The Virome in Primary and Secondary Immunodeficiency

Samuel Christopher Bruce Stubbs

This dissertation is submitted for the degree of

Doctor of Philosophy

Darwin College

June 2018

Declaration

I hereby declare that this dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as specified in the text

This dissertation contains less than 65,000 words including appendices, bibliography, footnotes, tables and equations, and has less than 150 figures.

Samuel Stubbs

June 2018

"We announce the discovery of this cure and that but make no mention of the new diseases which we have created en route"

Henry Miller, The Colossus of Maroussi

Acknowledgements

This work would not have been possible without the help and support of many people. In no particular order:

My PhD supervisor, Jonathan Heeney. During my time in his group, Jon has provided me with unconditional support, and the freedom to pursue my goals and develop my scientific thinking.

My second supervisor, Helen Baxendale who, despite great adversity, strove to continue providing me with the samples required for my work. Her clinical expertise was vital in formulating my research questions.

Post-docs and PIs of the LVZ group, particularly Gordon Daly, Hanna Dreja, and Bethany Dearlove. Who provided guidance and support. Gordon was my mentor when I first arrived in the lab in 2012 and provided me with much of the knowledge and skills I employed for this thesis. Dr Barbara Blacklaws who has been my first stop when I have need scientific advice, and always made herself available. Dr Simon Frost for his assistance with complex statistics and advice with regards to computational methods. Paul Tonks, for his untiring service to the department, keeping the equipment functioning, and for helping with extractions.

My fellow PhD and MSc students, all have contributed to make the experience an enjoyable one. Constanza, Sarah, Arden, and Giacomo have become great friends. Dr James Lester repeatedly took the time to help me produce many of the complex figures within this thesis; Dr Ed Greenwood has consistently provided me with sound advice regarding both science and life in general; and Dr Fabian Schmidt, who helped me when I most needed it.

The students I supervised over the course of my work: Luke Bibby, Alexandra Moskaluk, and Samuel Franklin. All of whom assisted in the lab and eased the ambitious burden I had set for myself.

Professor Ken Smith and Dr Paul Lyons. Who kindly provided me with samples from their own patient cohorts, saving me years of work, and their expert advice regarding said patients.

Danai, my parents, and the rest of my friends and family, who have encouraged me and provided me with unwavering support and understanding. In particular, my grandfather Alan Bruce, the scientist, who inspired me to pursue research in the first place.

Finally, I am also grateful to the Medical Research Council and Glaxo-Smith Kline (formerly Novartis Diagnostics) for their generous funding of this PhD studentship.

Abbreviations

AIDS	–	Acquired immunodeficiency syndrome
APC	–	Antigen presenting cell
ART	–	Anti-retroviral therapy
BCR	–	B-cell receptor
BKV	–	BK polyomavirus
cDNA	–	Complementary DNA
CMV	–	Cytomegalovirus
CVID	–	Common variable immunodeficiency
EBV	–	Epstein-Barr virus
ELISA	–	Enzyme-linked immunosorbant assay
GBV-C	–	GB-virus C
GWAS	–	Genome-wide association study
HBV	–	Hepatitis B virus
HCV	–	Hepatitis C virus
HHV8	–	Human herpesvirus 8
HIV	–	Human immunodeficiency virus
HLA	–	Human leukocyte antigen
HPgV	–	Human pegivirus
HSE	–	Herpes simplex encephalitis
HSV	–	Herpes simplex virus
IEL	–	Intra-epithelial lymphocyte
IFN	–	Interferon
Ig(A-E)	–	Immunoglobulin (isotype A-E)
MHC	–	Major histocompatibility complex
NGS	–	Next generation sequencing
NK cell	–	Natural killer cell
PBMC	–	Peripheral blood mononuclear cell
PCR	–	Polymerase chain reaction
PID	–	Primary immunodeficiency
PRR	–	Pattern recognition receptor
qPCR	–	Quantitative polymerase chain reaction
RCA	–	Rolling circle amplification
RT-PCR	–	Reverse-transcription PCR
rRNA	–	Ribosomal RNA
sCD14	–	Soluble CD14
SCID	–	Severe combined immunodeficiency
SID	–	Secondary immunodeficiency
SISPA	–	Sequence independent single primer amplification
TTMDV	–	Torque-teno midi virus
TTMV	–	Torque-teno mini virus
TTV	–	Torque-teno virus
XLA	–	X-linked agammaglobulinaemia

Summary

The afflictions suffered by immunocompromised individuals have historically been attributed to overt infections caused by bacterial and fungal pathogens. For this reason, treatment methods have focused on resolving these infections, with great success in terms of reducing short-term morbidity and mortality rates. This initial success only served to reinforce the dogmatic opinion that to 'cure' immunodeficiency, one needs only to resolve and prevent recurrence of bacterial and fungal infections. However, reports of long-term health problems in immunocompromised cohorts suggest that treatment of bacterial and fungal infections alone does not resolve all aspects of the disease, and that viruses may play a greater role than previously expected.

This thesis investigates whether viral infections do indeed have a significant impact in the immunocompromised patient. The overall prevalence of blood-borne viral infections in immunocompromised cohorts was determined through the combined use of unbiased, metagenomic sequencing and qPCR. The viral species detected were compared with patient records in order to determine whether there were any correlations between viral presence and patient outcome following treatment. Furthermore, by investigating a cross-section of cohorts with both inherited and acquired immunodeficiencies, commonalities and differences could be found in terms of the types of viruses that infect these cohorts and their abundance in patients with different types of immunodeficiency.

The findings of this work suggest that a large number of clinically undiagnosed viruses infect immunocompromised patients, however the prevalence of these viruses varies according to the form of immunodeficiency and, to a lesser extent, according to differences between individuals in the same cohort. Importantly, some of the more common viruses detected appear to be correlated with poor patient outcomes such as graft rejection and future infectious complications. Overall, these results suggest that viral infections do indeed play a larger role in the health of immunocompromised patients than has previously been thought although whether this is due to a direct cause or as a consequence is yet to be determined.

Contents

List of Figures

Chapter 4

Chapter 5

List of Tables

Appendices

Chapter 1

Introduction

1.1 Viruses

Viruses are obligate intracellular parasites that infect all living organisms including animals, plants, bacteria and archaea. At their most basic, viral particles consist of nucleic acid encapsulated within a protein shell known as the capsid, which is encoded by the viral genome. The capsid facilitates the persistence of the virus by delivering the genome to its site of replication and shielding it against the environment. A lipid membrane, known as the viral envelope, also coats some species of virus. The viral envelope is derived from the host cell membrane and can play a key role in cell attachment, cell entry and evasion of the host immune system.[1] Their simple genomic structure means that viruses do not encode all of the cellular machinery required to independently produce progeny. It is for this reason that viruses must infect and parasitise the cells of other organisms in order to make use of the cell's resources and many of its enzymatic functions.

Viruses show a great degree of variability. Unlike the genomes of cellular organisms, which are exclusively composed of double-stranded DNA molecules, virus genomes exist in a variety of nucleic acid forms. Viral genomes can be linear or circular molecules and composed of either RNA or DNA, which in turn can be double-stranded, single-stranded or a combination of the two (e.g. hepatitis B virus).[2] RNA genome polarity can be positive-sense, which may be directly translated by the host cell, or negative-sense, which must be first converted into a positive-sense polarity by RNA polymerase prior to translation.

Some virus genomes are ambisense, where the genome is in part positive and part negative, sharing features of each.[3] Viral genomes may also be composed of single molecules or exist as multiple segments. Segmented genomes facilitate the exchange of genes between individual viruses by re-assortment, speeding the rate of mutation.[4]

Viruses are thought to have evolved from multiple lineages.[5] Therefore, whilst viruses within clades can be classified by the conventional nucleotide similarity based methods used for all other living species, a lack of orthologous viral proteins present across distinct viral families means that the same methods cannot (always) be used for classification.[6] Instead, the Baltimore classification system, designed by the biologist David Baltimore,[7] is commonly used to group viruses based on the type of genome and method of nucleic acid replication (Figure 1.1). These features have implications with regards to the life cycle of a virus, which, in some cases, can impact on how it can be diagnosed and treated.

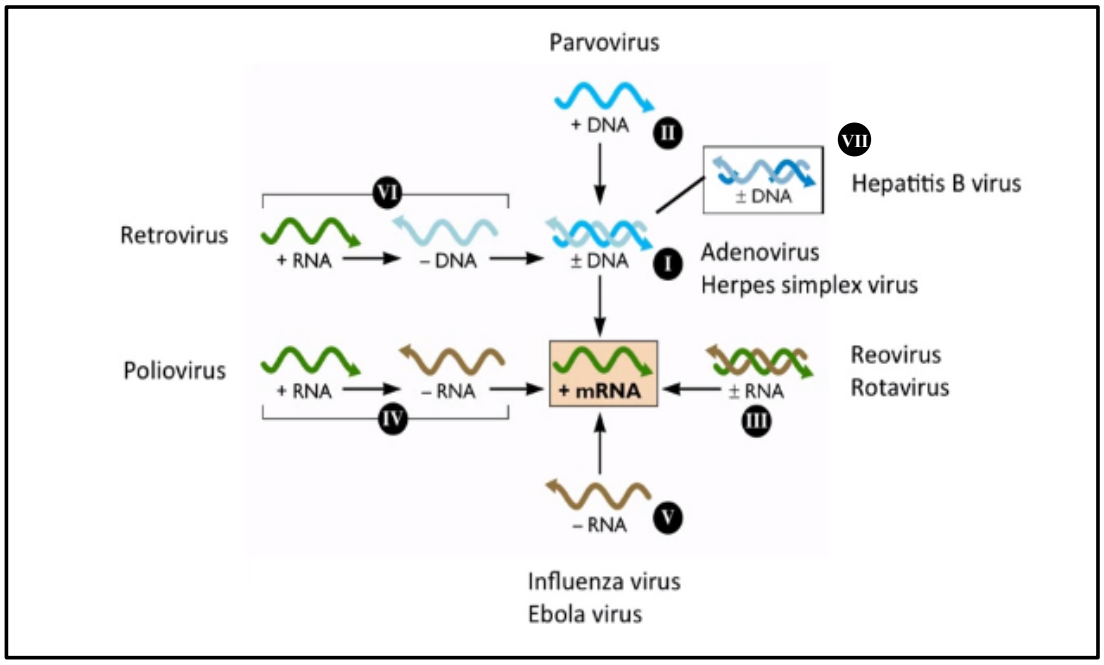


Figure 1.1. **The Baltimore classification of viruses.** Regardless of their genomic structure, all viruses must produce viral mRNA for the translation of viral proteins. The Baltimore classification system classifies viruses into seven groups according to their genomic architecture and method of replication: double-stranded DNA (group I) (e.g. adenoviruses and herpesviruses); single-stranded DNA (group II) (e.g. parvoviruses); double-stranded RNA (group III) (e.g. rotaviruses); positive-sense single-stranded RNA (group IV) (e.g. poliovirus); negative-sense single-stranded RNA (group V) (e.g. orthomyxoviruses and filoviruses); positive-sense single-stranded RNA which replicates through a DNA intermediate (group VI) (e.g. retroviruses); and double-stranded DNA which replicates through an RNA intermediate (group VII) (e.g. hepatitis B virus). Image reproduced from www.microbeonline.com

1.1.1 The General Life Cycle of Viruses

Each virus species has evolved its own specific mechanisms in order to efficiently infect cells and replicate its genome (Figure 1.2). However, most viruses broadly follow a similar course of infection.

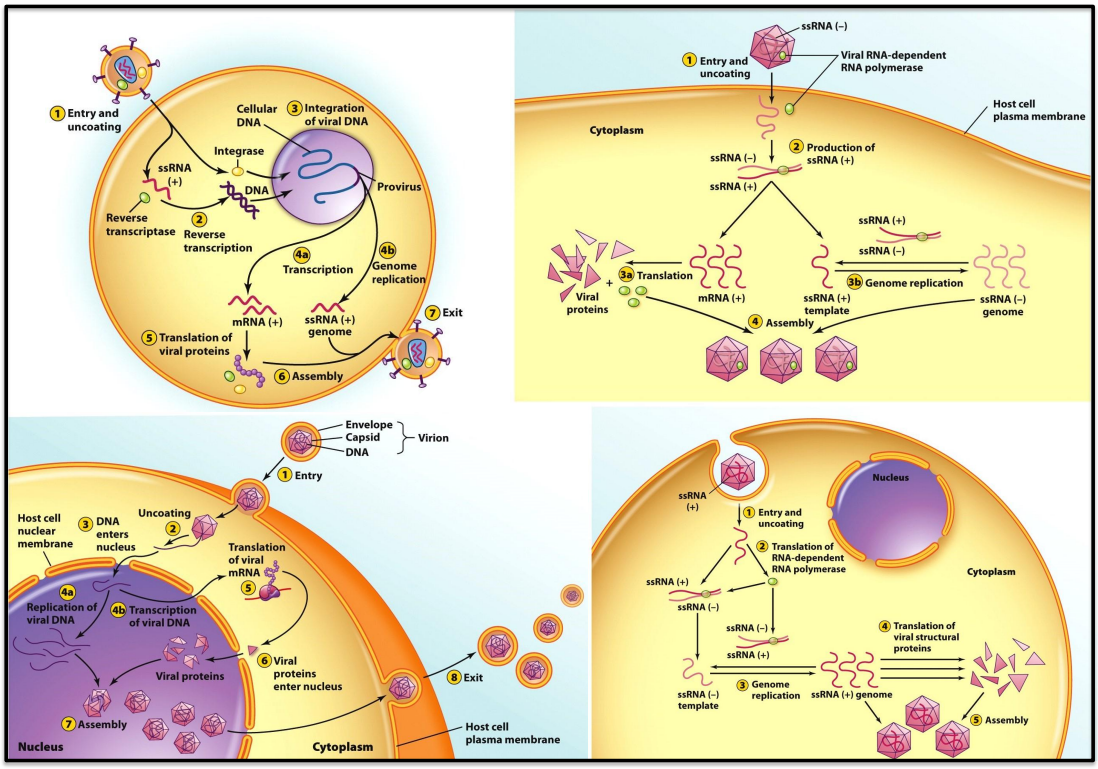


Figure 1.2. **Replication mechanisms of viruses.** Mechanisms of cell entry, cell exit, and replication differ between virus species. However, in order to reproduce, all viruses must produce viral mRNA for translation into viral proteins. Clockwise from top-left: enveloped retrovirus; non-enveloped single-stranded (-) RNA virus; non-enveloped single-stranded (+) RNA virus; and enveloped double-stranded DNA virus. Images reproduced from Virology: Principles and Applications, 2nd Edition.[8]

Upon exposure to a potential host, the virus must first breach the epithelial barrier to gain entry to the body. Viruses have evolved to utilise different transmission pathways to achieve this: influenza virus enters its host via inhalation of viral particles and infects epithelial cells of the respiratory tract,[9] while norovirus enters through the digestive tract on the surface of contaminated objects such as food.[10,11] Other viruses rely on direct exposure to the blood through mechanical breaches of the epithelia: hepatitis C virus transmission relies predominantly on the contamination of equipment used during blood transfusion and intravenous drug use,[12] while a large number of viruses, known as arboviruses, have evolved to be transmitted through the bite of blood-feeding insects.[13] Each of these transmission mechanisms require the virus in question to have evolved certain life-cycle characteristics, such as the ability to persist in the environment, or the ability to replicate in the salivary glands of mosquitoes. From these examples alone, it is clear that viruses represent an extraordinary level of diversity.

Following successful entry into the host, the virus must reach a cell it is able to infect. Some viruses such as influenza and norovirus come into direct contact with cells permissive to infection whereas many others, such as hepatitis C virus make use of the circulatory system to spread within the host and reach their primary site of replication. The circulatory system is the ideal transport system for viruses needing to reach their preferred site of infection, as the blood is required to reach all organs, in order to provide the nutrients necessary for cell function and survival. It is another testament to the diversity of viruses, that some have evolved to also make use of the lymphatic system,[14] and the nervous system as alternative forms of transport.[15–17]

Upon encountering a cell-type permissive to infection, molecules on the virion's surface are able to bind to the cell and trigger intra-cellular signalling pathways

to enter.[18] Cells permissive to infection are recognised by specific receptors on the cell surface, for example, influenza virus utilises α2,6-linked sialic acid expressed by cells located predominantly in the upper and lower respiratory tract of humans.[19] Within the cell, the virus un-packages it's genome and hi-jacks cellular machinery to replicate: copying its genomic material and building the protein structures needed to assemble new capsids. Viral particles, known as virions are assembled from the capsid proteins and the newly produced virions containing a copy of the viral genome are released by the cell either by cell lysis (most non-enveloped viruses) or membrane budding (most enveloped viruses).[20] The newly released infectious virions proceed to infect more cells, replicate further, and exit the body via various pathways, dependent on the virus species, to enter further susceptible individuals and repeat the cycle.

1.1.2 Productive, Abortive and Latent Infections

Entry into a cell does not necessarily lead to replication and production of viral progeny but can result in either a productive, abortive or latent infection. For an infection to be productive, the virus must replicate and produce further infectious particles capable of infecting new cells, whilst an abortive infection occurs when an infection does not result in the production of new infectious viral particles.[20] Both forms of infection are commonly observed in nature and are a determining factor as to whether an individual spontaneously recovers from an infection or goes on to develop disease.

A latent infection is a form of infection where the production of new viral particles does not occur immediately. Instead, the virus genome persists in the infected cell in silence: restricting viral gene expression and producing no viral particles. Cells infected with viruses in a latent state exhibit practically no sign of infection and therefore latent viruses go undetected by the host immune system

and result in next to no discernible symptoms.[21] Many viruses that establish latent infections are highly prevalent in the human population, such as herpes simplex virus 1 (HSV1) which is estimated to infect two thirds of the global population below the age of 50 through this mechanism of immune-evasion.[22] Following infection, this virus establishes life-long persistence with few overt symptoms in the immunocompetent host as re-activation events are rapidly cleared in most hosts.[23]

Latent virus infections are generally well tolerated and any reactivations controlled by a healthy immune system. However, in order to successfully reproduce and transmit, latent viruses must be able to reactivate fully on occasion. This reactivation often coincides with periods of reduced immune strength brought on by factors such as stress, aging, infection by other pathogens, and therapeutic immune suppression.[24] Periods of re-activation are often discernible by break-out symptoms of disease such as the development of shingles, often seen in the elderly, and caused by varicella zoster virus.[25]

1.1.3 Acute and Chronic Infections and Latency

Viral infections can also be acute or chronic. Acute infections are short lived and tend to result in either clearance of the virus or death of the host. Chronic infections are defined as infections that persist after the period of time by which an acute infection is usually cleared. Whilst acute infections typically represent a period of rapid shifts in terms of virus population size and intensity of the host immune response, chronic infections occur when a state of balance is reached between the two.[20]

Whilst some viruses exclusively cause one form of infection or the other (e.g. rabies virus infection is only ever acute and human immunodeficiency virus 1

(HIV-1) infection is almost always chronic), many viruses vary, causing acute disease followed by clearance of the virus in some individuals whilst resulting in a long-term, persistent infection in others. Hepatitis B virus is an example of this where the majority of cases spontaneously resolve but in cases where the virus is able to persist, patients can develop severe disease including fulminant hepatitis, cirrhosis and hepatocarcinoma.[26]

Despite a lack of overt symptoms during many chronic and latent infections, there is evidence that presence of virus and constant probing of the immune system leads to an immune state of "exhaustion", characterised by a population of virus specific cytotoxic T-cells (described in section 1.2.5) that are unable to proliferate or to produce cytokines upon stimulation by antigens.[27] The effects of chronic exposure to antigen can be significant: "exhaustion" of the immune system has been hypothesised to result in reduced immunity and facilitate viral persistence,[28] whilst chronic infection by cytomegalovirus has been found to reduce life expectancy by an average of almost 4 years.[29]

Why a virus may infect and persist in one individual whilst being rapidly cleared in another is thought to be predominantly due to a combination of genetic and acquired differences that affect how the host immune system and the virus interact. For example, norovirus infection usually results in acute gastroenteritis which lasts for an average of 12-72 hours before clearing.[30] However, individuals with a weak immune systems are more likely to develop chronic norovirus infection, which has been reported as more severe in some populations,[30] and asymptomatic in others.[31]

1.1.4 Disease and Virulence

The mechanisms by which different viruses cause disease are also varied: host cells may be killed directly through immune-mediated apoptosis,[32] by cell lysis following viral replication and escape,[33] or by the suppression of vital intracellular processes by viral proteins.[34] Some viruses elicit strong inflammatory responses which damage host tissue resulting in symptoms such as encephalitis and haemorrhagic fever,[35,36] whilst others such as rabies virus may interfere with cellular mechanisms. Indeed, despite its severe neurological symptoms, acute rabies virus infection results in only limited histopathological lesions and generation of a weak immune response.[37] Instead, it is cellular dysfunction, through the inhibition of protein synthesis in the cells of the central nervous system, which results in eventual death of the host.[37]

Recent metagenomic sequencing studies have surprisingly revealed the presence of high levels of viruses in asymptomatic individuals.[38] Even the blood of healthy individuals, once believed to be sterile, is now known to contain a diverse community of viruses.[39] Furthermore, there is even evidence that infection by viruses previously thought of as highly pathogenic are largely asymptomatic. One serological-based study of influenza virus found that only 23% of those shown to be exposed to the virus developed illness.[40] Why infection of some individuals can result in severe disease, while infection of others by the same virus can be asymptomatic, is largely due to differences in the immune response. These differences are governed by genetic variation in both the virus and the host.

A virus' capacity to cause disease in its host is known as its virulence. The greater a virus' virulence, the more likely it is to cause disease and the more likely that disease is to be severe. Ebolavirus is a clear example of a highly

virulent virus. Reports from the 2014 outbreak in West Africa described symptoms of fever in 87.1% of diagnosed cases, haemorrhagic symptoms in 18% and a mortality rate of 70.8% in those cases with definitive outcomes.[41] The virulence of Ebolavirus can be attributed to strains of the virus actively triggering the host immune system, leading to wide-spread cell damage caused by high levels of inflammation.[42] Conversely, torque-teno virus is an example of a virus with very low virulence. This virus is acquired early in life,[43,44] and establishes low-level chronic infections in close to 100% of the healthy population with no evidence of disease.[45–47] The virus is therefore thought to make up a natural part of the commensal flora that exists within our bodies.[48]

On the side of the host, there is mounting evidence that inter-individual variation plays a major role in the outcome of infection. Transcriptional analysis of patient genes following influenza virus infection has revealed striking differences between symptomatic and asymptomatic groups.[49] Furthermore, a genome-wide association study (GWAS) demonstrated a correlation between mutations in certain immune-related genes (LARGE and IL-21) and regions of endemicity for the highly pathogenic Lassa virus,[50] suggesting that evolution of the immune response in these regions has resulted in decreased pathogenicity. Overall, understanding why differences in the immune response can cause severe disease in one individual but be entirely asymptomatic in another could contribute to greatly reducing the immense burden of viral-driven morbidity and mortality worldwide.

1.2 The Anti-Viral Immune Response

1.2.1 Immune Recognition of Viral Infection

The initial response of the immune system to viral infection relies on the recognition of conserved 'viral signatures' that are common in virus genomes but rarely found in the host. The broad detection of such a diverse group of pathogens is achieved by receptor molecules known as pattern recognition receptors (PRRs).[20] The vast majority of virus specific PRRs detect viral nucleic acids, as these molecules are produced within the host cell as a necessary stage of viral replication and so cannot easily be hidden from the hosts' pathogen sensors.[51] Viral genome signatures detected by PRRs include: long, double-stranded RNA molecules which are detected by the Melanoma Differentiation-Associated protein 5 (MDA5);[52,53] 5'-triphosphate uncapped single-stranded and double-stranded RNA molecules detected by the retinoic acid-inducible gene I receptor (RIG-I)[54]; and un-methylated CpG repeat sequences in DNA molecules recognized by the Toll-like Receptor 9 (TLR9).[55]

1.2.2 The Innate Immune Response to Viruses

When a virus is detected by a PRR, the initial antiviral response, known as the innate response, is launched. Unlike the adaptive immune system discussed below (section 1.2.3), the innate immune system is not antigen specific. This allows it to respond rapidly to infections, although the slower adaptive immune response is often required to completely clear the infection as many viruses have evolved mechanisms of evading the conserved, innate response.[56]

One major component of the innate immune response is the complement system, which is made up of a number of small proteins. This system is activated by

specific PRRs, predominantly mannose binding lectin (MBL), which binds to carbohydrate residues found on the surface of virus infected cells, as well as directly to the surface of some viruses, including HIV-1 and Ebola virus.[57] Once bound, MBL triggers a series of series of reactions known as the complement cascade, which culminates in the formation of pores in the surface of the virus or infected cell. Formation of these pores results in destruction of the virus or cell by lysis.[58] Certain protein cleavage products (C3a, C4a and C5a) produced during the complement cascade also amplify the immune response by triggering local inflammation, increasing vascular permeability and promoting the migration of macrophages and lymphocytes to the site of infection.[59]

Other proteins of the complement system, C3b in particular, when deposited on the surface of virions or infected cells are also able to enhance phagocytic uptake by binding to MBL in a process known as opsonization.[60,61] MBL can also interact with cellular receptors to enhance phagocytosis. The complement-MBL complexes on the surface of a virus therefore stimulate phagocytes (macrophages are particularly important) to engulf the tagged virus, process the viral antigens, and display them on the cell surface in a fashion required to activate the adaptive immune response.

As part of the innate response to viral infection, infected cells also begin producing small protein molecules known as cytokines which perform two major roles: 1) restriction of viral replication and spread by inducing an antiviral state in the infected cell as well as neighboring un-infected cells and 2) induction of the inflammatory response which results in the stimulation of the adaptive immune system and destruction of virus and virus infected cells.[62]

The induction of an antiviral state is largely dependent on a class of cytokines known as type-I inferferons, so named because they ‘interfere’ with viral

replication.[63] The importance of these molecules on the outcome of viral infections is well demonstrated by mice deficient in signal transducer and activator of transcription 1 (STAT1), a major protein involved in the IFN signaling pathway. These animals display extremely high morbidity and mortality rates when exposed to normally non-lethal viruses such as vesicular stomatitis virus.[64] Autonomous type I interferon signaling by an infected cell promotes the production of many different proteins that interfere with the viral replication cycle, including: oligoadenylate synthetase (OAS) with RNase L which degrades viral mRNA;[65] Protein Kinase R which inhibits translation of mRNA (viral and cellular);[66] and Mx, a GTPase which targets viral nucleocapsid proteins.[67] Paracrine signaling by type-1 IFN 'primes' neighbouring cells to express the above proteins to protect them from infection, preventing spread of the virus.[68]

All cells have the potential to be infected and therefore have the ability to respond to infection through the activation of intrinsic PRRs,[51,69] and to display viral antigens (peptide epitopes) on their surface, bound to a molecule known as the major histocompatibility complex class I (MHC class I).[70]

Natural killer (NK) cells are cells of the innate immune system responsible for the destruction of virus-infected cells, inducing apoptosis by releasing cytotoxic granules and of producing important cytokines (IFN-gamma and TNF alpha) early in the immune response.[71] They bridge an important gap between the innate and the adaptive immune response, as they are able to respond quickly and in the absence of antigen-presenting MHC class I. This is particularly important for defense against some viruses, which have evolved mechanisms to down-regulate MHC class I receptors as these molecules are required for the recognition of infected cells by CD8+ cytotoxic T-lymphocytes of the adaptive immune system (discussed in section 1.2.5.).[72] NK cells play a vital role against

the viruses of the *Herpesviridae* family in particular as demonstrated by deadly herpes virus infections in NK cell deficient mice.[73]

Cells of the monocyte lineage are the final effector cells of the innate immune system that I will discuss. Monocytes circulate in the peripheral blood and migrate to the tissues in response to cytokine signals released by infected cells. Upon reaching the site of infection, monocytes amplify and perpetuate the inflammatory response through the secretion of large amounts of type I IFNs.[74] Upon activation these cells are able to differentiate into macrophages and dendritic cells.[75]

Macrophages perform the role of phagocytosis, safely disposing of dying cells, pathogens and other debris by engulfing and enzymatically degrading them whilst preventing the release of any viral particles contained within.[76] Macrophages are present in the majority of tissues and have also been shown to protect vulnerable cells against viral infection by permitting themselves to be infected, soaking up viral particles, and producing high levels of IFN to prime nearby cells in response.[77] They are also important antigen presenting cells at the peak of the T cell response in infected tissues.

Dendritic cells are the primary link between the innate and adaptive arms of the immune system. They patrol the body via the blood and enter peripheral tissues to sample their environment for the presence of pathogens. Upon encountering a virus, the pathogen is ingested by the cell and broken down into small molecules known as epitopes, which are then assembled into MHC class I and II receptor-bound complexes to be presented on the cell surface. The presence of inflammatory cytokines such as type 1 IFN and TNF alpha cause maturation of the dendritic cell; signaling its exit from the tissue in afferent lymph to the draining lymph node and increased expression of MHC molecules and co-

receptor molecules (CD80, CD86, and CD40L) required for activation of naïve T cells. Therefore dendritic cells play a vital role in the stimulation of the adaptive immune response.[78]

1.2.3 The Adaptive Immune Response to Viruses

The adaptive immune system is antigen-specific and is so named for its ability to adapt to recognise any pathogen. Through a process of genetic recombination, the cells of the adaptive immune system produce an extremely diverse repertoire of receptor molecules.[79] These receptors make the adaptive immune system far more specific than the innate immune system and far more difficult to evade, as it is capable of detecting pathogens with any conceivable structural conformation and alter its response to changes in the viral population. Compared to the innate immune response, the adaptive response to a previously un-encountered pathogen is slow, taking days rather than hours to respond.[80] However, following infection, the adaptive immune system develops a 'memory' of the pathogen, allowing it to mount a strong, rapid response if it encounters the pathogen again. This mechanism forms the basis of protection by vaccination.[79]

The adaptive immune system is comprised of cells called lymphocytes. Lymphocytes can be divided into B-lymphocytes and T-lymphocytes based on the presence of their cell-surface receptors known as B-cell receptors and T-cell receptor respectively. The receptor diversity of these cells means that there is a risk of an immune response being mounted against the host itself. Therefore, these cells undergo a stringent selection process and activation procedure before they can mount an immune response.[81]

1.2.4 B-lymphocytes

During viral infection, B lymphocytes are primarily responsible for the clearance of extracellular viral particles. This is achieved through the B-cell antigen receptor (BCR), which is capable of recognizing and binding to antigenic regions of whole viral particles in their natural, three-dimensional conformation. Upon activation of the lymphocytes, the BCR molecules can be secreted in a soluble form known as immunoglobulins (or antibodies) which are able to neutralize virus in a highly specific manner by binding to their target and mechanically blocking entry into cells.[82] Phagocytic cells of the immune system are also able to recognize immunoglobulin-virus complexes, triggering virus uptake and processing of antigens to be presented by through the MHC class II presentation pathway.[60]

Immunoglobulins can be expressed in different forms known as isotypes, each of which performs a different role in the immune response. Initially all antigen-naïve B-cells express immunoglobulin isotypes M (IgM) and D (IgD) on their surface.[83] IgM is the first isotype to be secreted in response to initial exposure to a virus and is capable of neutralizing viruses and activating the complement system.[84]

Upon successful binding of their surface-bound BCR to an antigen, B-cells undergo activation, with the help of a T-helper cell (discussed in section 1.2.5.), and differentiate into either a plasmablast or a memory B-cell. Plasmablasts are short-lived cells that rapidly divide in a clonal fashion to produce many immunoglobulin secreting cells capable of binding the virus.[79] Following proliferation, plasmablasts either die or differentiate into plasma cells which are long-lived and capable of secreting large amounts of immunoglobulin. Memory B-cells do not secrete immunoglobulin, but rather enter the circulation and are

able to proliferate rapidly in response to re-exposure to antigen, forming the basis of immunisation.[85]

Following activation B-cells undergo a process known as class-switching which enables them to produce immunoglobulin isotypes G, A and E.[83] IgG is secreted in large amounts by plasma cells and makes up the vast majority of immunoglobulin found in circulation. This isotype has a greater affinity to antigen than IgM and is rapidly produced by plasma cells and memory B-cells in large amounts following re-exposure to the antigen.[86,87] IgA is also produced by plasma cells, however this isotype is predominantly expressed at mucosal surfaces which is discussed in a later section (1.2.6). IgE plays a specific role in the immune response against extracellular parasites and will not be discussed here.

1.2.5 T-lymphocytes

T-lymphocytes are primarily responsible for the destruction of intracellular virus through the killing infected cells and therefore are involved in the clearance of chronic viral infections. Like B-cells, the T-cell receptor is highly specific, however the receptor is only able to recognize viral peptide antigens when presented by major histocompatibility complex (MHC) molecules displayed on the surface of infected cells or antigen presenting cells.[88]

T-lymphocytes are classically divided into two groups based on the presence of the cell surface receptors CD4 and CD8. T-cells expressing the CD8 receptor (CD8+) are known as cytotoxic T-cells. Successful binding to the MHC-antigen complex results in activation of the cytotoxic T-cell followed by release of a cocktail of enzymes, which penetrate the infected cell and trigger the death of the cell through the apoptosis pathway.[89] Induced apoptosis of infected cells is a

form of cell death that results in the destruction of the cell and all of the viral particles contained within. Therefore, as opposed to lytic infection of a cell, no virus is released.[90]

CD4+ cells are known as T-helper cells and play an important role in coordinating the anti-viral immune response (Figure 1.3). The presence of activated T-helper cells is required for the activation and proliferation of both B-cells and cytotoxic T-cells.[79] The importance of T-helper cells is perhaps best demonstrated through the effects of HIV-1 which selectively infects CD4+ lymphocytes, depleting this population of cells.[91] The result is acquired immunodeficiency syndrome (AIDS), which is characterised by frequent, life-threatening infections by opportunistic pathogens.[92]

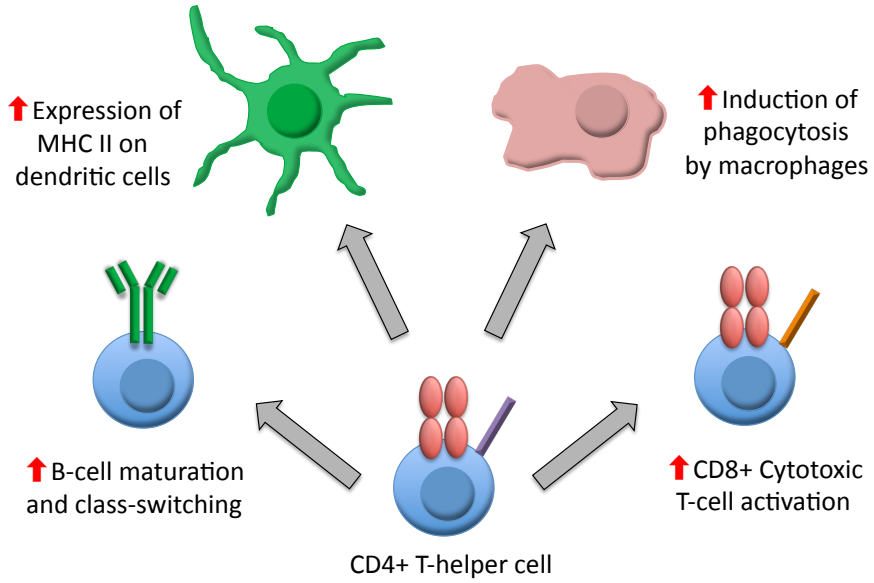


Figure 1.3. **The antiviral role of CD4+ T-helper cells.** CD4+ T-helper cells are responsible for the orchestration of various parts of the antiviral immune response including activation of B-cells and cytotoxic T-cell, as well as

stimulation of phagocytosis by macrophages. Adapted from Marshall NB and Swain SL (2011).[93]

1.2.6 Mucosal Immunity

As thin membranes, directly exposed to the environment, the mucosal surfaces are a major site of challenge by viruses and other pathogens. Mucosal immunity is a therefore a first defence against many viruses and a specialised arm of the immune system is responsible for its protection.

