371/journal.pone.0253440.

Günther, O.P. *et al.* (2019) 'Immunosignature Analysis of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS)', *Molecular Neurobiology*, 56(6), pp. 4249–4257. Available at: https://doi.org/10.1007/s12035-018-1354-8.

Guo, C. *et al.* (2023) 'Deficient butyrate-producing capacity in the gut microbiome is associated with bacterial network disturbances and fatigue symptoms in ME/CFS', *Cell Host & Microbe*, 31(2), pp. 288-304.e8. Available at: https://doi.org/10.1016/j.chom.2023.01.004.

Guo, J. *et al.* (2021) 'VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses', *Microbiome*, 9(1), p. 37. Available at: https://doi.org/10.1186/s40168-020-00990-y.

Guo, Q. *et al.* (2021) 'Dynamic changes of intestinal flora in patients with irritable bowel syndrome combined with anxiety and depression after oral administration of enterobacteria capsules', *Bioengineered*, 12(2), pp. 11885–11897. Available at: https://doi.org/10.1080/21655979.2021.1999374.

Guo, W. *et al.* (2021) 'Depletion of Gut Microbiota Impairs Gut Barrier Function and Antiviral Immune Defense in the Liver', *Frontiers in Immunology*, 12, p. 636803. Available at: https://doi.org/10.3389/fimmu.2021.636803.

Gurevich, A. *et al.* (2013) 'QUAST: quality assessment tool for genome assemblies', *Bioinformatics (Oxford, England)*, 29(8), pp. 1072–1075. Available at: https://doi.org/10.1093/bioinformatics/btt086.

Halkjær, S.I. *et al.* (2018) 'Faecal microbiota transplantation alters gut microbiota in patients with irritable bowel syndrome: results from a randomised, double-blind placebo-controlled study', *Gut*, 67(12), pp. 2107–2115. Available at: https://doi.org/10.1136/gutjnl-2018-316434.

Halpin, P. *et al.* (2017) 'Myalgic encephalomyelitis/chronic fatigue syndrome and gulf war illness patients exhibit increased humoral responses to the herpesviruses-encoded dUTPase: Implications in disease pathophysiology', *Journal of Medical Virology*, 89, pp. 1636–1645. Available at: https://doi.org/10.1002/jmv.24810.

Handley, S.A. (2020) 'Virus+ Sequence Masked Human Reference Genome (hg19)'. https://zenodo.org: Zenodo. Available at: https://doi.org/10.5281/zenodo.4116107.

Handley, S.A. and Virgin, H.W. (2019) 'Drowning in Viruses', *Cell*, 177(5), pp. 1084–1085. Available at: https://doi.org/10.1016/j.cell.2019.04.045.

Hansen, K.D., Brenner, S.E. and Dudoit, S. (2010) 'Biases in Illumina transcriptome sequencing caused by random hexamer priming', *Nucleic Acids Research*, 38(12), p. e131. Available at: https://doi.org/10.1093/nar/gkq224.

Hanson, M.R. (2023) 'The viral origin of myalgic encephalomyelitis/chronic fatigue syndrome', *PLOS Pathogens*, 19(8), p. e1011523. Available at: https://doi.org/10.1371/journal.ppat.1011523.

Hanson, M.R. and Germain, A. (2020) 'Letter to the editor of metabolites', *Metabolites*, 10(5), pp. 6–7. Available at: https://doi.org/10.3390/metabo10050216.

Hara, S., Terauchi, K. and Koike, I. (1991) 'Abundance of Viruses in Marine Waters: Assessment by Epifluorescence and Transmission Electron Microscopy', *Applied and Environmental Microbiology*, 57(9), pp. 2731–2734. Available at: https://doi.org/10.1128/aem.57.9.2731-2734.1991.

Hardcastle, S.L. *et al.* (2014) 'Analysis of the Relationship between Immune Dysfunction and Symptom Severity in Patients with Chronic Fatigue Syndrome/Myalgic Encephalomyelitis (CFS/ME)', *Journal of Clinical and Cellular Immunology*, 5(1), pp. 1–9. Available at: https://doi.org/10.4172/2155-9899.1000190.

Harris, P.A. *et al.* (2009) 'Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support', *Journal of Biomedical Informatics*, 42(2), pp. 377–381. Available at: https://doi.org/10.1016/j.jbi.2008.08.010.

Harris, P.A. *et al.* (2019) 'The REDCap consortium: Building an international community of software platform partners', *Journal of Biomedical Informatics*, 95, p. 103208. Available at: https://doi.org/10.1016/j.jbi.2019.103208.

Hata, D.J., Powell, E.A. and Starolis, M.W. (2023) 'Utility and Recommendations for the Use of Multiplex Molecular Gastrointestinal Pathogen Panels', *The Journal of Applied Laboratory Medicine*, 8(6), pp. 1148–1159. Available at: https://doi.org/10.1093/jalm/jfad009.

Hatfull, G.F. and Hendrix, R.W. (2011) 'Bacteriophages and their genomes', *Current Opinion in Virology*, 1(4), pp. 298–303. Available at: https://doi.org/10.1016/j.coviro.2011.06.009.

He, G. *et al.* (2023) 'Causal Effects between Gut Microbiome and Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: A Two-Sample Mendelian Randomization Study', *Frontiers in Microbiology*, 14. Available at: https://doi.org/10.3389/fmicb.2023.1190894.

Heaney, J. *et al.* (2015) 'Low-Density TaqMan® Array Cards for the Detection of Pathogens', in A. Sails and Y.-W. Tang (eds) *Methods in Microbiology*. Academic Press, pp. 199–218. Available at: https://doi.org/10.1016/bs.mim.2015.06.002.

Heider, S. and Metzner, C. (2014) 'Quantitative real-time single particle analysis of virions', *Virology*, 462–463, pp. 199–206. Available at: https://doi.org/10.1016/j.virol.2014.06.005.

Helsedirektoratet (2014) *Nasjonal veileder Pasienter med CFS/ME: Utredning, diagnostikk, behandling, rehabilitering, pleie og omsorg*. Helsedirektoratet. Available at: https://helsedirektoratet.no/retningslinjer/nasjonal-veileder-pasienter-med-cfsme-utredning-diagnostikk-behandling-pleie-og-omsorg (Accessed: 19 April 2024).

Hernandes, R.T. *et al.* (2009) 'An overview of atypical enteropathogenic Escherichia coli', *FEMS Microbiology Letters*, 297(2), pp. 137–149. Available at: https://doi.org/10.1111/j.1574-6968.2009.01664.x.

Hernandez, R. *et al.* (2019) 'Purification and Proteomic Analysis of Alphavirus Particles from Sindbis Virus Grown in Mammalian and Insect Cells', *Bio-Protocol*, 9(10), pp. 1–33. Available at: https://doi.org/10.21769/bioprotoc.3239.

Hickie, I. *et al.* (2006) 'Post-infective and chronic fatigue syndromes precipitated by viral and non-viral pathogens: Prospective cohort study', *British Medical Journal*, 333(7568), pp. 575–578. Available at: https://doi.org/10.1136/bmj.38933.585764.AE.

Hoel, F. *et al.* (2021) 'A map of metabolic phenotypes in patients with myalgic encephalomyelitis/chronic fatigue syndrome', *JCI Insight*, 6(16). Available at: https://doi.org/10.1172/jci.insight.149217.

Holmes, G.P. *et al.* (1988) 'Chronic fatigue syndrome: A working case definition', *Annals of Internal Medicine*, 108(3), pp. 387–389. Available at: https://doi.org/10.7326/0003-4819-108-3-387.

Holmfeldt, K. *et al.* (2012) 'Cultivated single-stranded DNA phages that infect marine bacteroidetes prove difficult to detect with DNA-binding stains', *Applied and Environmental Microbiology*, 78(3), pp. 892–894. Available at: https://doi.org/10.1128/AEM.06580-11.

Holster, S. *et al.* (2019) 'The Effect of Allogenic Versus Autologous Fecal Microbiota Transfer on Symptoms, Visceral Perception and Fecal and Mucosal Microbiota in Irritable Bowel Syndrome: A Randomized Controlled Study', *Clinical and Translational Gastroenterology*, 10(4), p. e00034. Available at: https://doi.org/10.14309/ctg.0000000000000034.

Holvoet, T. *et al.* (2021) 'Fecal Microbiota Transplantation Reduces Symptoms in Some Patients With Irritable Bowel Syndrome With Predominant Abdominal Bloating: Short- and Long-term Results From a Placebo-Controlled Randomized Trial', *Gastroenterology*, 160(1), pp. 145-157.e8. Available at: https://doi.org/10.1053/j.gastro.2020.07.013.

Hornig, M. *et al.* (2015) 'Distinct plasma immune signatures in ME/CFS are present early in the course of illness', *Science Advances*, 1(1), p. e1400121. Available at: https://doi.org/10.1126/sciadv.1400121.

Howard, C.J., Brownlie, J. and Clarke, M.C. (1987) 'Comparison by the neutralisation assay of pairs of non-cytopathic and cytopathic strains of bovine virus diarrhoea virus isolated from cases of mucosal disease', *Veterinary Microbiology*, 13, pp. 361–369.

Hoyles, L. *et al.* (2014) 'Characterization of virus-like particles associated with the human faecal and caecal microbiota', *Research in Microbiologoy*, 165(10), pp. 803–812. Available at: https://doi.org/10.1016/j.resmic.2014.10.006.

Hsieh, S.-Y. *et al.* (2021) 'Comparison of PCR versus PCR-Free DNA Library Preparation for Characterising the Human Faecal Virome', *Viruses*, 13(10), p. 2093. Available at: https://doi.org/10.3390/v13102093.

Hsieh, S.-Y. *et al.* (2023) 'Investigating the Human Intestinal DNA Virome and Predicting Disease-Associated Virus–Host Interactions in Severe Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS)', *International Journal of Molecular Sciences*, 24(24), p. 17267. Available at: https://doi.org/10.3390/ijms242417267.

Huang, H.L. *et al.* (2019) 'Relief of irritable bowel syndrome by fecal microbiota transplantation is associated with changes in diversity and composition of the gut microbiota', *Journal of Digestive Diseases*, 20(8), pp. 401–408. Available at: https://doi.org/10.1111/1751-2980.12756.

Hulo, C. *et al.* (2011) 'ViralZone: a knowledge resource to understand virus diversity', *Nucleic Acids Research*, 39(suppl_1), pp. D576–D582. Available at: https://doi.org/10.1093/nar/gkq901.

d'Humières, C. *et al.* (2019) 'A simple, reproducible and cost-effective procedure to analyse gut phageome: from phage isolation to bioinformatic approach', *Scientific Reports*, 9(1), p. 11331. Available at: https://doi.org/10.1038/s41598-019-47656-w.

Hwang, J.-H. *et al.* (2023) 'Evaluation of viral infection as an etiology of ME/CFS: a systematic review and meta-analysis', *Journal of Translational Medicine*, 21(1), p. 763. Available at: https://doi.org/10.1186/s12967-023-04635-0.

Hyman, P., Trubl, G. and Abedon, S.T. (2021) 'Virus-Like Particle: Evolving Meanings in Different Disciplines', *Phage*, 2(1), pp. 11–15. Available at: https://doi.org/10.1089/phage.2020.0026.

Imdad, A. *et al.* (2023) 'Fecal transplantation for treatment of inflammatory bowel disease', *Cochrane Database of Systematic Reviews* [Preprint], (4). Available at: https://doi.org/10.1002/14651858.CD012774.pub3.

Institute of Medicine (2015) *Beyond Myalgic Encephalomyelitis/Chronic Fatigue Syndrome*, *Beyond Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: Redefining an Illness*. Washington, D.C.: National Academies Press. Available at: https://doi.org/10.17226/19012.

International Human Genome Sequencing Consortium (2001) 'Correction: Initial sequencing and analysis of the human genome', *Nature*, 412(6846), pp. 565–566. Available at: https://doi.org/10.1038/35087627.

Iwasaki, A. and Pillai, P.S. (2014) 'Innate immunity to influenza virus infection', *Nature Reviews Immunology*, 14(5), pp. 315–328. Available at: https://doi.org/10.1038/nri3665.

Jang, H.B. *et al.* (2019) 'Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks', *Nature Biotechnology*, 37(6), pp. 632–639. Available at: https://doi.org/10.1038/s41587-019-0100-8.

Janovitz, T. *et al.* (2017) 'Parvovirus B19 integration into human CD36+ erythroid progenitor cells', *Virology*, 511, pp. 40–48. Available at: https://doi.org/10.1016/j.virol.2017.08.011.

Janowski, A.B. *et al.* (2017) 'Statoviruses, A novel taxon of RNA viruses present in the gastrointestinal tracts of diverse mammals', *Virology*, 504(October 2016), pp. 36–44. Available at: https://doi.org/10.1016/j.virol.2017.01.010.

Jason, L.A. *et al.* (1999) 'A Community-Based Study of Chronic Fatigue Syndrome', *Archives of Internal Medicine*, 159(18), pp. 2129–2137. Available at: https://doi.org/10.1001/archinte.159.18.2129.

Jason, L.A. *et al.* (2005) 'Chronic Fatigue Syndrome: The Need for Subtypes', *Neuropsychology Review*, 15(1), pp. 29–58. Available at: https://doi.org/10.1007/s11065-005-3588-2.

Jason, L.A. *et al.* (2010) 'The Development of a Revised Canadian Myalgic Encephalomyelitis Chronic Fatigue Syndrome Case Definition', *American Journal of Biochemistry and Biotechnology*, 6(2), pp. 120–135. Available at: https://doi.org/10.3844/ajbbsp.2010.120.135.

Jason, L.A. *et al.* (2014) 'Examining case definition criteria for chronic fatigue syndrome and myalgic encephalomyelitis', *Fatigue: Biomedicine, Health and Behavior*, 2(1), pp. 40–56. Available at: https://doi.org/10.1080/21641846.2013.862993.

Jason, L.A. *et al.* (2015) 'Chronic fatigue syndrome and myalgic encephalomyelitis: towards an empirical case definition', *Health Psychology and Behavioral Medicine*, 3(1), pp. 82–93. Available at: https://doi.org/10.1080/21642850.2015.1014489.

Jason, L.A. *et al.* (2016) 'Are Myalgic Encephalomyelitis and chronic fatigue syndrome different illnesses? A preliminary analysis', *Journal of Health Psychology*, 21(1), pp. 3–15. Available at: https://doi.org/10.1177/1359105313520335.

Jason, L.A. *et al.* (2021) 'Risks for Developing Myalgic Encephalomyelitis/Chronic Fatigue Syndrome in College Students Following Infectious Mononucleosis: A Prospective Cohort Study', *Clinical Infectious Diseases*, 73(11), pp. e3740–e3746. Available at: https://doi.org/10.1093/cid/ciaa1886.

Jason, L.A. *et al.* (2022) 'Predictors for Developing Severe Myalgic Encephalomyelitis/Chronic Fatigue Syndrome Following Infectious Mononucleosis', *Journal of rehabilitation therapy*, 4(1), pp. 1–5. Available at: https://doi.org/10.29245/2767-5122/2021/1.1129.

Jason, L.A. and Dorri, J.A. (2023) 'ME/CFS and Post-Exertional Malaise among Patients with Long COVID', *Neurology International*, 15(1), pp. 1–11. Available at: https://doi.org/10.3390/neurolint15010001.

Jason, L.A. and Johnson, M. (2020) 'Solving the ME/CFS criteria and name conundrum: the aftermath of IOM', *Fatigue: Biomedicine, Health and Behavior*, 0(0), pp. 1–11. Available at: https://doi.org/10.1080/21641846.2020.1757809.

Jason, L.A., Yoo, S. and Bhatia, S. (2022) 'Patient perceptions of infectious illnesses preceding Myalgic Encephalomyelitis/Chronic Fatigue Syndrome', *Chronic Illness*, 18(4), pp. 901–910. Available at: https://doi.org/10.1177/17423953211043106.

Jerse, A.E. *et al.* (1990) 'A genetic locus of enteropathogenic Escherichia coli necessary for the production of attaching and effacing lesions on tissue culture cells.', *Proceedings of the National Academy of Sciences*, 87(20), pp. 7839–7843. Available at: https://doi.org/10.1073/pnas.87.20.7839.

Johansen, J. *et al.* (2022) 'Genome binning of viral entities from bulk metagenomics data', *Nature Communications*, 13(1), p. 965. Available at: https://doi.org/10.1038/s41467-022-28581-5.

Johnsen, P.H. *et al.* (2018) 'Faecal microbiota transplantation versus placebo for moderate-to-severe irritable bowel syndrome: a double-blind, randomised, placebo-controlled, parallel-group, single-centre trial', *The Lancet Gastroenterology and Hepatology*, 3(1), pp. 17–24. Available at: https://doi.org/10.1016/S2468-1253(17)30338-2.

Johnston, S. *et al.* (2013) 'The prevalence of chronic fatigue syndrome/ myalgic encephalomyelitis: A meta-analysis', *Clinical Epidemiology*, 5(1), pp. 105–110.

Johnston, S.C., Staines, D.R. and Marshall-Gradisnik, S.M. (2016) 'Epidemiological characteristics of chronic fatigue syndrome/myalgic encephalomyelitis in Australian patients', *Clinical Epidemiology*, 8, pp. 97–107. Available at: https://doi.org/10.2147/CLEP.S96797.

Kaletta, J. *et al.* (2020) 'A rigorous assessment and comparison of enumeration methods for environmental viruses', *Scientific Reports*, 10(1), pp. 1–12. Available at: https://doi.org/10.1038/s41598-020-75490-y.

Kallies, R. *et al.* (2019) 'Evaluation of Sequencing Library Preparation Protocols for Viral Metagenomic Analysis from Pristine Aquifer Groundwaters', *Viruses*, 11(6), p. 484. Available at: https://doi.org/10.3390/v11060484.

Kang, D.D. *et al.* (2015) 'MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities', *PeerJ*, 3, p. e1165. Available at: https://doi.org/10.7717/peerj.1165.

Kang, D.D. *et al.* (2019) 'MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies', *PeerJ*, 7, p. e7359. Available at: https://doi.org/10.7717/peerj.7359.

Kang, D.W. *et al.* (2017) 'Microbiota Transfer Therapy alters gut ecosystem and improves gastrointestinal and autism symptoms: An open-label study', *Microbiome*, 5(1), pp. 1–16. Available at: https://doi.org/10.1186/s40168-016-0225-7.

Karlsson, O.E., Belák, S. and Granberg, F. (2013) 'The Effect of Preprocessing by Sequence-Independent, Single-Primer Amplification (SISPA) on Metagenomic Detection of Viruses', *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, 11(S1), pp. S227–S234. Available at: https://doi.org/10.1089/bsp.2013.0008.

Kasimir, F. *et al.* (2022) 'Tissue specific signature of HHV-6 infection in ME/CFS', *Frontiers in Molecular Biosciences*, 9. Available at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.1044964.

Kassambara, A. (2023) *ggpubr*, *ggpubr: 'ggplot2' Based Publication Ready Plots*. Available at: https://rpkgs.datanovia.com/ggpubr/. (Accessed: 1 September 2023).

Kassiotis, G. and Stoye, J.P. (2016) 'Immune responses to endogenous retroelements: taking the bad with the good', *Nature Reviews Immunology*, 16(4), pp. 207–219. Available at: https://doi.org/10.1038/nri.2016.27.

Kato, Y.H. *et al.* (2009) 'No apparent difference in the prevalence of parvovirus B19 infection between chronic fatigue syndrome patients and healthy controls in Japan', *Journal of Clinical Virology*, 44(3), pp. 246–247. Available at: https://doi.org/10.1016/j.jcv.2009.01.001.

Katz, B.Z. *et al.* (2009) 'Chronic fatigue syndrome after infectious mononucleosis in adolescents', *Pediatrics*, 124(1), pp. 189–193. Available at: https://doi.org/10.1542/peds.2008-1879.

Kenyon, J.N., Coe, S. and Izadi, H. (2019) 'A retrospective outcome study of 42 patients with Chronic Fatigue Syndrome, 30 of whom had Irritable Bowel Syndrome. Half were treated with oral approaches, and half were treated with Faecal Microbiome Transplantation', *Human Microbiome Journal*, 13, p. 100061. Available at: https://doi.org/10.1016/j.humic.2019.100061.

Kernbauer, E., Ding, Y. and Cadwell, K. (2014) 'An enteric virus can replace the beneficial function of commensal bacteria', *Nature*, 516(7529), pp. 94–98. Available at: https://doi.org/10.1038/nature13960.

Kerr, J.R. *et al.* (2002) 'Chronic fatigue syndrome and arthralgia following parvovirus B19 infection.', *The Journal of Rheumatology*, 29(3), pp. 595–602.

Kerr, J.R. (2005) 'Pathogenesis of Parvovirus B19 Infection: Host Gene Variability, and Possible Means and Effects of Virus Persistence', *Journal of Veterinary Medicine, Series B*, 52(7–8), pp. 335–339. Available at: https://doi.org/10.1111/j.1439-0450.2005.00859.x.

Kerr, J.R. *et al.* (2008) 'Gene expression subtypes in patients with chronic fatigue syndrome/myalgic encephalomyelitis', *Journal of Infectious Diseases*, 197(8), pp. 1171–1184. Available at: https://doi.org/10.1086/533453.

Kerr, J.R. *et al.* (2010) 'Antibody to parvovirus B19 nonstructural protein is associated with chronic arthralgia in patients with chronic fatigue syndrome/myalgic encephalomyelitis', *Journal of General Virology*, 91(4), pp. 893–897. Available at: https://doi.org/10.1099/vir.0.017590-0.

Kerr, J.R. and Mattey, D.L. (2008) 'Preexisting psychological stress predicts acute and chronic fatigue and arthritis following symptomatic parvovirus B19 infection', *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 46(9), pp. e83-87. Available at: https://doi.org/10.1086/533471.

Kieft, K. *et al.* (2022) 'vRhyme enables binning of viral genomes from metagenomes', *Nucleic Acids Research*, 50(14), p. e83. Available at: https://doi.org/10.1093/nar/gkac341.

Kieft, K., Zhou, Z. and Anantharaman, K. (2020) 'VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences', *Microbiome*, 8(1), p. 90. Available at: https://doi.org/10.1186/s40168-020-00867-0.