One mechanism of mucosal defence is the production of secretory immunoglobulins. Immunoglobulin A (IgA) is the most abundant immunoglobulin isotype in the body and two-thirds is in the form of a dimeric molecule with four antigen receptor sites, secreted by cells at the mucosal surface.[94] By binding viruses and bacteria, the immunoglobulin prevents their adherence to and penetration of the epithelial barrier.[95] Some individuals with selective IgA deficiency present with recurrent infections of the respiratory and gastrointestinal tract however most IgA-deficient patients do not display these symptoms.[94] Instead, they produce increased levels of a second immunoglobulin isotype, IgM, to compensate, demonstrating some redundancy between the two types.[96]

The other major component of mucosal immunity is the intraepithelial lymphocytes (IELs), a population of T-cells that reside in the mucosal tissue. IELs comprise of two cell types: tissue resident memory cells which are primed in the blood and migrate to the epithelia;[97] and double negative cells which migrate to the small intestinal epithelia directly following maturation.[98] IELs differ from other T-cells in that they are rapidly activated. Upon activation these cells generate a potent antiviral response, destroying infected cells and promoting

epithelial repair.[99] These cells also secrete high levels of interferon type I (α/β) and type III (λ). Type I interferon receptors are expressed by cells throughout the body, however interferon λ receptors are primarily expressed on the surface of epithelial cells, illustrating its important role in mucosal immunity.[100] Interferon λ upregulates the antiviral response, priming cells against infection by viruses through various pathways including Mx2.[101] Pre-treatment of mice with interferon λ has been found to protect against infection by a range of viruses that infect the gut and lung tissue including influenza virus, SARS coronavirus, respiratory syncytial virus, rotavirus and norovirus and has been used to resolve infection by dengue virus.[101–103]

1.2.7 The Major Histocompatibility Complex

The major histocompatibility complex (MHC) class I, and II molecules make up a group of proteins that play a vital role in the recognition of virus and other pathogens by the cells of the adaptive immune system. MHC class II is expressed predominantly by professional antigen presenting cells (APCs). Upon phagocytosis of a virus by an APC, broken down viral peptides are bound to the MHC class II complex and transported to the cell surface to be displayed. The recognition of viruses by T-helper cells relies on viral peptides presented in this manner. MHC class I molecules are present on all nucleated cells as well as erythrocytes. If a cell becomes infected, intracellular viral peptides are bound and displayed by MHC class I molecules in the same way as MHC class II. CD8+ cytotoxic T-cells recognise this as a marker of infection and trigger the death of the cell.[79]

As mentioned in section 1.1.5, host genetic background can have a strong influence on whether or not infection by a particular virus is successful.[104] A major source of variation between individuals can be seen in the form of the

human leukocyte antigen (HLA) genes that code for MHC molecules. Certain alleles of MHC class I have been found to determine the strength of immune response to hepatitis B virus vaccination,[105] whilst certain MHC class I and MHC class II alleles act as major determinants of whether hepatitis C virus infection persists or is cleared.[106] Humans have three main HLA loci (HLA-A, HLA-B and HLA-C) at which there are hundreds of variant alleles. Genetic inheritance of an allele of each HLA gene determines the structure of the MHC molecule peptide binding site and so the peptides presented to T cells, resulting in a great degree of diversity between individuals. The evolution of MHC heterogeneity is thought to have been at least partly driven by pathogens and protects communities from epidemic outbreaks of disease.[107]

1.2.8 Viral Evasion of the Immune System

Viruses reliance on the parasitism of cells to replicate has resulted in an evolutionary 'arms-race' between viruses and their hosts, each exhibiting a strong selective pressure on their counter-part.[108] Millions of years of co-evolution have led to the generation of our immensely complex immune system to protect our cells against viral infection resulting in components such as the MHC molecule,[109,110] interferons,[111] and restriction factors.[112] In response, viruses have evolved a range of mechanisms to counter the host immune response.[113]

One such mechanism that has evolved, particularly in RNA viruses, is an extremely high genomic mutation rate due to the enzymes responsible for genome replication (RNA dependent RNA polymerases) having a high rate of error.[114] These mutations give rise to the production of viral particles that the existing antibody response is incapable of binding. Driven by the selective pressure of antibodies, all virus species are constantly undergoing mutation at

the nucleotide and amino acid level, whilst B-cell populations must constantly adjust to produce antibodies capable of binding to new mutants. Furthermore, some viruses exist not as clonal populations but as quasispecies; genetically diverse populations which are able to evolve rapidly in response to selection pressure from the immune system and antiviral drugs.[115] Those viruses with segmented genomes, such as influenza, are also capable of a more drastic form of genome alteration known as antigenic shift.[116] This occurs when multiple viruses infect the same cell and mixed segments of the viruses are packaged together during replication to produce an entirely different subtype with a combination of surface antigens from the original, infecting viruses. Free viral particles of some viruses are able to further evade the antibody response through masking of highly conserved antigenic targets such as the masking of the gp120 surface protein by the envelope glycoprotein in HIV-1.[117]

Furthermore, evasion of the immune system continues once the virus has entered a cell. Virus genomes can interfere with signaling pathways crucial for viral detection such as the intracellular viral genome sensing pathways;[118] suppression of cell apoptosis [119,120] and down-regulation of T-cell activation pathways.[121]

Many viruses have not only developed ways to evade the immune system but also ways to weaken it, thus optimizing conditions for viral persistence and replication.[122] Cytokines are particularly common targets for viruses.[123] Upon infecting a cell, viruses such as HIV-1, hepatitis B virus and hepatitis C virus directly up-regulate anti-inflammatory pathways by stimulating production of cytokines such as IL-10.[122] Other viruses such as poxviruses and herpesviruses are able to produce proteins that mimic IL-10 with the same result,[124,125] and blocking of the IL-10 signaling pathway has been shown to lead to clearance of some viral infections.[126]

1.3 Immunodeficiencies

For individuals with functional immune systems, the majority of viral infections are short-lived and asymptomatic due to rapid clearance of the virus or suppression of the virus into a latent state. In immunosuppressed individuals however, these infections are more likely to be severe, recurrent, persistent and often life threatening.

Deficiencies of the immune system can be divided into two types: primary and secondary. Primary immunodeficiencies (PIDs) are caused by mutations in the genes coding for components of the immune system and therefore result in immune dysfunction from birth. Secondary immunodeficiencies (SID) are those acquired due to an event later in life. SIDs can be transient or permanent and are caused by a number of factors, some of which are discussed below. Importantly, individuals with both primary and secondary forms of immunodeficiency suffer from high rates of hospitalization due to severe and recurrent infectious complications and represent a significant economic burden to health-care services.[127–131]

1.3.1 Primary Immunodeficiencies

Genetic defects of the immune system are being identified at an increasing rate.[132] The immune system is complex and is governed by a large number of genes and a mutation in a single one of these genes has the potential to result in immunodeficiency. In recent years, large-scale, genome-wide association studies (GWAS) have been undertaken in an attempt to determine the genetic defects underlying a number of primary immunodeficiency diseases.[133,134] This approach has been particularly useful in identifying mutations associated with complex diseases such as common variable immunodeficiency (CVID).[135] CVID is a

particularly variable form of humoral immunodeficiency for which multiple genetic mutations have been implicated (Table 1.1). However, the mutations that have so far been identified account for only 10-15% of all CVID cases.[136] Thus, the disease is the result of a number of different mutations which all manifest in a similar, antibody deficient, phenotype.

Gene	Role	Impact of mutation
BAFF-R[137]	Receptor for B-cell survival and maturation	Impaired B-cell development and reduced numbers of mature B-cells
CD19[138]	B-cell co-receptor for antigen recognition (forms a complex with CD21, CD81, and CD225)	Impaired B-cell response to antigen stimulation
CD21[139]	B-cell co-receptor for antigen recognition (forms a complex with CD19, CD81, and CD225)	Disrupts CD19 complex formation leading to an impaired B-cell response to antigen stimulation
CD81[140]	B-cell co-receptor for antigen recognition (forms a complex with CD19, CD21, and CD225)	Disrupts CD19 complex formation leading to an impaired B-cell response to antigen stimulation
ICOS[141]	Co-receptor for B- and T-cell interaction, and for T-cell stimulation and proliferation	Reduced proliferation of T-cells, impaired B-cell activation and class switching
ITPKB[142]	Regulation of signalling associated with B-cell survival and development	Impaired B-cell development, survival and activation
LRBA[143]	Involved in transport of vesicles containing activated receptor complexes	Impaired B-cell activation, autophagy and survival
TACI[144]	B-cell survival, maturation and antibody production	Impaired B-cell proliferation and class-switching

Table 1.1. Genetic mutations known to be associated with approximately 10-15% of CVID cases. The genes listed are predominantly associated with the B-cell development pathways, but also includes genes in other pathways such as *ICOS,* which plays a role in the interaction of T-cell and B-cells required for B-cell activation.

Primary immunodeficiencies fall broadly into two categories: those of the adaptive immune system and those of the innate immune system. Adaptive immunodeficiencies affect lymphocytes and can therefore result in humoral immunodeficiencies such as CVID, T-cell deficiencies such as severe combined immunodeficiency (SCID) due to IL2R deficiency or a combined deficiency of both B- and T-lymphocytes such as SCID due to RAG1 deficiency, an enzyme involved in the recombination of both the T- and B-cell receptor genes in the early stages of lymphocyte development[145]. The cause of an adaptive immunodeficiency can occur at any point in the lymphocyte development and differentiation pathway (Figure 1.4). Therefore, some are identifiable due to lack of a certain cell type whereas others may have normal cell counts but these cells are not functional as they are unable to respond to stimulation in the normal way. CD4+ T-cell deficiencies can be particularly severe, as the cells of this lineage are responsible for activation of both the B-cell and cytotoxic T-cell responses (Figure 1.3).

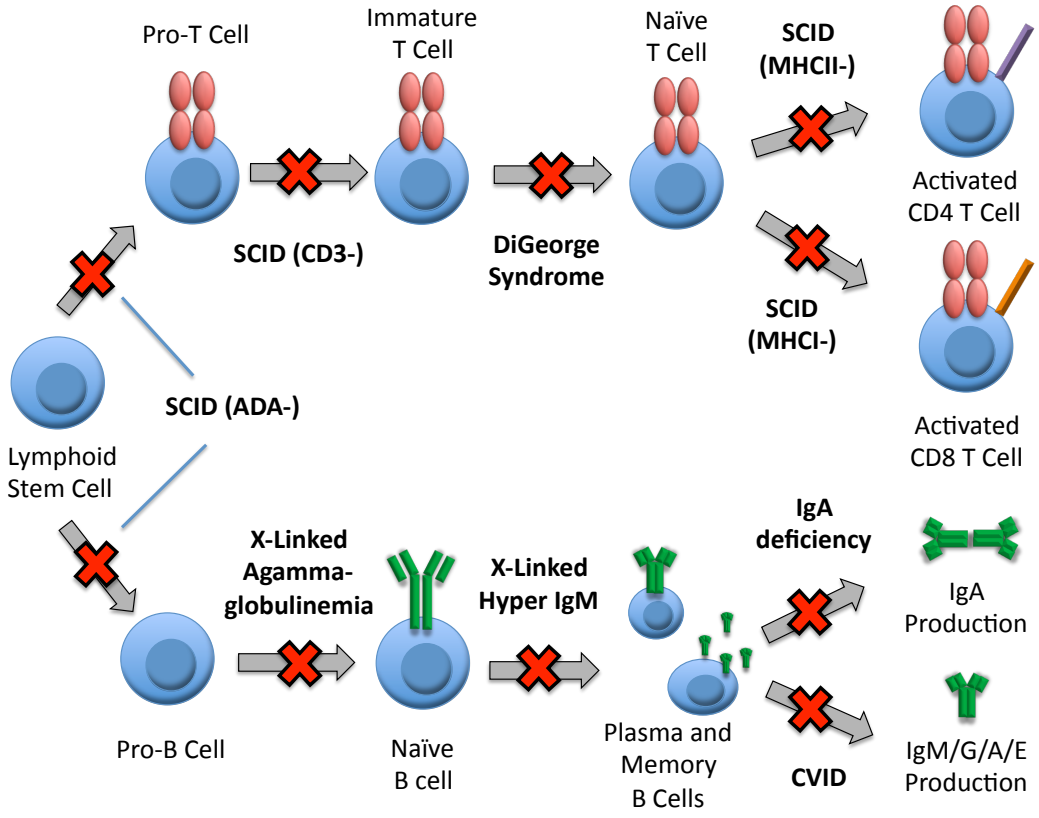


Figure 1.4. **Primary immunodeficiencies of the adaptive immune system.** Defects early in the lymphocyte development pathway can lead to combined B- and T-cell deficiencies known as severe combined immunodeficiency (SCID), (e.g. SCID caused by adenosine deaminase (ADA) deficiency). Depending on the point in the T-cell development pathway which it affects, a T-cell deficiency can result in loss of all T-cell subtypes (e.g. SCID caused by T-cell antigen receptor (CD3) deficiency) or specific subtypes such as T-helper cells (CD4+) or cytotoxic T-cells (CD8+). Similarly, defects in the B-cell development pathway can lead to deficiency in all antibody classes (e.g. XLA) or specific subclasses (e.g. CVID). Figure adapted from en.wikipedia.org/wiki/Primary_immunodeficiency.

Innate immunodeficiencies can affect the cells of the innate immune system such as NK-cells and phagocytes but also include defects in pathogen detection and signalling pathways such as interferon and Toll-like receptor deficiencies.

Primary immunodeficiencies are generally thought of as rare, however a 2007 study by Boyle and Buckley estimated that roughly 1 in 1200 people in the USA have been diagnosed with some form of PID.[146] The investigators suggested that this is likely to be an underestimation of the true number due to those people living with an undiagnosed PID. Lack of clinical awareness and mild, recurrent symptoms means that it is not uncommon for individuals suffering with some forms of PID, particularly those from minority groups, to go undiagnosed and continue to suffer from related health problems.[147]

1.3.2 Secondary Immunodeficiencies

Secondary immunodeficiencies can occur through a range of mechanisms including malnutrition, aging, environmental toxins, and infection by pathogens.[148–150] Some SIDs are also induced as an unintended side-effect of treatment for diseases such as cancer, or through intended therapeutic suppression of the immune system to treat patients in whom a strong immune response is undesirable such as transplant recipients and those with autoimmune disease.[151,152]

In the past decade solid organ transplantation has become a common treatment option for a range of life-threatening diseases. Organ transplant recipients are put on strong, lifelong, immunosuppressant drugs following their transplant in order to prevent immune-mediated rejection of the organ. NHS statistics reported performing over 4600 solid-organ transplant operations during the 2015/2016 financial year and an additional 6400 patients were on the list,

awaiting donors.[153] Therapeutic immunosuppression is also used to treat a large number of inflammatory diseases such as systemic lupus erythematosus, multiple sclerosis and vasculitis. Immunosuppressive regimes for these diseases are also generally life-long as the effects are palliative rather than curative. No studies have been undertaken to specifically assess the prevalence of autoimmune diseases in the UK, however a 2010 meta-analysis estimated an overall prevalence of between 7.6 – 9.4% for 29 auto-immune diseases in the USA, based upon the results of previous epidemiological studies performed in Europe and North America,[154] and it appears that the prevalence of autoimmune diseases is increasing.[155] Furthermore, the effects of therapeutic immunosuppression are often broad, affecting multiple lymphocyte populations (Figure 1.5), which can result in a marked susceptibility to infection.[156,157]

Increasing life spans combined with falling birth rates mean that the world's population is aging.[158] Within the next 10 years the proportion of the UK population over 65 is expected to exceed those under 15 years of age for the first time.[159] In the UK, almost 7,000,000 patients aged 65 and over were vaccinated against seasonal influenza between September 2016 and January 2017 as they were identified as at risk of severe infection.[160] It has long been recognised that infections and autoimmune diseases occur more frequently in the elderly. This is primarily due to a decline in immune function termed immunosenescence.[161,162] Studies of immunosenescence have demonstrated that cells of both the innate and the adaptive immune system are affected including lymphocytes, regulatory cells and NK cells.[163–166] This deterioration has been linked to a number of 'age-related' viral diseases such as West Nile virus disease and the mousepox virus model in aging mice,[167] suggesting that viral infections have a greater impact on health as the immune system declines.

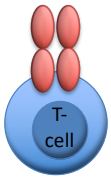
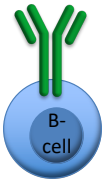
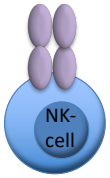
	Mechanism of Suppression				
Cell	Cyclosporin A	Tacrolimus	Prednisolone	Azathioprine	Mycophenolate
	↓ Activation / Cytokine Transcription [168]	↓ Activation / Cytokine Transcription [168]	↓ Activation / Cytokine Transcription [168]	↓ Proliferation / DNA Synthesis [168]	↓ Proliferation / DNA Synthesis [168]
	↓ Proliferation (Direct) [169]	↓ Proliferation (Indirect only) [170]	↓/↑ Variable [171]	↓ Proliferation [171]	↓ Activation and Proliferation [172]
	↓ Cytotoxicity and Proliferation [173]	↓ Cytotoxicity and Proliferation [174]	↓ Cytotoxicity and Proliferation [172]	↓ Cytotoxicity [175]	↓ Cytotoxicity and Proliferation [172]

Figure 1.5. **Reported effects of common immunosuppressive drugs.** Common immunosuppressive drugs used in transplantation have a broad effect across multiple immune cell subtypes including T-cells, B-cells and NK-cells. References: Ferron and Jusko (1998),[168] Hannam-Harris et al (1985),[169] Heidt et al (2010),[170] Tareyeva et al (1980),[171] Eickenberg et al (2012),[172] Meehan et al (2013),[173] Morteau et al (2010),[174] Cseuz and Panayi (1990).[175]

Many cancer patients are also immunocompromised due to the use of chemotherapy, which has a profound impact on the immune system as it targets quickly dividing cells of which leucocytes are a part. According to the systemic anti-cancer therapy (SACT) dataset, approximately 197,000 patients are currently receiving cancer chemotherapy through the NHS in England as of 2017.[176]

Finally, secondary immunodeficiencies can also be acquired through infection by a pathogen. The most infamous example is HIV-1 which depletes the T-helper

cell population, leading to a severe form of combined immunodeficiency.[177] HIV-1 infects approximately 96,000 people in the UK.[178] This number is rising as modern combination antiretroviral therapy (cART) has led to a growing number of survivors of both horizontal and vertical transmission.[179]

One aspect shared by every one of the immunodeficiencies described above, is that their prevalence is increasing, be it due to increased awareness and diagnosis of primary immunodeficiencies, increasing use of therapeutic immunosuppression, or ageing of the population. Considering this together with reports of transient immunosuppression in healthy individuals due to common events such as pregnancy,[154] stress,[155] previous infection by unrelated pathogens,[148] lack of sunlight,[180] time of day,[181] and gender,[182] makes it clear that the interaction between our immune systems and viruses needs to be better understood.

1.4 Viral Infections in Immunodeficient Individuals

All deficiencies of the immune system clinically manifest as a reduced ability to combat infections. However, they exist in a spectrum of form and severity. Susceptibility to viral infections has classically been associated with lymphocyte deficiencies, particularly deficiencies of CD8+ cytotoxic T-cells. However, a large number of other cell types contribute both directly and indirectly to the anti-viral response. The impact of opportunistic viral infections in immunodeficient individuals is best documented in PIDs, where the identity of the infectious agent can provide clues as to the underlying genetic defect.

1.4.1 Viral Infections in Primary Immunodeficiency

The compromised immune state suffered by immunodeficient patients puts them at increased risk of disease caused by pathogens that are innocuous in a healthy host. Some primary immunodeficiencies result in infection by a diverse range of pathogens, whereas others predispose individuals to severe infections from a specific pathogen only, with no other outward signs of immune dysfunction.[183]

An example of the latter is a mutation in the human TLR3 gene, a component of the interferon induction pathway, which results in herpes simplex encephalitis (HSE) caused by the ubiquitous HSV1. Case-studies of TLR3 deficient individuals suggest that the gene is redundant in the immune response against most pathogens but vital in the protection of the central nervous system against HSV1 resulting in severe inflammation of the brain.[184,185]

Another example of this is X-linked agammaglobulinemia (XLA). XLA is a humoral immunodeficiency caused by a mutation in the BTK gene which contributes to B-cell development.[186] XLA patients predominantly suffer from recurrent bacterial infections but have also been shown to have an increased susceptibility to enterovirus infection resulting in hepatitis and encephalitis which is often fatal.[187] It is interesting to note that XLA patients do not suffer infections by the highly prevalent Epstein-Barr virus, as the virus infects B-cells, which these patients are lacking.[188]

Severe combined immunodeficiency (SCID) is at the opposite end of the primary immunodeficiency spectrum. SCID is characterised as a defect in T-cell development pathways, which generally has a knock-on effect on other lymphocyte lineages such as B-cells and NK-cells.[189] The fact that this deficiency compromises multiple immune compartments means that these patients are

susceptible to a wider range of viruses. Indeed, a severe course of disease in SCID patients has been reported for a number of viruses including: respiratory syncytial virus,[190] parainfluenza virus,[190] paramyxovirus,[191] adenovirus,[191] Epstein-Barr virus,[192] rotavirus,[193] and astrovirus.[194] Overall, primary immunodeficiencies can often result in an increased susceptibility to viral infections. However, the precise nature of the deficiency is a major determinant of whether this susceptibility is to a single specific virus or to a range of virus species.

1.4.2 Viral Infections in Secondary Immunodeficiency

Similarly to those with SCID, individuals with secondary immunodeficiencies are at risk to a range of viral infections. This is particularly true for those with SIDs acquired through chemotherapy and immunosuppressive drugs, or brought about by aging, as these forms of immunodeficiency are non-specific and generally affect multiple immune cell populations (e.g. Figure 1.5.), which results in a broad gap in immunity. A recent report on emerging viral infections in the transplant setting, illustrates this point by citing multiple reports of newly acquired infections by a number of viruses including: hepatitis E virus; coronavirus; measles virus; and lymphocytic choriomeningitis virus.[195] Acquired viral infections also represent a significant cause of morbidity and mortality in chemotherapy patients and elderly adults.[196,197] In one US-based study, influenza virus and respiratory syncytial virus accounted for an average of 43 hospitalisations and 32 deaths per 1000 care home residents annually;[198] and meta-analysis study demonstrated a significant increase in both risk of death and risk of admission to hospital following seasonal influenza infection in those over 65 compared to younger individuals.[199]

Many of the infections mentioned above, are not uncommon in the healthy population, however the outcome tends to be far more severe in patients with secondary immunodeficiencies compared to healthy individuals. Perhaps the most obvious example of this is the occurrence of AIDS in HIV-1 infected individuals, in whom the development of immunodeficiency leads to life-threatening disease caused by everyday pathogens.[92] Indeed, human herpesvirus 8 (HHV8) infection of HIV positive patients results in the development of Kaposi's sarcoma, one of the defining diseases of AIDS.[200] Whereas HHV8 infections in healthy individuals are predominantly asymptomatic.[201]

Furthermore, the development of secondary immunodeficiencies in later life means that, unlike primary immunodeficiencies, patients frequently suffer from complications associated with the re-activation of viral infections that were acquired prior to the onset of immunodeficiency. Indeed, until the development of effective antiviral drugs which are now used as prophylaxis, reactivation of herpesvirus represented a major cause of morbidity and mortality in populations such as solid organ transplant recipients and the elderly.[202,203] The reactivation of previously resolved hepatitis B virus infections is also well documented in a range of immunosuppressed populations including patients undergoing chemotherapy treatment,[204] solid organ transplant recipients,[205] and the elderly.[206]

1.5 Virus Diagnostics

The first stage required to treat any form of infectious disease is diagnosis. Viral diagnostic assays typically detect either the presence of the virus itself, indicative of current infection, or the presence of an existing immune response to the virus, indicative of past or on-going exposure.

1.5.1 Viral Culture

Early attempts at viral diagnosis relied on the culture of virus *in vitro* through the inoculation of cultured cells with clinical samples and observation for cytopathic effect by microscopy over the following days. Immunofluourescent staining or electron microscopy would also be used to provide further confirmation of viral presence.[20] Viral culture is now rarely used for laboratory diagnosis as it is slow and lacks sensitivity and specificity compared to more modern nucleic acid and serology based techniques. Further disadvantages include the fact that culture assays require viable (replication competent) viral particles, which is not always possible when samples cannot immediately be stored appropriately, such as those taken in the field. Many viruses, such as HBV and HCV, have also proven difficult to culture using conventional methods.[207,208] Indeed, difficulty in culturing HBV was a major factor in hindering the discovery of the aetiologic agent of "catarrhal jaundice".[209]

1.5.2 Enzyme-Linked Immunosorbent Assay

Enzyme-linked immunosorbent assay (ELISA) is currently one of the most commonly used laboratory-based methods for viral diagnosis. The assay uses monoclonal antibodies capable of binding to a specific viral antigen, or to host antibodies that have themselves bound to the antigen.[210] The presence of the target is detected through the conjugation of enzymes to the monoclonal antibodies. After incubation with a sample from the patient, un-bound monoclonal antibodies are washed away and a substrate of the conjugated enzyme is added. If the target of the assay is present, the substrate is catalysed in the presence of the enzyme to produce a visible signal, the intensity of which is measured as a measure of target concentration.

Many diseases of unknown viral aetiology are thought to be caused by acute 'hit-and-run' infections, where the virus itself has been cleared from the host but symptoms of disease remain.[211] Therefore, ELISAs designed to detect an antibody response to a specific virus hold an advantage over many other diagnostic assays such as the commonly used polymerase chain reaction (PCR) (described below), as they do not require the virus to be present at the time of sampling. Instead, these assays indicate whether a patient has previously been exposed to a virus. IgM-specific ELISAs can be used to detect signs of recent exposure to antigen (2-3 months), however, the results of all ELISAs must be treated with caution, as past exposure cannot easily be implicated as the cause of disease in the present.

1.5.3 Polymerase Chain Reaction

The development of nucleic acid-based tests (NATs) simultaneously increased the speed, sensitivity and specificity of viral diagnostic methods. The polymerase chain reaction (PCR) is the prototypical NAT. Developed in the late 1980's, the method employs two synthetically manufactured oligonucleotide molecules known as primers, which are designed to specifically bind to the viral genome molecule by conventional base-pairing and instigate the amplification of a portion of the viral genome by the Taq DNA polymerase enzyme.[212] Following multiple rounds of DNA amplification which exponentially increase the amount of DNA present, the DNA is visualised by gel electrophoresis. Taq polymerase uses DNA as a template and therefore cannot directly detect the many viruses with RNA genomes. Therefore, for RNA virus detection, a technique known as reverse-transcription (RT) must be used to first transcribe the RNA into DNA prior to PCR amplification.[213]

PCR is highly sensitive, as it theoretically requires the presence of only one genome molecule in order to detect a virus. However, in reality factors such as the visualisation of PCR products and the imperfection of primer design restrict this. Further benefits of PCR over viral culture include its ability to be easily tailored to detect any virus or gene by altering the primers used as well as the fact that, unlike virus culture methods, the virus does not need to be replication competent to be detected.

1.5.4 Quantitative PCR and Digital PCR

Quantitative PCR (qPCR) is an adaptation of the classical PCR approach and uses fluorescent dyes to directly measure the amount of DNA amplification occurring in a PCR reaction. The resulting intensity of fluorescence is directly proportional to the amount of DNA template in the starting sample and can be used to compare the amount of virus in different samples. The actual number of viral genome copies present in a sample can also be calculated through comparison to the fluorescence generated by serial dilutions of a DNA standard of known concentration. Therefore qPCR is extremely useful for applications where changes in viral load are relevant, such as monitoring a HIV positive patient's viral load in response to therapy.[214]

Two forms of qPCR are commonly used for viral diagnosis: SYBRGreen and Taqman probe, both of which work following the same basic principles as conventional PCR, however both also generate a measurable fluorescent signal when PCR amplification occurs. SYBRgreen makes use of a fluorescent dye which intercollates with double-stranded DNA (dsDNA) molecules, causing an increase in fluorescent intensity as dsDNA molecules are produced by the PCR reaction.[215] The Taqman probe approach employs an additional oligonucleotide molecule known as the probe, which binds downstream of the primers. Attached to either

end of the probe are a fluorescent dye molecule and a quencher molecule that absorbs any light emitted from the dye whilst the two are in close proximity. The Taq polymerase enzyme cleaves the dye molecule through its exo-nuclease activity as it passes along the DNA strand, freeing the dye from its proximity to the quencher molecule and allowing it to fluoresce.[216] SYBR green assays are generally cheaper to perform as they do not require the manufacture of the Taqman probe however non-specific amplification and primer-dimer formation can generate false positives as the dye binds to any dsDNA molecule. Non-specific signal can however be identified by dissociation curve analysis.[217] The use of a probe can increase the specificity of the reaction but also reduces the assay's ability to detect variants of a virus. As well as its ability to quantify virus, qPCR is faster and generally more sensitive than traditional PCR as there is no need for an extra step to visualise the product and the results can be read in real-time for rapid diagnosis.

Digital PCR (dPCR) is a recently developed modification of the qPCR approach which works by performing thousands of individual qPCR reactions on a single sample. The sample is diluted so that each reaction contains either a single or zero viral genomes.[218] By calculating the number of positive to negative reactions and employing Poisson statistics to account for reactions that may contain greater than one viral genome, the viral load can be calculated without a standard. Foregoing use of a standard curve to estimate copy number has the ability to improve inter-laboratory variation in quantitation results. Furthermore, by performing many reactions on a single sample dPCR has the potential for greater sensitivity than that of qPCR and capable of detecting virus at low levels which have been suggested as significant in some patient groups.[219] Despite these factors, the greater cost of performing multiple reactions means that dPCR is unlikely to replace qPCR for everyday diagnostic assays.

1.5.5 Multiplex qPCR and Microarrays

One downfall of the methods described so far is that *a priori* knowledge or expectation of a specific virus target is required for the diagnosis to be meaningful (particularly regarding the design of PCR primers). By performing a single reaction containing multiple probes with dyes that emit light at different wavelengths, multiplex qPCR can be used to simultaneously detect several viruses at once. This is particularly useful when diagnosing patients with symptoms typical of multiple viruses such as hepatitis,[220] or when a single primer set is not sufficient to detect every virus sub-type.[221]

Multiplex qPCR is restricted by the need for primer sets with similar thermo-stabilities.[222] They are also generally limited to the detection of less than ten targets, due to limitations in the wavelengths of light that are detectable by the machine and also the number of available dyes that can be distinguished by the wavelength that they emit.

Taqman low-density arrays (TLDAs) and microarray technology have vastly increased the number of viruses that can be targeted by a single assay. ViroChip, a microarray capable of targeting over 1500 viruses simultaneously is particularly noteworthy.[223] Both TLDA and microarray assays work by attaching oligonucleotides, which are capable of complimentary binding to viral genomes, directly to a plate. TLDAs work by using the oligonucleotides as primers for a qPCR reaction,[224] whilst microarrays detect the hybridisation of fluorescently labelled DNA or cDNA to the oligonucleotides.[225] Both approaches are still limited by *a priori* expectation, as oligonucleotide probes must be designed to specifically bind the virus genomes of interest. Primer design for TLDAs is further complicated by the fact that all qPCR reactions must be optimised and validated to work at the same temperature. Both assays are also often expensive

as the production of plates is a specialised process and there is a trade-off between the number of targets on a single plate and the number of samples that can be run. Furthermore, microarrays generate a large amount of data which requires investment in computational infrastructure, informatics knowledge and time to analyse.

1.5.6 Viral Metagenomic Sequencing

Next generation sequencing (NGS) technologies have provided a means of overcoming the bias of sequence specific PCR-based diagnostic assays. By sequencing all of the DNA or RNA present in a sample, a method known as metagenomic sequencing, any virus genome can be detected with no need for primer design or the *a priori* knowledge of pathogens that may be present. This method can be used as an independent diagnostic assay, or as a tool to assess representative populations and provide guidance for the faster and cheaper directed diagnostic methods described above. Advances in metagenomics have been largely facilitated by the falling cost and increasing power of NGS over the past decade as technologies have improved, due in large part to the '1000 Genome Project'.[226]

Metagenomics-based viral diagnostics has facilitated the discovery of many new viruses,[227] as well as aided in the diagnosis of diseases caused by unsuspected viral pathogens,[228] and resolved the presence of co-infections.[229] However, whilst the method clearly has great potential as a tool for identifying viruses in patient samples, it also has several limitations. The foremost of these limitations is a lack of sensitivity, particularly when compared to targeted diagnostic assays such as (RT)-qPCR. An example of this flaw was recently observed by a research team responding to the 2015-16 South American Zika virus epidemic. The team performed metagenomic sequencing on clinical samples that had previously

tested positive for Zika virus by qRT-PCR. However, despite using a method designed to enrich for viral sequences (ribosomal RNA [rRNA] depletion), clinical samples containing lower levels of virus (10-48 genome copy equivalents per µl of extracted RNA) produced very few Zika specific reads (mean = 0.005% of total reads, range 0-0.011%). Hence no detectable viral reads were were detected in some samples, while others produced only low sequence coverage of the virus genome of interest (mean = 26.32%, range 0-79.2%).[230]

The genomes of human viruses are many orders of magnitude smaller than that of their host. Therefore, unless viral copy number is high (e.g. during the acute stage of infection), direct sequencing of nucleic acid extracted from an infected tissue sample will predominantly produce sequences derived from the host and very few from the infecting virus. Viral enrichment techniques, such as the rRNA depletion method mentioned above, have therefore been developed in order to address this imbalance. Ribosomal RNA is well conserved and highly abundant in cells, making it a relatively easy target for subtraction methods. In general, these methods aim to deplete host nucleic acid prior to sequencing in order to increase the overall proportion of reads generated from the infecting virus. By doing this, the sensitivity of the approach can be improved and the chance of generating sufficient reads to reconstruct the viral genome is increased.

Early methods of virus enrichment were developed before NGS was available. At the time they were used in combination with cloning and Sanger sequencing to produce viral sequences. However, the advent of NGS greatly increased their potential in terms of throughput. Representational difference analysis (RDA) was one of the first viral enrichment methods described.[231] RDA is a hybridisation-based method for detecting differences between two nucleic acid populations and was successfully applied in the discovery of a novel herpesvirus (human herpesvirus 8) from the lesions of AIDS patients.[232] The method works by

extracting total nucleic acid from the affected tissue of a patient with the disease of interest. The RNA in the extract is reverse transcribed and subjected to second-strand synthesis to make double-stranded DNA (dsDNA). An oligonucleotide adapter is then ligated to either end of the dsDNA molecules. This adapter-bound DNA is then mixed with an excess of DNA prepared from unaffected tissue of the same individual. The mixture is heated to denature the dsDNA molecules and cooled to allow re-annealing of complimentary sequences. This results in three distinct types of dsDNA molecule: 1) those consisting of one DNA strand derived from the affected tissue and another from the control tissue (e.g. the molecule is present in both diseased and healthy tissues); 2) those consisting of both strands from the control tissue and; 3) those consisting of both strands derived from the affected tissue. The latter group should be enriched for DNA molecules that are present exclusively in the diseased tissue and not in the healthy control tissue. Furthermore, this is the only group containing dsDNA with the adapter molecule attached to both strands. PCR amplification is performed on the resulting mixture, using a primer directed to the oligonucleotide adapter, enriching for these molecules. This method can be repeated multiple times to further enrich the sample for 'disease-specific sequences' before cloning or sequencing. The primary drawback of this method is the need for an unaffected control sample, usually from the same patient, which is not always possible. This requirement also means that non-solid tissue samples such as plasma or urine are not viable.

Virus discovery cDNA-amplified fragment length polymorphism (VIDISCA) is another method that has been successfully used to identify a previously unknown virus infecting a patient.[233] This method again begins by extracting total nucleic acid from the affected sample and performing reverse transcription and second-strand synthesis to convert RNA into dsDNA. The dsDNA is then subjected to restriction digest reactions using enzymes with common

recognition sites which are therefore likely to be present in viral molecules. Restriction enzyme specific 'anchors' are then ligated to the end of the digested products and used as primer-binding sites for PCR amplification. The resulting PCR fragments are then compared to those produced by an uninfected control, and those specific to the infected sample are cloned and sequenced.

Both RDA and VIDISCA were vital tools in the early stages of viral discovery. However, noticeably, both methods require well-matched uninfected tissue as a control and this resource is rarely available. Furthermore, the methods are complex, time-consuming, and low throughput, and are therefore not well suited to clinical settings. Instead, the most widely used viral enrichment methods are those that are simple, fast and affordable for most laboratories. Reported methods typically include a combination of: centrifugation and filtration of samples to remove bacteria and host cells;[234] nuclease degradation of un-protected host DNA (the viral genomic material is protected by the viral capsid); [235] and specific depletion of highly abundant host ribosomal RNA by hybridisation.[236]

Studies comparing these viral enrichment methods have yielded contrasting results. Rosseel et al (2015) tested combinations of sample filtration, DNase treatment, and rRNA depletion on a single RNA virus culture, artificially spiked into lung homogenate and serum.[237] The authors reported that a combination of DNase and rRNA depletion generated the greatest proportional increase of viral sequence reads from serum, whilst rRNA depletion alone performed best for the lung sample. Indeed, DNase treatment of the lung sample actually decreased the depth of sequencing coverage of the virus. These findings suggest that it is difficult to apply a single method to any given sample type. Furthermore, extrapolation of this result to other viruses and sample types may be unreliable as the method was only tested on a single RNA virus.

A further study by Hall et al (2013) tested the effect of 3 viral enrichment methods: low-speed centrifugation, filtration through a 0.45μm syringe filter, and nuclease (DNase and RNase) digestion on a selection of 3 viruses: two with RNA genomes (influenza A virus and enterovirus) and one with a DNA genome (adenovirus).[238] Virus cultures were mixed with aliquots of *E. coli* and human cell cultures to mimic a clinical sample. Each of the enrichment methods tested led to a reduction in both host and viral copy number, as determined by (RT)-qPCR. However, following sequencing the authors concluded that a combined 3-step treatment led to the greatest increase in viral reads as a proportion of the total reads produced. The proportion of enterovirus-specific sequences was increased from 0.2 % to 4.7 % of total reads by using the 3-step enrichment method. However, the increase in influenza virus sequence reads from 0.001 % to 0.0113 % was less impressive, and the proportion of adenovirus reads actually decreased from 0.002 % to 0.0006 % suggesting this method is not suitable for the identification of DNA viruses. Interestingly, despite being a part of the 3-step protocol, the use of nucleases alone led to a decrease in influenza virus reads. Furthermore, adenovirus read output was reduced even when using simple methods such as centrifugation and filtration, both of which one would not expect to specifically impact viruses with a DNA genome.

The comparisons of viral enrichment methods described above are largely focused on RNA viruses. RNA viruses are frequently the target of viral metagenomic studies because they are considered the viruses with the most "epidemic" potential. However, a truly unbiased metagenomic approach would have the ability to detect both RNA and DNA viruses simultaneously.

One study created a pool of 25 RNA and DNA viruses representing 9 distinct viral families to test the effect of filtration and nuclease treatment.[239] Both forms of enrichment led to a reduction in the number of viral sequences generated, as

well as a reduction in the total number of virus species detected. It should however be noted that this study was merely observing the effect of these methods on virus recovery. The material being tested did not contain any human cells and therefore these results cannot necessarily be extrapolated to predict how the method would perform in samples with high levels of background (e.g. clinical samples).

A further study using infected tissue samples aimed to develop a protocol for unbiased detection of both RNA and DNA viruses.[240] A range of published methods for tissue homogenisation, virus enrichment, nucleotide extraction, and PCR amplification were tested on tissue from embryonated chicken eggs, which had been experimentally infected with 1 of 4 viruses: 2 with DNA genomes (T3 reovirus and vaccinia virus) and 2 with RNA genomes (influenza A virus and Sendai virus). Each method was scored based on whether it lead to an increase or decrease in the proportion of viral to host sequences. A maximum of 4 points was awarded when the method proportionately increased the virus to host ratio for all 4 virus species. Three virus enrichment methods scored the maximum of 4 points, these were 1) purification of virions on a sucrose gradient followed by further ultracentrifugation to pellet the virions; 2) DNase digestion of host nucleotides; and 3) filtration through a 0.2 μm filter (interestingly, an alternative method using a larger 0.45μm filter scored -2 points). Another notable result was the score of only 1 point for rRNA depletion, a method commonly used in metagenomics studies.

Considering the results of the above studies, it is clear that, despite these methods having the potential to improve metagenomic sequencing sensitivity, the improvements are not universal for all viruses and may come at a cost to the unbiased approach depending on the virus and sample type. Furthermore, these methods can only partially deplete non-viral genetic material. Therefore,

sequencing of enriched samples still results in the production of predominantly non-viral reads.[238]

During the course of this thesis, a new approach named VirCapSeq was described, which appears to be a promising tool for vastly reducing background sequences and improving the sensitivity of viral metagenomic sequencing. The method employs approximately 2 million biotinylated oligonucleotide probes, which are mixed with the NGS library prior to sequencing, to specifically hybridise virally derived DNA.[241] The viral DNA - probe complex can then be retrieved from solution by using streptavidin-coated beads to bind the biotin-labelled probes, enriching the library for viral sequences. This library of oligonucleotide probes was computationally designed using a database of coding sequences representing all the members of viral taxa known to infect vertebrates. The 100 nucleotide-long oligomers were designed to bind at regular intervals with 25 - 50 bp overlaps to cover the entire coding region of viruses. The study reported that use of the array for enrichment resulted in a 100 to 1000-fold increase in depth of genome coverage, across the coding regions of a range of RNA and DNA viruses, compared to a standard virus sequencing approach (filtering and nuclease treatment). Despite being designed based on 207 viral taxa, this approach may still introduce bias against undiscovered viruses not represented in the sequence database, as the probes are designed on known viral sequences. However, to account for natural sequence variation, the authors included variants with greater than 10 % diversity. Production of this array is expensive if used for large, multi-sample, viral metagenomic studies. However, by reducing the depth of sequencing required to generate adequate viral coverage, higher levels of multiplexing are possible, thus reducing the cost per sample.

Ironically, one of the major strengths of metagenomic approaches is also one of its key weaknesses. Its non-specificity means that the approach is highly susceptible to contamination from both other samples and the laboratory environment, which can result in false-positive results. Contaminating viruses have been detected in a range of common laboratory reagents: a novel parvovirus-like virus was detected in nucleic acid purification columns;[242] a complete genome of bovine viral diarrhoea virus was generated from low level contamination of commercial foetal bovine serum, a reagent commonly used in cell culture methods;[243] and there have also been a number of reports of trace contamination by murine leukaemia virus in samples that had undergone reverse transcription[244] (likely due to the fact that reverse transcriptase enzymes are engineered from murine retroviruses).

One particularly interesting case of potential contamination involves the first ever report of a novel giant virus, giant blood Marseille virus (GBM), in human blood.[245] The virus was detected in a single blood sample when performing metagenomic sequencing of plasma from 10 asymptomatic donors in France. The presence of the virus was confirmed by PCR, electron microscopy, *in situ* hybridisation and virus culture methods. This result was surprising as relatives of this virus primarily infect amoeba and had never previously been isolated from any animal species. The same group went on to use PCR to screen the blood of a further 174 blood donors and 22 patients who had received multiple blood transfusions from southern France, reporting a GBM prevalence of 4% and 9.1% respectively.[246]

A second group wished to test these findings and recruited 339 patients including: 150 healthy blood donors, 50 patients who had received multiple blood transfusions, and another 139 blood donors, recruited from the southern region of France where the high prevalence of GBM was reported in the previous

study.[247] Using the PCR assay described in the initial report, this group was unable to detect a single positive patient, leading them to propose laboratory contamination as a possible reason behind the earlier results.

The initial group have since followed up with a report of GBM in the blood of a child suffering from adenitis, and further identified the virus was present in the lymph node by *in situ* hybridisation and immunohistochemistry.[248] They also showed that the virus was detectable in the lymph node of a woman diagnosed with Hodgkin's lymphoma (again by PCR, *in situ* hybridisation and immunohistochemistry).[249]

Interestingly the original authors suggest that these giant viruses are being missed by the majority of human viral metagenomics studies due to enrichment methodologies such as filtration which exclude larger structures.[250] However, reproducibility is important for novel findings, and until these results can be confirmed in other laboratories it is difficult to exclude contamination as the reason why this virus is being detected.

The potential of contamination is a major issue for NGS studies, particularly when using lower amounts of input RNA/DNA such as those generally seen when working with clinical samples.[251] Several methods have been suggested as a way to reduce the risk of false-positive results including the use of a negative control when preparing and sequencing samples,[252] setting a read count threshold below which sequences are not considered for analysis,[253] and confirming sequencing results by PCR of the original sample.[254] None of these methods are perfect individually and reliance on a single one can potentially lead to false-negative results. However, combined with appropriate sample handling techniques, unidirectional workflows and cleaning routines, they can dramatically reduce the potential for contamination.

Overall, an array of viral diagnostic methods is available, with variations in speed, sensitivity, specificity and cost. Next-generation metagenomic sequencing is a powerful tool for the detection of viruses in populations such as immunodeficient patients, where a wide-range of pathogens may be suspected, as no *a priori* assumptions are made as to what pathogens are present. However, it is expensive compared to current clinical diagnostics and the sensitivity of this approach is low, relying on clumsy enrichment methods to allow the generation of sufficient viral sequence reads for use. Furthermore, even if sufficient viral reads are produced, the method is then still reliant on suitable informatic methods and databases to identify these sequences as viral (discussed in chapter 3, section 3.1.7).

1.6. Conclusion and Aims

In summary, viruses are important human pathogens capable of causing disease and mortality through a wide-range of mechanisms. Immunodeficiencies weaken an individual's defences against all pathogens including viruses, which often result in severe and persistent infections. However, the routine, clinical investigatory method for viruses is typically restricted towards a small number of suspected pathogens and, once an agent is detected, it typically does not take the possibility of further sub-clinical infections or co-infections into account. Therefore, the true prevalence and impact of many of viral infections is likely to be underestimated. In particular, sub-clinical infections of immunodeficient populations require investigation as evidence suggests that they may play a key role in seeding new disease outbreaks.

The aim of this research was therefore to investigate the prevalence and effects of viral infections in different immunocompromised populations, based on the following three objectives:

1. To optimise an unbiased metagenomic sequencing method for determining the prevalence of viruses in clinical blood samples.

2. To survey the prevalence of viral infections in the blood of patients across a range of immunodeficiency types.

3. To determine whether any of the detected viruses are associated with specific clinical outcomes or disease.

Chapter 2

Methodology

2.1. Nucleic Acid Extraction and Preparation from Plasma

Total nucleic acid was extracted from 200 µL of plasma using the High-Pure Viral Nucleic Acid Kit (Roche, Basel, Switzerland) according to manufacturer's instructions and eluted in 40 µL of elution buffer.

An overview of the method used to prepare nucleic acid for sequencing is shown in Figure 2.1. The RNA contained in the nucleic acid extractions was reverse transcribed in three reactions per sample using 5 µL of total nucleic acid per reaction as input. 50 ng of random hexamer primers and 1 µL of 10 mM dNTPs were added and volume was adjusted by adding RNase-free water to a volume of 13 µL. The mixture was heated to 65°C for 5 minutes to denature secondary structure and anneal the primers. 200 U of Superscript III enzyme (Thermo Fisher Scientific, Waltham, Massachusetts, USA), 40 U of RNase OUT (Thermo Fisher Scientific, Waltham, Massachusetts, USA), 1X Superscript III Reaction buffer, and 1 µL of 0·1M dithiothreitol were added to the reaction mixture as advised in the enzyme manufacturers protocol, to a final volume of 20 µL. Reverse transcription was performed by incubating at 25°C for 5 minutes to extend the random primers, followed by cDNA synthesis for 50 minutes at 50°C. The reaction was terminated by heating to 94°C for 3 minutes, and then cooled to 4°C.

Second-strand synthesis of cDNA and single-stranded DNA was performed by adding 2·5 U of large klenow fragment (New England Biolabs, Ipswich,

Massachusetts, USA) to each reaction and incubating at 37°C for 60 minutes followed by 75°C for 10 minutes to inactivate the enzyme.

Reactions containing double-stranded DNA were combined for each sample to give a volume of 60 µL. The DNA was purified using AMPure XP beads (Beckman-Coulter, Brea, California, USA) and re-suspended in 30 µL of Tris-EDTA buffer (10mM Tris, 1mM EDTA, pH 8.0) (Life Technologies, Carlsbad, California, USA). A bead to sample ratio of 1·8 : 1 was used to recover the greatest possible range of DNA species whilst removing small fragments. Quality and quantity of double-stranded DNA was determined using the Bioanalyser 2100 high sensitivity DNA kit (Agilent, Santa Clara, California, USA) and Qubit high sensitivity dsDNA kit (Thermo Fisher Scientific) before library preparation.

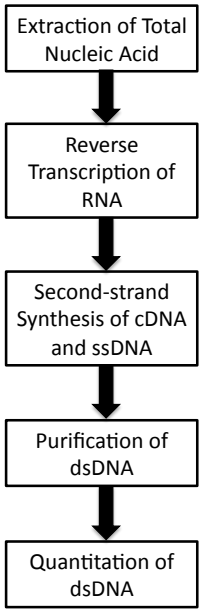


Figure 2.1. **Pre-library preparation workflow for total nucleic from plasma.** Total nucleic acid (RNA and DNA) was extracted from 200 µl of plasma. The combined nucleic acid extract was subjected to reverse-transcription to

transcribe RNA into cDNA, and second-strand synthesis to convert single-stranded DNA/cDNA into double-stranded DNA. Double-stranded DNA was then purified and quantified for library preparation.

2.2. Illumina Library Preparation and Sequencing

Library preparation and Illumina sequencing was performed by the Wellcome Sanger Institute unless otherwise stated. 100-200 ng of DNA was fragmented to ~ 500bp by sonication using the LE220 sonicator (Covaris Inc., Woburn, Massachusetts, USA). Individual Illumina sequencing libraries were prepared and indexed from the fragmented DNA using the Sure Select library preparation kit (Agilent, Santa Clara, California, USA). Index tagged samples were then amplified using the KAPA HiFi kit for 6 cycles (Roche, Basel, Switzerland), quantified using the LabChip GX 1k assay (Perkin Elmer, Waltham, Massachusetts, USA), and pooled together in an equimolar fashion to be multiplexed 48X on the Hi-Seq 2500 (Illumina, San Diego, California, USA), generating 100-125 base-pair, paired-end reads.

2.3. Informatic Analysis of Illumina Sequence Data

Sequences of viral origin were identified using the workflow depicted in Figure 2.2. Paired-read files were imported into CLC Genomics Workbench v 7.5.1 for analysis (Qiagen, Hilden, Germany). Low quality reads were trimmed using the CLC trimming algorithm with the recommended default threshold of 0.05. The algorithm uses a modified version of Mott's trimming algorithm to remove low quality bases from the ends of reads, by calculating the probability of an erroneous base-call based on Phred quality scores.[255] If a read contained > 1 ambiguous base-call, consisted entirely of low complexity repeats (e.g. polyA/T), or less than 50 bases (33% of the read) remained following trimming, the entire

read was discarded to avoid non-specific mapping in the steps to follow.

To speed analysis, sequence datasets were next reduced in size by digital subtraction, mapping reads to the Homo sapiens reference genome (GRCh37 build, grch37.ensembl.org/), the human CCDS database (GCF_000001405.26, NCBI annotation release 106, Ensembl annotation release 76, www.ncbi.nlm.nih.gov/ projects/CCDS),[256] and an additional database of human mitochondrial and ribosomal RNA sequences. Mapping was performed using the optimal subtractive parameters: requiring matches of 90% similarity over 80% of the read length as described by Daly et al.[257]

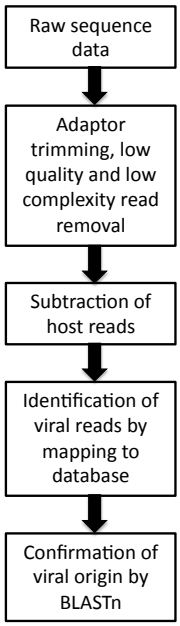


Figure 2.2. **Overview of the informatic workflow for the detection of viral reads in Illumina sequencing data.** Raw sequence data was imported into CLC genomics workbench where low quality and low complexity reads were removed. The remaining reads were first mapped to the human reference genome to remove host sequences and then mapped to the viral reference

database. Reads successfully mapped to the viral database were interrogated against the complete non-redundant nucleotide database by BLASTn to confirm their viral identity.

Reads that remained un-mapped to the human reference dataset were interrogated by mapping to a custom database of virus genomes containing 1420 genomes generated from the NCBI RefSeq database (www.ncbi.nlm.nih.gov/genome/viruses). This database was constructed by manual removal of viruses that exclusively infect non-vertebrate hosts (eg. viruses of the order *Caudovirales*, the family *Virgaviridae,* and the genus *Cypovirus,* which exclusively infect bacteria, plants and arthropods respectively) (see Appendix for a list of viral genomes included in the final database). Reads were mapped to the vertebrate virus database using a match threshold of 80 % similarity over 80 % of the read. These parameters meant that a maximum divergence of 36% between sequenced viruses and the reference genomes was allowed for. The high level of viral genome variability even within clades means that it is possible that more lenient mapping parameters could have led to the identification of further viral sequences. However, parameters that are over-lenient can result in large numbers of sequences being incorrectly identified as viral, particularly when dealing with short reads such as those produced by Illumina sequencing. The resulting mapped reads and contiguous sequences were confirmed as viral by BLASTn alignment to the non-redundant nucleotide database (https://blast.ncbi.nlm.nih.gov/ Blast.cgi).[258] BLAST results were sorted by e-value and sequences whose best match was either non-viral or ambiguously viral/non-viral were excluded.

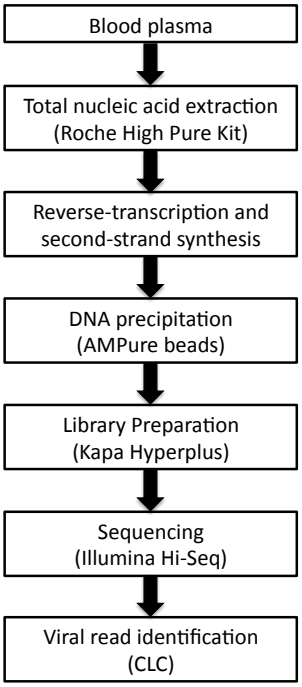


Figure 2.3. **Overview of the complete method to be used for viral metagenomic sequencing of blood as determined by the experiments described in chapter 3.** Plasma was selected over PBMCs for sequencing due to its propensity to result in greater viral coverage. This is likely due to the lower abundance of host genetic material in plasma. The high pure viral nucleic extraction kit was selected for nucleic acid extraction due to its combined use of chemical and enzymatic denaturation methods which has been shown to increase viral nucleic acid recovery. RNA and single-stranded DNA was converted to double-stranded DNA by reverse transcription followed by second-strand synthesis, however no PCR amplification method was to be used so as to avoid amplification bias. AMPure XP beads were selected to purify double-stranded DNA prior to library preparation as the method is comparable to ethanol precipitation with the benefit of removing small (<100 nt) DNA fragments. The Kapa hyperplus kit was selected for library preparation as it is

suitable for low amounts of input DNA (< 100 ng), which are produced when PCR amplification is foregone. Illumina sequencing platform was selected as it produces the greatest number of reads which is likely to benefit the metagenomic approach being taken. Finally, it was determined that the CLC mapping algorithm was most sensitive for the detection of viral reads.

2.4. Contamination and Confirmation of Results

To reduce chances of cross-contamination between samples and from the environment, all manipulations of samples and reagents were performed wearing disposable gloves, and each stage of sample preparation was performed in contained cabinets, following a linear workflow. Cabinets, work surfaces and equipment were cleaned regularly between uses with 10% bleach, or DNase, and RNase-removal solutions. Total nucleic acid extraction was performed in a class II biosafety cabinet contained within a room dedicated to nucleic acid extraction and the manipulation of non-amplified nucleic acid. Reaction mastermixes for reverse transcription, second-strand synthesis and PCR assays were prepared in a dead-air PCR cabinet, containing dedicated pipettes and fitted with a UV light for sterilisation between uses in order to avoid contaminating the reagents. PCR products were handled exclusively in a post-PCR room.

Despite the measures described above, potential cross-sample contamination was thoroughly investigated. Typical patterns of contamination such as the same virus occurring in multiple consecutive samples in a lane were confirmed by nucleotide alignments and PCR testing of the original sample. Detected contamination events are described in each individual chapter when applicable.

2.5. Anellovirus Quantification

Anellovirus was quantified by dual-labelled qPCR using the primers and probe described by Vasilyev et al [47] (Table 2.1) and the QuantiTect Virus + ROX Vial Kit (Qiagen, Hilden, Germany). Triplicate reactions contained: 2 µL of nucleic acid; 5 µL of QuantiTect Virus Mastermix (5X); primers to a final concentration of 0·4 µM and; dual-labelled probe to a final concentration of 0.2 µM. RNase-free water was added to a final volume of 20 µL.

Name	Sequence
TTV sense	5'-CCT GCA CTT CCG AAT GGC TGA GTT-3'
TTV antisense	5'-CTT GAC TGC GGT GTG TAA ACT CAC-3'
TTV probe	5'-FAM-CCT CCG GCA CCC GCC CTC GGG ACG-BHQ1-3'

Table 2.1. **qPCR primer and probe sequences used for the quantitation of anellovirus**, targeting the conserved UTR region of the genome. Primer and probe sequences were designed by Vasilyev et al.[47]

The assays were performed on the Rotor-Gene 6000 (Qiagen) following the manufacturer's recommended 2-step cycling with a combined annealing/extension at 60°C for 40 cycles. Gain settings were optimised automatically prior to first acquisition and fluorescence data was acquired during the annealing / extension step of the cycling. Threshold levels were set manually.

PCR product from a positive sample was purified by agarose gel electrophoresis, extracted, quantified using the Qubit HS dsDNA kit (Thermo Fisher Scientific) and serially diluted for use as a standard. The assay's limit of detection was

designated by the concentration at which detection of the standard was stochastic (equivalent to 1000 copies / mL of plasma).

2.6. Statistical Methods

Basic statistical comparisons throughout all chapters of this thesis included: paired t-tests to compare the means of two paired groups; Fisher's exact tests to compare proportions between two groups; Mann-Whitney U tests to compare the medians of two groups; McNemar's tests to compare proportions between two paired groups; Wilcoxon-signed rank tests to compare medians between two paired groups; linear regression analysis to test for a relationship between two or more variables; ANOVA to compare the means of more than two groups and Tukey's multiple comparison test to make individual group comparisons; and Mantel-Cox tests to compare the survival distribution between two groups. Analyses were implemented in R v.3.3,[259] (McNemar's test and linear regression analysis) and GraphPad Prism v.5.0 for Mac OSX (GraphPad Software, La Jolla California USA, www.graphpad.com).

Chapter 3

Development of a Metagenomic Sequencing Method for the Detection of Blood-Borne Viruses

3.1. Introduction

Following the completion of the human genome project 15 years ago, next generation sequencing (NGS) has become an instrumental tool in practically all areas of biology.[260] The impact of metagenomic sequencing on the field of virology is perhaps best demonstrated by the rate of novel viruses discovered since the release of the first NGS machines in 2008, as the development of these machines was accompanied by falling costs and greater accessibility to sequencing technology (Figure 3.1). Some further examples of viral metagenomic sequencing success include: its use complementing routine clinical diagnostics;[261,262] the identification of novel virus species in cases where routine diagnostics have failed;[263–266] the safety screening of vaccines for adventitious pathogens;[267] and monitoring of changes in the virome that are associated with diseases such as inflammatory bowel disease and AIDS.[268–270]

The ability to sequence at increasing depths has allowed researchers to begin focusing on entire communities of organisms without culturing them, in a method known as metagenomics. Of particular relevance to this study is the impact of metagenomics on our understanding of the communities of bacteria, yeast, fungi and viruses that inhabit the human body. In its totality, this community is known as the microbiome, which includes the most variable and poorly defined component, the 'virome'.[271]

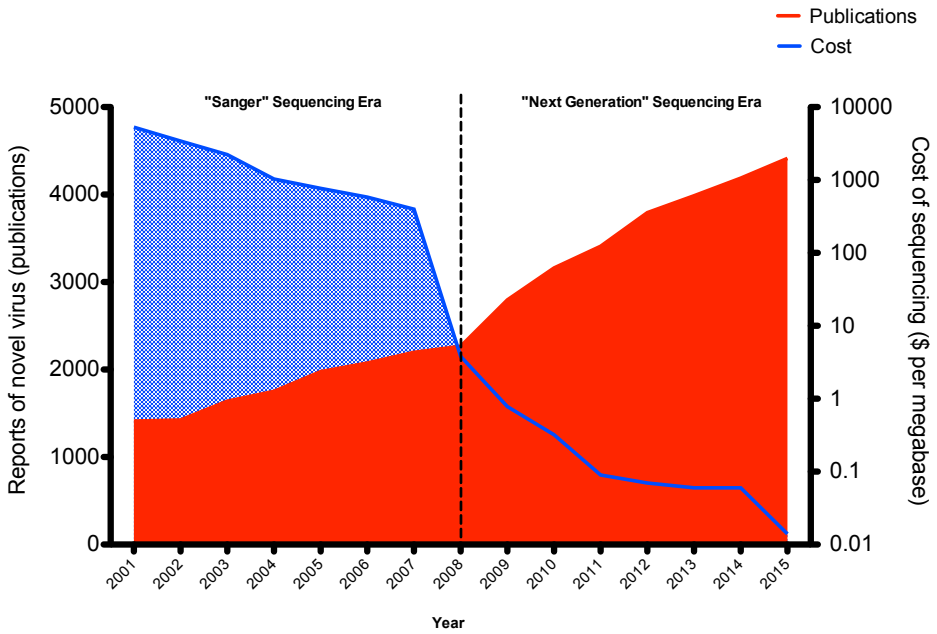


Figure 3.1. **Rate of viral discovery and sequencing costs in the "Next Generation" sequencing era.**

The release of next generation sequencing technology in January 2008 coincided with a fall in the cost of sequencing and a rise in reports of novel viruses. Publication number was calculated using the search term 'novel virus' in pubmed. Cost of sequencing data was obtained from an NIH report written by Wetterstrand (2016).[342] As costs were reported multiple times per year, the months of September and October were used.

As discussed in the introduction to this thesis, the major benefit of metagenomics is that it can be used to detect viruses without *a priori* knowledge regarding which species are likely to be present. However, the approach lacks sensitivity compared to PCR-based methods. To account for this low sensitivity, it is essential to optimise the steps of the metagenomic pipeline, as each represents a potential bottleneck that can result in sample loss (Figure 3.2).

3.1.1 Sample selection

Sample selection is an important factor when designing a metagenomic study, as it determines which viruses it will be possible to detect. The life cycle of viruses means that many can be found at high levels in the blood, as they use the circulatory system as a means of transport to and from the tissue in which they replicate. Furthermore, some viruses, such as EBV and HIV-1, infect the blood cells themselves.[272,273] Unlike viruses found in the digestive or respiratory tracts, those found in the blood have crossed the epithelial barrier. Therefore, their presence is indicative of systemic infection and less likely to be coincidental. Finally, in the clinical setting, blood samples are routinely collected from patients to be used in a range of diagnostic tests. Repositories of these samples provide researchers with access to cohorts of patients that would take years to recruit prospectively, stored in the form of frozen plasma and peripheral blood mononuclear cells (PBMCs).

Metagenomic studies examining the blood virome have only recently begun to be published and therefore methodologies are extremely variable. Whilst the majority have used blood plasma,[39,245,274,275] some have opted to sequence PBMCs or whole blood.[276,277] The majority of methods have tended to mimic early metagenomic studies of the faecal virome by focusing on DNA viruses,[39,245,275,278] therefore excluding the large number of viral families with

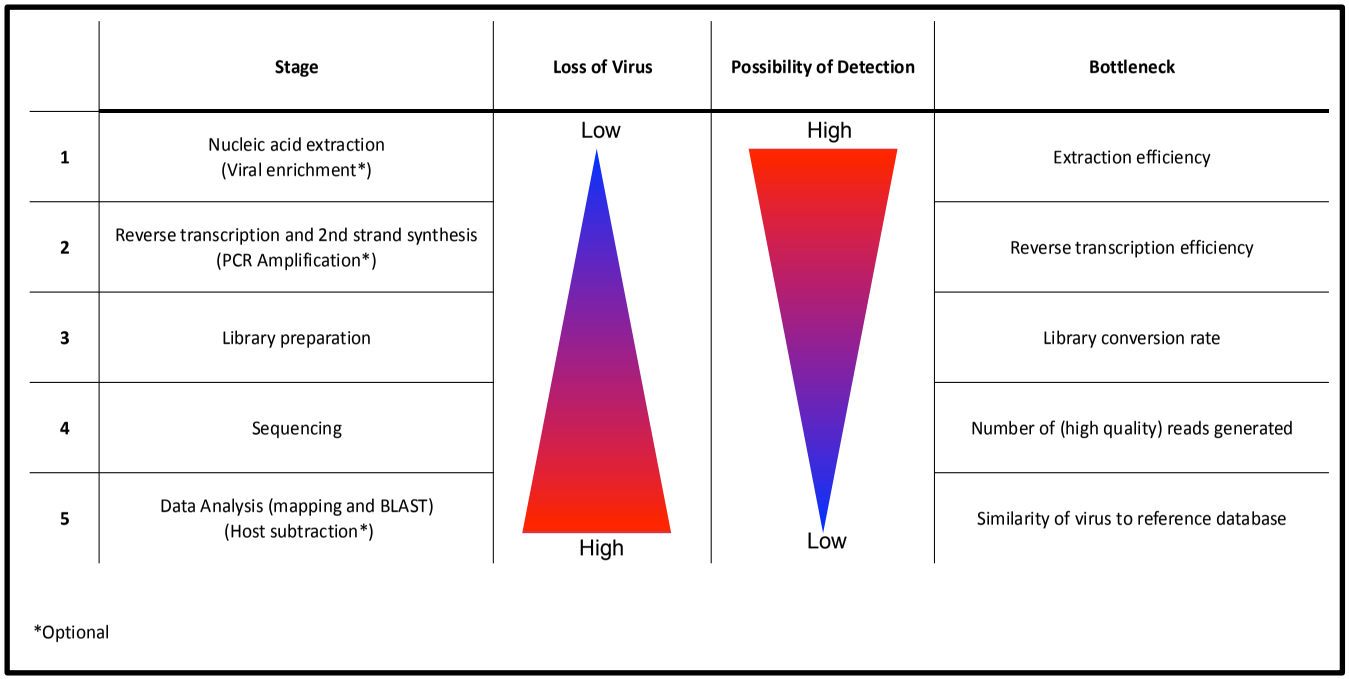


Figure 3.2. **Sample loss bottlenecks in the metagenomic workflow**.

Overview of the viral metagenomic sequencing methodology stages at which loss of sample can occur. Each stage represents a potential risk of sample loss due to the bottleneck described. As more sample (including virus) is lost, the likelihood of detecting low abundance species such as viruses is decreased. Therefore, it is necessary to optimise each stage to maximise sensitivity. Stages marked as optional are those used in some published studies but not in others.

RNA genomes. Conversely, others have excluded DNA entirely and focused exclusively on RNA viruses.[279,280] All of these methodologies follow a similar workflow but vary in terms of sample type, method of nucleic acid extraction, use of viral enrichment and PCR amplification, and method of viral sequence identification. At present no attempt has been made to compare these methods, and therefore the optimal approach for viral metagenomics of blood samples has been unclear.

3.1.2. Nucleic acid extraction

Following the collection of samples, nucleic acid must be extracted from within cells and viral capsids before it can be sequenced. Choice of nucleic acid extraction method can have a major effect on the efficiency of viral nucleic acid recovery and therefore sequencing success. It has previously been shown that extraction methods which combine both chemical and enzymatic denaturation of proteins perform best, while use of carrier RNA is also beneficial.[281] For these reasons, the high pure viral nucleic acid kit (Roche Diagnostics) was selected for this study, as this method uses both a chaotropic agent and proteinase K for protein disruption, and a poly(A) carrier.

3.1.3. Viral enrichment

Nucleic acid extraction results in the recovery of RNA and DNA from all organisms present within a sample e.g. virus, bacteria, and host. The abundance of non-viral sequences can easily mask the proportionally low level of virus; making virus nucleic acid difficult to sequence directly. Furthermore, unlike bacteria, viruses do not share homologous genomic regions, such as the bacterial 16S ribosomal gene, across families. Therefore, targeted amplification and sequencing of a single (or even multiple) gene(s) to detect all virus species is not

feasible. Instead, a 'brute-force' approach is required to generate as many reads as possible and increase the likelihood of detecting low abundance species.

To this end, methodologies have been developed which remove host and bacterial genomic material in order to 'enrich' samples for viral sequences.[238] Common approaches include: filtering,[282] selective exclusion of highly abundant host species such as rRNA,[283] nuclease degradation of unprotected host DNA and RNA,[235] and hybridization probe-based capture of viral genomes[284]. These methods have all been shown to improve the overall virus to host ratio in some instances however it has also been suggested that the most commonly used treatments such as filtering and nuclease digestion, followed by PCR amplification, can bias against the detection of some viruses.[285] For example, filtration-based methods, which work by excluding larger organisms such as host cells and bacteria, can also result in loss of some of virus species with larger capsid structures, such as the herpesviruses and mimivirus.[286] Nuclease-based viral enrichment methods can also have a variable effect depending on the virus in question. One study reported that nuclease treatment resulted in a significant increase in the proportion of reads generated for adenovirus, influenza and poliovirus, whilst having a negligible effect on the proportion of reads generated for CMV.[287] Furthermore, nuclease-based methods generally result in the recovery of only a few picograms of DNA, particularly when working with samples which contain low concentrations of DNA to begin with, such as plasma. Therefore, use of this method often relies on random PCR amplification post-enrichment in order to generate sufficient material for library preparation and sequencing. Such amplification methods have been shown to introduce bias in metagenomic studies, primarily due to differences in GC-content between species, which can mean that some viral genomes are preferentially amplified and therefore over-represented in the final sequencing results at the expense of other species.[285,288] Therefore, several enrichment methods were tested to

determine whether viral read number and coverage results could be improved without negatively affecting the recovery of one or more virus species.

3.1.4. PCR amplification

Sequence-independent single primer amplification (SISPA) is a PCR-based method developed for the random amplification of unknown DNA templates,[289,290] and has been used in many viral metagenomic studies, particularly those for viral discovery, in order to increase sensitivity.[235,282,291,292] However, SISPA has also been found to preferentially amplify particular templates, introducing bias intro libraries and leading to mixed genome coverage.[293,294] Alarmingly, one study found that the use of different amplification methods prior to metagenomic sequencing results in the detection of entirely different species of viruses.[288] Furthermore, PCR amplification bias can lead to high levels of duplicate reads and PCR artefacts,[293] which may further dilute virally-derived DNA in the sequencing libraries and reduce sensitivity of the approach. For these reasons, it was decided to not amplify DNA prior to library preparation for this study.

3.1.5. Library preparation

Following nucleic acid extraction, RNA and single-stranded DNA must be converted to double-stranded DNA, and oligonucleotide adapters must be ligated to the double-stranded DNA molecules, in preparation for sequencing. A range of commercial kits is available for Illumina library preparation. For this study the KAPA Hyperplus kit (Roche) was selected. For the purpose of viral metagenomics, this kit has a number of advantages over its competitors including: 1) superior library conversion efficiency and library yield which is

extremely important for the detection of rare species and accurate representation of metagenomes;[295] 2) the kit is able to accept a wide-range of DNA input amounts (1ng – 1 μg), which negates the need for random amplification prior to library preparation; and 3) the kit allows for library preparation (from fragmentation to adapter ligation) to be completed in a single-tube without the need for clean-up steps, which can improve DNA recovery and allows for lower input requirements compared to many commercial kits. It achieves this by using a fragmentase enzyme, which also has a lower GC bias compared to tagmentation methods, allowing for more uniform sequencing coverage, and single-tube preparation as opposed to sonication. The fragmentase enzyme also produces blunt-ended fragments which do not require an additional end-repair step prior to adaptor ligation. Furthermore, a high library conversion rate means that PCR amplification-free workflows are possible even when a low DNA input is used (e.g. 100 ng). This is useful as some polymerases can introduce artificial mutations and amplification bias.[296,297]

3.1.6. Choice of sequencing platform

Studies have found that choice of sequencing platform does not have a major impact on the make-up of viral metagenomes.[298,299] Therefore, the main factors to consider when selecting a platform are: the number of reads required, read length, read accuracy and cost. At time of writing, the vast majority of recently published viral metagenomics studies have opted for the Illumina platform,[39,274,275,300] due to its superior sequencing output and low error rate compared to its main competitors: Roche 454 and ION torrent.[301,302] Illumina sequencing produces a huge number of reads, which are short compared to those produced by competing technologies. However, Illumina libraries can also be prepared to generate 'paired' reads from either end of a single DNA molecule which somewhat compensates for their short length when performing tasks such

as *de novo* assembly.[303] Overall, the Illumina platform suits metagenomics studies, which tend to favour a higher number of reads to increase the potential for detection of rare species. Furthermore, the Illumina platform allows multiplexing which can reduce the costs of studies with large numbers of samples.

3.1.7. Informatic identification of viral sequences

Finally, the sequences produced by metagenomic sequencing must be identified computationally. A single lane of the Illumina HiSeq 2500 can produce files containing up to four billion paired-end reads (> 1 TB of data),[304] and so a number of computational pipelines and methodologies have been developed with the aim of rapidly identifying viral reads with high accuracy.[305] Despite the development of the aforementioned pipelines, the vast majority of published viral metagenomic studies have tended to use independently validated methods for viral read identification.[39,280,306,307] This is most likely due to the desire to freely modify and optimise the pipeline as required, and is made possible by the fact that the necessary software packages such as BLAST are free to acquire.[258] Clearly, as with sample preparation, informatic methodologies are varied. Despite this, most take the approach of first reducing the dataset to a more computationally manageable size by removing the contaminating host reads, followed by alignment of the remaining sequences to a database of viral genomes.[257,308,309] Major differences between the most commonly used methods include the use of *de novo* assembly to create contiguous sequences, the databases selected for alignment, and the method and stringency of alignments.