Kim, K.-H. and Bae, J.-W. (2011) 'Amplification Methods Bias Metagenomic Libraries of Uncultured Single-Stranded and Double-Stranded DNA Viruses', *Applied and Environmental Microbiology*, 77(21), pp. 7663–7668. Available at: https://doi.org/10.1128/AEM.00289-11.

Kim, M.-S. *et al.* (2011) 'Diversity and Abundance of Single-Stranded DNA Viruses in Human Feces', *Applied and Environmental Microbiology*, 77(22), pp. 8062–8070. Available at: https://doi.org/10.1128/AEM.06331-11.

Kirchberger, P.C., Martinez, Z.A. and Ochman, H. (2022) 'Organizing the Global Diversity of Microviruses', *mBio*, 13(3), pp. e00588-22. Available at: https://doi.org/10.1128/mbio.00588-22.

Kitami, T. *et al.* (2020) 'Deep phenotyping of myalgic encephalomyelitis/chronic fatigue syndrome in Japanese population', *Scientific Reports*, 10(1), p. 19933. Available at: https://doi.org/10.1038/s41598-020-77105-y.

Klasse, P.J. (2015) 'Molecular Determinants of the Ratio of Inert to Infectious Virus Particles', in *Progress in Molecular Biology and Translational Science*. Elsevier Inc., pp. 285–326. Available at: https://doi.org/10.1016/bs.pmbts.2014.10.012.

Kleiner, M., Hooper, L.V. and Duerkop, B.A. (2015) 'Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes', *BMC Genomics*, 16(1), p. 7. Available at: https://doi.org/10.1186/s12864-014-1207-4.

Kluyver, T. *et al.* (2016) 'Jupyter Notebooks – a publishing format for reproducible computational workflows', in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. IOS Press, pp. 87–90. Available at: https://doi.org/10.3233/978-1-61499-649-1-87.

Knezevic, P., Adriaenssens, E.M. and ICTV Report Consortium (2021) 'ICTV Virus Taxonomy Profile: Inoviridae', *Journal of General Virology*, 102(7), p. 001614. Available at: https://doi.org/10.1099/jgv.0.001614.

Knight, R. *et al.* (2018) 'Best practices for analysing microbiomes', *Nature Reviews Microbiology*, 16, pp. 410–422. Available at: https://doi.org/10.1038/s41579-018-0029-9.

Kodani, M. *et al.* (2011) 'Application of TaqMan low-density arrays for simultaneous detection of multiple respiratory pathogens', *Journal of Clinical Microbiology*, 49(6), pp. 2175–2182. Available at: https://doi.org/10.1128/JCM.02270-10.

Koelle, D.M. *et al.* (2002) 'Markers of Viral Infection in Monozygotic Twins Discordant for Chronic Fatigue Syndrome', *Clinical Infectious Diseases*, 35(5), pp. 518–525. Available at: https://doi.org/10.1086/341774.

Kogelnik, A.M. *et al.* (2006) 'Use of valganciclovir in patients with elevated antibody titers against Human Herpesvirus-6 (HHV-6) and Epstein—Barr Virus (EBV) who were experiencing central nervous system dysfunction including long-standing fatigue', *Journal of Clinical Virology*, 37, pp. S33–S38. Available at: https://doi.org/10.1016/S1386-6532(06)70009-9.

Komaroff, A.L. and Bateman, L. (2021) 'Will COVID-19 Lead to Myalgic Encephalomyelitis/Chronic Fatigue Syndrome?', *Frontiers in Medicine*, 7. Available at: https://doi.org/10.3389/fmed.2020.606824.

Komaroff, A.L. and Lipkin, W.I. (2023) 'ME/CFS and Long COVID share similar symptoms and biological abnormalities: road map to the literature', *Frontiers in Medicine*, 10. Available at: https://doi.org/10.3389/fmed.2023.1187163.

König, R.S. *et al.* (2022) 'The Gut Microbiome in Myalgic Encephalomyelitis (ME)/Chronic Fatigue Syndrome (CFS)', *Frontiers in Immunology*, 12. Available at: https://www.frontiersin.org/articles/10.3389/fimmu.2021.628741 (Accessed: 10 July 2023).

Koonin, E.V., Krupovic, M. and Agol, V.I. (2021) 'The Baltimore Classification of Viruses 50 Years Later: How Does It Stand in the Light of Virus Evolution?', *Microbiology and Molecular Biology Reviews*, 85(3). Available at: https://doi.org/10.1128/mmbr.00053-21.

Körner, E. and Lorentz, A. (2023) 'Fecal microbiota transplantation in patients with irritable bowel syndrome: an overview of current studies', *Journal of Applied Microbiology*, 134(3), p. lxad044. Available at: https://doi.org/10.1093/jambio/lxad044.

Kotenko, S.V. *et al.* (2003) 'IFN-λs mediate antiviral protection through a distinct class II cytokine receptor complex', *Nature Immunology*, 4(1), pp. 69–77. Available at: https://doi.org/10.1038/ni875.

Kramberger, P. *et al.* (2012) 'Evaluation of nanoparticle tracking analysis for total virus particle determination', *Virology Journal*, 9(1), p. 265. Available at: https://doi.org/10.1186/1743-422X-9-265.

Kramná, L. and Cinek, O. (2018) 'Virome Sequencing of Stool Samples', in A. Moya and V. Pérez Brocal (eds) *The Human Virome*. New York, NY: Springer New York (Methods in Molecular Biology), pp. 59–83. Available at: https://doi.org/10.1007/978-1-4939-8682-8_6.

Krupp, L.B. *et al.* (1989) 'The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus', *Archives of Neurology*, 46(10), pp. 1121–1123. Available at: https://doi.org/10.1001/archneur.1989.00520460115022.

Kutter, E. (2009) 'Phage Host Range and Efficiency of Plating', in M.R.J. Clokie and A.M. Kropinski (eds) *Bacteriophages: Methods and Protocols, Volume 1: Isolation, Characterization, and Interactions*. Totowa, NJ: Humana Press, pp. 141–149. Available at: https://doi.org/10.1007/978-1-60327-164-6_14.

Lacasa, M. *et al.* (2023) 'Unsupervised Cluster Analysis Reveals Distinct Subtypes of ME/CFS Patients Based on Peak Oxygen Consumption and SF-36 Scores', *Clinical Therapeutics*, 45(12), pp. 1228–1235. Available at: https://doi.org/10.1016/j.clinthera.2023.09.007.

Lahtinen, P. *et al.* (2020) 'Randomised clinical trial: faecal microbiota transplantation versus autologous placebo administered via colonoscopy in irritable bowel syndrome', *Alimentary Pharmacology & Therapeutics*, 51(12), pp. 1321–1331. Available at: https://doi.org/10.1111/apt.15740.

Lam, S. *et al.* (2022) 'Roles of the gut virome and mycobiome in faecal microbiota transplantation', *The Lancet Gastroenterology & Hepatology*, 7(5), pp. 472–484. Available at: https://doi.org/10.1016/S2468-1253(21)00303-4.

Langmead, B. *et al.* (2019) 'Scaling read aligners to hundreds of threads on general-purpose processors', *Bioinformatics*, 35(3), pp. 421–432. Available at: https://doi.org/10.1093/bioinformatics/bty648.

Langmead, B. (2020) 'Human / GRCh38 no-alt analysis set'. Available at: https://genome-idx.s3.amazonaws.com/bt/GRCh38_noalt_as.zip (Accessed: 23 February 2023).

Langmead, B. and Salzberg, S.L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods*, 9(4), pp. 357–359. Available at: https://doi.org/10.1038/nmeth.1923.

Lavialle, C. *et al.* (2013) 'Paleovirology of "syncytins", retroviral env genes exapted for a role in placentation', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1626), p. 20120507. Available at: https://doi.org/10.1098/rstb.2012.0507.

Lee, J.-S. *et al.* (2021) 'Salivary DNA Loads for Human Herpesviruses 6 and 7 Are Correlated With Disease Phenotype in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome', *Frontiers in Medicine*, 8. Available at: https://doi.org/10.3389/fmed.2021.656692.

Legoff, J. *et al.* (2017) 'The eukaryotic gut virome in hematopoietic stem cell transplantation: New clues in enteric graft-versus-host disease', *Nature Medicine*, 23(9), pp. 1080–1085. Available at: https://doi.org/10.1038/nm.4380.

Lepage, P. *et al.* (2008) 'Dysbiosis in inflammatory bowel disease: A role for bacteriophages?', *Gut*, 57(3), pp. 424–425. Available at: https://doi.org/10.1136/gut.2007.134668.

Lerner, A.M. *et al.* (2007) 'Valacyclovir Treatment in Epstein-Barr Virus Subset Chronic Fatigue Syndrome: Thirty-six Months Follow-up', *In Vivo*, 21(5), pp. 707–713.

Lerner, A.M. *et al.* (2012) 'Antibody to Epstein-Barr Virus Deoxyuridine Triphosphate Nucleotidohydrolase and Deoxyribonucleotide Polymerase in a Chronic Fatigue Syndrome Subset', *PLOS ONE*, 7(11), p. e47891. Available at: https://doi.org/10.1371/journal.pone.0047891.

Letellier, L. *et al.* (2004) 'Main features on tailed phage, host recognition and DNA uptake', *Front Biosci*, 9(1), pp. 1228–1339.

Levine, M.M. and Robins-Browne, R.M. (2012) 'Factors That Explain Excretion of Enteric Pathogens by Persons Without Diarrhea', *Clinical Infectious Diseases*, 55(suppl_4), pp. S303–S311. Available at: https://doi.org/10.1093/cid/cis789.

Li, D. *et al.* (2015) 'MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph', *Bioinformatics*, 31(10), pp. 1674–1676. Available at: https://doi.org/10.1093/bioinformatics/btv033.

Li, D. *et al.* (2016) 'MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices', *Methods*, 102, pp. 3–11. Available at: https://doi.org/10.1016/j.ymeth.2016.02.020.

Li, J. *et al.* (2022) 'Advances and challenges in cataloging the human gut virome', *Cell Host & Microbe*, 30(7), pp. 908–916. Available at: https://doi.org/10.1016/j.chom.2022.06.003.

Li, L. *et al.* (2015) 'Comparing viral metagenomics methods using a highly multiplexed human viral pathogens reagent', *Journal of Virological Methods*, 213, pp. 139–146. Available at: https://doi.org/10.1016/j.jviromet.2014.12.002.

Li, W. and Godzik, A. (2006) 'Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences', *Bioinformatics*, 22(13), pp. 1658–1659. Available at: https://doi.org/10.1093/bioinformatics/btl158.

Li, Y., Handley, S.A. and Baldridge, M.T. (2021) 'The dark side of the gut: Virome–host interactions in intestinal homeostasis and disease', *Journal of Experimental Medicine*, 218(5), p. e20201044. Available at: https://doi.org/10.1084/jem.20201044.

Liang, G. and Bushman, F.D. (2021) 'The human virome: assembly, composition and host interactions', *Nature Reviews Microbiology*, 19(8), pp. 514–527. Available at: https://doi.org/10.1038/s41579-021-00536-5.

Lim, E.J. *et al.* (2020) 'Systematic review and meta-analysis of the prevalence of chronic fatigue syndrome/myalgic encephalomyelitis (CFS/ME)', *Journal of Translational Medicine*, 18(1), pp. 1–15. Available at: https://doi.org/10.1186/s12967-020-02269-0.

Lim, E.-J. and Son, C.-G. (2020) 'Review of case definitions for myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)', *Journal of Translational Medicine*, 18(1), p. 289. Available at: https://doi.org/10.1186/s12967-020-02455-0.

Lim, E.S. *et al.* (2015) 'Early life dynamics of the human gut virome and bacterial microbiome in infants', *Nature Medicine*, 21(10), pp. 1228–1234. Available at: https://doi.org/10.1038/nm.3950.

Lima-Junior, D.S. *et al.* (2021) 'Endogenous retroviruses promote homeostatic and inflammatory responses to the microbiota', *Cell*, 184(14), pp. 3794-3811.e19. Available at: https://doi.org/10.1016/j.cell.2021.05.020.

Liu, F. and Zhou, Z.H. (2007) 'Comparative virion structures of human herpesviruses', in A. Arvin et al. (eds) *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis*. Cambridge: Cambridge University Press. Available at: http://www.ncbi.nlm.nih.gov/books/NBK47399/ (Accessed: 11 December 2023).

Liu, J. *et al.* (2013) 'A laboratory-developed taqman array card for simultaneous detection of 19 enteropathogens', *Journal of Clinical Microbiology*, 51(2), pp. 472–480. Available at: https://doi.org/10.1128/JCM.02658-12.

Liu, L. *et al.* (2019) 'Commensal viruses maintain intestinal intraepithelial lymphocytes via noncanonical RIG-I signaling', *Nature Immunology*, 20(12), pp. 1681–1691. Available at: https://doi.org/10.1038/s41590-019-0513-z.

Liu, Q., Mak, J.W.Y., *et al.* (2022) 'Gut microbiota dynamics in a prospective cohort of patients with post-acute COVID-19 syndrome', *Gut*, 71(3), pp. 544–552. Available at: https://doi.org/10.1136/gutjnl-2021-325989.

Liu, Q., Zuo, T., *et al.* (2022) 'Longitudinal Evaluation of Gut Bacteriomes and Viromes after Fecal Microbiota Transplantation for Eradication of Carbapenem-Resistant Enterobacteriaceae', *mSystems*, 7(3), pp. e01510-21. Available at: https://doi.org/10.1128/msystems.01510-21.

Liu, X. *et al.* (2023) 'Resurrection of endogenous retroviruses during aging reinforces senescence', *Cell*, 186(2), pp. 287-304.e26. Available at: https://doi.org/10.1016/j.cell.2022.12.017.

Loebel, M. *et al.* (2017) 'Serological profiling of the EBV immune response in Chronic Fatigue Syndrome using a peptide microarray', *PLOS ONE*, 12(6), pp. 1–17. Available at: https://doi.org/10.1371/journal.pone.0179124.

Lu, J. *et al.* (2022) 'Metagenome analysis using the Kraken software suite', *Nature Protocols*, 17(12), pp. 2815–2839. Available at: https://doi.org/10.1038/s41596-022-00738-y.

Lupo, G.F.D. *et al.* (2021) 'Potential role of microbiome in Chronic Fatigue Syndrome/Myalgic Encephalomyelits (CFS/ME)', *Scientific Reports*, 11(1), pp. 1–18. Available at: https://doi.org/10.1038/s41598-021-86425-6.

Macchietto, M.G., Langlois, R.A. and Shen, S.S. (2020) 'Virus-induced transposable element expression up-regulation in human and mouse host cells', *Life Science Alliance*, 3(2), pp. 1–14. Available at: https://doi.org/10.26508/lsa.201900536.

Maes, M. *et al.* (2012) 'Evidence for inflammation and activation of cell-mediated immunity in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS): Increased interleukin-1, tumor necrosis factor-α, PMN-elastase, lysozyme and neopterin', *Journal of Affective Disorders*, 136(3), pp. 933–939. Available at: https://doi.org/10.1016/j.jad.2011.09.004.

Maes, M. *et al.* (2014) 'Evidence for the existence of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) with and without abdominal discomfort (irritable bowel) syndrome', *Neuroendocrinology Letters*, 35(6), pp. 445–453.

Maes, M. (2015) 'A new case definition of Neuro-Inflammatory and Oxidative Fatigue (NIOF), a neuroprogressive disorder, formerly known as chronic fatigue syndrome or Myalgic Encephalomyelitis: results of multivariate pattern recognition methods and external validation by neuro-immune biomarkers', *Neuroendocrinology Letters*, 36(4), pp. 320–329.

Maes, M. and Leunis, J.C. (2008) 'Normalization of leaky gut in chronic fatigue syndrome (CFS) is accompanied by a clinical improvement: Effects of age, duration of illness and the translocation of LPS from gram-negative bacteria', *Neuroendocrinology Letters*, 29(6), pp. 902–910.

Maes, M., Mihaylova, I. and Leunis, J.C. (2007) 'Increased serum IgA and IgM against LPS of enterobacteria in chronic fatigue syndrome (CFS): Indication for the involvement of gram-negative enterobacteria in the etiology of CFS and for the presence of an increased gut-intestinal permeability', *Journal of Affective Disorders*, 99(1–3), pp. 237–240. Available at: https://doi.org/10.1016/j.jad.2006.08.021.

Maes, M., Twisk, F.N.M. and Johnson, C. (2012) 'Myalgic Encephalomyelitis (ME), Chronic Fatigue Syndrome (CFS), and Chronic Fatigue (CF) are distinguished accurately: Results of supervised learning techniques applied on clinical and inflammatory data', *Psychiatry Research*, 200(2), pp. 754–760. Available at: https://doi.org/10.1016/j.psychres.2012.03.031.

Magnus, P. *et al.* (2015) 'Chronic fatigue syndrome/myalgic encephalomyelitis (CFS/ME) is associated with pandemic influenza infection, but not with an adjuvanted pandemic influenza vaccine', *Vaccine*, 33(46), pp. 6173–6177. Available at: https://doi.org/10.1016/j.vaccine.2015.10.018.

Majewska, J. *et al.* (2019) 'Induction of Phage-Specific Antibodies by Two Therapeutic Staphylococcal Bacteriophages Administered per os', *Frontiers in Immunology*, 10. Available at: https://doi.org/10.3389/fimmu.2019.02607.

Mandarano, A.H. *et al.* (2018) 'Eukaryotes in the gut microbiota in myalgic encephalomyelitis/chronic fatigue syndrome', *PeerJ*, 6, p. e4282. Available at: https://doi.org/10.7717/peerj.4282.

Mare, A.D. *et al.* (2021) 'Enteropathogenic Escherichia coli—A Summary of the Literature', *Gastroenterology Insights*, 12(1), pp. 28–40. Available at: https://doi.org/10.3390/gastroent12010004.

Marine, R. *et al.* (2014) 'Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome', *Microbiome*, 2(1), p. 3. Available at: https://doi.org/10.1186/2049-2618-2-3.

Martín, F. *et al.* (2023) 'Increased gut permeability and bacterial translocation are associated with fibromyalgia and myalgic encephalomyelitis/chronic fatigue syndrome: implications for disease-related biomarker discovery', *Frontiers in Immunology*, 14, p. 1253121. Available at: https://doi.org/10.3389/fimmu.2023.1253121.

Martín, R. *et al.* (2023) 'Faecalibacterium: a bacterial genus with promising human health applications', *FEMS Microbiology Reviews*, 47(4), p. fuad039. Available at: https://doi.org/10.1093/femsre/fuad039.

McCormick, W. and Mermel, L.A. (2021) 'The basic reproductive number and particle-to-plaque ratio: comparison of these two parameters of viral infectivity', *Virology Journal*, 18(1), p. 92. Available at: https://doi.org/10.1186/s12985-021-01566-4.

McMurdie, P.J. and Holmes, S. (2013) 'phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data', *PLOS ONE*, 8(4), p. e61217. Available at: https://doi.org/10.1371/journal.pone.0061217.

McNab, F. *et al.* (2015) 'Type I interferons in infectious disease', *Nature Reviews Immunology*, 15(2), pp. 87–103. Available at: https://doi.org/10.1038/nri3787.

Metzger, R.N., Krug, A.B. and Eisenächer, K. (2018) 'Enteric virome sensing—its role in intestinal homeostasis and immunity', *Viruses*, 10(4). Available at: https://doi.org/10.3390/v10040146.

Mirdita, M. *et al.* (2021) 'Fast and sensitive taxonomic assignment to metagenomic contigs', *Bioinformatics*, 37(18), pp. 3029–3031. Available at: https://doi.org/10.1093/bioinformatics/btab184.

Mirdita, M., Steinegger, M. and Söding, J. (2019) 'MMseqs2 desktop and local web server app for fast, interactive sequence searches', *Bioinformatics*, 35(16), pp. 2856–2858. Available at: https://doi.org/10.1093/bioinformatics/bty1057.

Mirin, A.A., Dimmock, M.E. and Jason, L.A. (2022) 'Updated ME/CFS prevalence estimates reflecting post-COVID increases and associated economic costs and funding implications', *Fatigue: Biomedicine, Health & Behavior* [Preprint]. Available at: https://www.tandfonline.com/doi/abs/10.1080/21641846.2022.2062169 (Accessed: 3 April 2024).

Mirzaei, M.K. *et al.* (2021) 'Challenges of Studying the Human Virome – Relevant Emerging Technologies', *Trends in Microbiology*, 29(2), pp. 171–181. Available at: https://doi.org/10.1016/j.tim.2020.05.021.

Mirzaei, M.K. and Maurice, C.F. (2017) 'Ménage à trois in the human gut: Interactions between host, bacteria and phages', *Nature Reviews Microbiology*, 15(7), pp. 397–408. Available at: https://doi.org/10.1038/nrmicro.2017.30.

Missailidis, D., Annesley, S.J. and Fisher, P.R. (2019) 'Pathological mechanisms underlying myalgic encephalomyelitis/chronic fatigue syndrome', *Diagnostics*, 9(3). Available at: https://doi.org/10.3390/diagnostics9030080.

Mohamed, A.Z. *et al.* (2023) 'Objective sleep measures in chronic fatigue syndrome patients: A systematic review and meta-analysis', *Sleep Medicine Reviews*, 69, p. 101771. Available at: https://doi.org/10.1016/j.smrv.2023.101771.

Montoya, J.G. *et al.* (2013) 'Randomized clinical trial to evaluate the efficacy and safety of valganciclovir in a subset of patients with chronic fatigue syndrome', *Journal of Medical Virology*, 85(12), pp. 2101–2109. Available at: https://doi.org/10.1002/jmv.23713.

Morris, G. *et al.* (2019) 'Myalgic encephalomyelitis or chronic fatigue syndrome: how could the illness develop?', *Metabolic Brain Disease*, 34(2), pp. 385–415. Available at: https://doi.org/10.1007/s11011-019-0388-6.

Mosby, C.A. *et al.* (2022) 'Interaction with mammalian enteric viruses alters outer membrane vesicle production and content by commensal bacteria', *Journal of extracellular vesicles*, 11(1), p. e12172. Available at: https://doi.org/10.1002/jev2.12172.

Mozhgani, S.-H. *et al.* (2021) 'Human Herpesvirus 6 Infection and Risk of Chronic Fatigue Syndrome: A Systematic Review and Meta-Analysis', *Intervirology*, 65(1), pp. 49–57. Available at: https://doi.org/10.1159/000517930.