The downside to this approach is that identification of viral reads is clearly dependent on successful alignment to the database. If the database does not contain a closely related reference genome, identification of the viral read is

impossible. Translation of the nucleic acid sequences into amino acids for alignment using tBLASTx can account for some of the high level of variability in viral genomes but still requires viral reads to be somewhat related to those in the database.[310]

For this reason, attempts have been made to develop non-alignment-based methods of viral identification. These tend to rely on recognising 'features' specific to viral genomes such as the frequency of particular oligomer sequences, or the order of genes within cassettes.[311,312] These methods do have the potential to be used for the detection of unknown viruses, distantly related to those in the database. However, their sensitivity is not comparable to alignment-based methods. This is likely due to the fact that, unlike bacterial and mammalian genomes, many viral genomes are too short to produce meaningful k-mer frequencies. Furthermore, viruses, particularly those with DNA and RNA genomes, are thought to be descendants of multiple discrete lineages,[313,314] making them difficult to classify using any method which relies on species having a shared evolutionary past.

3.1.8. Conclusion

Overall, viral metagenomic sequencing is a useful tool for the un-biased detection of viruses in clinical samples, however the approach has limitations in terms of sensitivity, and variations in methodology can have a significant impact on results. It is clear that the studies that have been published in this relatively new field vary hugely and are poorly validated. This is particularly true for samples derived from blood, as an attempt has been made to standardise an approach for faecal samples.[286]

Therefore, in order to proceed with my investigation into viral infections in immunocompromised cohorts, I first wished to develop a suitable metagenomic methodology capable of simultaneously detecting both RNA and DNA viruses in clinical blood samples by comparing methods for those steps that have not yet been sufficiently evaluated in the literature.

Here, the effect of sequencing PBMC versus plasma samples was investigated. This included comparing the effect of enrichment techniques on sensitivity and viral genome coverage, as well as methods for the computational detection of viruses in sequencing data. Finally, the overall performance of the method was tested in terms of robustness and sensitivity in comparison to qPCR.

3.2. Methods

3.2.1. Comparison of plasma and PBMCs for viral metagenomics

3.2.1.1. Samples and ethics

To test whether a difference in virome is detectable in plasma and PBMCs, blood samples were taken from three cohorts of patients identified as at high risk of viral infection (Table 3.1).

Cohort	Form of risk	n
Intravenous drug users	Exposure to blood-borne pathogens	9
Common variable immunodeficiency	Primary immunodeficiency	8
Kidney transplant recipients	Secondary immunodeficiency	7

Table 3.1. **Overview of cohorts for comparison of plasma and PBMC sequencing.** Paired plasma and PBMC samples were obtained from each patient (24 in total) and each was sequenced separately using a metagenomic approach.

Kidney transplant patient samples were kindly provided by Prof Ken Smith and Dr Elaine Jolly (Addenbrooke's Hospital, Cambridge, UK) with informed written consent and the approval of the Cambridge Local Research Ethical committee (reference **08/H0308/176**). Common variable immunodeficiency patient samples were kindly provided by Dr Helen Baxendale (Papworth Hospital, Cambridge, UK) with informed written consent under NHS REC (**12/WA/0148**) following an amendment to support this work given on 10th January 2013. Intravenous drug user samples were kindly provided by Dr Emma Thompson (University of Glasgow, UK) with informed written consent and the approval of the West of Scotland Research Ethics Service, NHS REC (**12/WS/0002**).

3.2.1.2. Separation of plasma and PBMCs

PBMC and plasma separation from whole blood was performed using Leucosep tubes (Griener Bio One) according to the manufacturer's instructions. 4 mL of whole blood collected in EDTA was diluted 1:1 with PBS and transferred to a LeucoSep tube. The tube was centrifuged at 600 x g at room temperature for 10 min with the brake off. 5 mL of diluted plasma was aspirated, transferred to cryovials and stored at -150 °C. The remaining plasma and PBMCs were decanted into a new tube, washed twice with PBS, re-suspended in 1 mL of PBS and PBMCs were counted using a C-Chip haemocytometer (DigitalBio). PBMCs were diluted to 1 x 10^7 cells/mL in PBS supplemented with 0.5 % BSA (Ambion) and aliquots of 2 x 10^6 cells were stored at -150 °C.

3.2.1.3. Nucleic acid extraction and preparation from plasma

Nucleic acid was extracted and prepared from plasma as described in section 2.1 of this thesis.

3.2.1.4. Nucleic acid extraction and preparation from PBMCs

Metagenomic sequencing of virus from cellular samples is extremely difficult without enrichment due to the abundance of host genomic DNA.[235] Therefore, an adapted version of the enrichment methodology described by Daly et al (2011) was used to aid detection of viruses from PBMCs.[235] Approximately 2 x 10^6 PBMCs were pelleted at 600 x g, the buffer was removed and the cells re-suspended in 200 µl of RNase-free water. Samples were subjected to 3 freeze-thaw cycles, performed by placing each sample directly on dry ice until frozen and quickly removing to thaw. Additional light mechanical shearing using a pipette was used to help break open the cell membranes. Samples were then

centrifuged at 600 x g to pellet the cell nuclei and other debris. Un-protected cytoplasmic DNA was digested for 30 minutes at 37°C, using 2 U of TURBO DNase (Thermo Fisher) according to manufacturer's instructions. Total nucleic acid was then extracted from the nuclease treated sample as described for plasma.

For nucleic acid extracted from PBMCs, random PCR was performed using a sequence-independent single primer amplification (SISPA) method due to extremely low concentrations of nucleic acid remaining following nuclease treatment.[289] For the SISPA protocol, reverse-transcription was performed as for plasma (see section 2.2) which the exception that the SISPA primer FR26RV-N, composed of a random hexamer with an additional PCR adaptor sequence attached, was used rather than random hexamers (Table 3.2).[290] Following second-strand synthesis, six SISPA reactions were set-up for each sample. Each reaction mix contained 5 µl of DNA template from the RT reaction (1.5), 2 U of Advantage 2 DNA Polymerase (Takara-Clontech), 5 µl of Advantage 2 PCR Buffer (10 x), 1 µl of 10mM dNTP mix (Fisher) and primer FR20RV to a final concentration of 0.08 µM (FR20RV is the same sequence as the 5' adapter portion of FR26RV-N, table 3.2) with RNase-free water added to 50 µl total. PCR reactions were performed using the Mastercycler Epgradient S (Eppendorf). DNA polymerase was activated by heating to 94°C for 4 minutes, followed by 94°C for 1 minute, 64°C for 1 minute and 72°C for 2 minutes for 8 cycles. Double-stranded DNA was purified using AMPure XP beads as described in section 2.2.

Primer	Sequence (5'-3')
FR26RV-N	GCC GGA GCT CTG CAG ATA TCN NNN NN
FR20RV	GCC GGA GCT CTG CAG ATA TC

Table 3.2 **Primers used for sequence-independent single primer amplification (SISPA)** as described by Reyes et al [289]. This method facilitates non-specific amplification of RNA and DNA populations. The random hexamer sequence at the 3' end of FR26RV-N is first used to randomly prime for reverse transcription/second-strand synthesis. Sequence independent amplification is then performed by using the 5' adapter sequence of FR26RV as a primer binding extension sequence.

3.2.2. Comparison of DNA purification methods

DNA samples were split into equal 60µl aliquots and purified by either AMPure XP beads (Beckman Coulter) or ethanol purification. The results of each purification method were run side-by-side on the Bioanalyser 2100 (Agilent Technologies), for comparison using the high sensitivity double-stranded DNA kit. AMPure XP bead purification was performed following the manufacturers protocol, using a bead to sample ratio of 1.8 : 1 to recover the greatest possible range of DNA species whilst removing small fragments. DNA was re-suspended in 30µL of TriS-HCl (10mM Tris, 1mM EDTA, pH 8.0) (Thermo Fisher Scientific). For ethanol precipitation, 1 µL of glycogen (Thermo Fisher Scientific), 0.1 volumes of 3M sodium acetate (Ambion) and 2.5 volumes of ice cold 100% ethanol (Thermo Fisher Scientific) were added to the DNA sample and vortexed to mix. Samples were incubated at -20 °C overnight, removed, and centrifuged at 16,000 x g for 30 mins. The liquid was aspirated, the pellet washed with ice cold 75% ethanol three times and then let stand to dry for 2 minutes. The DNA pellet

was then re-suspended in 30µl of TriS-HCl (10mM Tris, 1mM EDTA, pH 8.0) and gently dissolved by pipetting.

3.2.3. Library Preparation and Sequencing

Library preparation and sequencing was outsourced to Cambridge Genome Services (samples used for comparison of plasma and PBMCs, and comparison of enrichment strategies), the SMCL Next Generation Sequencing Hub (samples for informatics pipeline testing), and the Wellcome Sanger Institute (samples for limit of detection testing). All Illumina sequencing libraries were prepared and indexed using the KAPA HyperPlus kit (Roche). Libraries sequenced at CGS and the SMCL Hub were pooled to equimolar amounts and multiplexed 12x on the Mi-Seq v2 (300-cycle) (Illumina) to produce 150 base-pair, paired-end reads. Libraries sequenced at the Sanger Institute were pooled to equimolar amounts and multiplexed 48x on the Hi-Seq 2500 v4 (250-cycle) (Illumina) to produce 125 base-pair, paired-end reads in order to duplicate the high-throughput approach used for the experiments in chapters 4 and 5 of this thesis.

3.2.4. Analysis of sequence data

Sequencing data was analysed using CLC Genomics Workbench (Qiagen) as described in section 2.3 of this thesis unless stated otherwise. All viruses were identified by alignment down to a species level. However, in many cases due to low read counts and poor coverage, more detailed identification of the viral isolates was not possible.

3.2.5. Comparison of Enrichment Techniques

3.2.5.1. Samples for viral enrichment testing

To test the effect of enrichment methodologies on virus recovery torque-teno virus, BK virus, cytomegalovirus and hepatitis C virus positive plasma samples from naturally infected patients in the cohorts described in Table 3.1 were selected. These represent viruses from a range of Baltimore classification groups (single-stranded DNA, double-stranded DNA and RNA genomes) as well as viruses with and without an envelope. Each sample was extracted in triplicate using 3 different methods (filtering, nuclease treatment and extraction only) (overview of method shown in Figure 3.3.). Following extraction, viral load was quantified using published probe-based qPCR assays in order to test the different methods' effect on viral recovery (see Table 3.3 for primer sequences).

3.2.5.2. Viral enrichment by filtering

For the filter-based enrichment of viral particles, 300µl of plasma sample was passed consecutively through a 0.45 µm followed by a 0.22 µm syringe filter (Merck Millipore). 22 µl of RNase-free water (Ambion) was added to 200 µl of the filtrate to account for the volume of the nuclease treated samples (due to the addition of DNase buffer and enzyme as described below).

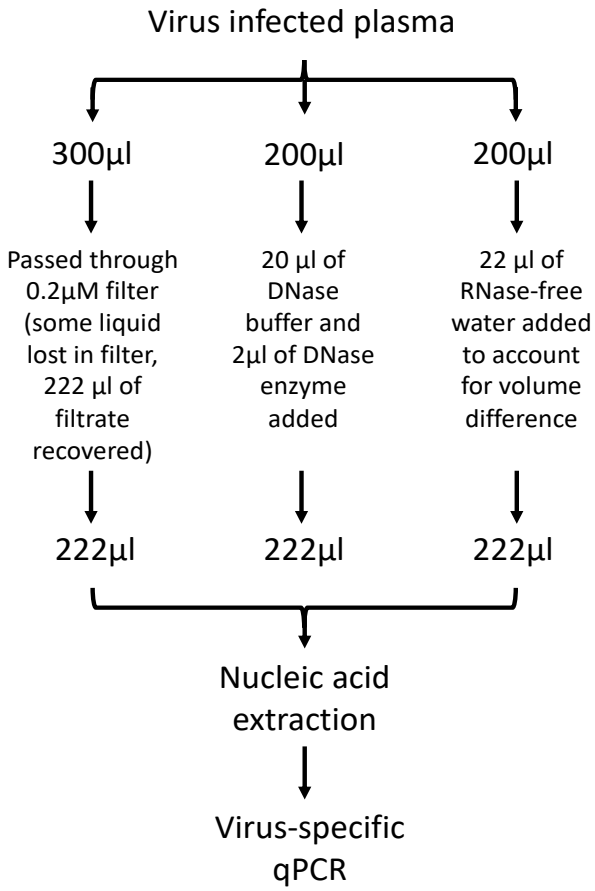


Figure 3.3. **Overview of the method used to compare the effect of viral enrichment methods on virus recovery**.

Four plasma samples naturally infected by a single virus (TTV, BKV, CMV, or HCV) were divided into three. Each sample was then subjected to three separate treatment methods (filtering, nuclease digestion, and no treatment) prior to extraction. The amount of virus recovered was quantified by qPCR. In order to compare the methods fairly, volumes used for extraction were kept equal (222µl), by the addition of nuclease free-water to the 'non-enriched' sample.

Virus	Primer	Sequence (5'-3')	Source
CMV	Forward	CAA GCG GCC TCT GAC AAC CAC MV	315
	Reverse	ACT AGG AGA GCA GAC TCT CAG AGG AT	
	Probe	FAM-TGC ATG AAG GTC TTT GCC CAG TAC ATT CT-BHQ1	
BKV	Forward	TGC TAG CCG AGT AGY GTT GG	316
	Reverse	ACT CGC AAG CAC CTT ATG AG	
	Probe	JOE-ACC ACA AGG CCT TTC GAG A-BHQ1	
HCV	Forward	AGC AGG CAA GGG TTC TAT TAC TAA AT	317
	Reverse	GAA GCA ACA GCA GAT TCT CAA CA	
	Probe	FAM-AGG ACC CTA AAG ACT TTC CCT CTG ATC TAC ACC AGT TT-TAMRA	
TTV	Forward	CCT GCA CTT CCG AAT GGC TGA GTT	47
	Reverse	CTT GAC TGC GGT GTG TAA ACT CAC	
	Probe	FAM-CCT CCG GCA CCC GCC CTC GGG ACG-BHQ1	

Table 3.3. **Primer and probe sequences used to quantify viral load following the testing of enrichment methods.**

3.2.5.3. Viral Enrichment by nuclease treatment

For the nuclease enrichment of viral particles, 200 µl of plasma sample was mixed with 2 U of TURBO DNase enzyme and 20 µl of 10X buffer (Ambion). Samples were incubated at 37°C for 60 minutes and DNase was inactivated using the inactivation reagent provided, following manufacturer's instructions, and 222 µl of nuclease treated sample was recovered for nucleic acid extraction.

3.2.5.4. No viral enrichment

22 µl of RNase-free water (Ambion) was added to 200 µl of un-treated plasma to account for the additional volume of the nuclease treated samples.

3.2.5.5. Extraction

Nucleic acid was extracted from 222 µl of each sample using the High Pure viral nucleic acid extraction Kit (Roche Diagnostics) and eluted in 40 µl of elution buffer as described for plasma in thesis section 1.4.

3.2.5.6. Virus quantification

Virus quantification was performed by qPCR using the QuantiTect Virus + ROX Vial Kit (Qiagen) and the dual-labelled primer and probe sets described in Table 3.3 (synthesised by Sigma-Aldrich). Reactions contained: 2 µl of nucleic acid; 5 µl of QuantiTect Virus Mastermix (5 X); primers to a final concentration of 0.4 µM and; Taqman probe to a final concentration of 0.2 µM. Mixes for RNA virus assays also contained 0.2 µl of QuantiTect Virus RT Mix (100 X). RNase-free water was added to all reactions to a final volume of 20 µl.

One-step (RT-)qPCR assays were performed on the Rotor-Gene 6000 (Qiagen). For RNA viruses, a reverse-transcription step was performed at 50°C for 20 minutes. All reactions then underwent a DNA polymerase activation step at 95°C for 5 minutes followed by 2-step cycling of denaturation at 95°C for 15 seconds and combined annealing/extension at 60°C for 45 seconds for 40 cycles. Gain settings were optimised automatically prior to first acquisition and fluorescence data was acquired during the annealing / extension step of the cycling. Threshold levels were set manually, ensuring they passed through the exponential stage of amplification.

3.2.6. Viral Enrichment Testing by NGS

3.2.6.1 Samples

To test whether the nuclease enrichment methodology improves the detection or coverage of viruses by sequencing we selected a further three plasma samples naturally infected by the viruses torque-teno virus, cytomegalovirus and human pegivirus.

3.2.6.2 Extraction, sequencing and data analysis

Plasma samples were split into two 200 µl aliquots and extracted, prepared, sequenced, and analysed in duplicate using the finalised metagenomic protocol detailed in the methods section of this thesis (Figure 2.3).

3.2.7. Comparison of Mapping Algorithms for Viral Read Identification

In order to compare the sensitivity of different mapping algorithms in terms of viral read identification, a cohort of plasma samples with clinically diagnosed viral hepatitis were prepared and sequenced using the finalised metagenomic protocol detailed in the methods section of this thesis (Figure 2.3). Following sequencing, low-quality reads were discarded and low-quality base-calls were trimmed from the ends of reads using Trimmomatic v.0.32.[318] The remaining, trimmed sequences were aligned to the human reference genome build GRCh37.p13 (www.ncbi.nlm.nih.gov/grc/human/data?asm=GRCh37), in order to subtract host-derived sequences. Unmapped reads were then aligned to a curated version of the Viral RefSeq database containing viruses of vertebrates (see thesis section 2.3).

Three alignment algorithms software packages were compared in terms of their ability to detect viral sequences in the trimmed dataset: Bowtie2 v.2.0.0,[319] SMALT v.0.7.4,[320] and the mapping algorithm of CLC Genomics Workbench v.7.5.1.[321] Bowtie2 was run using the default settings for both the host subtraction and virus read identification steps. Additionally, Bowtie2 was run using the "--very sensitive" option for the viral read identification step to determine whether sensitivity could be improved. SMALT was run using the default settings except for the sampling step size reduced to 3 in order to increase sensitivity. The CLC mapping algorithm was run using parameters requiring matches of 90% similarity over 80% of the read length for host read subtraction, and 80% similarity over 80% of the read length for virus read identification.[257] Mapped reads were confirmed by BLASTn to the nr database (https://blast.ncbi.nlm.nih.gov/ Blast.cgi).[322]

3.2.8. Estimation of the Limit of Detection

To determine the limit of detection of the methodology for RNA viruses, dilution series of the international standards for HIV-1 and HCV were used. The 3rd HIV-1 International standard (NIBSC code 10/152) was re-suspended in 1 mL of RNase-free water (Ambion) following manufacturer's instructions to give a concentration of 185,000 IU / mL. The 4th HCV International standard (NIBSC catalogue number: 06/102) was re-suspended in 1 mL of RNase-free water (Ambion) following manufacturer's instructions to give a concentration of 260,000 IU / mL.

Serial dilutions were made in duplicate using healthy human plasma confirmed as negative for HIV-1 and HCV by qPCR. Dilutions were extracted and sequenced, and viral reads were identified using the method described for plasma in section 2.1. of this thesis. This method is a slight adaptation of the one developed here

and was used in order to precisely replicate the high-throughput approach to be used for clinical samples sequenced in the following two chapters of this thesis. Conversion of international units into viral copy number for each standard was based on previous calculations.[323,324]

To estimate the limit of detection of the methodology for DNA viruses, I made use of an existing dataset consisting of paired sequencing and qPCR data, generated from 200 patient plasma samples (kidney transplant recipients, vasculitis patients and primary immunodeficiency patients described in chapters 4 and 5), to run a side-by-side comparison. To do this, a generalised linear model was constructed in order to test the effect of anellovirus genome copy number as determined by qPCR, on the likelihood of anellovirus detection by metagenomic sequencing.

3.2.9. Statistical Analysis

Basic statistical analysis was performed using the methods described in section 2.6 of this thesis.

3.3. Results and Discussion

3.3.1 Comparison of viruses detected in plasma and PBMCs

Given that many viruses are present predominantly in specific tissues, namely those in which they replicate, it was hypothesised that distinct viruses would be detected in the cellular and cell-free components of the same blood sample.

To test this hypothesis, paired plasma and PBMC samples were selected from three cohorts of patients at high risk of infection (Table 3.1). Un-biased, meta-genomic sequencing was performed for the samples from each patient and viruses were identified following the protocols described in the methodology section.

For all cohorts, the vast majority of sequencing reads from both plasma and PBMC samples mapped to the human reference genome (Figure 3.4). Human reads made up a significantly smaller proportion of the total reads from plasma samples compared to their PBMC counterparts (paired t-test, $p = 0.0114$) suggesting that detection of viruses in plasma may not require as high a sequencing depth. Viral reads were detected more often in plasma than PBMC samples (Fisher's exact test, $p = 0.0199$). However, this did not translate to a larger proportion of viral reads in plasma samples overall (paired t-test, $p = 0.1064$). In all samples, a proportion of reads could not be classified as either host or virus. These unclassified reads were not analysed further, however, it is likely they predominantly represent bacteriophage viruses and bacterial sequences, which have previously been shown to be abundant in the blood,[39,325] but were not tested for here. This group may also contain unknown viruses which are yet to be discovered and are therefore not detectable by sequence alignment to the virus reference database.

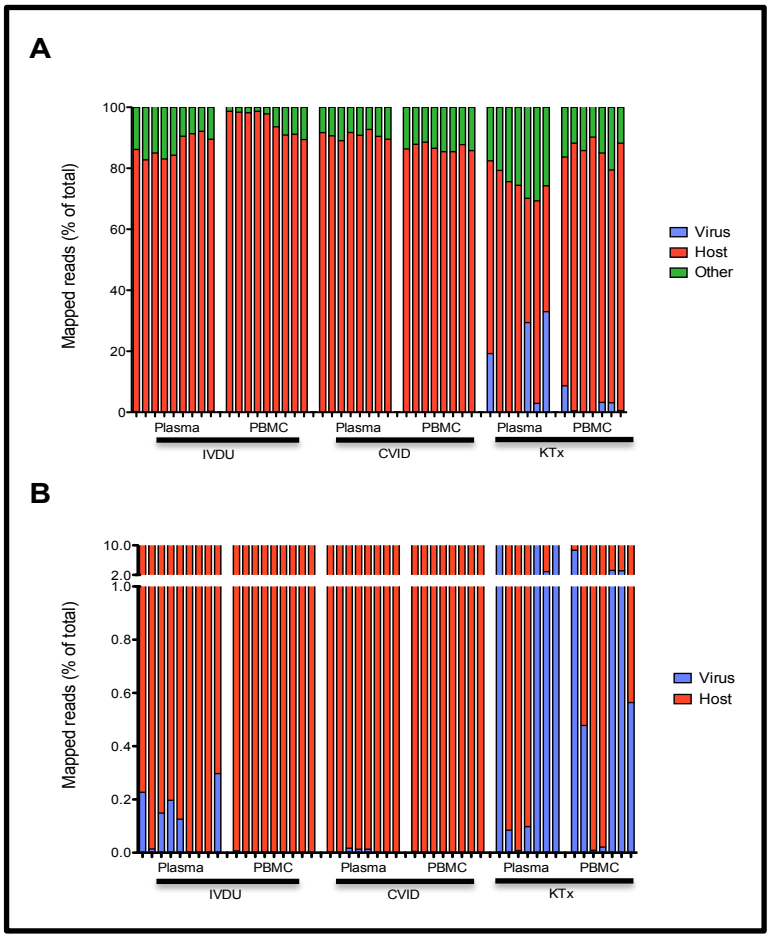


Figure 3.4. **Taxonomic distribution of sequencing reads from paired plasma and PBMC samples taken from three patient cohorts.**

Trimmed Illumina sequencing reads were identified as host, viral or 'other' by alignment to human and viral reference genomes. A) Overall taxonomic distribution. The x axis refers to individual patients, separated by cohort (IVDU: intravenous drug users; CVID: common variable immunodeficiency; and KTx: kidney transplant recipients). Plasma and PBMC pairs appear in the same order. 'Other' refers to the percentage of unclassified reads remaining following quality control, and mapping to the human and virus reference sequences. These sequences were not analysed further, but are likely to represent common organisms such as bacteria, bacteriophages which were not considered in this experiment. B) Enlargement of the y-axis of figure A to demonstrate the difference in viral read percentage between plasma and PBMC samples.

I next compared the virus species that were detected in paired plasma and PBMC samples. In total, ten different viruses across seven families were detected. Viral species were split between those that were: a) present at approximately equal amounts in both sample types or; b) predominantly detected in a single sample type (Figure 3.5.A). Of the viruses that were detected in plasma and PBMC samples from the same patient, greater depth and uniformity of coverage was seen from the plasma (Figure 3.5 B). This may simply be due to higher viral loads in the plasma. However, another possible explanation is that the greater abundance of host DNA in the cellular samples meant that host DNA made up too high a proportion of the NGS library, compared to virus, and therefore restricted the number of viral reads that could be generated.

Of the viruses detected in both cellular and cell-free samples, the anelloviruses TTV, TTMDV and TTMV were most abundant. Each of these viruses was detected roughly equally between the two compartments. They have previously been shown to replicate in mononuclear cells undergoing cell division,[326] and are commonly found at high levels in plasma and serum, particularly in immunodeficient individuals.[327,328] Parvovirus B19 was detected in both the plasma and PBMCs of a single patient, this virus has been shown to preferentially infect erythroid progenitor cells in the bone marrow,[329] however it can also persist in the blood and other tissues.[330] Human pegivirus (HPgV) was detected in both plasma and PBMC samples predominantly from the IVDU cohort, although it had a significantly higher prevalence in the plasma (Fisher's Exact Test, $p = 0.0226$). HPgV RNA has previously been detected in the plasma, semen, saliva, liver, spleen and bone marrow and studies to determine whether PBMCs are a site of HPgV replication have yielded contrasting results.[331–333] If HPgV is present in the PBMCs of these patients, it is possible that, without viral enrichment, metagenomic sequencing is not sufficiently sensitive for detection of the virus in cellular samples. Low levels of HPgV may have been 'drowned out' by

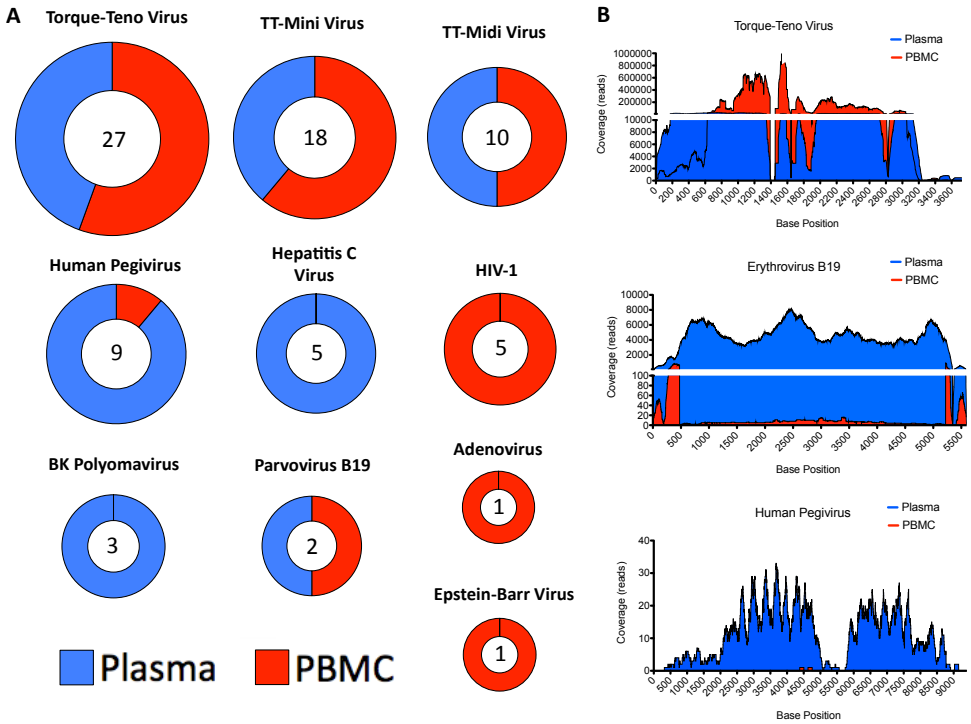


Figure 3.5. **Origin of viruses detected by metagenomic sequencing of plasma and PBMCs**.

The proportion of viral sequences detected in paired plasma and PBMC samples from IVDU, CVID and KTx cohorts by metagenomic sequencing. Presence of a virus was defined by mapping of one or more NGS read to the virus reference genome. A) Overview of all viruses detected in paired plasma and PBMC samples. Pie charts display the proportion of the total number of times a virus was detected across both plasma and PBMCs (n = 48). E.g. if a virus was equally detectable in both plasma and PBMCs, one would expect to see an equal proportion (e.g. TT midi virus and parvovirus B19). The size of each individual pie chart represents the total number of times that a virus was detected across all samples (both plasma and PBMCS) combined. The total frequency of detection is stated at the centre of each chart. B) Examples of the difference in viral genome coverage depth between paired plasma and PBMC samples.

the overwhelming abundance of host genomic sequence while the virus was persistently detected at high levels in the plasma of patients during the course of this study.

Two viruses were detected exclusively in plasma samples: HCV, and BKV. HCV is primarily transmitted through exposure to infected blood, and largely replicates in hepatocytes but can persist in B-cells and is found in the plasma at high levels during chronic infection.[334] BKV was detected in 2 renal transplant recipients. The virus predominantly exhibits tropism for kidney cells although replication in the brain, liver, lung, eye, and salivary glands has been demonstrated suggesting that the virus travels to many tissues during infection.[335–337] BKV is generally acquired early in life and exists at a high prevalence in the population, re-activating following immune suppression such as that experienced by transplant recipients.

A further three viruses were detected only in PBMC samples. All three of these viruses are known to replicate in cells of the immune system present in the blood.

Epstein-Barr virus was detected in the PBMCs of a single patient. EBV is highly prevalent worldwide, and the vast majority of individuals are infected at an early age.[338] However, following the development of immunity EBV persists in a latent state. Whilst latent, detection of the virus by non-specific methods is likely to be difficult as viral replication ceases and transcription is limited.[21] However, upon deterioration of the immune system such as that seen during ageing or drug-induced immunosuppression, the virus re-activates and proceeds to replicate at detectable levels unless immunity is restored.[202] Here, the low prevalence of the virus in the immunosuppressed kidney transplant cohort is likely due to the use

of antiviral drugs to protect these patients against the adverse effects of herpesvirus reactivation.

HIV-1 was exclusively detected in the PBMCs of five patients. The virus replicates in CD4+ T-lymphocytes and during an un-treated infection, plasma viral loads can reach high levels as the infection progresses. Here, the patients in which HIV-1 was detected were undergoing anti-retroviral therapy (ART) to treat the infection. During ART the virus is restricted to infected cells and plasma viral load is effectively reduced to zero, a finding reflected in the results.

Finally, adenovirus type 2 was detected in the PBMCs of a single immunodeficient (CVID) patient. Adenovirus type 2 is predominantly associated with symptomatic respiratory infections,[339] and so its detection in an antibody deficient patient with weakened mucosal immunity is unsurprising. The virus primarily infects cells of the lung, however certain types have been observed infecting primary lymphocytes and monocytes.[340] It has also been detected in a latent state in the T-cells of the tonsils and adenoid tissues in healthy individuals.[341] Overall, adenoviral infection of PBMCs in the immunocompetent host appears to be rare: previous PCR-based studies have reported a prevalence as low as 0 % in the PBMCs of healthy adult volunteers,[342,343] and so its presence here is suggestive of a heightened susceptibility to viral infection.

Overall, these results demonstrate that plasma samples generally yield a greater proportion of viral reads compared to PBMCs, likely due to the low level of genetic material derived from the host. However, it is clear that different types of viruses are detected depending on the type of sample sequenced. Whilst some viruses appear to be equally detectable in both compartments of the blood, HPgV, HCV and BKV were far more prevalent in the plasma. Conversely, it is possible that EBV and AdV are detectable only in cellular samples. This suggests

that a combined approach sequencing both plasma and PBMCs is potentially the best approach for detecting viruses in the blood however this would double the cost of sequencing. All aspects considered, it was decided to proceed with plasma samples as: a) on average they yield a greater proportion of viral reads; b) the lower amount of host genomic material in plasma should theoretically increase the sensitivity for viral detection and; c) plasma samples are more readily available in clinical bio-repositories.

3.3.2. Viral enrichment

In order to test the effect of viral enrichment methods, three approaches (filtration, nuclease digestion, and no enrichment) were applied to plasma from four patients naturally infected by viruses representing groups across the Baltimore classification: HCV (enveloped, +ssRNA), CMV (enveloped, dsDNA), BKV (non-enveloped, dsDNA) and TTV (non-enveloped, ssDNA).

Recovery of all viruses was reduced following the filtering and the nuclease enrichment protocols (Table 3.4). The effect was relatively minor with regard to both HCV and TTV, likely due to their small capsid sizes easily passing through the 0.22 μM filter and their respective RNA and single-stranded DNA genomes not being at risk of digestion by DNase. Conversely, the enrichment treatments had noticeable effects on the recovery of both CMV and BKV. Despite protection from the capsid, the double-stranded DNA genomes of both viruses are likely to have both suffered some degradation during the DNase treatment. This suggests that DNase treatment is not an appropriate method for enriching viruses with double-stranded DNA genomes.

CMV and BKV recovery was also affected by filtration. The CMV capsid is the largest of those tested and whilst it can theoretically pass through a 0.22 μM

filter, it is likely that some viral particles were impeded, resulting in reduced recovery. The reason for a markedly low recovery of BKV following filtering is less clear as the virus is a similar size to HCV, which passed through the filter mostly unhindered. Furthermore, it should be noted that the experiment was performed in triplicate and so the result is unlikely to be an anomaly. One possible cause is that the virus was adsorbed to the filter membrane which would prevent it passing through. This may be due to the biophysical properties of the capsid (e.g. hydrophobic interaction with the filter membrane) or due to the properties of the plasma, such as an increased salt concentration or low pH (e.g. due to kidney damage), both of which have been shown to increase virus adsorption to nitrocellulose filter membranes.[344] One further possible explanation is the presence of α-defensin, which is produced by the innate immune system and is known to interact directly with BKV to form aggregates which are unable to infect cells.[345] If the virions had formed these aggregates then they may be impeded from passing through the filter.

Whilst enrichment methods appear to have a detrimental effect on virus yield post-extraction, it is possible that this may be outweighed by the beneficial reduction of host genomic material, improving detection overall. Therefore, I selected enriched and non-enriched extracts from a further three naturally infected patients (HPgV (enveloped, +ssRNA), TTV and CMV) for metagenomic sequencing.

Once again, a decrease in CMV derived sequence following nuclease enrichment was detected. Indeed, CMV reads were solely detectable in the non-enriched sample. In the cases of HPgV and TTV, nuclease treated samples generated more viral reads than the un-treated samples (Figure 3.6.A). However, this did not correlate with viral coverage, which was increased for the non-enriched samples (Figure 3.6.B). The increased number of viral reads without a corresponding

Virus	Genome	Envelope	No Treatment (Ct)	Filter (Ct)	Nuclease (Ct)
BK Polyomavirus	dsDNA	No	20.46	37.13	26.75
Cytomegalovirus	dsDNA	Yes	35.81	37.46	N/A
Hepatitis C Virus	(+)ssRNA	Yes	32.56	33.99	34.51
Torque-Teno Virus	ssDNA	No	29.53	30.62	30.55

Table 3.4. **The effect of viral enrichment techniques on virus recovery**.

Plasma samples naturally infected by different viral species were subjected to three different viral enrichment methods (DNase digestion, filtering and no treatment). Following enrichment, nucleic acid was extracted and the amount of virus was quantified by qPCR in triplicate to determine the effects of enrichment on viral load. Average cycle threshold (Ct) value is shown.