Nacul, L. *et al.* (2020) 'How Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) Progresses: The Natural History of ME/CFS', *Frontiers in Neurology*, 11. Available at: https://doi.org/10.3389/fneur.2020.00826.

Nacul, L.C. *et al.* (2011) 'Prevalence of myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) in three regions of England: A repeated cross-sectional study in primary care', *BMC Medicine*, 9, pp. 1–12. Available at: https://doi.org/10.1186/1741-7015-9-91.

Naess, H. *et al.* (2010) 'Postinfectious and Chronic Fatigue Syndromes: Clinical Experience from a Tertiary-referral Centre in Norway', *In Vivo*, 24(2), pp. 185–188.

Naess, H. *et al.* (2012) 'Chronic fatigue syndrome after Giardia enteritis: clinical characteristics, disability and long-term sickness absence', *BMC Gastroenterology*, 12, p. 13. Available at: https://doi.org/10.1186/1471-230X-12-13.

Nagy-Szakal, D. *et al.* (2017) 'Fecal metagenomic profiles in subgroups of patients with myalgic encephalomyelitis/chronic fatigue syndrome', *Microbiome*, 5(1), p. 44. Available at: https://doi.org/10.1186/S40168-017-0261-Y.

Navaneetharaja, N. *et al.* (2016) 'A Role for the Intestinal Microbiota and Virome in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS)?', *Journal of Clinical Medicine*, 5(6), p. 55. Available at: https://doi.org/10.3390/jcm5060055.

Naviaux, R.K. *et al.* (2016) 'Metabolic features of chronic fatigue syndrome', *Proceedings of the National Academy of Sciences of the United States of America*, 113(37), pp. E5472–E5480. Available at: https://doi.org/10.1073/pnas.1607571113.

Nayfach, S., Camargo, A.P., *et al.* (2021) 'CheckV assesses the quality and completeness of metagenome-assembled viral genomes', *Nature Biotechnology*, 39(5), pp. 578–585. Available at: https://doi.org/10.1038/s41587-020-00774-7.

Nayfach, S., Páez-Espino, D., *et al.* (2021) 'Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome', *Nature Microbiology*, 6(7), pp. 960–970. Available at: https://doi.org/10.1038/s41564-021-00928-6.

Nayfach, S. (2023) 'uhgv-tools'. Available at: https://github.com/snayfach/UHGV/blob/main/CLASSIFY.md (Accessed: 16 March 2024).

Neri, U. *et al.* (2022) 'Expansion of the global RNA virome reveals diverse clades of bacteriophages', *Cell*, 185(21), pp. 4023-4037.e18. Available at: https://doi.org/10.1016/j.cell.2022.08.023.

Neurath, M.F., Uberla, K. and Ng, S.C. (2021) 'Gut as viral reservoir: Lessons from gut viromes, HIV and COVID-19', *Gut*, 70(9), pp. 1605–1608. Available at: https://doi.org/10.1136/gutjnl-2021-324622.

Newberry, F. *et al.* (2018) 'Does the microbiome and virome contribute to myalgic encephalomyelitis/chronic fatigue syndrome?', *Clinical Science*, 132(5), pp. 523–542. Available at: https://doi.org/10.1042/CS20171330.

NICE (2021) *Myalgic encephalomyelitis (or encephalopathy)/chronic fatigue syndrome: diagnosis and management*. London: National Institute for Health and Care Excellence (NICE) (National Institute for Health and Care Excellence: Guidelines). Available at: http://www.ncbi.nlm.nih.gov/books/NBK579533/ (Accessed: 26 May 2024).

Noble, R.T. and Fuhrman, J.A. (1998) 'Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria', *Aquatic Microbial Ecology*, 14(2), pp. 113–118. Available at: https://doi.org/10.3354/ame014113.

Nunes, J.M., Kell, D.B. and Pretorius, E. (2024) 'Herpesvirus Infection of Endothelial Cells as a Systemic Pathological Axis in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome', *Viruses*, 16(4), p. 572. Available at: https://doi.org/10.3390/v16040572.

Nurk, S. *et al.* (2017) 'metaSPAdes: a new versatile metagenomic assembler', *Genome Research*, 27(5), pp. 824–834. Available at: https://doi.org/10.1101/gr.213959.116.

Oakes, B. *et al.* (2013) 'Human endogenous retrovirus-K18 superantigen expression and human herpesvirus-6 and human herpesvirus-7 viral loads in chronic fatigue patients', *Clinical Infectious Diseases*, 56(10), pp. 1394–1400. Available at: https://doi.org/10.1093/cid/cit086.

Odendall, C. *et al.* (2014) 'Diverse intracellular pathogens activate type III interferon expression from peroxisomes', *Nature Immunology*, 15(8), pp. 717–726. Available at: https://doi.org/10.1038/ni.2915.

Oka, T. *et al.* (2015) 'Comprehensive Review of Human Sapoviruses', *Clinical Microbiology Reviews* [Preprint]. Available at: https://doi.org/10.1128/cmr.00011-14.

Ondov, B.D., Bergman, N.H. and Phillippy, A.M. (2011) 'Interactive metagenomic visualization in a Web browser', *BMC Bioinformatics*, 12(1), p. 385. Available at: https://doi.org/10.1186/1471-2105-12-385.

O'Neal, A.J. and Hanson, M.R. (2021) 'The Enterovirus Theory of Disease Etiology in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: A Critical Review', *Frontiers in Medicine*, 8(June), pp. 1–17. Available at: https://doi.org/10.3389/fmed.2021.688486.

Orji, N. *et al.* (2022) 'Prevalence of myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) in Australian primary care patients: only part of the story?', *BMC Public Health*, 22(1), p. 1516. Available at: https://doi.org/10.1186/s12889-022-13929-9.

Ortmann, A.C. and Suttle, C.A. (2009) 'Determination of virus abundance by epifluorescence microscopy.', *Methods in molecular biology (Clifton, N.J.)*, 501, pp. 87–95. Available at: https://doi.org/10.1007/978-1-60327-164-6_10.

Ott, S.J. *et al.* (2017) 'Efficacy of Sterile Fecal Filtrate Transfer for Treating Patients With Clostridium difficile Infection', *Gastroenterology*, 152(4), pp. 799-811.e7. Available at: https://doi.org/10.1053/j.gastro.2016.11.010.

Ovejero, T. *et al.* (2020) 'Activation of Transposable Elements in Immune Cells of Fibromyalgia Patients', *International Journal of Molecular Sciences*, 21(4), p. 1366. Available at: https://doi.org/10.3390/ijms21041366.

Palarea-Albaladejo, J. (2024) 'zCompositions'. Available at: https://github.com/Japal/zCompositions (Accessed: 19 April 2024).

Pan, W. *et al.* (2014) 'DNA polymerase preference determines PCR priming efficiency', *BMC Biotechnology*, 14(1), p. 10. Available at: https://doi.org/10.1186/1472-6750-14-10.

Pandolfo, M. *et al.* (2022) 'MetaPhage: an Automated Pipeline for Analyzing, Annotating, and Classifying Bacteriophages in Metagenomics Sequencing Data', *mSystems*, 7(5), pp. e00741-22. Available at: https://doi.org/10.1128/msystems.00741-22.

Parasa, S. *et al.* (2020) 'Prevalence of Gastrointestinal Symptoms and Fecal Viral Shedding in Patients With Coronavirus Disease 2019', *JAMA Network Open*, 3(6), p. e2011335. Available at: https://doi.org/10.1001/jamanetworkopen.2020.11335.

Parent, K.N. *et al.* (2010) 'P22 Coat Protein Structures Reveal a Novel Mechanism for Capsid Maturation: Stability without Auxiliary Proteins or Chemical Crosslinks', *Structure*, 18(3), pp. 390–401. Available at: https://doi.org/10.1016/j.str.2009.12.014.

Parras-Moltó, M. *et al.* (2018) 'Evaluation of bias induced by viral enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses', *Microbiome*, 6(1), p. 119. Available at: https://doi.org/10.1186/s40168-018-0507-3.

Pendergrast, T. *et al.* (2016) 'Housebound versus nonhousebound patients with myalgic encephalomyelitis and chronic fatigue syndrome', *Chronic Illness*, 12(4), pp. 292–307. Available at: https://doi.org/10.1177/1742395316644770.

Perez-Sepulveda, B.M. *et al.* (2021) 'An accessible, efficient and global approach for the large-scale sequencing of bacterial genomes', *Genome Biology*, 22(1), p. 349. Available at: https://doi.org/10.1186/s13059-021-02536-3.

Pfeiffer, J.K. and Virgin, H.W. (2016) 'Transkingdom control of viral infection and immunity in the mammalian intestine', *Science*, 351(6270), p. aad5872. Available at: https://doi.org/10.1126/science.aad5872.

Pietsch, C. and Liebert, U.G. (2019) 'Intrahost viral evolution during chronic sapovirus infections', *Journal of Clinical Virology*, 113, pp. 1–7. Available at: https://doi.org/10.1016/j.jcv.2019.02.001.

Pinto, Y. *et al.* (2023) 'Phage-inclusive profiling of human gut microbiomes with Phanta', *Nature Biotechnology*, pp. 1–12. Available at: https://doi.org/10.1038/s41587-023-01799-4.

Podlesny, D. *et al.* (2022) 'Identification of clinical and ecological determinants of strain engraftment after fecal microbiota transplantation using metagenomics', *Cell Reports Medicine*, p. 100711. Available at: https://doi.org/10.1016/j.xcrm.2022.100711.

Porter, A.F. *et al.* (2021) 'Metagenomic Identification of Viral Sequences in Laboratory Reagents', *Viruses*, 13(11), p. 2122. Available at: https://doi.org/10.3390/v13112122.

Pott, J. *et al.* (2011) 'IFN-λ determines the intestinal epithelial antiviral host defense', *Proceedings of the National Academy of Sciences of the United States of America*, 108(19), pp. 7944–7949. Available at: https://doi.org/10.1073/pnas.1100552108.

Proal, A. and Marshall, T. (2018) 'Myalgic encephalomyelitis/chronic fatigue syndrome in the era of the human microbiome: Persistent pathogens drive chronic symptoms by interfering with host metabolism, gene expression, and immunity', *Frontiers in Pediatrics*, 6(373), pp. 1–19. Available at: https://doi.org/10.3389/fped.2018.00373.

Ramsay, A.M. (1986) *The saga of royal free disease.* 1st edn. London, UK: Gower Publishing Corporation.

Randolph, C. *et al.* (1998) 'The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): Preliminary Clinical Validity', *Journal of Clinical and Experimental Neuropsychology*, 20(3), pp. 310–319. Available at: https://doi.org/10.1076/jcen.20.3.310.823.

Rangel-Pineros, G. *et al.* (2023) 'VIRify: An integrated detection, annotation and taxonomic classification pipeline using virus-specific protein profile hidden Markov models', *PLOS Computational Biology*, 19(8), p. e1011422. Available at: https://doi.org/10.1371/journal.pcbi.1011422.

Rao, A.V. *et al.* (2009) 'A randomized, double-blind, placebo-controlled pilot study of a probiotic in emotional symptoms of chronic fatigue syndrome', *Gut Pathogens*, 1(1), p. 6. Available at: https://doi.org/10.1186/1757-4749-1-6.

Rasa, S. *et al.* (2018) 'Chronic viral infections in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)', *Journal of Translational Medicine*, 16(1), pp. 1–25. Available at: https://doi.org/10.1186/s12967-018-1644-y.

Rasa-Dzelzkaleja, S. *et al.* (2023) 'The persistent viral infections in the development and severity of myalgic encephalomyelitis/chronic fatigue syndrome', *Journal of Translational Medicine*, 21(1), p. 33. Available at: https://doi.org/10.1186/s12967-023-03887-0.

Razizadeh, M.H., Khatami, A. and Zarei, M. (2022) 'Global molecular prevalence and genotype distribution of Sapovirus in children with gastrointestinal complications: A systematic review and meta-analysis', *Reviews in Medical Virology*, 32(3), p. e2302. Available at: https://doi.org/10.1002/rmv.2302.

Ren, J. *et al.* (2020) 'Identifying viruses from metagenomic data using deep learning', *Quantitative Biology*, 8(1), pp. 64–77. Available at: https://doi.org/10.1007/s40484-019-0187-4.

Reno, M.L., Cox, C.R. and Powell, E.A. (2022) 'Parvovirus B19: a Clinical and Diagnostic Review', *Clinical Microbiology Newsletter*, 44(12), pp. 107–114. Available at: https://doi.org/10.1016/j.clinmicnews.2022.06.003.

Reyes, A. *et al.* (2010) 'Viruses in the faecal microbiota of monozygotic twins and their mothers', *Nature*, 466(7304), pp. 334–338. Available at: https://doi.org/10.1038/nature09199.

Reyes, G.R. and Kim, J.P. (1991) 'Sequence-independent, single-primer amplification (SISPA) of complex DNA populations', *Molecular and Cellular Probes*, 5(6), pp. 473–481. Available at: https://doi.org/10.1016/S0890-8508(05)80020-9.

Rivas, J.L. *et al.* (2018) 'Association of T and NK cell phenotype with the diagnosis of myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)', *Frontiers in Immunology*, 9(MAY), pp. 1–13. Available at: https://doi.org/10.3389/fimmu.2018.01028.

Rodrigues, L.S. *et al.* (2019) 'HERV-K and HERV-W transcriptional activity in myalgic encephalomyelitis/chronic fatigue syndrome', *Autoimmunity Highlights*, 10(1), pp. 8–12. Available at: https://doi.org/10.1186/s13317-019-0122-8.

Rodrigues, T. *et al.* (2023) 'Procedures in Fecal Microbiota Transplantation for Treating Irritable Bowel Syndrome: Systematic Review and Meta-Analysis', *Journal of Clinical Medicine*, 12(5), p. 1725. Available at: https://doi.org/10.3390/jcm12051725.

Rosseel, T. *et al.* (2013) 'The Origin of Biased Sequence Depth in Sequence-Independent Nucleic Acid Amplification and Optimization for Efficient Massive Parallel Sequencing', *PLOS ONE*, 8(9), p. e76144. Available at: https://doi.org/10.1371/journal.pone.0076144.

Roth, A.N., Grau, K.R. and Karst, S.M. (2019) 'Diverse Mechanisms Underlie Enhancement of Enteric Viruses by the Mammalian Intestinal Microbiota', *Viruses*, 11(8), p. 760. Available at: https://doi.org/10.3390/v11080760.

Roux, S. *et al.* (2016) 'Towards quantitative viromics for both double-stranded and single-stranded DNA viruses', *PeerJ*, 4, p. e2777. Available at: https://doi.org/10.7717/peerj.2777.

Roux, S. *et al.* (2017) 'Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity', *PeerJ*, 5, p. e3817. Available at: https://doi.org/10.7717/peerj.3817.

Roux, S. *et al.* (2019) 'Minimum information about an uncultivated virus genome (MIUVIG)', *Nature Biotechnology*, 37(1), pp. 29–37. Available at: https://doi.org/10.1038/nbt.4306.

Rubio-Mora, E., Carrascoso, G.R. and Rodríguez, J.G. (2024) 'Sapovirus infection as another cause of persistent viral diarrhea: case series and review of the literature', *European Journal of Clinical Microbiology & Infectious Diseases*, 43(1), pp. 55–59. Available at: https://doi.org/10.1007/s10096-023-04695-6.

Rueden, C. *et al.* (2021) 'PyImageJ: Python wrapper for ImageJ2'. Available at: https://github.com/imagej/pyimagej.

Sabina, J. and Leamon, J.H. (2015) 'Bias in Whole Genome Amplification: Causes and Considerations', in T. Kroneis (ed.) *Whole Genome Amplification: Methods and Protocols*. New York, NY: Springer (Methods in Molecular Biology), pp. 15–41. Available at: https://doi.org/10.1007/978-1-4939-2990-0_2.

Salonen, T. *et al.* (2023) 'Randomized, double-blinded, placebo-controlled pilot study: efficacy of faecal microbiota transplantation on chronic fatigue syndrome', *Journal of Translational Medicine*, 21(1), p. 513. Available at: https://doi.org/10.1186/s12967-023-04227-y.

Santiso-Bellón, C. *et al.* (2022) 'Replication of Human Norovirus in Mice after Antibiotic-Mediated Intestinal Bacteria Depletion', *International Journal of Molecular Sciences*, 23(18), p. 10643. Available at: https://doi.org/10.3390/ijms231810643.

Sato, E., Furuta, R.A. and Miyazawa, T. (2010) 'An Endogenous Murine Leukemia Viral Genome Contaminant in a Commercial RT-PCR Kit is Amplified Using Standard Primers for XMRV', *Retrovirology*, 7(1), p. 110. Available at: https://doi.org/10.1186/1742-4690-7-110.

Schneider, C.A., Rasband, W.S. and Eliceiri, K.W. (2012) 'NIH Image to ImageJ: 25 years of image analysis', *Nature Methods*, 9(7), pp. 671–675. Available at: https://doi.org/10.1038/nmeth.2089.

Schreiner, P. *et al.* (2020) 'Human Herpesvirus-6 Reactivation, Mitochondrial Fragmentation, and the Coordination of Antiviral and Metabolic Phenotypes in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome', *ImmunoHorizons*, 4(4), pp. 201–215. Available at: https://doi.org/10.4049/immunohorizons.2000006.

Schult, D. *et al.* (2022) 'Gut bacterial dysbiosis and instability is associated with the onset of complications and mortality in COVID-19', *Gut Microbes*, 14(1), p. 2031840. Available at: https://doi.org/10.1080/19490976.2022.2031840.

Seishima, M. *et al.* (2008) 'Chronic Fatigue Syndrome after Human Parvovirus B19 Infection without Persistent Viremia', *Dermatology*, 216(4), pp. 341–346. Available at: https://doi.org/10.1159/000116723.

Selvakumar, T.A. *et al.* (2017) 'Identification of a predominantly interferon-λ-induced transcriptional profile in murine intestinal epithelial cells', *Frontiers in Immunology*, 8(OCT). Available at: https://doi.org/10.3389/fimmu.2017.01302.

Sepúlveda, N. *et al.* (2019) 'Myalgic Encephalomyelitis/Chronic Fatigue Syndrome as a Hyper-Regulated Immune System Driven by an Interplay Between Regulatory T Cells and Chronic Human Herpesvirus Infections', *Frontiers in Immunology*, 10, p. 2684. Available at: https://doi.org/10.3389/fimmu.2019.02684.

Sepúlveda, N. *et al.* (2022) 'Revisiting IgG Antibody Reactivity to Epstein-Barr Virus in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome and Its Potential Application to Disease Diagnosis', *Frontiers in Medicine*, 9. Available at: https://doi.org/10.3389/fmed.2022.921101.

Seton, K.A. (2022) *Investigating Immune Reactivity to the Intestinal Microbiome in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome.* Doctoral thesis. University of East Anglia. Available at: https://ueaeprints.uea.ac.uk/id/eprint/90862/.

Seton, K.A. *et al.* (2023) 'Investigating Antibody Reactivity to the Intestinal Microbiome in Severe Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS): A Feasibility Study', *International Journal of Molecular Sciences*, 24(20), p. 15316. Available at: https://doi.org/10.3390/ijms242015316.

Seton, K.A. *et al.* (2024) 'Advancing Research and Treatment: An Overview of Clinical Trials in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) and Future Perspectives', *Journal of Clinical Medicine*, 13(2), p. 325. Available at: https://doi.org/10.3390/jcm13020325.

Shen, W. *et al.* (2016) 'SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation', *PLOS ONE*, 11(10), p. e0163962. Available at: https://doi.org/10.1371/journal.pone.0163962.

Shen, W. and Ren, H. (2021) 'TaxonKit: A practical and efficient NCBI taxonomy toolkit', *Journal of Genetics and Genomics*, 48(9), pp. 844–850. Available at: https://doi.org/10.1016/j.jgg.2021.03.006.

Shikova, E. *et al.* (2020) 'Cytomegalovirus, Epstein-Barr virus, and human herpesvirus-6 infections in patients with myalgic encephalomyelitis/chronic fatigue syndrome', *Journal of Medical Virology*, (January), pp. 1–7. Available at: https://doi.org/10.1002/jmv.25744.

Shkoporov, A.N. *et al.* (2018) 'Reproducible protocols for metagenomic analysis of human faecal phageomes', *Microbiome*, 6(1), p. 68. Available at: https://doi.org/10.1186/s40168-018-0446-z.

Shkoporov, A.N. *et al.* (2019) 'The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific', *Cell Host & Microbe*, 26(4), pp. 527-541.e1-e5. Available at: https://doi.org/10.1016/j.chom.2019.09.009.

Shkoporov, A.N. *et al.* (2022) 'Viral biogeography of the mammalian gut and parenchymal organs', *Nature Microbiology*, 7(8), pp. 1301–1311. Available at: https://doi.org/10.1038/s41564-022-01178-w.

Shkoporov, A.N. and Hill, C. (2019) 'Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome', *Cell Host & Microbe*, 25(2), pp. 195–209. Available at: https://doi.org/10.1016/j.chom.2019.01.017.

Shukla, S.K. *et al.* (2015) 'Changes in gut and plasma Microbiome following exercise challenge in Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)', *PLOS ONE*, 10(12), pp. 1–15. Available at: https://doi.org/10.1371/journal.pone.0145453.

Skjevling, L. *et al.* (2024) 'Faecal microbiota transplantation (FMT) in Norwegian outpatients with mild to severe myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS): protocol for a 12-month randomised double-blind placebo-controlled trial', *BMJ Open*, 14(6), p. e073275. Available at: https://doi.org/10.1136/bmjopen-2023-073275.

Skjevling, L.K. *et al.* (2023) 'Colonic distribution of FMT by different enema procedures compared to colonoscopy – proof of concept study using contrast fluid', *BMC Gastroenterology*, 23(1), p. 363. Available at: https://doi.org/10.1186/s12876-023-02979-x.

Smith, R.A. (2010) 'Contamination of clinical specimens with MLV-encoding nucleic acids: implications for XMRV and other candidate human retroviruses', *Retrovirology*, 7(1), p. 112. Available at: https://doi.org/10.1186/1742-4690-7-112.

Smith, S.E. *et al.* (2022) 'Emerging technologies in the study of the virome', *Current Opinion in Virology*, 54, p. 101231. Available at: https://doi.org/10.1016/j.coviro.2022.101231.