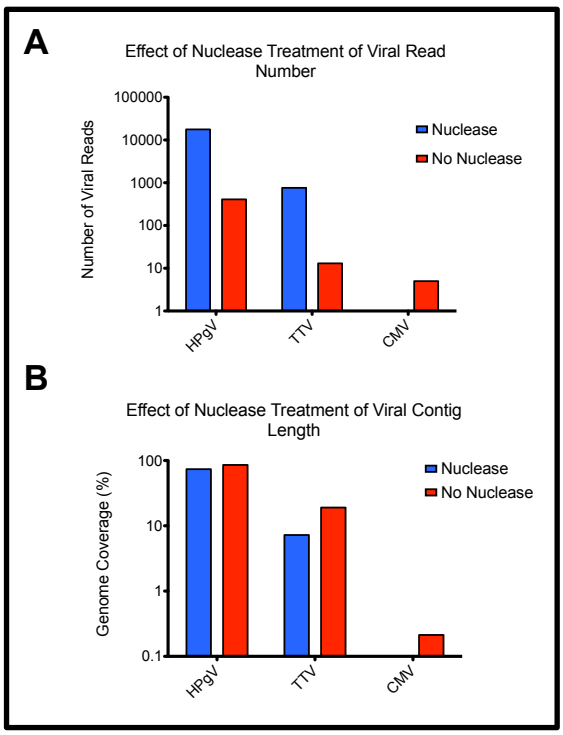


Figure 3.6. **Effect of nuclease enrichment on viral read number and virus genome coverage.**

Plasma samples from naturally infected patients were subjected to viral enrichment using DNase digestion, or no enrichment. Total nucleic acid was then extracted, sequenced and the sequencing data analysed. **A)** the effect of nuclease enrichment on the total number of viral reads generated. Viral sequences were identified through mapping to the virus genome database as described in the methodology chapter of this thesis. **B) t**he effect of nuclease enrichment on viral genome coverage. Viral genome coverage was calculated as the proportion of the reference genome length following mapping.

increase in coverage may be due to the use of random amplification of the treated samples prior to sequencing, although the number of PCR cycles was kept to a minimum in attempt to avoid amplification bias as much as possible. Amplification is necessary to generate sufficient input for library preparation after a large proportion of DNA is degraded following nuclease treatment however the amplification can lead to bias resulting in non-uniform coverage.

Overall, these results suggest that use of filtering or nuclease enrichment techniques that are vital for increasing sensitivity and viral coverage in solid tissue samples are not necessarily beneficial for the sequencing of viruses in cell-free samples such as plasma. It should be noted that hybridization-capture methods for viral enrichment, which have shown much promise, are noticeably absent from this experiment. The first of these to gain attention was VirCapSeq which was reported only after these experiments were completed.[241] Considering the reported results, it is likely that a capture-based viral enrichment method would significantly out-perform those tested here. However, it is also important to note that production of the oligonucleotide arrays is expensive, particularly for studies involving hundreds of samples, and the method is still inherently biased, enriching only for known viral sequences and their relatives.[346]

3.3.3. DNA purification

Following extraction (with optional enrichment), and conversion to double-stranded DNA for library preparation, DNA must be purified to concentrate samples to an appropriate level and remove excess salts that may interfere with quantitation and later steps in the library preparation protocol.

Classically, DNA purification is performed by ethanol precipitation. This approach is the extremely affordable however it also has disadvantages which make it not ideal for the purpose of high-throughput preparation of low abundance DNA samples such as those extracted from plasma. The disadvantages are as follows: variable DNA recovery, especially with low levels of DNA where a carrier such as glycogen is required and; the inefficient removal of contaminants such as dNTPs, primers, primer-dimers, salts and sub-100bp fragments of DNA that can interfere with the accurate measurement of concentration, which is required for calculating the appropriate input for library preparation.

Therefore, I wished to compare methods of purifying DNA samples prior to library preparation. Total nucleic acid extracts from 5 plasma samples were converted to cDNA and randomly amplified for 5 cycles using the SISPA protocol in order to create sufficient material for this experiment. Each PCR reaction was split in two, and purified either by ethanol precipitation, or by Ampure XP beads, a polyethylene glycol-based method. The polyethylene glycol-based approach is more expensive as it requires a kit but has the benefits of being quicker due to a shorter incubation period, and more consistently sensitive for lower levels of DNA. Visualisation of the resulting DNA using the high-sensitivity Bioanalyser chip (Agilent Technologies) revealed that Ampure XP beads are effective at removing contaminating, small, sub-100 bp DNA fragments (Figure 3.7) with no discernible loss of larger species and therefore this method was selected for the study.

3.3.4. Informatic identification of viral sequences

The sensitivity of viral metagenomic sequencing is also reliant upon the method used to detect any viral reads generated. The most straightforward approach to

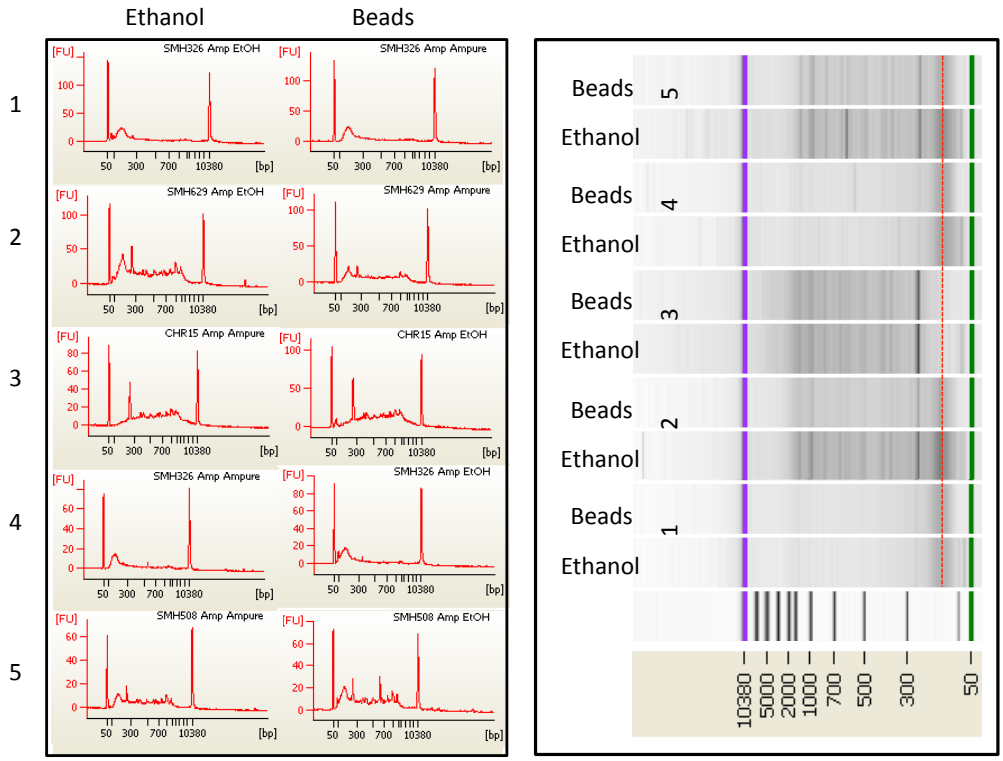


Figure 3.7. **Comparison of DNA purification methods prior to library preparation.**

Electropherogram and corresponding gel for randomly amplified DNA from five plasma samples (labelled 1-5) purified using either ethanol precipitation or polyethylene glycol (AMPure XP beads). Cut-off size below which DNA fragments are removed by PEG is indicated by a red dotted line on the gel.

viral read identification is through the alignment of sequences to a viral reference database. However, there are a number of alignment programmes available, and very few comparisons have been performed for these programmes in the viral metagenomics setting. Importantly, the demands of viral sequence identification in this setting are unconventional, as highly accurate alignments are likely to be too specific, and therefore detrimental to the detection of viruses which can evolve rapidly. This is particularly true for viruses with RNA genomes, which tend to have high rates of mutation compared to DNA viruses.[347] However, it has also been shown that some viruses with single-stranded DNA genomes, such as bacteriophage M13, have mutation rates closer to that observed for RNA viruses than those with double-stranded DNA genomes.[348]

Therefore, I used four sets of previously generated metagenomic data known to contain viral reads (HBV, HCV, HDV, and HPgV) to compare the sensitivity of three alignment algorithms: Bowtie2,[319] SMALT,[320] and the mapping algorithm of CLC Genomics Workbench (Qiagen).[321] These algorithms were selected as they form the basis of several viral identification pipelines: Bowtie2 is used by VirusSeeker and PATHSeq,[309,349] SMALT is used by VirMet and READSCAN,[287,350] whilst CLC has previously been used in this laboratory for the development of a viral discovery pipeline.[257] These algorithms make use of different alignment methods. Bowtie2 uses a Burrows-Wheeler transform (aka FM-index) method (also used by BWA and SOAP2), whereas SMALT uses a hash-index method. The CLC mapping algorithm is proprietary and information regarding its workings have not been made public. In order to compare these algorithms, plasma samples with clinically diagnosed viral infections were prepared and sequenced using the finalised protocol detailed in the methods section of this thesis (Figure 2.3).

The output of each algorithm was compared in terms of viral read number and genome coverage for a range of viruses, in order to determine the optimal method for the detection of viral reads (Figure 3.8). The specificity of each viral sequence detected was confirmed by BLASTn to the non-redundant nucleotide database and excluded if suspected to be of non-viral origin.

There was no significant difference between the algorithms in terms of the average number of reads aligned for any given virus ($p = 0.1303$, Two-way ANOVA) (Figure 3.8.A). However, on average, the CLC algorithm detected the most reads for all viruses but anellovirus, for which SMALT performed best. The three algorithms generated comparable levels of genome coverage for HBV (Figure 3.8.B). However, Bowtie2 generated the lowest coverage for the remaining 3 viruses: HPgV, HDV and HCV (Figures 3.8.C-E). CLC performed particularly well in terms of mapping HCV reads compared to the other algorithms. This discrepancy between algorithms is likely due to differences in their ability to tolerate mismatches brought about by viral mutations (e.g. insertions, deletions and substitutions) or artificially introduced by sequencing errors. Burrows-wheeler transform methods were primarily designed to speed up computationally expensive alignments to mammalian genomes, which are many times the size of the average viral genome. To achieve this speed, the algorithm makes compromises which result in a reduced ability to detect mismatches (such as those one would expect when aligning a mutated viral isolate against a reference strain). An inability to align mismatches would result in lower coverage compared to methods that can deal with them (e.g. hash-index methods). This is particularly obvious in the more rapidly evolving viruses such as HCV, where divergence from the reference strain is likely to be greater.[351]

Viral read identification from metagenomic datasets requires an alignment tool with the ability to detect highly divergent sequences based on a limited database.

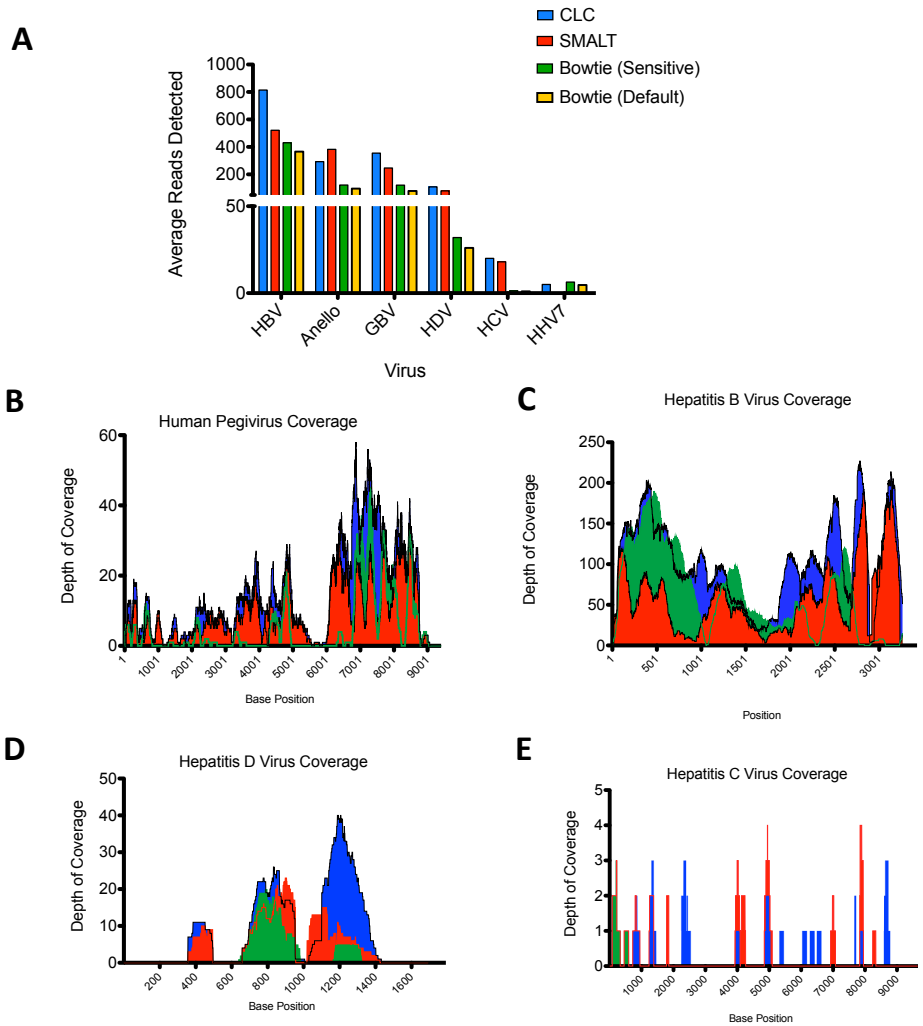


Figure 3.8. **Comparison of mapping algorithms for detection of viral sequences in metagenomic data.**

The performance of three mapping software algorithms: CLC, SMALT and Bowtie2, were compared in terms of their ability to detect viral reads in a metagenomic dataset using six real Illumina metagenomic sequencing datasets each known to contain virus specific sequences. All algorithms were run using default settings. However, Bowtie2 was tested using both the default and "very sensitive" options. **A)** Comparison of the average number of reads aligned to each virus revealed no significant difference between the algorithms ($p = 0.1303$, Two-way ANOVA). However, on average, the CLC algorithm performed best for all but one of the viruses. **B-E)** Viral genome coverage was varied, however SMALT and CLC generally achieved superior coverage to Bowtie2. This is particularly true for HCV.

Overall, these findings demonstrate that, when presented with the same dataset as input, SMALT and CLC are capable of consistently producing viral coverage greater than that of Bowtie2, particularly for viruses with a high mutation rate such as HCV. Of the three pipelines tested, CLC displayed the greatest sensitivity in terms of the number of viral sequences detected. The success of CLC is likely due to the fact that its mapping algorithm is able to align reads containing the greatest degree of sequence variation.

3.3.5. Limit of detection: RNA viruses

The performance of the finalised methodology was assessed in terms of both sensitivity and the utility of viral read number as a measure of viral load. Two RNA virus biological reference material standards were serially diluted in clean plasma in duplicate, extracted, and sequenced. The standards used were: heat inactivated HIV-1 diluted in pooled human plasma (NIBSC code 10/152) and; HCV genotype 1a RNA positive plasma, diluted in pooled human plasma (NIBSC code 06/102). Both standards were supplied lyophilised and re-suspended in 0.5mL of RNase-free water as recommended by the suppliers. One exception to the validated protocol used for this experiment was that library preparation and sequencing were outsourced to the Wellcome Sanger Institute and were therefore not validated. However, this modified approach was used in order to precisely replicate the high-throughput method to be applied to all samples sequenced in the following two chapters of this study.

At the highest concentration, HIV-1 was detected in 100% of samples (n = 2) (Table 3.5.A). However, lower concentrations produced no viral reads and HCV was not detected at any of the concentrations tested (Table 3.5.B). This suggests a limit of detection for RNA viruses of between 7,000 and 3,000 viral copies per

library preparation input (equivalent to between 96,000 and 42,000 virus copies per mL of plasma).

The precise detection limit of a metagenomic sequencing method is likely to be highly protocol-dependent. However, in terms of library preparation input, the detection limit found here is comparable to reports of previous Illumina-based metagenomic sequencing studies, which have described a limit of between 1 x 10^4 - 5 x 10^4 RNA virus genome copies.[352,353] It should be noted that the reported methods differ from the approach tested here in that they both employed a protocol which enriched the sample for viral nucleic acid by removing DNA. Therefore, these approaches can be expected to have a slightly increased sensitivity for RNA viruses, however this comes at the expense of excluding DNA viruses entirely.

Intriguingly, sequences from another RNA virus, HPgV were detected in the HCV standard dilutions during informatic analysis. While HCV was not detectable at any dilution, the volume of HCV standard used as input directly correlated with number of HPgV sequences produced (Figure 3.9). This suggests that the plasma provided as an HCV international reference standard is contaminated with HPgV. Unfortunately, calculation of the HPgV load was not possible as none of the standard remained, however this outcome does demonstrate that the method is able to detect RNA viruses and that viral read number is correlated with viral load as expected.

3.3.6. Limit of detection: DNA viruses

Despite the optimisation described in the sections above, no step in the sample preparation pathway is 100% efficient, and viral genomes must successfully pass through multiple bottlenecks in order to be successfully sequenced (Figure 3.2).

A

Sample	Starting Plasma Concentration (copies/mL)*	Library Prep Input (copies)*	% Virus Detected by NGS (n = 2)
HCV_A	42,100	3,158	0
HCV_B	21,050	1,579	0
HCV_C	10,525	789	0
HCV_D	5,263	395	0
HCV_E	2,630	197	0
HCV_F	1,315	99	0
HCV_G	658	49	0
HCV_H	329	25	0
HCV_I	163	12	0
HCV_J	81	6	0
HCV_K	40	3	0
Negative	0	0	0

*Converted from international units (1 IU = 2.5 copies) (Pawlotsky et al 2000)

B

Sample	Starting Plasma Concentration (copies/mL)*	Library Prep Input (copies)*	% Virus Detected NGS (n = 2)
HIV_A	96,656	7,249	100
HIV_B	24,164	1,812	0
HIV_C	6,041	453	0
HIV_D	1,510	113	0
HIV_E	378	28	0
Negative	0	0	0

*Converted from international units (1 IU = 1.74 copies) (Luft et al 2011)

Table 3.5. **Sequencing results for serial dilutions of RNA viruses.** In order to test the limit of detection of the metagenomic method, serial dilutions of international reference standards for HCV (A) and HIV-1 (B) were made in duplicate using healthy (PCR confirmed HCV and HIV-free) plasma. Each dilution was extracted and sequenced in duplicate using the method described in the methodology section of this thesis. As per the finalised method, library preparation input was equivalent to 75 µL of plasma (200 µL of plasma was extracted and 15 / 40 µL of eluate was used for library preparation). International units were converted to viral copy number using previously calculated estimations. [276,277]

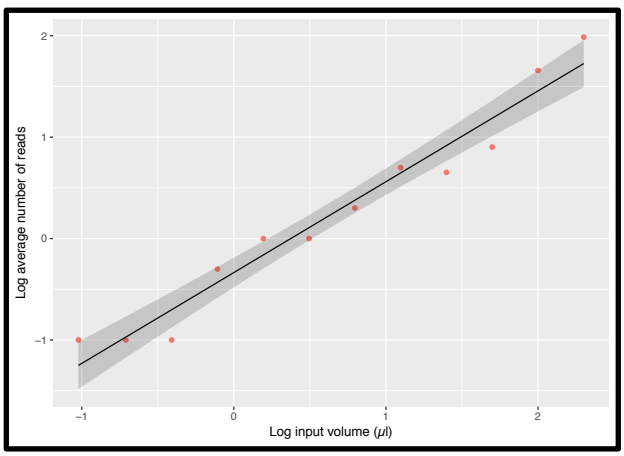


Figure 3.9. **Detection of human Pegivirus in serial dilutions of HCV international standard reference material.** Human Pegivirus sequences were detected in the dilutions of the hepatitis C virus international standard reference material. The number of HPgV sequences correlated with the dilution of the standard, suggesting contamination of the standard rather than the plasma used as a diluent. Number of HPgV reads is shown as the mean of the two duplicate samples prepared and sequenced separately for each dilution.

One major bottleneck for those viruses with RNA genomes is the efficiency of reverse-transcription of RNA to cDNA, and so one would expect the limit of detection for RNA and DNA viruses to differ.

In order to determine the limit of detection for DNA viruses, I took advantage of an existing dataset of paired qPCR and sequencing data generated from 200 patient plasma samples (renal transplant recipients, vasculitis patients, and primary immunodeficiency patients described in chapters 4 and 5 of this thesis). For each sample, the number of TTV reads generated by metagenomic sequencing was compared with TTV viral load as estimated by qPCR. The data showed a strong agreement between qPCR and NGS with regards to TTV status (positive or negative) ($p > 0.001$, Fisher's exact test), and a generalised linear model demonstrated a strong correlation between TTV load as measured by qPCR and number of TTV reads by NGS (normalised between patients by calculating the number of TTV reads per million human reads) ($p > 0.001$) (Figure 3.10.A). This indicates that normalised read number is a reasonable estimate of viral load when comparing between samples.

With regards to the sensitivity of the approach, TTV was detected by NGS in 82% (55/67) of samples with viral loads greater than 1,000 copies per ml of plasma (equivalent to a library preparation input of 75 copies). This increased to 90% (36/40) for samples with viral loads greater than 10,000 copies per ml of plasma. A generalised linear model was generated to model the effect of viral load on likelihood of virus detection by NGS (Figure 3.10.B). The model predicts an increasing probability of NGS detection as viral load increases, rising from 57% (95% c.i. 48-66%) for 1,000 copies/ml of plasma, to 74% (95% c.i. 62-82%) and 85% (c.i. 74-92%) for 10,000 and 100,000 copies/ml respectively. This is less sensitive than the qPCR assay, which has a limit of approximately 1000 copies/ml of plasma. However, reduced sensitivity should be expected for an

assay with such a broad specificity and the result is reasonable given that 1,000 - 100,000 copies/ml represent inputs of approximately 75 - 7,500 copies/library preparation. Furthermore, whilst there was a clear increase in sensitivity with increasing viral load, this model predicts a less clear-cut off value compared to the results of the RNA virus dilution experiment. This is likely to reflect the variability of sequencing with regards to the number of reads produced and library preparation efficiency.

3.3.7. Contamination

Practical measures including the use of hoods, and separation of pre- and post-amplification samples, were taken to minimise the risk of contamination (described in section 2.5). However, the nature of metagenomic sequencing means that care must be taken to identify and exclude contaminants in order to avoid false positive results and false associations.[354] Multiple studies have also reported the detection of bacteria and other environmental contaminants in common laboratory reagents such as DNA extraction kits and purification columns.[252,355] A novel parvovirus traced to contaminated extraction columns shows that this problem is not necessarily limited to bacteria.[356] In order to minimise the influence of environmental contamination, this study did not consider bacteriophages or viruses with a non-vertebrate host.

During this study a number of contamination events were detected. Clear instances of reagent contamination included bovine viral diarrhoeal virus in all PBMC samples containing bovine serum albumin, the contamination of which has been previously reported,[357] and various murine retroviruses in samples that had undergone reverse-transcription using Superscript III reverse transcriptase (the enzyme is derived from reverse engineering of the murine moloney leukemia virus).

Low levels of inter-sample contamination have also been shown to be somewhat inevitable during library preparation and sequencing.[251,358] During this study, it was found that contamination generally followed a clear pattern: low levels of viral reads across consecutive samples that overlapped temporally and spatially in terms of library preparation and sequencing. In agreement with previous studies, instances of contamination were mostly associated with the centre at which library preparation and sequencing was performed, rather than with sample extraction and preparation in the lab.[358] Some cases were particularly obvious, such as the detection of Ebola virus sequences in 24 kidney transplant patient plasma samples. These samples happened to share both a 96-well plate during library preparation, and a lane of a Hi-Seq for sequencing, with samples from African patients with undefined fever. A similar occurrence was detected for HIV-1, where 42 of 48 patient samples sharing a library preparation and NGS lane were positive for the virus (Figure 3.10.A). The likelihood of contamination was further highlighted by the lack of HIV-1 reads in patient samples from the same cohort that were extracted at the same time but prepared and sequenced separately.

Other contamination events were more difficult to claim certainty regarding. For instance, a suspicious pattern of influenza B virus and Epstein-Barr virus was detected in groups of vasculitis patient samples. The presence of these viruses alone is feasible, however their presence in multiple consecutive samples alerted me to the possibility of contamination. In all cases a combination of sequence alignment (where possible) (Figure 3.11.B) and PCR-based assays of the original sample (Figure 3.11.C) was used to confirm or reject contamination. For the purpose of this study, I was able to take the conservative approach of removing all sequences suspected of contamination. However, this issue of contamination must be taken seriously, as in the clinical setting this could lead to the dismissal of important infections in some patients.

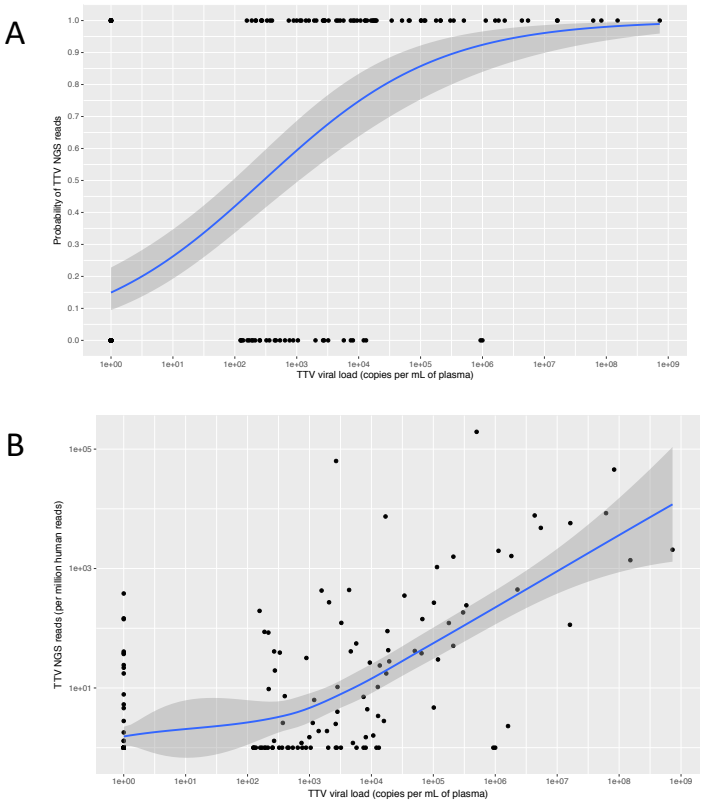


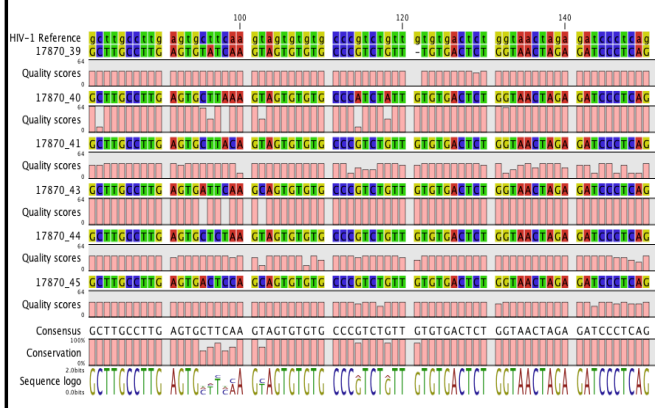
Figure 3.10. **Estimation of the sequencing limit of detection for DNA viruses.**

In order to estimate the limit of detection for DNA viruses, an existing dataset of paired qPCR and sequencing data generated from 200 patient plasma samples (renal transplant recipients, vasculitis patients, and primary immunodeficiency patients described in chapters 4 and 5 of this thesis). For each sample, the number of TTV reads generated by metagenomic sequencing was compared with TTV viral load as estimated by qPCR. **A)** The two methods agreed on the TTV status (-/+) of the samples 76% of the time (152/200) ($p < 0.001$, Fisher's exact test), demonstrating that overall there was significant agreement between the approaches. However, 15.5% (31/200) of samples positive for TTV by qPCR did not produce any TTV sequences by NGS, suggesting that the non-specific NGS approach is not as sensitive as the directed qPCR assay for the detection of TTV. There was also a strong correlation between TTV copy number (as determined by qPCR) and normalised viral read count (as determined by NGS) ($p < 0.001$, generalised linear model). **B)** Using the same data-set, a second linear model was fitted to estimate the effect of viral load (as measured by qPCR) on the likelihood of TTV detection by NGS. The model describes a positive correlation between TTV load and probability of detection by NGS. The model predicts a 57% probability of detection (95% c.i. 48-66%) for viral loads of 1,000 copies/ml of plasma, rising to 74% (95% c.i. 62-82%) and 85% (c.i. 74-92%) for 10,000 and 100,000 copies/ml respectively.

A

Hi-Seq Lane	Index Number	Library Prep	Library Prep Well	Sample Type	HIV-1 Reads
17870	49	399231	A7	CVID	3
17870	48	399231	H6	CVID	0
17870	47	399231	G6	CVID	2
17870	46	399231	F6	CVID	9
17870	45	399231	E6	CVID	1
17870	44	399231	D6	CVID	2
17870	43	399231	C6	CVID	2
17870	42	399231	B6	CVID	0
17870	41	399231	A6	CVID	7
17870	40	399231	H5	CVID	4
17870	39	399231	G5	CVID	4
17870	38	399231	F5	CVID	0
17870	37	399231	E5	CVID	3
17870	36	399231	D5	CVID	10
17870	35	399231	C5	CVID	6
17870	34	399231	B5	CVID	5
17870	33	399231	A5	CVID	7
17870	32	399231	H4	CVID	6
17870	31	399231	G4	CVID	3
17870	30	399231	F4	CVID	4
17870	29	399231	E4	CVID	0
17870	28	399231	D4	CVID	2
17870	27	399231	C4	CVID	3
17870	26	399231	B4	CVID	27
17870	25	399231	A4	CVID	2
17870	24	399231	H3	CVID	4
17870	23	399231	G3	CVID	6
17870	22	399231	F3	CVID	1
17870	21	399231	E3	CVID	3
17870	20	399231	D3	CVID	1
17870	19	399231	C3	CVID	15
17870	18	399231	B3	CVID	2
17870	17	399231	A3	CVID	4
17870	16	399231	H2	CVID	2
17870	15	399231	G2	CVID	3
17870	14	399231	F2	Kidney Transplant	0
17870	13	399231	E2	Kidney Transplant	0
17870	12	399231	D2	Kidney Transplant	3
17870	11	399231	C2	Kidney Transplant	4
17870	10	399231	B2	Kidney Transplant	3
17870	9	399231	A2	Kidney Transplant	5
17870	8	399231	H1	Kidney Transplant	21
17870	7	399231	G1	Kidney Transplant	0
17870	6	399231	F1	Kidney Transplant	3
17870	5	399231	E1	Kidney Transplant	7
17870	4	399231	D1	Kidney Transplant	8
17870	3	399231	C1	Kidney Transplant	18
17870	2	399231	B1	Kidney Transplant	2
17870	1	399231	A1	Primary Immunodeficiency	10

B



C

Virus	Assay Type	Assay Source
BK Polyomavirus	Probe-based qPCR	Dumoulin and Hirsch (2011)
Cytomegalovirus	Probe-based qPCR	Public Health England*
Epstein-Barr Virus	Probe-based qPCR	Public Health England*
Human Immunodeficiency Virus 1	Probe-based qPCR	Yun et al (2002)
Human Papillomavirus	Conventional PCR	Fujinaga et al (1991)
Human Pegivirus	Nested (RT-)PCR	Xiang et al (1998)
Influenza B Virus	Probe-based qPCR	Public Health England*
Parvovirus B19	Probe-based qPCR	Baylis et al (2004)
Simian Virus 40	Probe-based qPCR	Murata et al (2009)
Torque-Teno Virus	Probe-based qPCR	Vasilyev et al (2009)
Vesicular Stomatitis Virus	RT-PCR	Rodriguez et al (1993)

*Performed by the laboratory of Dr M Curran using assays optimised and validated in house

Figure 3.11. **Overview of methods used to detect contamination. A**) An example of the pattern of low level contamination often observed between groups of samples prepared separately (sample types were extracted and prepared on separate days) but sharing a sequencing lane and library preparation plate (Hi-Seq lane and library prep columns). **B)** An alignment of the HIV-1 sequences from the samples highlighted in red in figure A show clear similarities suggestive of a common contaminant. Mismatches generally correspond with low-quality base-calls as shown in the histogram. The displayed region was selected as one of the few with sufficient overlap for comparison of sequences by alignment. **C)** The PCR assays used to confirm or reject the presence of virus detected throughout this thesis.

3.4. Conclusion

In the relatively new field of viral metagenomics, studies have reported highly varied sample preparation, sequencing and analysis protocols. Here, each stage of the viral metagenomic protocol was considered via a combination of literature review and experimentation. From this, a full protocol was decided upon, with the overall aim of detecting all potential viral species with sensitivity and without bias (see thesis methods section, Figure 2.3).

It should be noted that due to time, sample, and cost constraints, not all options were tested exhaustively. For example, ideally the protocol should be tested on a wider range of viruses to confirm that none are being excluded. Furthermore, previous reports of the effectiveness of viral enrichment methods have been varied, however the general consensus is that these approaches are necessary to achieve the sensitivity required for viral sequencing. Therefore, the decision to forego viral enrichment methods is controversial and testing of more forms of viral enrichment and optimising these methods for plasma may have improved the outcome. However, the level of host contamination in plasma is low compared to most other tissues and the above results clearly suggest a benefit in terms of sensitivity and genome coverage for those viruses tested by not performing enrichment. It is also likely that an approach combining separate techniques for the detection of RNA and DNA viruses, or for simultaneously sequencing paired plasma and PBMC samples, would provide additional data and improve sensitivity. However, cost of sequencing and the time taken to prepare samples would effectively be doubled if either of these approaches were used.

Overall, there is a clear requirement for a further work to fully optimise and validate a method for the metagenomic detection of viruses from blood samples. In order to be truly useful in a clinical or diagnostic setting this method must be

sensitive and specific but should not require costly equipment or reagents, or long, complex preparation methods.

Chapter 4

The Virome During Therapeutic Immunosuppression

4.1. Introduction

Improved immunosuppressive regimes following solid organ transplantation have led to a vast reduction in the incidence of acute rejection and improved graft survival rates in the short-term.[359] However, despite these advances, survival rates in the mid- to long-term have not improved proportionately.[360]

Long-term immunosuppressive regimes have been associated with high rates of morbidity and mortality due to increased risk of infection, cardiovascular disease and malignancy.[361] Opportunistic infections are common following transplantation,[362] and are a major cause of repeated hospital visits which frequently result in hospitalization and long-stays.[363] Quality of life studies have found that recipients of lung, heart, liver and kidney transplants display a spectrum of reduced health factors compared to the general population.[364]

4.1.2. Clinical viral diagnostics for solid organ transplant patients

Routine clinical screening in organ donors and recipients comprises of fungal and bacterial culture as well as targeted PCR assays for "high risk" viral pathogens including HIV, HBV and HCV, and the herpesviruses CMV and EBV.[365,366] In renal transplant patients, the polyomavirus BKV is also monitored, as it is responsible for severe nephropathy for which the only available treatment is to reduce immunosuppression immediately, risking immune-mediated graft rejection.[367] Further diagnostics tend to rely on clinical suspicion,

however recognition of infection is difficult due to immunosuppression which may mask the symptoms of infection.

Previous qPCR-based studies have investigated the dynamics of the most clinically-relevant viruses following transplantation.[368] However, prior to the development of low-cost, high-throughput sequencing, no studies considered the entire spectrum of viral infections to which immunosuppressed patients are at risk. Known, but poorly monitored infections of transplant patients include skin infections by papillomavirus and molluscum contagiosum virus, gastrointestinal infections, and respiratory infections by viruses such as influenza virus, adenovirus and rhinovirus.[369–371] The need for this work has been further highlighted by recent studies demonstrating the persistence of viruses such as norovirus and rhinovirus in immunosuppressed patients, and the isolation of a previously unknown poxvirus from a renal transplant patient in the United States.[372–374] Immunosuppressed carriers of these pathogenic viruses may act as reservoirs with the potential to seed new outbreaks.

4.1.3. The virome in secondary immunodeficiency

Several recent studies have demonstrated that perturbations of the immune system can have a marked effect on both the pathogenic and commensal communities of bacteria and viruses that inhabit our bodies.[375,376] These communities are known collectively as the microbiome, of which, the viral component is known as the virome.

Studies of the virome rely on a non-targeted approach such as metagenomic sequencing, which only truly became possible with the development of affordable high-throughput sequencing technologies (as discussed in chapter 3). Therefore, at the time of starting this thesis, publications in the field with regards

to immunosuppression were somewhat sparse. However, in 2013, the first year of this work, a seminal paper on the effect of immunosuppression on the virome was published.[274] The study by De Vlaminck et al investigated the response of the blood plasma DNA microbiome to immunosuppressive therapy in heart and lung transplant recipients, and found that increasing immunosuppression was found to correlate with higher viral loads whereas the bacterial component remained relatively constant.[274] Varying levels of immunosuppression also had a dramatic effect on the DNA virome's composition; at low dosage, viruses of the *Herpesviridae* family tended to dominate, whilst higher dosages resulted in an increased proportion of *Anelloviridae*.