Smits, S.L. *et al.* (2014) 'Assembly of viral genomes from metagenomes', *Frontiers in Microbiology*, 5. Available at: https://doi.org/10.3389/fmicb.2014.00714.

Smuts, H. *et al.* (2014) 'Novel Hybrid Parvovirus-Like Virus, NIH-CQV/PHV, Contaminants in Silica Column-Based Nucleic Acid Extraction Kits', *Journal of Virology*, 88(2), pp. 1398–1398. Available at: https://doi.org/10.1128/jvi.03206-13.

do Socorro Fôro Ramos, E. *et al.* (2021) 'Composition of Eukaryotic Viruses and Bacteriophages in Individuals with Acute Gastroenteritis', *Viruses*, 13(12), p. 2365. Available at: https://doi.org/10.3390/v13122365.

Soler, N. *et al.* (2015) 'Membrane vesicles in natural environments: a major challenge in viral ecology', *The ISME Journal*, 9(4), pp. 793–796. Available at: https://doi.org/10.1038/ismej.2014.184.

Sommerfelt, K., Schei, T. and Angelsen, A. (2023) 'Severe and Very Severe Myalgic Encephalopathy/Chronic Fatigue Syndrome ME/CFS in Norway: Symptom Burden and Access to Care', *Journal of Clinical Medicine*, 12(4), p. 1487. Available at: https://doi.org/10.3390/jcm12041487.

Soria-Villalba, A. *et al.* (2024) 'Comparison of Experimental Methodologies Based on Bulk-Metagenome and Virus-like Particle Enrichment: Pros and Cons for Representativeness and Reproducibility in the Study of the Fecal Human Virome', *Microorganisms*, 12(1), p. 162. Available at: https://doi.org/10.3390/microorganisms12010162.

Stallmach, A. *et al.* (2024) 'The gastrointestinal microbiota in the development of ME/CFS: a critical view and potential perspectives', *Frontiers in Immunology*, 15, p. 1352744. Available at: https://doi.org/10.3389/fimmu.2024.1352744.

Steinegger, M. and Söding, J. (2017) 'MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets', *Nature Biotechnology*, 35(11), pp. 1026–1028. Available at: https://doi.org/10.1038/nbt.3988.

Steinegger, M. and Söding, J. (2018) 'Clustering huge protein sequence sets in linear time', *Nature Communications*, 9(1), p. 2542. Available at: https://doi.org/10.1038/s41467-018-04964-5.

Stevens, S. *et al.* (2018) 'Cardiopulmonary Exercise Test Methodology for Assessing Exertion Intolerance in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome', *Frontiers in Pediatrics*, 6. Available at: https://doi.org/10.3389/fped.2018.00242.

Stewart, A.L., Hays, R.D. and Ware, J.E.J. (1988) 'The MOS Short-form General Health Survey: Reliability and Validity in a Patient Population', *Medical Care*, 26(7), p. 724.

Stockdale, S.R. *et al.* (2023) 'Interpersonal variability of the human gut virome confounds disease signal detection in IBD', *Communications Biology*, 6(1), pp. 1–10. Available at: https://doi.org/10.1038/s42003-023-04592-w.

Strand, E.B. *et al.* (2019) 'Myalgic encephalomyelitis/chronic fatigue Syndrome (ME/CFS): Investigating care practices pointed out to disparities in diagnosis and treatment across European Union', *PLOS ONE*, 14(12), pp. 1–12. Available at: https://doi.org/10.1371/journal.pone.0225995.

Strawbridge, R. *et al.* (2019) 'Inflammatory proteins are altered in chronic fatigue syndrome—A systematic review and meta-analysis', *Neuroscience and Biobehavioral Reviews*, 107, pp. 69–83. Available at: https://doi.org/10.1016/j.neubiorev.2019.08.011.

Su, Q. *et al.* (2023) 'Gut microbiome signatures reflect different subtypes of irritable bowel syndrome', *Gut Microbes*, 15(1), p. 2157697. Available at: https://doi.org/10.1080/19490976.2022.2157697.

Sun, J. *et al.* (2020) 'Rapid Detection of Diarrheagenic Escherichia coli by a New Multiplex Real-Time Quantitative PCR Assay', *Applied Biochemistry and Microbiology*, 56(6), pp. 748–757. Available at: https://doi.org/10.1134/S0003683820060174.

Sun, Q. *et al.* (2023) 'Autoantibodies to selenoprotein P in chronic fatigue syndrome suggest selenium transport impairment and acquired resistance to thyroid hormone', *Redox Biology*, 65, p. 102796. Available at: https://doi.org/10.1016/j.redox.2023.102796.

Sunnquist, M. *et al.* (2017) 'A Comparison of Case Definitions for Myalgic Encephalomyelitis and Chronic Fatigue Syndrome.', *Journal of chronic diseases and management*, 2(2), pp. 1–11.

Sutton, T.D.S. *et al.* (2019) 'Choice of assembly software has a critical impact on virome characterisation', *Microbiome*, 7(1). Available at: https://doi.org/10.1186/s40168-019-0626-5.

Sutton, T.D.S., Clooney, A.G. and Hill, C. (2020) 'Giant oversights in the human gut virome', *Gut*, 69(7), pp. 1357–1358. Available at: https://doi.org/10.1136/gutjnl-2019-319067.

Sweere, J.M. *et al.* (2019) 'Bacteriophage trigger antiviral immunity and prevent clearance of bacterial infection', *Science*, 363(6434), p. eaat9691. Available at: https://doi.org/10.1126/science.aat9691.

Tapparel, C. *et al.* (2013) 'Picornavirus and enterovirus diversity with associated human diseases', *Infection, Genetics and Evolution*, 14, pp. 282–293. Available at: https://doi.org/10.1016/j.meegid.2012.10.016.

Taylor, L.J. *et al.* (2022) 'The enigmatic roles of *Anelloviridae* and *Redondoviridae* in humans', *Current Opinion in Virology*, 55, p. 101248. Available at: https://doi.org/10.1016/j.coviro.2022.101248.

Telatin, A. (2022) 'cleanup'. Quadram Institute. Available at: https://github.com/telatin/cleanup.

Teunisse, G.M. (2022) 'Fantaxtic - Nested Bar Plots for Phyloseq Data'. Available at: https://github.com/gmteunisse/Fantaxtic (Accessed: 3 January 2024).

Thaiss, C.A. *et al.* (2016) 'The microbiome and innate immunity', *Nature*, 535(7610), pp. 65–74. Available at: https://doi.org/10.1038/nature18847.

Thapaliya, K. *et al.* (2022) 'Alteration of Cortical Volume and Thickness in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome', *Frontiers in Neuroscience*, 16. Available at: https://doi.org/10.3389/fnins.2022.848730.

Thapaliya, K. *et al.* (2023) 'Brainstem volume changes in myalgic encephalomyelitis/chronic fatigue syndrome and long COVID patients', *Frontiers in Neuroscience*, 17. Available at: https://doi.org/10.3389/fnins.2023.1125208.

Tomaru, Y. and Nagasaki, K. (2007) 'Flow cytometric detection and enumeration of DNA and RNA viruses infecting marine eukaryotic microalgae', *Journal of Oceanography*, 63(2), pp. 215–221. Available at: https://doi.org/10.1007/s10872-007-0023-8.

Toogood, P.L. *et al.* (2021) 'Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS): Where will the drugs come from?', *Pharmacological Research*, 165, p. 105465. Available at: https://doi.org/10.1016/j.phrs.2021.105465.

Troeger, C. *et al.* (2017) 'Estimates of global, regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of Disease Study 2015', *The Lancet Infectious Diseases*, 17(9), pp. 909–948. Available at: https://doi.org/10.1016/S1473-3099(17)30276-1.

Trubl, G. *et al.* (2020) 'Coming-of-Age Characterization of Soil Viruses: A User's Guide to Virus Isolation, Detection within Metagenomes, and Viromics', *Soil Systems*, 4(2), p. 23. Available at: https://doi.org/10.3390/soilsystems4020023.

Tsai, S.-Y. *et al.* (2014) 'Increased risk of chronic fatigue syndrome following herpes zoster: a population-based study', *European Journal of Clinical Microbiology & Infectious Diseases*, 33(9), pp. 1653–1659. Available at: https://doi.org/10.1007/s10096-014-2095-x.

Tsai, S.-Y. *et al.* (2019) 'Increased risk of chronic fatigue syndrome in patients with inflammatory bowel disease: a population-based retrospective cohort study', *Journal of Translational Medicine*, 17(1), p. 55. Available at: https://doi.org/10.1186/s12967-019-1797-3.

Tschopp, R. *et al.* (2023) 'Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS): A preliminary survey among patients in Switzerland', *Heliyon*, 9(5), p. e15595. Available at: https://doi.org/10.1016/j.heliyon.2023.e15595.

Turzynski, V. *et al.* (2021) 'Imaging Techniques for Detecting Prokaryotic Viruses in Environmental Samples', *Viruses*, 13(11), p. 2126. Available at: https://doi.org/10.3390/v13112126.

Twisk, F. (2018) 'Myalgic Encephalomyelitis, Chronic Fatigue Syndrome, and Systemic Exertion Intolerance Disease: Three Distinct Clinical Entities', *Challenges*, 9(1), p. 19. Available at: https://doi.org/10.3390/challe9010019.

Underhill, R.A. (2015) 'Myalgic encephalomyelitis, chronic fatigue syndrome: An infectious disease', *Medical Hypotheses*, 85(6), pp. 765–773. Available at: https://doi.org/10.1016/j.mehy.2015.10.011.

Ungaro, F. *et al.* (2019) 'Metagenomic analysis of intestinal mucosa revealed a specific eukaryotic gut virome signature in early-diagnosed inflammatory bowel disease', *Gut Microbes*, 10(2), pp. 149–158. Available at: https://doi.org/10.1080/19490976.2018.1511664.

Unger, E.R. *et al.* (2024) 'Heterogeneity in Measures of Illness among Patients with Myalgic Encephalomyelitis/Chronic Fatigue Syndrome Is Not Explained by Clinical Practice: A Study in Seven U.S. Specialty Clinics', *Journal of Clinical Medicine*, 13(5), p. 1369. Available at: https://doi.org/10.3390/jcm13051369.

Usuku, S. *et al.* (2008) 'An Outbreak of Food-Borne Gastroenteritis Due to Sapovirus among Junior High School Students', *Japanese Journal of Infectious Diseases*, 61(6), pp. 438–441. Available at: https://doi.org/10.7883/yoken.JJID.2008.438.

Vaes, A.W. *et al.* (2023) 'Symptom-based clusters in people with ME/CFS: an illustration of clinical variety in a cross-sectional cohort', *Journal of Translational Medicine*, 21(1), p. 112. Available at: https://doi.org/10.1186/s12967-023-03946-6.

Valcarce, M.D. *et al.* (2021) 'Global distribution of sporadic sapovirus infections: A systematic review and meta-analysis', *PLOS ONE*, 16(8), p. e0255436. Available at: https://doi.org/10.1371/journal.pone.0255436.

Valdez, A.R. *et al.* (2019) 'Estimating prevalence, demographics, and costs of ME/CFS using large scale medical claims data and machine learning', *Frontiers in Pediatrics*, 6(JAN), pp. 1–14. Available at: https://doi.org/10.3389/fped.2018.00412.