Further metagenomic studies of various patient groups have revealed that the *Anelloviridae* family, which consists of a highly genetically diverse group of viruses with small (2 - 3.8kb), single-stranded DNA genomes, dominate the immunosuppressed virome.[328,377,378] This group includes torque-teno virus (TTV), and the smaller torque-teno midi and mini viruses (TTMDV and TTMV). At present, these species are thought to be non-pathogenic and predominantly commensal, as they have been detected at a high prevalence in healthy, asymptomatic individuals worldwide.[45–47] During the course of this thesis, several studies have demonstrated a close correlation between increased anellovirus loads and increased levels of immunosuppression.[379,380] Two further studies have investigated the plasma microbiome of HIV/AIDS patients: comparing those with high and low CD4 T cell counts as a measure of progression to AIDS;[275] and against healthy controls.[381] The first of these studies found that the plasma of treatment naive HIV/AIDS patients carried more bacteria (5.36 % of total sequenced nucleotides) than healthy controls (0.0084 %).[381] The DNA virome of the HIV/AIDS cohort was dominated by bacteriophages and endogenous retroviral elements whereas the virome of the health cohort exclusively contained anellovirus. This study did not try to detect

RNA viruses in the plasma but a follow-up study was able to detect HIV-1, GB virus C (GBV-C), hepatitis B virus (HBV), hepatitis C virus (HCV), anellovirus and human endogenous retrovirus (HERV) sequences from similar patients.[275] Interestingly, this association between anellovirus and the strength of the immune system has led to the virus being proposed as a marker of immune reconstitution following stem-cell transplantation as well as a measure of under-suppression of the immune system following solid organ transplant (and acute rejection as a consequence).[274]

Due to the field being relatively young, the impact of the changes described above is not well understood. However mounting evidence suggests that both the commensal and pathogenic components of the virome play an important immunomodulatory role in the host; altering the nature of the immune response and contributing to both health and disease.[38,271,382]

At the beginning of this thesis, metagenomic sequencing studies of immunosuppressed patients had focused entirely on DNA viruses in AIDS patients and transplant recipients.[306,307,383] Disruption of the virome can have marked health implications and therefore further investigation of the immunosuppressed virome was warranted;[368,382] RNA viruses in particular had been understudied. Furthermore, unbiased surveillance of the viruses that infect immunosuppressed populations is strategically important if we are to direct appropriate diagnostics and also to understand the potential role of immunodeficient individuals as carriers of pathogens.

4.1.4 Aims

Here, I combined directed PCR assays and an unbiased metagenomic approach to detect both RNA and DNA viruses in blood over the course of

immunosuppressive therapy and investigated whether particular viruses are associated with poor clinical outcomes such as acute rejection or opportunistic infections. Furthermore, in order to test whether the results could be applied to other groups of immunosuppressed patients, metagenomic sequencing was also performed on samples from a cohort of vasculitis patients. This enabled the first ever virome comparison between two forms of long-term therapeutic immunosuppression, however further analysis of the vasculitis cohort awaits follow-up clinical data.

4.2. Methods

4.2.1. Ethics and Cohort Overview

All samples were obtained with informed written consent and the approval of the Cambridge Local Research Ethical committee (reference **08/H0308/176**).

Two cohorts of secondary immunodeficiency patients were used in this study, both undergoing therapeutic immunosuppression. The first cohort consisted of 50 renal transplant recipients, receiving life-long immunosuppressive therapy following kidney transplantation in order to prevent immunological rejection of the graft. Transplant patients enrolled in the study all received the same induction therapy to control for pre-study differences between individuals. The maintenance immunosuppressive regime for transplant recipients consisted of an antimetabolite (azathioprine or mycophenolate mofetil) which inhibits lymphocyte proliferation by inhibiting purine synthesis,[384] a calcineurin inhibitor (tacrolimus or cyclosporin A) which inhibits T-cell differentiation and effector functions,[385] and a corticosteroid (prednisolone) which inhibits regulation of pro-inflammatory genes.[386] The second cohort consisted of individuals diagnosed with anti-neutrophil cytoplasmic autoantibody (ANCA)-associated vasculitis. These patients suffer from chronic inflammation resulting in damage to the small blood vessels. The inflammation is caused by autoantibodies targeting one of two proteins (PR3 or MPO), both of which are predominantly found in the cytoplasm of neutrophils.[387] Treatment for ANCA-associated vasculitis is long-term therapeutic immunosuppression.[388] Unlike post-transplant immunosuppression, treatment for these patients is highly varied between individuals (see table 4.1). Drug choices are dependent on both the severity and stage of disease, and the use of newer drugs such as mycophenolate mofetil is not well standardised, due, at least in part, to the rarity of the disease hindering the development of an evidence based, uniform

protocol.[389] Unfortunately, a lack of available clinical information for the vasculitis cohort made it impossible to test for meaningful associations between the virome and patient outcome. Therefore, the data generated from the vasculitis cohort was restricted to being used to make basic comparisons of the virome between the two immunosuppressed cohorts suffering from distinct diseases.

Blood samples from renal transplant recipients were taken on the day of transplant and 90 days post-transplant. Blood samples from vasculitis patients were taken on the first day of treatment, 90 days, and 12 months post-treatment. Blood samples were also taken from healthy volunteers to be used as controls.

Cohort	Sampling time-points (days post commencement of immunosuppression)	Immunosuppressive drugs used for treatment
Kidney transplant	0 (n = 50), 90 (n = 50)	Prednisolone, mycophenolate mofetil OR azathioprine, and tacrolimus OR cyclosporin A
ANCA-associated vasculitis	0 (n = 43), 90 (n = 36), 365 (n = 24)	Various combinations of: tacrolimus, azathioprine, mycophenolate mofetil, cyclosporine A, infliximab, adalimumab, rituximab, cyclophosphamide, and 15-deoxyspergualin

Table 4.1. **Overview of the secondary immunodeficiency cohorts.** Two cohorts of secondary immunodeficiency patients were studied: kidney transplant recipients and ANCA-associated vasculitis patients. Patients in both cohorts receive long-term therapeutic immunosuppression to treat their respective diseases. However, the types of drug used to treat vasculitis patients are more varied, making comparisons difficult for this group. Paired blood samples were taken at various time-points onset on immunosuppression in order to assess the longitudinal impact of immunosuppression on the blood virome.

4.2.2. Sample Preparation and Sequencing

Nucleic acid was extracted from plasma samples, prepared, sequenced and analysed as described in sections 2.1-2.3 of this thesis. An overview of the sequencing statistics for each group are shown in section 2.4.

4.2.3. Contamination and Confirmation of Results

Given the low abundance of nucleic acid in plasma, potential cross-sample contamination was thoroughly investigated. Typical patterns of contamination such as the same virus occurring in multiple consecutive samples in a lane were confirmed by nucleotide alignments and PCR testing of the original sample. To ensure no false associations were made, all viral sequences with any evidence of contamination were excluded from analysis. In particular, HIV-1 derived reads in samples 23-36 (day 0) from lane number 17870_1#2-12 was excluded.

4.2.4. Anellovirus Quantification

Plasma anellovirus load was quantified by qPCR as described in section 2.5.

4.2.5. MinION Sequencing of Anellovirus Populations

A protocol was devised in order to further assess the diversity of anellovirus populations in the plasma of immunosuppressed patients. To begin, circular DNA, which includes anellovirus genomes, was specifically amplified from nucleic extracted from plasma by rolling circle amplification (RCA) using random hexamer primers. RCA was performed on samples from 10 randomly selected renal transplant recipients using the Templiphi kit (GE Healthcare) according to the manufacturer's recommendations. Briefly, 15 µl of extracted total nucleic

acid was heat denatured at 95°C for 3 mins and cooled to 4°C. Following denaturation, the circular DNA was amplified using the Phi29 DNA polymerase enzyme and random hexamers for primers in a 30 µl volume. Reactions were incubated at 30°C for 18 hours. RNase-free water was used as a negative control and a plasmid (pUC19) was provided as a positive control in the kit. Agarose gel electrophoresis was used to confirm the presence of an extremely high molecular-weight product, indicating a successful amplification.

The amplified product from each sample was split into five, 5µl aliquots, and each product was treated using a different restriction enzyme capable of cutting multiple anellovirus species based on reference sequence alignments (Table 4.2.). The restriction digest products were visualised on a 2% agarose gel under a UV light. Products approximately the size of an anellovirus genome (~3-4 Kb) were cut from the gel and purified using the QIAquick Gel extraction kit (Qiagen).

Preparation of the DNA products for sequencing was performed using the MinION 2D Native Barcoding Kit (Oxford Nanopore Tech) but adapted for low input amounts following Josh Quick's protocol (www.zibraproject.org/blog/protocol-low-input-native-barcoding-protocol, full protocol in appendices). Briefly, 100ng of amplicon was diluted to 30 µl for each sample. End-repair was performed using the Ultra II End-Repair Kit, according to manufacturer's instructions. Barcode oligomers NB01-NB09 were ligated for each sample in individual reactions according to the manufacturer's protocol, and barcoded samples were then combined. Sequencing adaptor ligation was performed on the combined, barcoded samples in a single reaction and the final library was quantitated by Qubit high sensitivity double-stranded DNA kit (Thermo Fisher Scientific). 100 ng of the resulting library was loaded into a vR9 Flow-cell 1B (Oxford Nanoporetech) and the 2D basecalling software was run for 19 hours on the MinION sequencer Mk 1B (Oxford Nanoporetech).

Enzyme	Manufacturer	Conditions	Recognised Sequence	TTV (Reference) Species Cut
SmaI (FD)	Thermo Fisher	37°C for 20 mins	CCC^GGG	1,2,4,6,7,10,12,16,19,25,26,2 7,28,29
Acc65I (FD)	Thermo Fisher	37°C for 20 mins	G^GTACC	1,2,3,4,6,7,8,10,12,15,16,19,2 6,27,29
AccI (FD)	Thermo Fisher	37°C for 20 mins	GT^MKAC	1,2,3,4,6,7,8,10,12,15,16,19,2 7,28,29
SacI	Roche	37°C for 120 mins	G_AGCT^C	1,2,3,4,6,7,8,10,12,15,16,19,2 6,27,28,29*
ApaI	NEB	37°C for 120 mins	G_GGCC^C	1,2,3,4,6,7,8,10,12,15,16,19,2 6,27,28,29*

Table 4.2. **Enzymes used for restriction digest of rolling circle amplification products**. Restriction enzyme recognition sequences were compared to the reference database of 16 TTV genomes. Enzymes that have recognition sequences present within all of the TTV species are denoted by an asterisk.

FASTQ files from MinION sequencing were mapped to a FASTA file composed of 18 TTV reference sequences using GraphMap v0.5.1,[390] allowing for circular genomes. GraphMap is designed to quickly and accurately align long reads and is also able to robustly handle the higher rates of error often observed in nanopore sequencing reads. Default settings were used, including an e-value alignment score threshold of 1e0 required for successful alignment, reads scoring above this threshold were called unmapped. Reads mapping ambiguously to more than one TTV reference genome were also called unmapped, although the high level of inter-species variation combined with long reads meant this was a rare occurrence. The number of reads mapping to each reference was then extracted

using samtools idxstats,[391] and visualised using ggplot2 in R.[392]

4.2.6. Statistical Methods

Basic statistical analysis of sequencing and qPCR results was performed using the methods described in section 2.6 of this thesis.

Dr Simon Frost constructed a marginal structural model for a continuous exposure from the data, in order to estimate the association between anellovirus load and infectious disease complications, controlling for anti-metabolite.[393] This method is more robust than a standard logistic regression at controlling for the effects of other variables (e.g. antimetabolite used). Stabilised inverse probability weights (IPWs) were calculated using either linear regressions of log10-transformed anellovirus viral load (using lm in R), to generate normal IPWs or a cumulative logistic regression of deciles of anellovirus viral load (using polr from the MASS library in R).[394] These weights were used in a weighted logistic regression as implemented in the survey library in R.[395,396] The quantile IPWs gave similar estimates of the odds ratio of complications per log10 TTV load as the normal IPWs, and only the results for the normal IPWs are presented.

4.3. Results

4.3.1 Overview of Illumina sequencing data

Analysis of the Illumina sequencing data revealed that raw read count varied significantly across the patient and control groups ($p < 0.0001$, One-way ANOVA, Fig 4.1.A.). However, multiple comparison tests showed that time-points within the same cohorts were well matched: within the kidney transplant recipient cohort, average raw read count did not significantly differ between the day 0 time-point and the day 90 time-point ($p = 0.99$, Tukey's multiple comparison test) and the same was observed between the three vasculitis cohort time-points (p values = 0.21 - 0.98, Tukey's multiple comparison test).

On average, the healthy control group produced fewer sequencing reads than the patient groups. In particular, the number of reads produced by the healthy control samples was significantly lower than the number generated by the vasculitis cohort at day 0 and day 90, ($p < 0.05$, Tukey's multiple comparison test). A low read number may reduce the potential to detect viruses and therefore result in an underestimation of viral prevalence in the control group. However, it should be noted that during the preparation of the samples, those from the healthy group were split to be prepared and sequenced across multiple sequencing runs. Therefore, the consistently low number of reads produced by these samples may be due to an inherent quality of the samples rather than the result of multiple poor library preparations or sequencing runs. For example, it has been reported that blood plasma levels of circulating cell-free DNA, which is derived from dead and dying cells, are generally low in the blood of healthy individuals, but increased in patients with diseases including sepsis, arthritis, lupus, stroke and cancer.[4] Therefore, it is possible that the difference in read count between the healthy and patient cohorts in this study is due to a lower level of circulating cell-free DNA in healthy plasma.

Across all groups, a high proportion of reads were lost during the quality control stage of computational analysis ($\bar{x}$ = 24.7 – 52.8%, Fig 4.1.B.), suggesting that overall quality of sequence data was poor. Quality control of sequence data was performed using CLC software, which uses a modified version of Mott's trimming algorithm to remove low quality bases, by calculating the probability of an erroneous base-call based on Phred quality scores.[5] If a read contained > 1 ambiguous base-call, consisted of low complexity repeats (e.g. polyA/T), or less than 50 bases (33% of the read) remained following trimming, the entire read was discarded to avoid non-specific mapping in the steps to follow.

Following quality control, host-derived sequences were subtracted from the dataset by mapping to the human reference genome, in order to reduce the number of reads and speed up later analysis. As expected, a high percentage of the reads from all groups were of human origin (70.1 - 87.2%, Fig 4.1.C.).

The average number of reads remaining following quality control and host read subtraction ranged from 10^3 - 10^7 across groups (Fig 4.1.D.). Similar to the raw sequence dataset described above, there was a significant difference in average read number between groups ($p < 0.0001$, one-way ANOVA), but no significant difference was observed between patient groups within the same cohort (p values = 0.68 – 1.00, Tukey's multiple comparison test). Therefore, despite overall differences in read count and read quality between groups, and so comparisons of viral prevalence at different time-points within the same cohort should be largely unaffected.

4.3.2. Viruses Detected by Sequencing in Renal Transplant Patient Plasma

Marked changes were observed in the plasma virome following renal transplantation and 3 months of immunosuppression (Figure 4.2). There was a

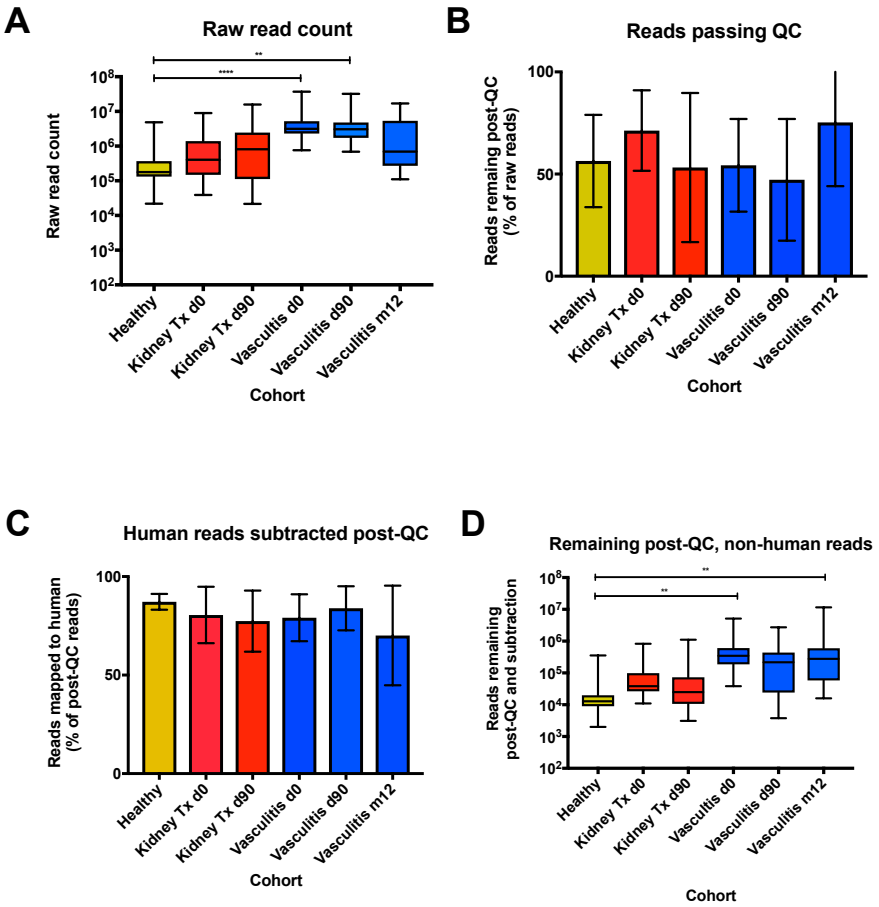


Figure 4.1. **Illumina sequencing data statistics.** A) Total number of sequencing reads generated for each cohort, grouped by treatment time-point. Comparing all groups simultaneously, there was a significant difference in average raw read count ($p < 0.001$, one-way ANOVA). Comparisons of individual groups showed that average read counts differed significantly between two of the vasculitis patient groups and the healthy control group ($p < 0.0001$ (vasculitis day 0) and $p = 0.0037$ (vasculitis day 90), Tukey's multiple comparison test). B) Trimming and removal of reads containing low quality and ambiguous bases resulted in the loss of an average of 24.7 – 52.8% of sequences. C) Host-derived reads that passed quality filters were subtracted by mapping to the human reference genome, removing an average of 70.1 – 87.2% of the reads from each group. D) Following QC and host-subtraction there remained an overall significant difference in average read count between groups ($p < 0.001$, one-way ANOVA). However, this was again due to a difference between the vasculitis patient groups and the control groups, ($p < 0.0017$ (day 0) and $p = 0.0095$ (month 12), Tukey's multiple comparison test). Mean and standard deviation is shown for all figures.

significant increase in the prevalence of the anelloviruses: TTV (McNemar's test, $p < 0{\cdot}0001$) and TTMDV (McNemar's test, $p < 0{\cdot}0055$), as well as the flavivirus HPgV (McNemar's test, $p < 0{\cdot}0026$). Clinically relevant viruses such as BK polyomavirus, simian virus 40 (SV40) polyomavirus, EBV, and other human herpesviruses were also detected but at a low prevalence which was not significant. Presence of each virus was confirmed by PCR to compensate for the low sensitivity afforded by the broad NGS approach. Specific PCR assays revealed two additional BK Polyomavirus infections in patient samples that were not picked up by NGS screens of plasma (marked in grey).

4.3.3. Effect of Immunosuppression on Anellovirus Load in Renal Transplant Patients

Given the high prevalence of anellovirus, a qPCR assay was employed to quantify anellovirus load and discriminate better between the virome of patients. Following transplantation and the commencement of immunosuppression there was a highly significant increase in anellovirus load (Wilcoxon test, $p < 0.0005$) whilst there was no difference in anellovirus load between patients prior to transplantation and healthy volunteers (Figure 4.3.A).

A significant increase in both anellovirus species and genogroups was detected following immunosuppression (Figure 4.3.B, paired t-test, $p < 0.0001$). Regression analysis using a general linear model showed a strong positive correlation between anellovirus load and the number of anellovirus species detected from the NGS data (Figure 4.3.C, paired t-test, $p < 0.0001$).

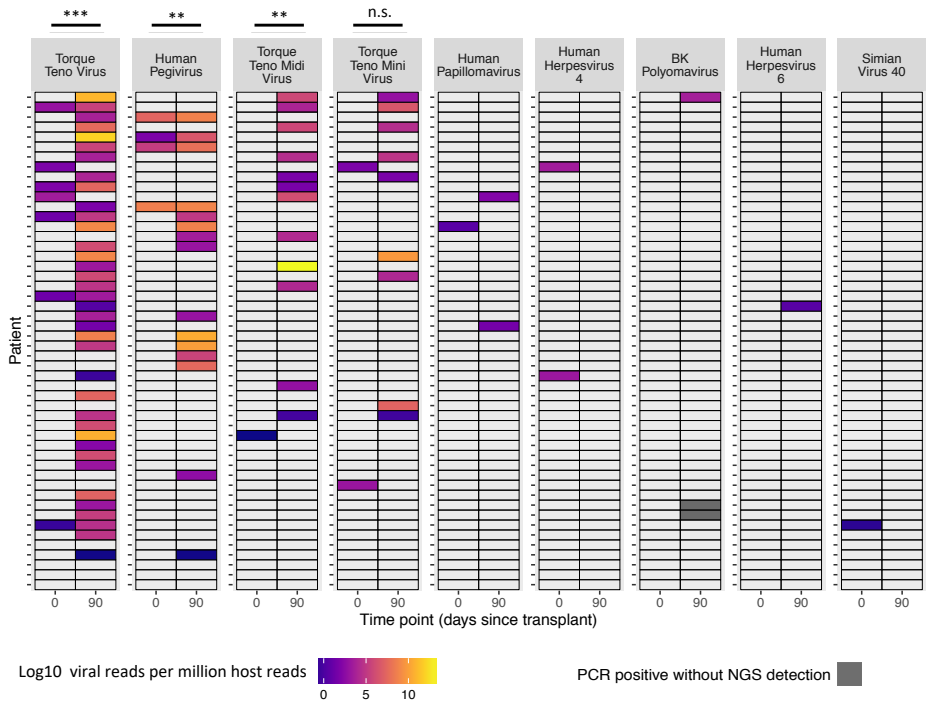


Figure 4.2. **Viruses detected by metagenomic Illumina sequencing of renal transplant patient plasma.** There was a significant increase in the prevalence of the anelloviruses: TTV (McNemar's test, p < 0·0001) and TTMDV (McNemar's test, p < 0·0055), as well as the flavivirus HPgV (McNemar's test, p < 0·0026). Clinically relevant viruses such as BK polyomavirus, simian virus 40 (SV40) polyomavirus, EBV, and other human herpesviruses were also detected but at a low prevalence which was not significant. Sampling time points represent the day of transplantation (prior to commencement of maintenance immunosuppression) and 90-days post transplant. Viral read count is shown as a ratio of host-derived reads in order to normalize between individuals. Virus presence was confirmed by PCR. In two separate patients, the polyomavirus BK virus (BKV) was detected by qPCR in samples that had no detectable BKV sequencing reads by NGS. These are identified in grey in the figure.

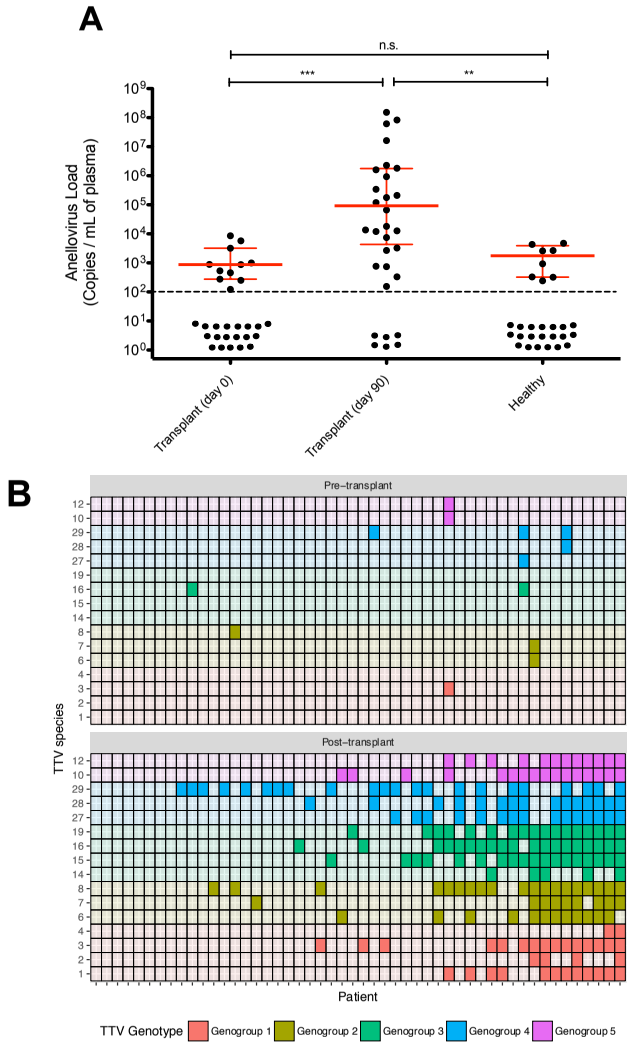


Figure 4.3. **The effect of immunosuppression of renal transplant patients on anellovirus load and diversity. A)** Anellovirus load was measured by qPCR in paired plasma samples from renal transplant recipients pre- and 90 days post-immunosuppression (n = 30) and healthy volunteers (n = 28). Viral load was significantly increased following 90 days of immunosuppression compared to the same patients at day 0 ($p < 0.0005$, Wilcoxon test) and healthy controls ($p < 0.0005$, Mann-Whitney test). Limit of detection for the assay is indicated by the dotted line. Error bars show median. **B)** A model of anellovirus load measured by qPCR against the number of anellovirus species detected by Illumina sequencing demonstrated a positive correlation between the two. Only reads that aligned exclusively to a single reference sequence were considered. **C)** A significant increase in both anellovirus species and genotype diversity post-transplant (paired t-test, $p < 0.0001$) was detected.

4.3.4. Torque-Teno Virus Correlates with Opportunistic Infections in Renal Transplant Patients

Given its association with immunosuppression, I next wished to determine whether anellovirus load could be used to identify those patients more susceptible to infection following transplant. In order to do this, I examined individual's patient records for infectious complications in the year following sampling (Table 4.3.). Infectious complications were defined as follows: bacterial infections requiring the prescription of >= 4 courses of oral antibiotics within the first-year post transplant or, hospital admission and intravenous antibiotic for severe, culture-positive bacterial infection; active viral infection as diagnosed by PCR-based assays during routine post-transplant surveillance; fungal infections as diagnosed by culture methods.

Simon Frost performed a logistic regression analysis to identify correlates of infectious complication, measuring the predictive ability of these models by calculating the area-under-the-curve (AUC) of a receiver-operator characteristic (ROC) curve. ROC curves can be used to quantify how accurately a diagnostic test differentiates between two patient states (e.g. infected or not), with the highest possible AUC of 1.00 meaning that a test correctly identifies the true state 100% of the time.[397] Analysis revealed that anellovirus was significantly associated with the incidence of infectious complication in a dose-dependent manner; for each log10 increase in viral load, the odds ratio (OR) for infectious complications increased by 1.44 ($p = 0.0143$, 95% confidence intervals = 1.10 - 1.98). Given that anellovirus load varied by over seven orders of magnitude, this is a strong effect, with an AUC of 0.72 (Figure 4.4.A). Presence of the virus alone was not sufficient to predict infectious events: when the information on anellovirus load was omitted, and only anellovirus infection status considered, the statistical significance and AUC fell (overall OR = 7.0, $p = 0.0803$, AUC = 0.60). The

correlation between anellovirus viral load and infectious complications was further demonstrated using the Kaplan–Meier method (Cox-rank test, $p = 0.004$), based on an arbitrary 'high viral load' threshold of 1,000 copies per mL of plasma (Figure 4.4.B).

The logistic regression found that infectious complications were also associated ($p < 0.05$) with the antimetabolite used (OR = 22.75, 95% c.i. = 5.0-166.6), and delayed graft function (OR = 5.0, 95% c.i = 1.37-21.5), with a trend ($p < 0.1$) for patients who underwent peritoneal dialysis (OR=6.88), who experienced rejection (OR = 3.89) or were treated with pulse steroid therapy (OR = 3.18) (Table 4.4.). The model for anellovirus had the second highest predictive performance for infectious complications (AUC = 0.72) following the use of antimetabolite (AUC = 0.78). Interestingly, the antimetabolite used was also associated with anellovirus load. However, even after controlling for the antimetabolite, anellovirus load was independently associated with a higher odds for infectious complication (OR = 1.4, 95% c.i. = 1.06-1.85 per log10).

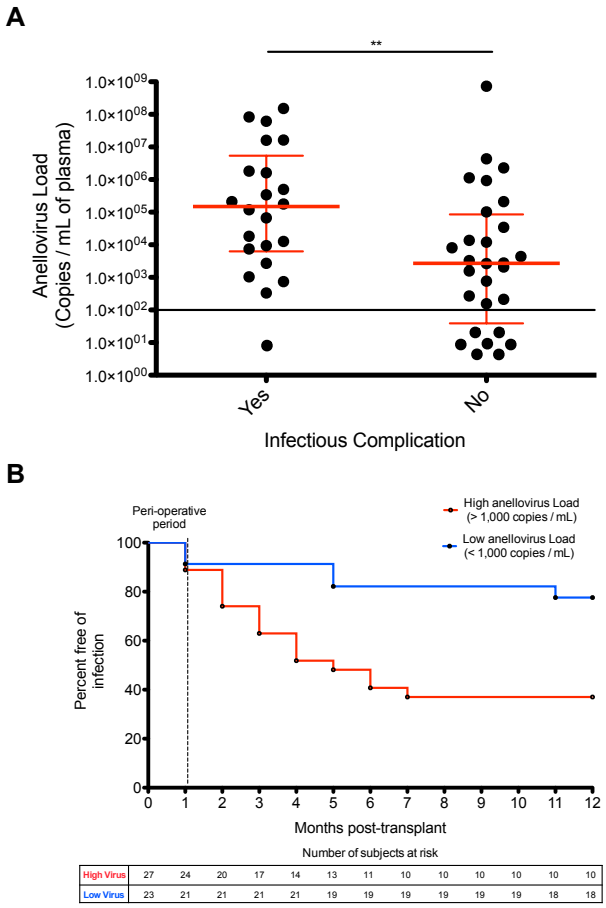


Figure 4.4. **Anellovirus load correlates with future infectious complications in renal transplant recipients. A)** Anellovirus load at 90 days post transplant was significantly increased in patients that went on to suffer from an infectious complication in the following year (n = 24) compared to those that did not (n = 26) (p = 0.0046, Mann-Whitney test. Infectious complications were defined as follows: bacterial infections requiring the prescription of >= 4 courses of oral antibiotics within the first-year post transplant or, hospital admission and intravenous antibiotic for severe, culture-positive bacterial infection; active viral infection as diagnosed by PCR-based assays during routine post-transplant surveillance; fungal infections as diagnosed by culture methods. Defined infectious complications. Error bars show median. **B)** Kaplan-Meier plot demonstrating the incidence of infectious complications in the year following transplant for patients with high Anellovirus load (n = 27) and low Anellovirus load (n = 23). (p = 0.0044, Mantel-Cox Test). Significance increases if infections in the peri-operative period, which are generally associated with surgery rather than immunosuppression, are excluded (p = 0.0021, Mantel-Cox test).

Patient	AVL (copies / plasma mL)	Primary diagnosis	Donor type	Form of dialysis	Infection (months post-transplant)[a]	Description of infections (chronological order)	ACR therapy (days post-transplant)	Calcineurin inhibitor	Anti-metabolite
1	7289512	Reflux	HB	Haemodialysis				FK	MMF
2	151930600	Lithium nephrotoxicity	LUR	Haemodialysis	5	BKV, recurrent urinary and chest sepsis	39	CyA	MMF
3	83181820	ADPKD	NHB	Haemodialysis	7	CMV, recurrent infectious complications (inc fungal)	7	FK	MMF
4	61521940	Diabetes	NHB	Haemodialysis	7	CMV, tooth infection, acute EBV		CyA	Aza
5	16306980	Reflux	LUR	Pre-dialysis	5	Recurrent UTIs, intermittant E coli	7	FK	MMF
6	1606284	N/K	NHB	Pre-dialysis	4	CMV, PCP pneumonia		CyA	MMF
7	4328282	ADPKD	LUR	Haemodialysis			10	CyA/FK	Aza
8	2283750	IgAN	LRT	Haemodialysis			6	CyA	Aza
9	1820243	Membranous	HB	Haemodialysis	8	BKV		FK	MMF
10	1607811	Membranous	LRT	Haemodialysis	6	BKV, recurrent LRTIs	10	FK	MMF
11	1128093	Hypertension	LUR	Haemodialysis			35	CyA/FK	Aza
12	931814	Membranous	LRT	Pre-dialysis				CyA/FK	Aza
13	499931	MCGN Type 1	HB	Haemodialysis	3	UTI, LRTIs	32	FK	MMF
14	341668	ADPKD	NHB	Pre-dialysis	3	BKV, CMV, EBV, PCP pneumonia, UTI and LRTIs	49	CyA/FK	MMF
15	210777	Hypertension	HB	Peritoneal	1	UTIs		FK	Aza
16	209786	ADPKD	HB	Haemodialysis				CyA/FK	Aza
17	178212	SLE	NHB	Haemodialysis	2	CMV, LRTIs		FK	MMF
18	118813	Urinary stone disease	LUR	Haemodialysis	3	CMV		FK	Aza
19	102175	Diabetes	LRT	Haemodialysis				CyA	Aza
20	66508	ADPKD	LRT	Peritoneal	1	Post-op wound infection		FK	Aza
21	34127	Hypertension	LRT	Haemodialysis				FK	Aza
22	18142	N/K	LRT	Haemodialysis	2	CMV, BKV		CyA	MMF
23	13645	N/K	LRT	Peritoneal				CyA/FK	Aza
24	12680	Reflux	LRT	Haemodialysis	3	BKV	8	FK	Aza
25	12063	Congenital renal dysplasia	NHB	Pre-dialysis				FK	Aza
26	9426	N/K	NHB	Haemodialysis	1	Post-op wound infection		CyA	MMF
27	8115	ADPKD	HB	Haemodialysis				CyA	Aza
28	7458	Hypertension	NHB	Haemodialysis	1	Post-op LRTI		FK	Aza
29	4362	IgAN	LRT	Haemodialysis				CyA	Aza
30	3265	Posterior urethral valves	LRT	Haemodialysis				CyA	Aza

Table 4.3. **Overview of transplant patient metadata (part 1).**

Patient	AVL (copies / plasma mL)	Primary diagnosis	Donor type	Form of dialysis	Infection (months post-transplant)[a]	Description of infections (chronological order)	ACR therapy (days post-Tx)	Calcineurin inhibitor	Anti-metabolite
31	2757	IgAN	LRT	Pre-dialysis				CyA/FK	Aza
32	2686	FSGS	HB	Peritoneal	3	BKV	8	FK	MMF
33	2659	IgAN	LRT	Pre-dialysis			7	FK	Aza
34	2069	IgAN	LRT	Pre-dialysis				FK	Aza
35	1566	ADPKD	LRT	Pre-dialysis				FK	Aza
36	1043	FSGS	NHB	Peritoneal	12	CMV		CyA/FK	MMF
37	Below detection[b]	ADPKD	LRT	Pre-dialysis				FK	Aza
38	Below detection[b]	ADPKD	NHB	Haemodialysis	1	UTIs before stent removed	8	CyA	MMF
39	Below detection[b]	ADPKD	LUR	Pre-dialysis	6	BKV		FK	Aza
40	Below detection[b]	Wegner's granulomatosis	LUR	Haemodialysis				FK	Aza
41	Below detection[b]	Chronic TIN	LRT	Pre-dialysis				CyA	Aza
42	Below detection[b]	Obstructive uropathy	LRT	Haemodialysis				CyA	Aza
43	Below detection[b]	N/K	LRT	Pre-dialysis				CyA	Aza
44	Below detection[b]	Chronic Pyelonephritis	HB	Haemodialysis				FK	MMF
45	Below detection[b]	Reflux	LRT	Pre-dialysis			9	CyA	Aza
46	Below detection[b]	MCGN Type 2	LRT	Peritoneal	6	CMV		FK	Aza
47	Below detection[b]	IgAN	LUR	Haemodialysis				FK	Aza
48	Below detection[b]	IgAN	LRT	Pre-dialysis				CyA	Aza
49	Below detection[b]	GN	LRT	Peritoneal				CyA	Aza
50	Below detection[b]	Reflux	LRT	Haemodialysis				CyA/FK	Aza

Table 4.3. **Overview of transplant patient metadata (continued).** Metadata was used for logistic regression analysis in order to identify correlates of infectious complication. Anellovirus load was quantitated by qPCR. Only infectious complications reported within 12 months post-surgery were considered. Calcineurin inhibitors were either tacrolimus (FK) or cyclosporin A (CyA). Anti-metabolites were either mycophenolate mofetil (MMF) or azathioprine (Aza). [a] Infections were defined as follows: Bacterial infections requiring the prescription of>= 4 courses of oral antibiotics within the first-year post transplant or; hospital admission and intravenous antibiotics for severe, culture-positive bacterial infection; viral infection was diagnosed on the detection of active viral replication in protocol post-transplant viral surveillance blood samples by PCR methods; fungal infections were diagnosed by culture. [b] Limit of detection for the qPCR assay was equivalent to 1000 viral genome copies per mL of plasma.
Abbreviations: ACR, acute chronic rejection; ADPKD, autosomal dominant polycystic kidney disease; AVL, anellovirus load; AZA, Azathioprine; BKV, BK polyomavirus; CMV, cytomegalovirus; CyA, Cyclosporin A; EBV, Epstein-Barr virus; FK, Tacrolimus; FSGS, focal segmental glomerulosclerosis; GN, glomerulonephritis; HB, heart-beating donor; IgAN, IgA nephropathy; LRT, living related transplant; LRTI, lower respiratory tract infection; LUR, living unrelated donor; MCGN, mesangiocapillary glomerulonephritis; MMF, Mycophenolate mofetil; NHB, non-heart beating donor; N/K, not known; PCP, Pneumocystis pneumonia; SLE, systemic lupus erythematosus; TIN, tubulointerstitial nephritis; UTI, urinary tract infection.

Variable	Odds ratio	AUC	P-value
Anti-metabolite	22.75	0.78	0.0036
TTV load (per log10)	1.44	0.72	0.0143
TTV load (10^1 copies per mL)	1.44	0.72	0.0143
TTV load (10^8 copies per mL)	11.52	0.72	0.0143
Delayed graft function	4.99	0.66	0.0195
Peritoneal dialysis	6.87	0.58	0.0588
Acute graft rejection	3.89	0.61	0.0753
Treatment for rejection	3.18	0.62	0.0778
TTV status	7.00	0.60	0.0803
Calcineurin inhibitor	0.43	0.40	0.1700
Haemodialysis (type)	2.38	Not tested	0.2121

Table 4.4. Logistic regression analysis results for the prediction of infectious complications. Correlates of infectious disease complication were identified using a logistic regression. Odd's ratio for TTV load (copies per mL of plasma) is cumulative per log10 increase. TTV status refers to the presence or absence of TTV by qPCR. Calcineurin inhibitors were either tacrolimus (n = 26) or cyclosporin A (n = 24). Anti-metabolites used were azathioprine (n = 38) or mycophenolate mofetil (n = 12).

4.3.5. Anellovirus Increased Following Acute Rejection Episodes in Renal Transplant Patients

I next set out to test whether any of the viruses detected by NGS were associated with patient treatment history or transplant outcome. Analysis revealed increased anellovirus load in patients who had previously suffered acute rejection episodes (Mann-Whitney U test, p = 0.0138) (Figure 4.5.A).

Anellovirus load is strongly correlated with the level of immunosuppression, demonstrated by the increase between time point 0 and 90 (Figure 4.3.A). This led me to hypothesize that rather than the rejection event itself, pulse steroid therapy (used to treat acute rejection episodes) may have triggered increased anellovirus load. To address this, I included four patients in the analysis who had been treated with pulse-steroid therapy prior to the day 90 time point but had not suffered an acute rejection episode. The association remained strong following the addition of these patients (Figure 4.5.B) (Mann Whitney U test, $p = 0.0170$). Importantly, there was no significant disparity in baseline levels of immunosuppression between the groups at any time point and so baseline maintenance dosages are unlikely to be a factor influencing the viral load (Figure 4.6).

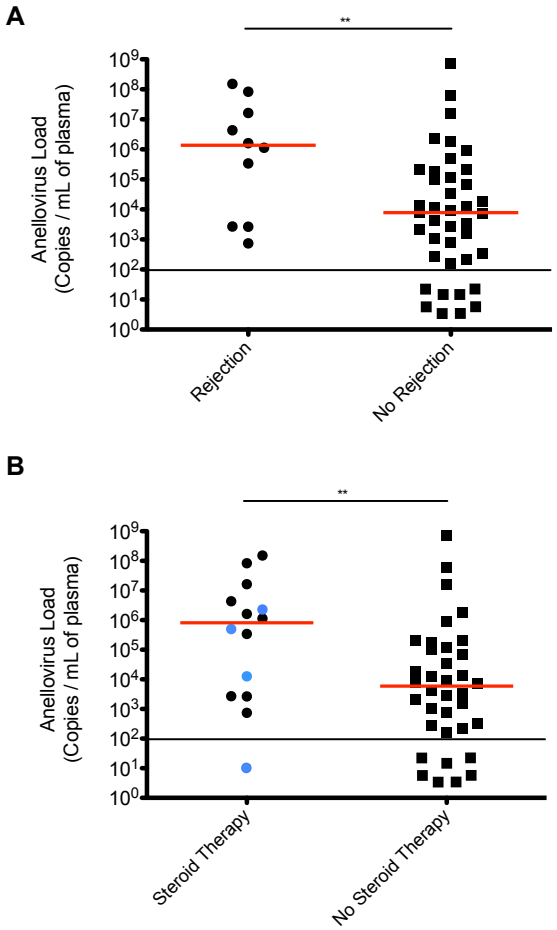


Figure 4.5. **Anellovirus load correlates with previous rejection episodes in renal transplant recipients.**

A) Anellovirus load at 90 days post-transplant was significantly increased in patients who had suffered from acute rejection episode (n = 10) compared to those that had not (n = 40) (p = 0.0138, Mann-Whitney test). Limit of detection for the assay is indicated by the dotted line. Error bars show median. **B)** Anellovirus load at 90 days post-transplant was significantly increased in patients who had been treated with pulse steroid therapy (regardless of rejection) (n = 14) compared to those that had not been treated (n = 36) (p = 0.0170, Mann-Whitney test). Patients who were treated but did not experience rejection are highlighted in blue. Error bars show median.

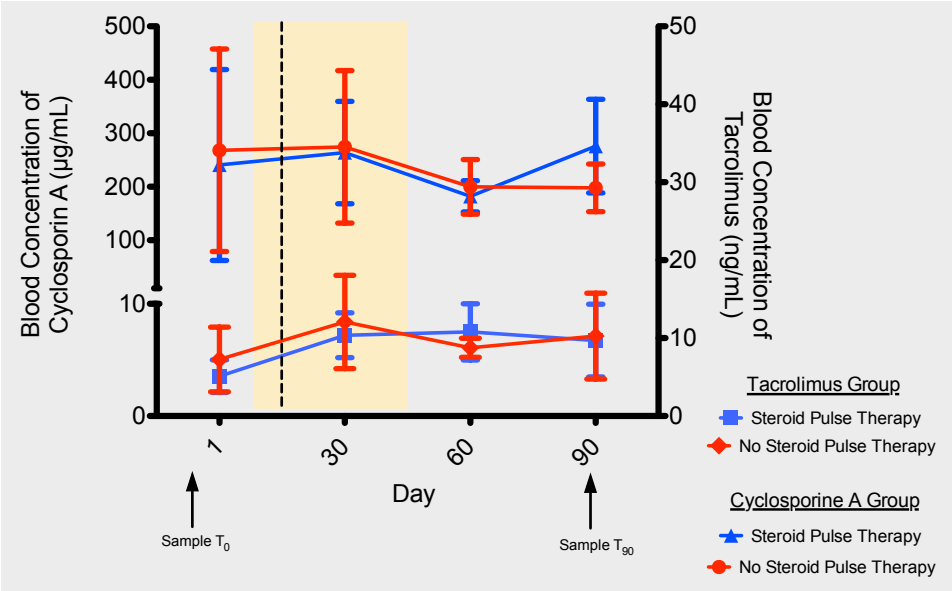


Figure 4.6. **Timeline of maintenance immunosuppression dosage in rejecting and non-rejecting renal transplant recipients.**

A comparison of baseline immunosuppression blood-levels between patients who received pulse steroid therapy and those who did not demonstrated no difference between the two. Patients received either tacrolimus (right y-axis) or cyclosporine A (left y-axis). Shaded area denotes the time range during which pulse steroid therapy was given, mean time-point of therapy is denoted by the dotted line (day 16).

4.3.6. Impact of Other Viruses in Renal Transplant Patients

No further association was found between the viruses detected by NGS and patient treatment or outcome. In particular, no association was found between the presence of HPgV, BKV, TTMDV or TTMV at day 90, and the occurrence of acute graft rejection, infectious complications, anellovirus load or blood transfusion.

4.3.7. Viruses Detected by Sequencing in Vasculitis Patient Plasma

Metagenomic sequencing of plasma from a cohort of patients receiving long-term immunosuppressive therapy for ANCA-vasculitis again revealed that 90 days of immunosuppression results in a significant increase in the prevalence of TTV (McNemar's test, $p = 0.027$) (Figure 4.7.). By one year, the prevalence of TTV had fallen, however the decrease between day 90 and one year was not significant (McNemar's test, $p = 1$). Clinically relevant viruses were also detected at a low prevalence. These included Merkel-cell polyomavirus, SV40, CMV, HSV1 and vesicular stomatitis virus. Aside from TTV, no significant change in virus prevalence was seen following immunosuppression.

4.3.8. Comparison of Viral Prevalence Between Renal Transplant and Vasculitis Patients

Prior to immunosuppression, there was no significant difference between the vasculitis and transplant cohorts in terms of viral prevalence. However, when all viruses were considered together, both cohorts did display an increased overall prevalence compared to the healthy volunteers (Fisher's exact test, $p < 0.001$ and 0.037 respectively) (Figure 4.8.A). If each virus was considered independently at

day 0, TTV and HPV prevalence in the vasculitis cohort differed significantly from the healthy volunteers ($p = 0.001$ and 0.018 respectively) (Figure 4.8. B-H).

Following 90 days of immunosuppression, the prevalence of the anelloviruses TTV, TTMDV and TTMV increased in both cohorts (Figure 4.8. B, C and D), however the prevalence of TTV (Fisher's exact test, $p = 0.012$) and HPgV (Fisher's exact test, $p < 0.001$) increased to significantly greater levels in the transplant cohort (Figure 4.8. A, B and F).

Herpesviruses were also detected across all cohorts, including the healthy volunteers (Figure 4.8.G). CMV and HSV 1 were detected solely in the vasculitis patient cohort, whilst Epstein-Barr virus was detected in a single patient from the transplant cohort at day 0. There was a trend towards increased cytomegalovirus prevalence in the vasculitis cohort across day 0 and day 90, however this was not significant (Fisher's exact test, $p = 0.075$).

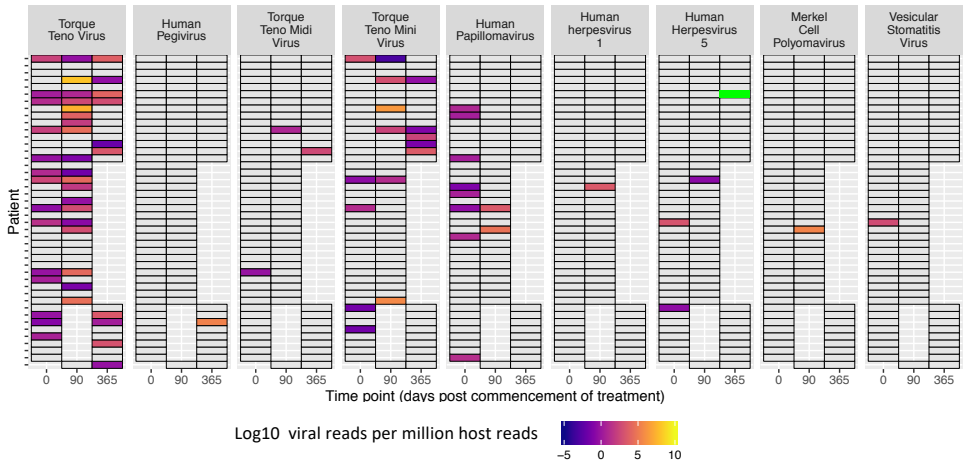


Figure 4.7. **Viruses detected by metagenomic Illumina sequencing of plasma from patients with vasculitis, pre- and post-commencement of maintenance immunosuppression.**

Metagenomic sequencing of plasma from vasculitis patients was performed and virus-specific reads were detected by alignment to a viral genome database. Sampling time points represent day of treatment (prior to commencement of immunosuppression), 90-days, and 365-days post treatment. 90 days of immunosuppression resulted in a significant increase in the prevalence of TTV (McNemar's test, $p = 0.027$). By one year, the prevalence of TTV had fallen, however the decrease between day 90 and one year was not significant (McNemar's test, $p = 1$). Viral read count is shown as a ratio of host-derived reads in order to normalize between individuals. Virus presence was confirmed by PCR. In one patient, human herpesvirus 5 (cytomegalovirus/CMV) was detected by qPCR in a sample that had no detectable CMV sequencing reads by NGS. This patient is identified in green in the figure.

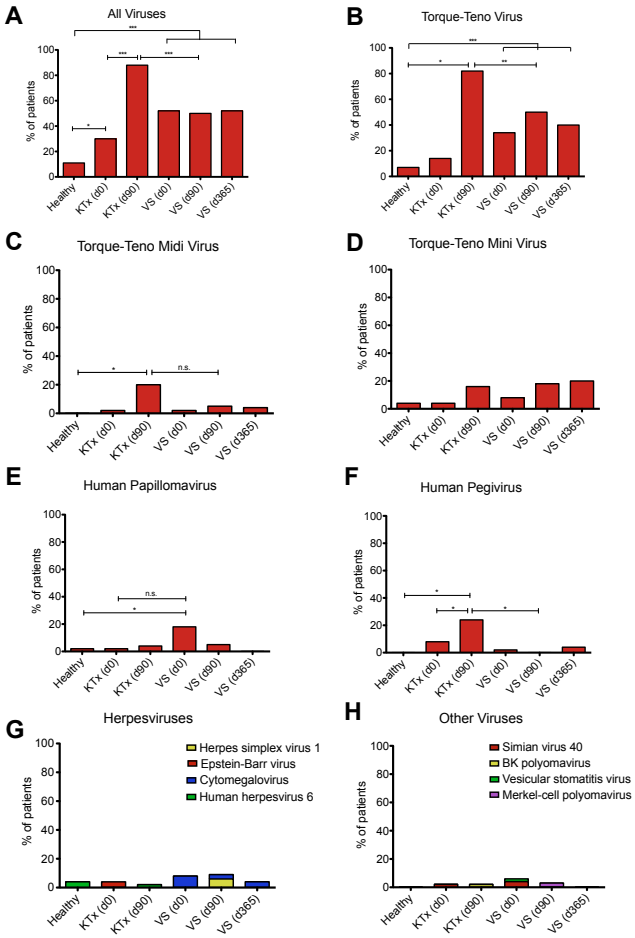


Figure 4.8. **Virus prevalence in renal transplant and vasculitis patients undergoing maintenance immunosuppression.** The prevalence of each of the viruses detected by metagenomic sequencing was calculated for healthy volunteers, transplant recipients and vasculitis patients at various time points, beginning at day 0 when therapeutic immunosuppression is initiated. Graphs show the proportion of patients positive for: **A)** Any virus; **B)** torque-teno virus; **C)** torque-teno midi virus; **D)** torque-teno mini virus; **E)** human papillomavirus; **F)** human pegivirus; **G)** herpes viruses; and **H)** other viruses. Following 90 days of immunosuppression, the prevalence of the anelloviruses TTV, TTMDV and TTMV increased in both cohorts (Figures B, C and D), however the prevalence of TTV (Fisher's exact test, $p = 0.012$) and HPgV (Fisher's exact test, $p < 0.001$) increased to significantly greater levels in the transplant cohort. Differences in prevalence between patient groups were tested by Fisher's exact test. Differences in prevalence within the same group over time was tested by McNemar's test.

4.3.9. Comparison of Anellovirus Load and Diversity Between Renal Transplant and Vasculitis Patients

At day zero, average anellovirus load of the vasculitis and transplant cohorts did not differ (Mann-Whitney U test, p = 0.412) (Figure 4.9.A). Following 90 days of immunosuppression, the average anellovirus load had significantly increased in the vasculitis cohort (Wilcoxon signed rank test, p = 0.027). However, the anellovirus load in vasculitis patients at day 90 was significantly lower than that of the transplant cohort (Mann-Whitney U test, p = 0.005). By one year, the average viral load of the vasculitis cohort had fallen to a level that did not significantly differ from that of day zero (Wilcoxon signed rank test, p = 0.063) or day 90 (Wilcoxon signed rank test, p = 0.375).

In the vasculitis cohort, the number of anellovirus species detected by metagenomic sequencing increased following 90 days of immunosuppression (Mann-Whitney U test, p = 0.024) and had reverted to a similar level to day zero by one year (p = 0.503) (Figure 4.9. B and C). The average number of anellovirus species was greater in the vasculitis cohort than the transplant cohort at day zero (Mann-Whitney U test, p = 0.031), whereas the average number of species was greater in the transplant cohort following 90 days of immunosuppression (Mann-Whitney U test, p = 0.024) (Figure 4.9.C).

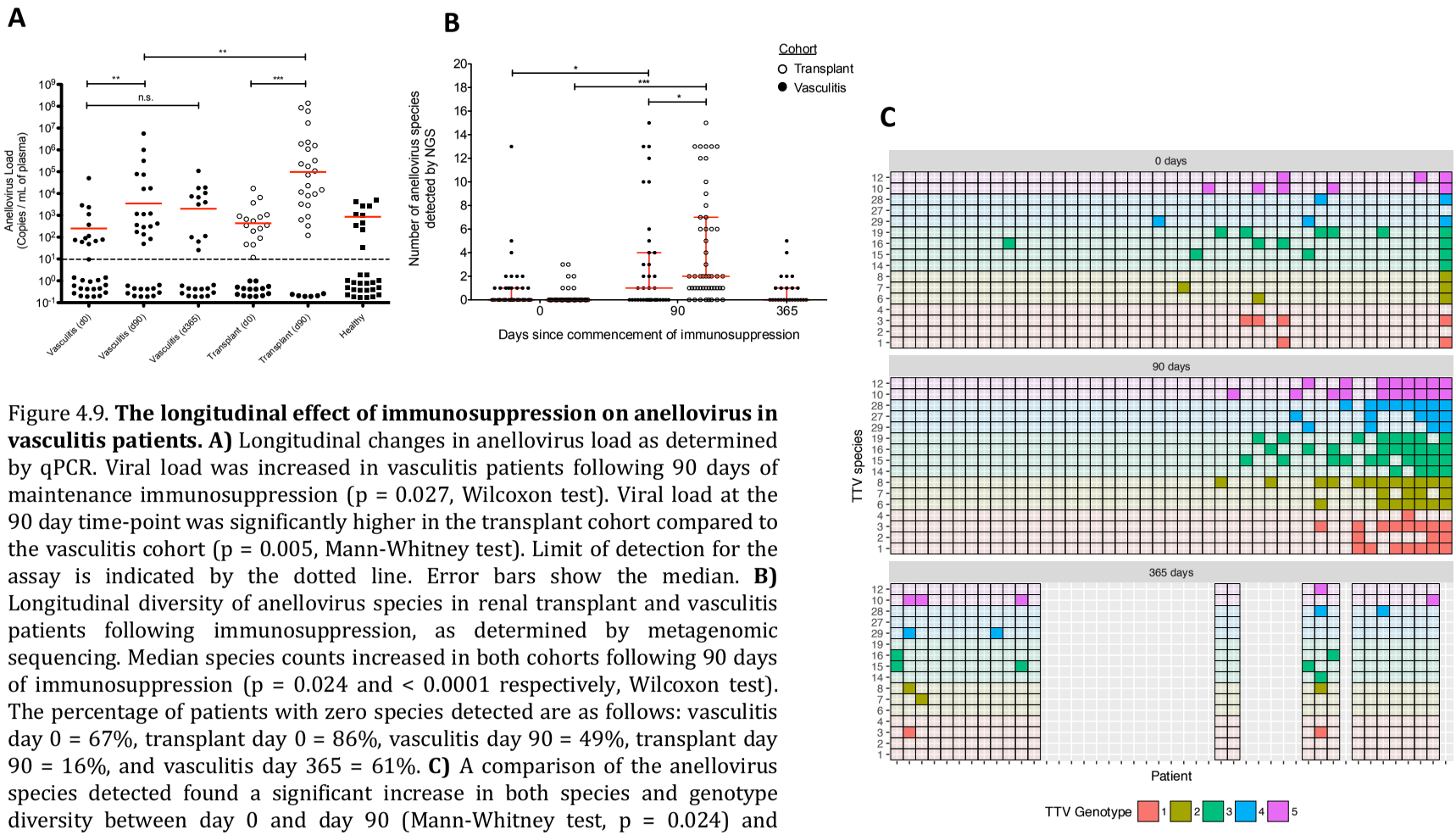


Figure 4.9. **The longitudinal effect of immunosuppression on anellovirus in vasculitis patients. A)** Longitudinal changes in anellovirus load as determined by qPCR. Viral load was increased in vasculitis patients following 90 days of maintenance immunosuppression (p = 0.027, Wilcoxon test). Viral load at the 90 day time-point was significantly higher in the transplant cohort compared to the vasculitis cohort (p = 0.005, Mann-Whitney test). Limit of detection for the assay is indicated by the dotted line. Error bars show the median. **B)** Longitudinal diversity of anellovirus species in renal transplant and vasculitis patients following immunosuppression, as determined by metagenomic sequencing. Median species counts increased in both cohorts following 90 days of immunosuppression (p = 0.024 and < 0.0001 respectively, Wilcoxon test). The percentage of patients with zero species detected are as follows: vasculitis day 0 = 67%, transplant day 0 = 86%, vasculitis day 90 = 49%, transplant day 90 = 16%, and vasculitis day 365 = 61%. **C)** A comparison of the anellovirus species detected found a significant increase in both species and genotype diversity between day 0 and day 90 (Mann-Whitney test, p = 0.024) and decrease between day 90 and day 365. Only reads that aligned to a single TT virus species were considered.

4.3.10. Full Genome Sequencing of Anellovirus Populations

As previously noted, the identification of viral reads in metagenomic datasets is limited by the need for alignment against a reference sequence. This is particularly true for anellovirus species as they have highly divergent genomes and therefore may not have sufficient homology to the sequences in the reference database for successful alignment. To address this, a method for non-specific, whole genome sequencing of anellovirus was devised (see methods section 4.2.5). This method was then used to assess the true diversity of anellovirus populations in a subset of ten transplant patient plasma samples.

Rolling circle amplification (RCA) using random hexamer primers can be used to selectively amplify circular DNA (see Figure 4.10). Therefore, this method was employed to preferentially enrich for circular anellovirus genomes in total nucleic acid extracted from plasma. RCA reaction success was confirmed by gel electrophoresis, which demonstrated the presence of an extremely high molecular mass product (constituting concatermerised anellovirus-derived DNA) (Figure 4.11 A).

RCA products were then divided into five, 5µl aliquots, and each aliquot was subjected to digestion by a different restriction enzyme. Restriction enzymes were selected based on their ability to cut multiple anellovirus species, as determined by reference sequence alignments (see Table 4.2). Restriction digest products were visualised on a 2% agarose gel under UV light (Figure 4.11 B-E), and those between 500-4,000 bp were excised and purified using the QIAquick gel extraction kit (Qiagen). Purified DNA from restriction digest reactions was pooled for each individual sample.

The MinION was selected to sequence the digested RCA products, as the platform is capable of producing single reads that cover the entire length of the 500 – 4,000 bp products. This is preferable to short-read platforms such as Illumina as the high level of variability in anellovirus genomes and the presence of multiple co-infecting anellovirus species can make non-reference-based assembly of short reads difficult.

Analysis of the sequence data revealed that every one of the patients harboured multiple species of anellovirus (Figure 4.12.A). Furthermore, on average 25% of sequences from each patient could not be mapped to the anellovirus reference genome database (range 3% - 71%) (Figure 4.12.B). Interrogation of the unmapped sequences by BLASTn revealed that they were most similar to other isolates of anellovirus not included in the reference database, suggesting that total anellovirus population diversity is under-represented by aligning sequences to the reference genome database alone.

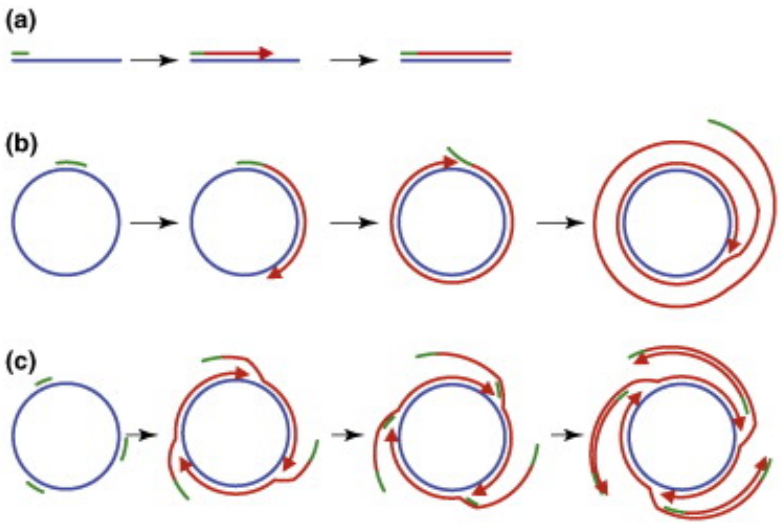


Figure 4.10. **Mechanism for non-specific enrichment of circular DNA using rolling circle amplification.**

Rolling circle amplification using random hexamer primers can be used to specifically amplify circular DNA species in a metagenomic sample containing both circular and linear DNA, therefore enabling enrichment of anellovirus genomic DNA in a nucleic extract containing primarily linear (non-anellovirus) DNA. **A)** For a linear template, a complementary strand is generated, however no further amplification occurs. **B)** Using a specific primer for a circular template also results in a complementary strand being generated. However, upon reaching the site of primer binding, the synthesised strand is displaced by the polymerase enzyme and DNA synthesis continues, producing a single-stranded, concatermeric product. **C)** Using random primers for a circular template allows for primers to bind and initiate DNA synthesis at multiple sites. Upon displacement, further random primers can bind to the displaced strand and DNA synthesis is also initiated on these strands. If the reaction is performed for a long period of time (e.g. 12-16 hours) the result is a high-molecular weight, double-stranded DNA product. Image reproduced from Johne et al.[390]

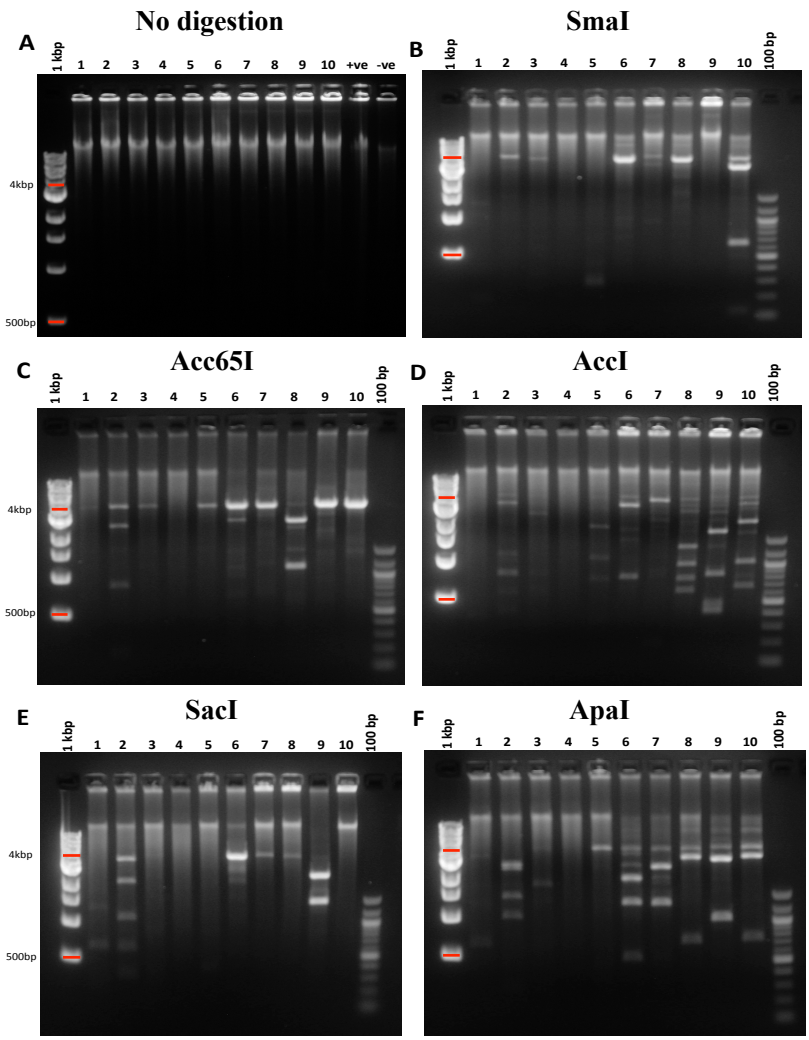


Figure 4.11. **Restriction digestion of rolling circle amplification products.** Rolling circle amplification (RCA) was performed on total nucleic extracts from the plasma of ten renal transplant recipients (day 90 time point) in order to amplify all circular DNA species, resulting in a high-molecular mass DNA product **(A).** RCA products were treated with five restriction enzymes: SmaI **(B)**, Acc65I **(C)**, AccI **(D)**, SacI **(E)**, and ApaI **(F)**. All bands equal to or smaller than the 3.8 kbp anellovirus genome (within the red demarcations) were extracted for sequencing on the MinION platform.

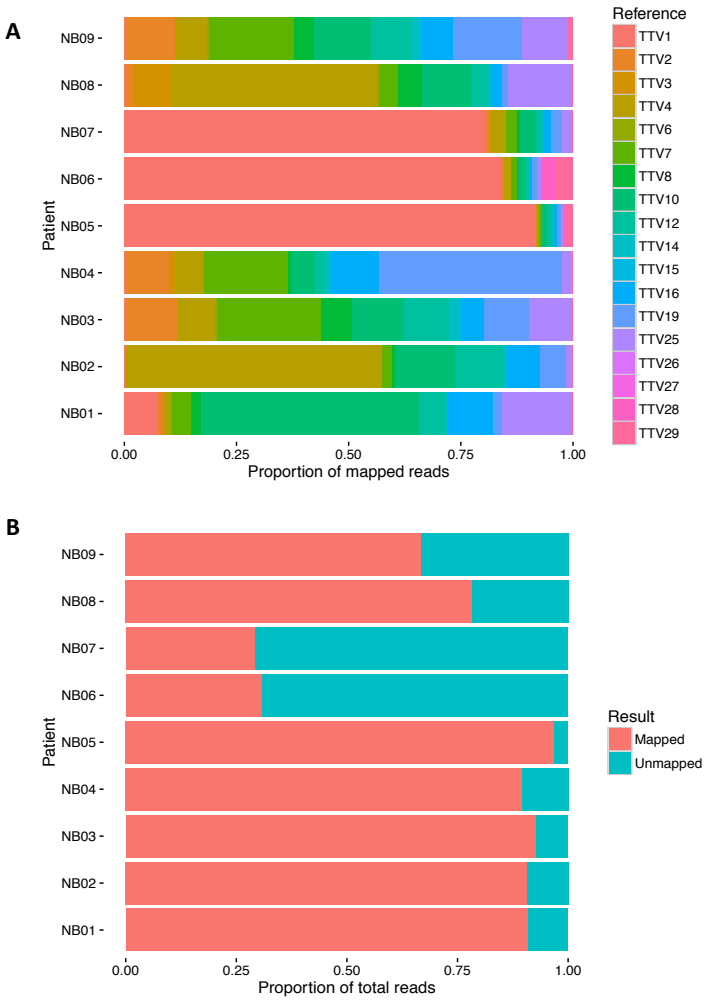


Figure 4.12. **Whole genome sequencing of anellovirus populations in transplant patient plasma.** 2D sequencing of RCA products was performed on the MinION platform. MinION sequencing reads were mapped to all 18 reference sequences available for anellovirus using GraphMap, allowing for circular genomes. Default settings were used, including an e-value alignment score threshold of 1e0, above which reads were called unmapped. Ambiguous alignments matching multiple TTV species were excluded from the analysis **(A)**. On average 25% of sequences from each patient could not be mapped to the anellovirus reference genome database (range 3% - 71%) **(B)**, despite BLASTn of the read confirming anellovirus as the best match.

4.4. Discussion

4.4.1. Anellovirus as a Prognostic Marker of Infection in Renal Transplant Patients

Here, I demonstrate marked changes in both the DNA and RNA plasma virome following immunosuppressive therapy. In agreement with previous studies examining the immunosuppressed DNA virome, a striking increase in anellovirus variants was detected.[274,377] Of the variables measured, total anellovirus load was found to be the best predictor of infectious complications in renal transplant patients, following the antimetabolite used. The significance of anellovirus load in renal transplant patients was even greater when infections occurring in the peri-operative period were excluded from the analysis. An argument could be made for the exclusion of infections during this period as they are rarely opportunistic; instead they are typically associated with surgical complications and the re-activation of pre-existing infections of the host or donor.[362]

Studies have previously demonstrated a strong correlation between anellovirus and immune status. Low viral loads have been associated with the high levels of immune activation which occur during acute graft rejection.[274] Here, I show that this association between increased viral load and immune suppression can also be translated into a measure of susceptibility to pathogens in the transplant cohort. Considered together, the results of this study support the notion that anellovirus accurately represents an individual's immune state, with regards to under-suppression as indicated by rejection, and over-suppression as indicated by infection. Therefore, patient prognosis and the tailoring of immunosuppressive drug doses may be improved through the use of anellovirus as a biomarker.

However, further work is required to optimise this assay and strengthen its predictive power. One possible method of achieving this is by targeting one or more specific anellovirus species that best correlate(s) with future infections. Examination of entire anellovirus populations by full genome sequencing revealed that the within host diversity is extremely large. Indeed, a number of reads that were identified as anellovirus by BLAST could not be aligned to any of the species in the anellovirus reference genome database. This finding suggests that the diversity of anellovirus populations is not fully represented in this study, and many species may have gone undetected by both sequencing, due to the use of a reference sequence database, and the qPCR assay. Therefore, it will be important to determine whether including any of these species can contribute in the search for an improved qPCR assay.

4.4.2. The Association Between Anellovirus and Graft Rejection

The current literature regarding anellovirus would primarily point towards increased levels of immunosuppression as the cause of the increased anellovirus load in patients who had experienced acute rejection. Additional doses of the immunosuppressive drugs used to treat acute rejection may sufficiently weaken anti-anellovirus immune mechanisms and facilitate increased replication of anellovirus. No difference in base-line blood level of immunosuppressive drugs was found between patients who experienced rejection and those that did not at any time point during the study. Therefore, the primary difference in immunosuppression between the two groups lies with the intense, short-term, pulse-steroid intervention given to treat the rejection episodes. However, it should be noted that rejection episodes and steroid interventions occurred on average two months prior to sampling. Therefore, this hypothesis would appear to suggest a long-lasting immunomodulatory effect on the virome, immune-competence, and susceptibility to infection by association.

One alternative explanation is that overall higher levels of anellovirus antigen or specific species of anellovirus (shown here to increase with load) may drive immune activation, which may promote graft rejection. At present there is no proven association between anellovirus and host versus graft pathology, however there is evidence for the virus eliciting an immune response. However, serological studies have confirmed the presence and persistence of antibodies against putative antigenic regions located in the ORF1 gene.[398] Indeed, virus-bound antibody complexes have been shown to account for the vast majority of TT virus in circulation,[399] which may themselves be deleterious for glomeruli. Anellovirus has also been shown to be capable of eliciting a cellular response. Clones of CD4+ T-cells isolated from the cerebral spinal fluid of a multiple sclerosis patient were found to respond to poly-arginine motifs present on both the virus and cells of the central nervous system.[400] Intriguingly, this implicates the virus as a potential cause of the host-directed immune response and may explain the association between acute rejection and increased viral loads. However, the ubiquity of this virus and reports of lower levels of anellovirus as a strong predictor of on-going rejection events would appear in conflict with this hypothesis.