VanElzakker, M.B. (2013) 'Chronic fatigue syndrome from vagus nerve infection: A psychoneuroimmunological hypothesis', *Medical Hypotheses*, 81(3), pp. 414–423. Available at: https://doi.org/10.1016/j.mehy.2013.05.034.

Vignuzzi, M. and López, C.B. (2019) 'Defective viral genomes are key drivers of the virus–host interaction', *Nature Microbiology*, 4(7), pp. 1075–1087. Available at: https://doi.org/10.1038/s41564-019-0465-y.

Villafane, R. (2009) 'Construction of Phage Mutants', in M.R.J. Clokie and A.M. Kropinski (eds) *Bacteriophages: Methods and Protocols, Volume 1: Isolation, Characterization, and Interactions*. Humana Press, a part of Springer Science+Business Media, pp. 223–238. Available at: doi.org/10.1007/978-1-60327-164-6.

Virgin, H.W. (2014) 'The virome in mammalian physiology and disease', *Cell* [Preprint]. Available at: https://doi.org/10.1016/j.cell.2014.02.032.

Vogl, T. *et al.* (2022) 'Systemic antibody responses against human microbiota flagellins are overrepresented in chronic fatigue syndrome patients', *Science Advances*, 8(38), p. eabq2422. Available at: https://doi.org/10.1126/sciadv.abq2422.

Vyas, J. *et al.* (2022) 'Impact of myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) on the quality of life of people with ME/CFS and their partners and family members: an online cross-sectional survey', *BMJ Open*, 12(5), p. e058128. Available at: https://doi.org/10.1136/bmjopen-2021-058128.

Walitt, B. *et al.* (2024) 'Deep phenotyping of post-infectious myalgic encephalomyelitis/chronic fatigue syndrome', *Nature Communications*, 15(1), p. 907. Available at: https://doi.org/10.1038/s41467-024-45107-3.

Walters, W.A. *et al.* (2023) 'Longitudinal comparison of the developing gut virome in infants and their mothers', *Cell Host & Microbe*, 31(2), pp. 187-198.e3. Available at: https://doi.org/10.1016/j.chom.2023.01.003.

Wang, G. *et al.* (2023) 'Optimization and evaluation of viral metagenomic amplification and sequencing procedures toward a genome-level resolution of the human fecal DNA virome', *Journal of Advanced Research*, 48, pp. 75–86. Available at: https://doi.org/10.1016/j.jare.2022.08.011.

Wang, J.-H. *et al.* (2024) 'Clinical evidence of the link between gut microbiome and myalgic encephalomyelitis/chronic fatigue syndrome: a retrospective review', *European Journal of Medical Research*, 29(1), p. 148. Available at: https://doi.org/10.1186/s40001-024-01747-1.

Warwick-Dugdale, J. *et al.* (2019) 'Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands', *PeerJ*, 7(2), p. e6800. Available at: https://doi.org/10.7717/peerj.6800.

Watt, T. *et al.* (2012) 'Response to valganciclovir in chronic fatigue syndrome patients with human herpesvirus 6 and Epstein-Barr virus IgG antibody titers', *Journal of Medical Virology*, 84(12), pp. 1967–1974. Available at: https://doi.org/10.1002/jmv.23411.

Wei, H. *et al.* (2023) 'Serum from Myalgic encephalomyelitis/chronic fatigue syndrome patients causes loss of coherence in cellular circadian rhythms', *Journal of Neuroimmunology*, 381, p. 578142. Available at: https://doi.org/10.1016/j.jneuroim.2023.578142.

Wen, K., Ortmann, A.C. and Suttle, C.A. (2004) 'Accurate estimation of viral abundance by epifluorescence microscopy', *Applied and Environmental Microbiology*, 70(7), pp. 3862–3867. Available at: https://doi.org/10.1128/AEM.70.7.3862-3867.2004.

White, P.D. (2019) 'A perspective on causation of the chronic fatigue syndrome by considering its nosology', *Journal of Evaluation in Clinical Practice*, 25(6), pp. 991–996. Available at: https://doi.org/10.1111/jep.13240.

Wick, R.R. *et al.* (2015) 'Bandage: Interactive visualization of de novo genome assemblies', *Bioinformatics*, 31(20), pp. 3350–3352. Available at: https://doi.org/10.1093/bioinformatics/btv383.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag. Available at: https://ggplot2.tidyverse.org.

Wickham, H. *et al.* (2019) 'Welcome to the Tidyverse', *Journal of Open Source Software*, 4(43), p. 1686. Available at: https://doi.org/10.21105/joss.01686.

Wilen, C.B. *et al.* (2018) 'Tropism for tuft cells determines immune promotion of norovirus pathogenesis', *Science*, 360(6385), pp. 204–208. Available at: https://doi.org/10.1126/science.aar3799.

Williams, M.V. *et al.* (2019) 'Epstein-Barr Virus dUTPase Induces Neuroinflammatory Mediators: Implications for Myalgic Encephalomyelitis/Chronic Fatigue Syndrome', *Clinical Therapeutics*, 41(5), pp. 848–863. Available at: https://doi.org/10.1016/j.clinthera.2019.04.009.

Wirusanti, N.I., Baldridge, M.T. and Harris, V.C. (2022) 'Microbiota regulation of viral infections through interferon signaling', *Trends in Microbiology*, 30(8), pp. 778–792. Available at: https://doi.org/10.1016/j.tim.2022.01.007.

Wong, T.L. and Weitzer, D.J. (2021) 'Long COVID and Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS)-A Systemic Review and Comparison of Clinical Presentation and Symptomatology', *Medicina (Kaunas, Lithuania)*, 57(5), p. 418. Available at: https://doi.org/10.3390/medicina57050418.

Wood, D.E., Lu, J. and Langmead, B. (2019) 'Improved metagenomic analysis with Kraken 2', *Genome Biology*, 20(1), p. 257. Available at: https://doi.org/10.1186/s13059-019-1891-0.

Woods Acevedo, M.A. and Pfeiffer, J.K. (2021) 'Microbiota-immune system interactions and enteric virus infection', *Current Opinion in Virology*, 46, pp. 15–19. Available at: https://doi.org/10.1016/j.coviro.2020.08.005.

World Medical Association (2013) 'World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects', *JAMA*, 310(20), pp. 2191–2194. Available at: https://doi.org/10.1001/jama.2013.281053.

Wu, J., Lv, L. and Wang, C. (2022) 'Efficacy of Fecal Microbiota Transplantation in Irritable Bowel Syndrome: A Meta-Analysis of Randomized Controlled Trials', *Frontiers in Cellular and Infection Microbiology*, 12. Available at: https://www.frontiersin.org/articles/10.3389/fcimb.2022.827395 (Accessed: 1 October 2023).

Xiao, N. (2023) *ggsci: Scientific Journal and Sci-Fi Themed Color Palettes for 'ggplot2'*. Available at: https://nanx.me/ggsci/ (Accessed: 1 September 2023).

Xiong, R. *et al.* (2023) 'Multi-'omics of gut microbiome-host interactions in short- and long-term myalgic encephalomyelitis/chronic fatigue syndrome patients', *Cell Host & Microbe*, 31(2), pp. 273-287.e5. Available at: https://doi.org/10.1016/j.chom.2023.01.001.

Yamashita, Y. *et al.* (2010) 'Molecular characterization of sapovirus detected in a gastroenteritis outbreak at a wedding hall', *Journal of Medical Virology*, 82(4), pp. 720–726. Available at: https://doi.org/10.1002/jmv.21646.

Yan, A. *et al.* (2023) 'Multiomic spatial analysis reveals a distinct mucosa-associated virome', *Gut Microbes*, 15(1), p. 2177488. Available at: https://doi.org/10.1080/19490976.2023.2177488.

Yang, T. *et al.* (2019) 'The clinical value of cytokines in chronic fatigue syndrome', *Journal of Translational Medicine*, 17(1), p. 213. Available at: https://doi.org/10.1186/s12967-019-1948-6.

Yaron, J.R. *et al.* (2020) 'Immune protection is dependent on the gut microbiome in a lethal mouse gammaherpesviral infection', *Scientific Reports*, 10(1), p. 2371. Available at: https://doi.org/10.1038/s41598-020-59269-9.

Yeoh, Y.K. *et al.* (2021) 'Gut microbiota composition reflects disease severity and dysfunctional immune responses in patients with COVID-19', *Gut*, 70(4), pp. 698–706. Available at: https://doi.org/10.1136/gutjnl-2020-323020.

Yilmaz, S., Allgaier, M. and Hugenholtz, P. (2010) 'Multiple displacement amplification compromises quantitative analysis of metagenomes', *Nature Methods*, 7(12), pp. 943–944. Available at: https://doi.org/10.1038/nmeth1210-943.

Yoshida, T. *et al.* (2009) 'Characterization of sapoviruses detected in gastroenteritis outbreaks and identification of asymptomatic adults with high viral load', *Journal of Clinical Virology*, 45(1), pp. 67–71. Available at: https://doi.org/10.1016/j.jcv.2009.03.003.

Żaczek, M. *et al.* (2016) 'Antibody Production in Response to Staphylococcal MS-1 Phage Cocktail in Patients Undergoing Phage Therapy', *Frontiers in Microbiology*, 7. Available at: https://doi.org/10.3389/fmicb.2016.01681.

Zakrzewska, K. *et al.* (2023) 'Parvovirus B19: Insights and implication for pathogenesis, prevention and therapy', *Aspects of Molecular Medicine*, 1, p. 100007. Available at: https://doi.org/10.1016/j.amolm.2023.100007.

Zhang, F. *et al.* (2021) 'Longitudinal dynamics of gut bacteriome, mycobiome and virome after fecal microbiota transplantation in graft-versus-host disease', *Nature Communications*, 12(1), p. 65. Available at: https://doi.org/10.1038/s41467-020-20240-x.

Zhang, L. *et al.* (2010) 'Microbial infections in eight genomic subtypes of chronic fatigue syndrome/myalgic encephalomyelitis', *Journal of Clinical Pathology*, 63(2), pp. 156–164. Available at: https://doi.org/10.1136/jcp.2009.072561.

Zhang, Y. *et al.* (2020) 'Composition of the murine gut microbiome impacts humoral immunity induced by rabies vaccines', *Clinical and Translational Medicine*, 10(4), p. e161. Available at: https://doi.org/10.1002/ctm2.161.

Zigmond, A.S. and Snaith, R.P. (1983) 'The Hospital Anxiety and Depression Scale', *Acta Psychiatrica Scandinavica*, 67(6), pp. 361–370. Available at: https://doi.org/10.1111/j.1600-0447.1983.tb09716.x.

Zuo, T. *et al.* (2018) 'Bacteriophage transfer during faecal microbiota transplantation in Clostridium difficile infection is associated with treatment outcome', *Gut*, 67(4), pp. 634–643. Available at: https://doi.org/10.1136/gutjnl-2017-313952.

Zuo, T. *et al.* (2019) 'Gut mucosal virome alterations in ulcerative colitis', *Gut*, 68(7), pp. 1169–1179. Available at: https://doi.org/10.1136/gutjnl-2018-318131.

APPENDICES

Appendix 1: Preprint of manuscript based on chapter 2, submitted to MDPI *Viruses*.