A further possibility is that lymphocyte proliferation elicited by the immune response against the graft may simply facilitate anellovirus replication. The virus has been shown to replicate in stimulated peripheral blood cells,[326] and may be using the increased number of actively dividing cells to replicate. Indeed, studies have shown that anellovirus DNA levels increase following immune stimulation by vaccination.[401]

4.4.3. Other Viruses Detected by Metagenomic Sequencing of the Renal Transplant Cohort

HpGV was detected in 32% of patients post-transplantation. However, no association was found between the presence of HpGV and patient outcomes including other infectious complications, acute rejection episodes or poor graft score up to one year post-transplant. At present, the virus has not been associated with any form of pathology and is thought to infect individuals in a cryptic fashion.[402] Infections in healthy individuals are generally resolved, but persistence has been demonstrated in HIV-infected individuals.[403] This suggests that immunosuppressed patients could act as a reservoir for the virus and is a possibility supported by the high prevalence observed in this study. HpGV has been shown to down-regulate T-cell response pathways which are associated with reduced inflammation,[404] however here no association was found between the presence of HpGV or increased anellovirus load. Therefore, I suggest that infection with HpGV is unlikely to interfere with the potential use of anellovirus load as a marker of immunosuppression.

BKV was detected in 6% of patients over the 90-day period. Reactivation of BKV infection is a major cause of graft pathology in renal transplant patients. No association was found between BKV and patient outcome however it should be noted that all three BKV positive patients experienced rejection episodes either during or following the study period. This result is largely due to the small group size. Data from a recent study appeared to suggest that levels of BKV and anellovirus in the urine of renal transplant recipients were inversely correlated, leading the authors to hypothesise that there is competition between the two viruses for the cellular resources required to replicate.[307] I tested the data for the same inverse correlation in the plasma of this cohort, and found it was not the

case. However, a larger study with greater numbers of BKV positive individuals may be required.

4.4.4. Comparison of the Plasma Virome in Two Therapeutically Immunosuppressed Cohorts

Comparison of the two cohorts revealed a similar increase in viral prevalence and species diversity, as well as anellovirus load, following onset of immunosuppression. However, the increase in anellovirus in the vasculitis cohort was not as marked as the increase in the transplant cohort. Given the previously reported correlation between anellovirus and immunosuppression, this difference suggests a greater intensity of immunosuppression in the transplant cohort. Unfortunately, it has not yet been possible to obtain the clinical records for the vasculitis patients and so a comparison of type of immunotherapeutic and dosage could not be performed to determine whether this was indeed the case. Lack of clinical information also prevented testing of whether a similar association between anellovirus and infectious complications was detectable in these patients.

Further slight differences were detected between the two cohort's viromes. HPgV was significantly more abundant in the transplant cohort, whilst HPV and CMV showed signs of increased prevalence in the vasculitis cohort. Each of these differences could be due to disparities in disease, form of treatment or immunosuppressive regime. For example, Human Pegivirus is thought to be primarily transmitted via blood products and so is likely to have been introduced to the transplant patients during surgery or transfusions.[279] TTMDV prevalence appears to mirror TTV but to a lower degree, and so it is unsurprising that it's prevalence was slightly greater in the transplant cohort. The fact that HPV prevalence is slightly increased in the vascultitis patients at day 0 suggests that it may be associated with a facet of the untreated disease such as its aetiology, exposure, or activation caused by the inflammatory environment.

Interestingly, HPV was no longer detectable following immunosuppression. Finally, reactivation of CMV is highly associated with graft rejection in transplant recipients.[405] For this reason, these patients are treated with prophylactic antiviral drugs for a minimum of 3 months following transplant. It appears that the same drugs are not given to vasculitis patients, presumably due to the apparently lower intensity or type of immunosuppression, and lack of rejection risk. Due to the low prevalence of these viruses, larger studies would be required to determine whether they are common infections of these cohorts or merely artifacts.

The increase in overall prevalence of viruses at day zero in both patient groups compared to healthy volunteers is likely due to induction therapy. The fact that anellovirus load is not significantly increased in the patients at this time point illustrates that the level of immunosuppression during induction is minimal compared to the maintenance regime. Interestingly, viral prevalence did not appear to differ significantly between transplant and vasculitis patients at the day zero time point. This suggests that during the induction stage of treatment, both patient groups are susceptible to viral infections compared to healthy individuals, but neither cohort is significantly more susceptible than the other, and it is not until the maintenance regimes commence that the viromes begin to diverge.

Finally, an additional time point for the vasculitis patient cohort shows anellovirus load and prevalence falling after 1 year of immunosuppression. This trend has been previously reported in a longitudinal study of solid organ transplant recipients and was found to be highly associated with the tapering of immunosuppressive dose.[278] Therefore I would expect to see the same in pattern in the transplant patients after a year.

4.5. Conclusion

Overall, perhaps the most important clinical finding is that increased anellovirus at day 90 post-transplantation has not only the potential to be used as a measure of functional immunosuppression as has been previously suggested, but also to predict the risk of severe infectious complications in transplant patients. Opportunistic infections are a major cause of hospitalisation following solid organ transplantation and for many other immune-deficient conditions. Therefore, clinical monitoring of anellovirus load in plasma may become an important clinical management tool for fine-tuning of immunotolerance to address this risk. Larger, multi-centre studies are warranted in order to determine the robustness of these findings against a broader clinical setting. The comparison between the transplant and vasculitis cohorts revealed both commonalities and differences in the virome. Therefore, extrapolation of these results with regards to immunosuppressed patients outside of the transplant setting should be approached with caution without further validation as they may not be applicable to cohorts on differing immunosuppressive regimes. Further molecular studies of anellovirus including, the relevance of specific species, how it is acquired and its response to immunosuppression are also required to better understand the role of different anelloviruses in human health and disease.

Chapter 5

The Virome in Primary Antibody Deficiency

5.1. Introduction

5.1.1. Evidence for an impaired antiviral immune response in antibody deficient patients.

Early reports of patients with primary antibody deficiencies described severe bacterial infections as the primary cause of morbidity and mortality.[406–408] This observation, reinforced by the lack of appropriate or sensitive tests for viruses at the time, has led to the long-standing, dogmatic view that the impact of antibody deficiency is predisposition to bacterial, rather than viral, infections.[409] Indeed, the majority of antibody deficient patients still present with a history of recurrent bacterial infections. This has been supported by the therapeutic response to antibiotics combined with the introduction of intravenous immunoglobulin therapy. Together, these treatments have proved effective at preventing bacterial infections, leading to a pronounced decline in mortality rates.[410,411] From this, it has generally been concluded that bacteria are major pathogens for antibody deficient patients, reflecting the clinical focus of treatment on preventing and/or resolving bacterial infections.

However, observations in murine models have demonstrated that antibodies also play an important role in the resolution of both chronic and acute viral infections. For example, high rates of mortality have been described in B-cell deficient mice chronically infected with gammaherpesvirus, whilst immunocompetent animals were able to survive by controlling the virus, maintaining its latent state.[412] A further study found that B-cell deficient mice are

also far more susceptible to acute influenza virus infection compared to their immunocompetent counterparts; the immunodeficient mice also demonstrated a lack of immunological memory following re-challenge by an attenuated strain of the virus.[413] In humans, patients with antibody deficiency have been recently reported to suffer chronic infections by viruses usually responsible for acute disease.[414] It should be noted that despite the clear importance of antibodies with regards to viral infection, in the absence of T-cells, presence of antibody alone is not sufficient for viral clearence.[415] Regardless, the evidence above suggests that individuals lacking antibodies are likely to have an impaired anti-viral immune response.

5.1.2. B- and T-cell defects in antibody deficient patients

Of the primary antibody deficiency disorders, common variable immunodeficiency (CVID) is the most frequently diagnosed in adults, with an estimated prevalence of 1:10,000 – 50,000.[416] It may manifest in childhood but is generally not diagnosed until the second or third decade of life, often on the basis of recurrent bacterial infections of the upper and lower respiratory tract. CVID is primarily characterised by a severe reduction of at least two immunoglobulin isotypes, a poor response to vaccination and the exclusion of differential diagnoses. This relaxed definition has resulted in an extremely heterogeneous disease, and a diverse range of immune pathway defects are likely to be involved given the immense complexity of the adaptive immune system. Indeed, at time of writing, the specific underlying genetic defects in CVID have been identified for only 10-15% of cases.[417,418]

Unsurprisingly, the majority of CVID patients display irregularities in their B-cell compartment, in terms of both quantity and function, which are presumably due to defects at different stages of the B-cell development pathway. Initially, B-cell

immunophenotyping was used to classify heterogeneous cohorts of CVID patients into discrete groups which could be used to predict the likelihood of disease-related complications and survival rates.[410] However, despite humoral deficiency being the definitive defect for CVID, it was soon discovered that the B-cells of many CVID patients are in fact capable of both proliferation, and the secretion of normal amounts of immunoglobulin if stimulated appropriately in vitro.[419–423] This finding lead to the hypothesis that other cells of the immune system were involved in at least some cases.

Investigations have since found that 40-50% of CVID patients also possess T-cell abnormalities.[424,425] In two large studies (n = 285 and 313), the most frequently observed T-cell abnormality was an inverted CD4:CD8 ratio, characterised by both a significant reduction in naïve $CD4^+$ T-cells and a significant increase in effector $CD8^+$ T-cells compared to healthy blood donors.[426,427] It has been suggested that the reduction in the number of circulating naïve T-cells is possibly due to a diminished thymic output, whilst the increased number of effector T-cells is likely driven by chronic antigenic-stimulation, such as that seen during viral re-activation events.[428] Further studies have also demonstrated various functional defects in the T-cells of some CVID patients, including a reduced proliferative response to mitogens and failure to respond to both novel and recall antigen.[419,429] These findings have led to T-cell phenotyping being implemented as an important, additional tool for the classification of CVID patients.

5.1.3. T-cell abnormalities in antibody deficient patients appear to be associated with inflammatory disease

Intriguingly, the presence of T-cell abnormalities in CVID appears to be closely linked to idiopathic inflammatory complications. A European study investigating

a cohort of 313 CVID patients found a strong association between a reduced percentage of naïve $CD4^+$ T-cells, an increased percentage of effector $CD8^+$ T-cells, and increased disease severity in the form of lymphoproliferative disorders, autoimmune diseases and gastric complications.[427] The same abnormality was also shown to be significantly associated with reduced lung function,[430] splenomegaly,[426] presence of granulomata,[431] and gastrointestinal symptoms[432] in CVID. Another study of over 500 antibody deficient patients found that those with a reduced CD4+ naïve T-cell count suffered an increased rate of mortality and inflammatory disease.[433] Furthermore, this phenotype of antibody deficiency combined with T-cell abnormalities appears to be associated with a more severe form of disease as these patients accounted for 77% of the total deaths during the study period despite representing only 20% of the cohort. Elevated markers of immune activation (HLA-DR and CD38) on CD8 T-cells have also been associated with incidence of splenomegaly, autoimmune disease and lymphoproliferative disorders in a number of studies.[434–436]

5.1.4. The inflammatory phenotype in primary antibody deficiency

Chronic immune activation affects approximately 30-60% of CVID patients.[410,437,438] Patients who experience these symptoms represent a distinct phenotype which does not alter over time.[439] These patients also experience a range of secondary immune-related complications, including: enteropathy, hepatomegaly, splenomegaly, granulomatous disease of the lung, liver and spleen, autoimmune disease and lymphoproliferative diseases.[440]

The introduction of intravenous immunoglobulin (IVIg) replacement therapy for CVID patients, whilst proving an effective treatment for the prevention of bacterial infections, does not appear to have fully addressed the inflammatory complications described above, and the general health of CVID patients is still

poorer than that of the average adult.[441] These complications are now the major cause of morbidity and death in CVID patients,[410,442,443] and account for an 11-fold increase in morbidity and mortality rate compared to CVID subjects who do not experience these complications.[444] The cause of the immune activation is unclear, however recurrent and chronic infections at mucosal surfaces have been implicated due to the fact that instigation of intravenous immunoglobulin (IVIg) therapy can sometimes ameliorate symptoms in these patients.[445] Conversely, some markers of activation (Ki67 and HLA-DR) and $CD8^+$ T-cell counts do not return to normal levels following treatment, suggesting that an underlying stimulus remains.[446]

5.1.5. Evidence for a viral aetiology in inflammatory disease of primary antibody deficient patients

Despite research and treatment continuing to ignore viral infections in antibody deficient patients, a number of small clinical cohort studies have highlighted a potential association between viral aetiologies and some forms of inflammatory complication in CVID patients, suggesting that viral infections in CVID may prove to be a greater burden than previously thought.

Intriguingly, the CD4/CD8 T-cell phenotype described above resembles that seen in patients suffering from chronic viral reactivation events.[428] Furthermore, increased levels of 'exhausted', senescent T-cells ($CD57^+$ $CD279^+$) have been reported in multiple CVID cohort studies.[427,447–449] This phenotype is also suggestive of antigen-driven, persistent immune activation caused by chronic infections.

One study reported a strong association between the presence of CMV in CVID patients and systemic inflammatory disease with active replication of CMV at

sites of inflammation demonstrated by immunohistochemistry.[450] A further study reported the detection of CD8+ T-cells specific for CMV and EBV epitopes in approximately half of the CVID patients studied.[451] Additionally, one of these patients suffered from severe gastroenteritis, which was resolved by treatment with Valganciclovir, providing further evidence for a herpes virus such as CMV in at least some cases of disease.

Aside from CMV, it has also been observed that CVID patients accidentally infected by hepatitis C virus (HCV) during immunoglobulin replacement therapy suffered an aggressive clinical course of hepatitis.[452] Finally, a possible association between another complication of CVID, granulomatous-lymphocytic interstitial lung disease (GLILD), and human herpes virus 8 (HHV8) infection has been reported, however the small study size meant that the results were not statistically significant (HHV8+ in 6/9 with GLILD and 1/21 without GLILD).[453] Overall, both RNA and DNA viruses have been detected at the site of inflammatory complications suffered by CVID patients, however, causation remains unproven, and, for the majority of cases, the underlying mechanism remains unclear.

5.1.6. Aims

Here, I investigated the hypothesis that viral prevalence is increased in the blood plasma of antibody deficient patients, particularly those with T-cell abnormalities. I further investigated whether any of the viral infections detected were associated with the inflammatory complications responsible for the high level of morbidity and mortality observed in CVID patients, and if administration of intravenous immunoglobulin therapy correlated with the resolution of some of these viral infections.

5.2. Methodology

5.2.1. Study design and participants

The experiments in this study made use of two separate cohorts of patients diagnosed with primary antibody deficiency. A distinct cohort of healthy volunteers was used as a control for each of the patient cohorts.

Cohort 1: Anonymised, fresh whole blood samples taken from patients with primary hypogammaglobulinemia and healthy volunteers were obtained with written consent under the approval of the University of Cambridge Human Biology Research Ethics Committee (reference **HBREC.2014.24)** and Cambridge Local Research Ethical Committee (reference **08/H0308/176**) respectively. Whole blood samples were stained for flow cytometry (see below) and the remaining sample was separated into plasma and PBMCs using Leucocep centrifuge tubes (Kremsmünster, Austria) for later experiments. Patient samples were predominantly, but not exclusively, diagnosed with CVID (see Table 5.1).

Cohort 2: Anonymised, archived serum samples from patients diagnosed with CVID. Two samples were taken per patient, representing two time-points: pre- and post-commencement of antibody replacement therapy. Both the patient samples and archived samples from healthy volunteers were obtained with written consent and the approval of the NHS ethical review committee (reference **12/WA/0148**).

Diagnosis of common variable immunodeficiency was suspected when a patient had a history of recurrent infections. The diagnosis was made when blood tests showed low levels of at least two of the immunoglobulin isotypes IgG, IgA and IgM, and the exclusion of differential diagnoses such as Bruton agammaglobulinemia, X-Linked agammaglobulinemia, and secondary causes of immunodeficiency.

Patient ID	Diagnosis	Anellovirus status	Pegivirus status	CD4+ naïve (%)	CD8+ effector (%)	Treatment status	Inflammatory disease	Other Clinical Complications
1	CVID	+	-	32.4	27.1	POST	Yes	
2	CVID	+	-	0.7	79	POST	No	Recurrent lung infection
3	CVID	+	-	16.3	71	POST	Yes	
4	PID/specific Ab deficiency	+	+	0.9	92	POST	Unknown	Recurrent infection, obliterative bronchiolitis
5	PID/specific Ab deficiency	+	-	2.0	60.1	POST	Unknown	JIA, Blackfan Diamond Anaemia
6	CVID	+	-	13.1	38.1	POST	No	ITP, AIHA
7	PID/Hypogammaglobulinemia	+	+	11.5	63	PRE	No	COPD
8	PID/Hypogammaglobulinemia	+	-	0.4	71.7	POST	Yes	
9	CVID	+	-	0.4	79.3	POST	Yes	
10	CVID	-	-	8.6	55	POST	Yes	ITP (AI)
11	CVID	-	-	5.1	36.2	POST	Yes	ITP (AI)
12	CVID	-	-	16.7	56.1	POST	Yes	Splenomegaly, adenopathy, GLILD
13	CVID	-	-	5.1	53.3	POST	Yes	Human Papillomavirus
14	CVID	-	-	77.2	3.7	PRE	No	Sjorgrens
15	CVID	-	-	32.5	40.1	PRE	Unknown	Hepatitis, Hepatitis E Virus
16	CVID	-	-	43	32.4	PRE	No	Recurrent lung infection
17	PID/Hypogammaglobulinemia	-	-	28.1	71.1	PRE	No	Recurrent lung infection
18	SPAD	-	-	44.4	38.2	PRE	No	Recurrent lung infection
19	PID/Hypogammaglobulinemia	-	-	67.7	49.4	POST	No	HPV, Candida
20	CVID	-	-	20.3	78.1	POST	Unknown	
21	CVID	-	-	48.9	23.7	PRE	Unknown	

Table 5.1. **Overview of Cohort 1. Prospective whole blood samples taken from antibody deficient patients at a single time-point.**

The majority of patients in cohort 1 were diagnosed with CVID, however the cohort also included other primary immunodeficiency patients (PID) with antibody deficiencies (see diagnosis). Anellovirus status was defined by qPCR and Human Pegivirus status was defined by metagenomic sequencing and confirmed by PCR (virus positive status is highlighted in red). CD4 CD45RA+ $CCR7^{+}$ naïve and $CD8^{+}$ $CD45RA^{-}$ $CCR7^{-}$ effector T-cell levels are shown as percentages of $CD3^{+}$ $CD4^{+}$ T-cells and $CD3^{+}$ $CD8^{+}$ T-cells respectively. The majority of patients suffered from additional infectious and inflammatory complications, which are described, however the available clinical dataset remained incomplete for some patients.

5.2.2. Sample preparation and sequencing

Nucleic acid was extracted from plasma samples, prepared, sequenced and analysed as described in sections 2.1-2.3 of this thesis. An overview of the sequencing statistics for each group are shown in section 2.4.

5.2.3. Contamination and confirmation of results

Given the low abundance of nucleic acid in plasma, the possibility of cross-sample contamination was thoroughly investigated. Typical patterns of contamination such as the same virus occurring in multiple consecutive samples in a run were confirmed on the basis of alignments and PCR testing of the original sample. All viral sequences with any evidence of contamination were excluded from analysis, to ensure no false associations were made. In particular, HIV-1 reads were excluded from samples across three runs (17870_1, 17150_1 and 18174) and Epstein-Barr virus reads from samples in a single run (18456_7). In both cases contamination was suspected due to the presence of the virus at a low level across consecutive samples, including those sequenced as part of separate studies. Absence of the viruses was confirmed by PCR of selected samples.

5.2.4. Flow cytometry staining and analysis

Flow cytometry was performed using the panels of antibodies shown in Table 5.2. (purchased from Thermo Fisher Scientific, Waltham, Massachusetts, USA).

Panel	FL1 (FITC)	FL2 (PE)	FL3 (PERCP + EF700)	FL4 (APC)
T-cell	CD3	CCR7	CD45RA	CD4 or CD8
T-cell activation	CD3	HLA-DR	CD38	CD4 or CD8
T-cell dysfunction	CD3	CD57	CD279 (PD-1)	CD4 or CD8
Count	CD3	CD14	CD20	CD335 (NkP46)
FMO CCR7-	CD3	-	CD45RA	CD4
FMO CD45RA-	CD3	CCR7	-	CD4
FMO HLA-DR-	CD3	-	CD38	CD8
FMO CD38-	CD3	HLA-DR	-	CD8
FMO CD57-	CD3	-	CD279 (PD-1)	CD8
FMO CD279-	CD3	CD57	-	CD8

Table 5.2. **Antibody panels used for flow cytometry-based phenotyping of lymphocyte subsets.** T-cells, B-cells, NK cells and macrophages were defined as $CD3^+$, $CD20^+$, $NKp46^+/CD3^-$ and $CD14^+$ respectively. Counts were performed using Accucheck Counting Beads (Invitrogen) according to manufacturer's instructions. CCR7 and CD45RA expression was used to define $CD4^+$ and $CD8^+$ T-cells as naïve ($CD45RA^+$ $CCR7^+$), central memory ($CD45RA^-$ $CCR7^+$), effector memory ($CD45RA^-$ $CCR7^-$) and terminally differentiated ($CD45RA^+$ $CCR7^-$) subsets.[454] HLA-DR and CD38 expression were used as measures of T-cell activation.[455] CD57 and CD279 expression were used as measures of T-cell senescence / dysfunction.[428] Fluorescence minus one (FMO) staining was used to aid gating for markers when required.[456]

Fresh, whole blood samples were stained for 15 min in the dark at room temperature with a mixture of antibodies at optimal concentrations determined by titration (see Table 5.2.) followed by red blood cell lysis using FACS Lyse (Thermo Fisher Scientific). Acquisition was performed using an AccuriC6 flow cytometer (Becton Dickinson, Franklin Lakes, New Jersey, USA). Counts were

performed using Accu-check counting beads (Thermo Fisher Scientific) according to manufacturer's instructions. Data were analyzed using the C6 Analysis software (Becton Dickinson).

5.2.5. Human Endogenous Retrovirus K quantitation

Total nucleic acid from 21 patients and 19 healthy controls was extracted from approximately 2 x 10^6 PBMCs using the high-pure viral nucleic acid extraction kit (Roche) and eluted in 40 µl of elution buffer.

Quantitation of human endogenous retrovirus K was performed using primers and a dual-labelled probe targeting the *gag* gene described by Depil et al (2002).[457] One-step qPCR assays were performed on the Rotor-Gene 6000 (Qiagen): a reverse-transcription step at 50°C for 30 min followed by DNA Polymerase activation at 95°C for 5 min and 2-step cycling of denaturation at 95°C for 15 s and combined annealing and extension at 60°C for 45 s for 40 cycles. Reactions were set up using the Quantitect Virus Kit (Qiagen) following manufacturer's instructions for 20 µl reactions with 2 µl of template. To ensure detection of retroviral transcripts only, genomic DNA was removed by treatment with 2 U of Turbo DNase (Thermo Fisher Scientific) for 30 minutes at 37°C. qPCR reactions omitting the reverse-transcription enzyme were performed to ensure full degradation of DNA. Any sample positive for HERV-K in a reaction without the RT enzyme was subjected to a further round of DNase treatment and re-tested.

Primer sets for three house-keeping genes (GAPDH, β-actin and HPRT) were tested using the Norm-finder software (v0.953) to determine the optimal genes for normalization of measurements between samples and a combination of GAPDH and β-actin was selected as the most stable.[458] Quantitation of house-

keeping genes was performed using the primers described by Chege et al (2010)[459] and the Rotor-Gene SYBR Green RT-PCR kit (Qiagen) following manufacturer's instructions for 25 µl reactions with 2 µl of template. One-step RT-PCR was performed at 55°C for 10 min followed by DNA Polymerase activation at 95°C for 5 min and 2-step cycling of denaturation at 95°C for 5 s and combined annealing and extension at 60°C for 10 s for 40 cycles. All reactions were performed in triplicate and the mean value used.

Relative levels of Human Endogenous Retrovirus-K expression were calculated using the following formula:

$$\Delta HERV-Kexpression = (\frac{(1+HK1^{eff})^{\overline{x}(HK1Ct)}}{(1+HERV^{eff})^{\overline{x}(HERVCt)}} x \frac{(1+HK2^{eff})^{\overline{x}(HK2Ct)}}{(1+HERV^{eff})^{\overline{x}(HERVCt)}})^{0.5}$$

Where amplification efficiency (eff) was calculated by performing the assay on serial dilutions in triplicate. Amplification efficiencies for the HERV-K, β-actin, and GAPDH genes were 96%, 98%, and 89% respectively.

5.2.6. IgM and sCD14 Enzyme-Linked Immunosorbent Assays

Soluble CD14 (sCD14) and immunoglobulin M (IgM) levels were measured in plasma samples paired with the PBMC samples described above. sCD14 was used as a marker of bacterial translocation into the blood as previously described.[460] sCD14 was quantitated in triplicate using the human sCD14 Quantikine ELISA kit (Bio-Techne, Minneapolis, Minnesota, USA), and IgM was quantitated in triplicate using the Human IgM total ELISA Ready-SET-Go! Kit (Thermo Fisher Scientific, Waltham, Massachusetts, USA) both according to manufacturer's instructions.

5.2.7 Statistical Analysis

Basic statistical analysis was performed using the methods described in section 2.6 of this thesis.

5.3. Results

5.3.1 Overview of Illumina sequencing data

Raw read counts varied significantly across the patient and control groups of both cohorts ($p < 0.0001$, one-way ANOVA, Fig 5.1.A.). In both cohort 1 and 2, significantly fewer reads were generated by the healthy control groups compared to the patient groups (both $p < 0.0001$, Tukey's multiple comparison test). However, comparing across the cohorts, the number of reads generated by the two healthy control groups did not significantly differ ($p = 0.99$, Tukey's multiple comparison test) and neither did the number of reads generated by the three CVID patient groups ($p = 0.51 – 0.66$, Tukey's multiple comparison test). As discussed in chapter 4, a low read number in the control group may result in an underestimation of viral prevalence. However, a similarly low read count was observed for the healthy control group in both cohort 1 and 2. This suggests that the difference in read counts between the healthy and patient groups may be due to an inherent property of the samples (such as that described in chapter 4), rather than an issue with the quality of the samples or sequencing data.

Across all groups, a high proportion of reads were lost during the computational quality control stages ($\bar{x}$ = 15.0 – 43.6%, Fig 5.1.B.). Mapping of the dataset to the

human reference genome further reduced the remaining reads by 73.4 - 91.1% (Fig 5.1.C.).

The number of remaining high quality, non-human reads ranged from 10^3 - 10^7 across the groups (Fig 5.1.D.). There was a significant difference in average remaining read number when all groups were considered ($p < 0.0001$, one-way ANOVA). However, comparisons of individual groups revealed significant differences between only the patient and healthy control groups of cohort 2. As with the raw read counts, there was no significant difference in remaining reads between the two healthy control groups ($p = 0.1$, Tukey's multiple comparison test) or the patient groups across the two cohorts ($p = 0.19 – 0.43$, Tukey's multiple comparison test).

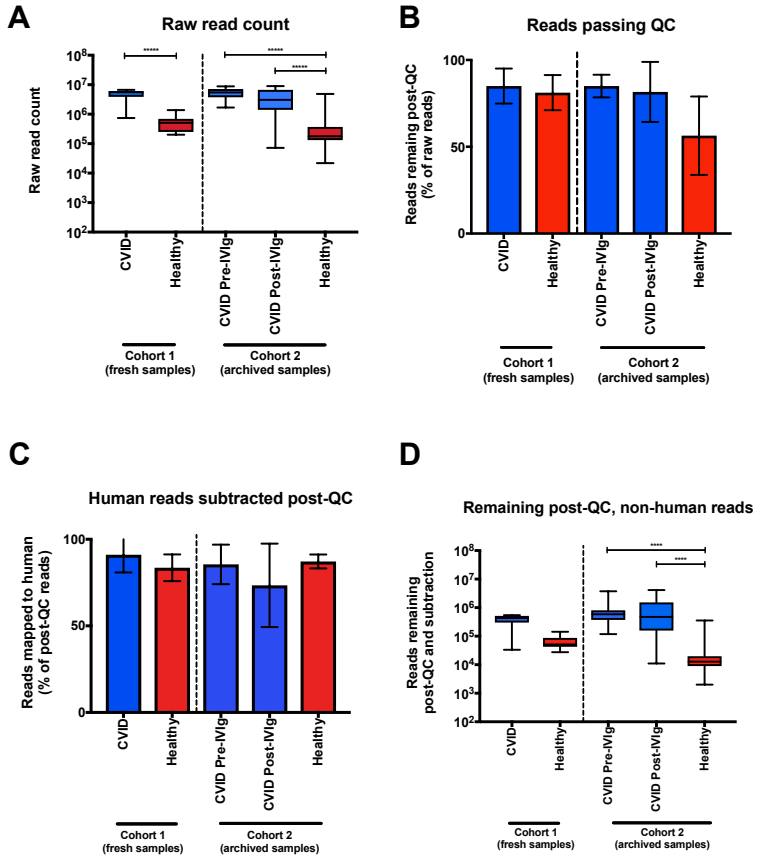


Figure 5.1. **Illumina sequencing data statistics for fresh blood (cohort 1) and archived plasma samples (cohort 2).** A) Average number of sequencing reads generated for each sample type. There was a significant difference in average raw read count between the 5 groups ($p < 0.001$, one-way ANOVA). Individual group comparisons showed that average read counts differed significantly between the patient and control groups in both cohorts ($p < 0.0001$, Tukey's multiple comparison test) but there was no significant difference when comparing the patient groups ($p = 0.51 – 0.67$) or the control groups ($p = 0.99$) across the two cohorts. B) Trimming and removal of low quality reads resulted in the loss of an average of 15.0 – 43.6% of sequences. C) Subtraction of human sequences resulted in the loss of 73.4 – 91.1% of the remaining high-quality reads from each group. D) Following QC and host-subtraction there remained a significant difference in average read count between the 5 groups ($p < 0.001$, one-way ANOVA). For cohort 1, there was no significant difference in the number of remaining reads between the patient and control group ($p = 0.18$, Tukey's multiple comparison test). However, for cohort 2, there were significantly more reads remaining in the patient groups than the control group ($p = 0.007 – < 0.0001$). Comparing across cohorts there was no significant difference in the number of reads remaining in the patient groups ($p = 0.19 – 0.43$) nor the control groups ($p = 1.0$). Mean and standard deviation is shown for all figures.

5.3.2. T-cell phenotype is correlated with the blood virome in antibody deficient patients

Flow cytometry was used to measure specific T-cell subsets and activation markers in whole-blood samples from the antibody deficient patients and healthy volunteers of cohort 1, in order to group patients by the T-cell phenotypes described above. Total B-cell, T-cell, NK cell and macrophage counts were also calculated. Virus prevalence was measured in plasma samples using un-biased metagenomic sequencing and a more sensitive, directed qPCR assay was used to specifically quantify anellovirus. The full data set, including: flow cytometry results, virus infection status, and records of inflammatory complications, are reported in Table 5.1.

A multivariate analysis of the data revealed that $CD4^+$ naïve and $CD8^+$ effector T-cell levels were highly correlated with the detection of anellovirus by qPCR ($p = 0.002$, permutational MANOVA). Patients positive for anellovirus had, on average, a significantly lower percentage of circulating $CD4^+$ naïve T cells compared to those patients negative for the virus ($p = 0.0062$, Mann-Whitney U test), and compared to healthy controls ($p = 0.0002$, Mann-Whitney U test) (Figure 5.2.A). Furthermore, those patients positive for anellovirus also had a significantly higher percentage of circulating $CD8^+$ effector T cells compared to patients negative for TTV ($p = 0.0428$, Mann-Whitney U test) and healthy controls ($p = 0.0139$, Mann-Whitney U Test) (Figure 5.2.B). The analysis detected no association between anellovirus status or load and total numbers of B-cells, T-cells, NK cells or macrophages.

There was an overall difference in the proportion of $CD8^+$ cells displaying the activation marker HLA-DR between groups ($p = 0.0232$, Kruskal Wallis test). $HLA\text{-}DR^+$ $CD8^+$ T-cell levels were increased in the anellovirus positive group

compared to both the anellovirus negative group ($p = 0.037$, Mann Whitney test) and healthy volunteers ($p = 0.009$, Mann Whitney test) (Figure 5.3.A). However, there was no difference between groups in terms of CD38 expression by $CD8^+$ cells ($p = 0.098$, Kruskal Wallis test) (Figure 5.3.B). Nor was there a difference between groups in terms of HLA-DR and CD38 expression of the $CD4^+$ population ($p = 0.135$, and $p = 0.02579$ respectively, Kruskal Wallis test) (Figures 5.3. C and D). There was also no overall difference in the expression of $CD8^+$ T-cell markers for immune senescence (CD57) or dysfunction (CD279) between the groups ($p = 0.088$ and 0.141 respectively, Kruskal Wallis test) (Figures 5.3. E and F).

On average, patients with anellovirus levels detectable by qPCR were younger than those without ($p = 0.013$, means = 43 and 60, Mann-Whitney test) but did not significantly differ in age to the healthy volunteer group ($p = 0.067$, mean = 37, Mann-Whitney test). Unfortunately, a comparison of viral load between groups was not possible due the viruses' low prevalence in patients with normal levels of $CD8^+$ and $CD4^+$ T-cells (n = 1) and healthy volunteers (n = 1).

Metagenomic sequencing results further supported the aforementioned trend (Figure 5.4.). Viruses were detected more often in the patients with simultaneously reduced $CD4^+$ naïve cells and increased $CD8^+$ effector cells compared to those patients with normal levels of one or both of these cell types ($p = 0.009$, Fisher's exact test). This was also true for patients with high levels of $CD8^+$ effector cells alone ($p = 0.031$) and was close to significant for patients with reduced levels of $CD4^+$ naïve cells ($p = 0.057$, Fisher's exact test). Individually, TTV and TT Mini Virus were both more prevalent in the combined $CD8^+$ high, $CD4^+$ low group ($p = 0.039$ and 0.047 respectively, Fisher's exact test). It should be noted that a single patient with both reduced $CD4^+$ naïve cells and increased $CD8^+$ effector cells did not produce any viral reads by NGS, however anellovirus was detectable in this patient by the more sensitive qPCR assay.

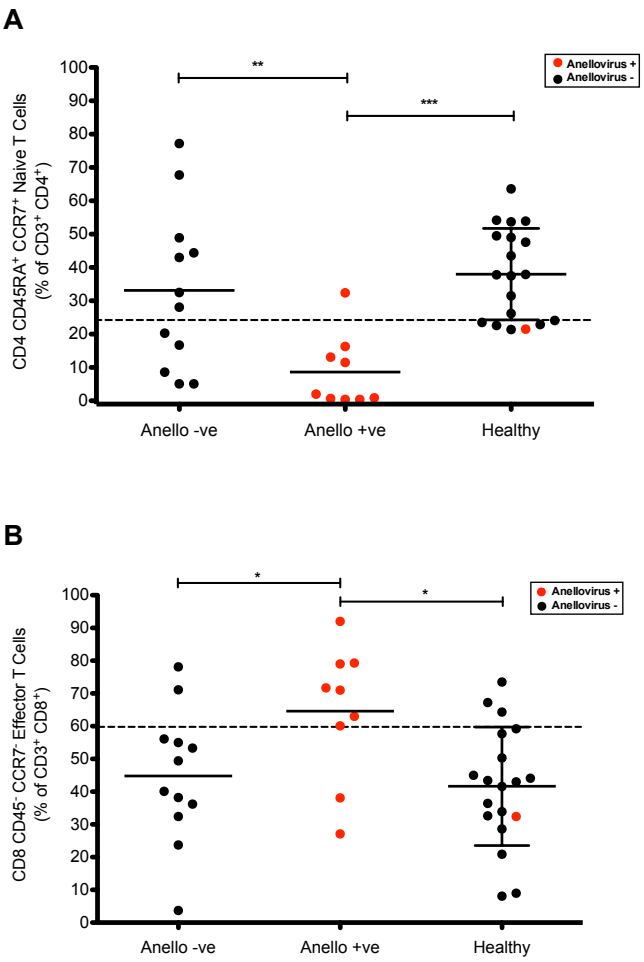


Figure 5.2. **Correlation of T-cell population phenotype and anellovirus status.**

Prospective whole blood samples taken from antibody deficient patients and healthy volunteers at a single time-point (cohort 1). A) Antibody deficient patients positive for anellovirus positive by qPCR displayed significantly lower levels of CD4 CD45RA$^+$ CCR7$^+$ naïve T cells compared to anellovirus negative patients ($p = 0.0062$, Mann-Whitney Test) and healthy volunteers ($p = 0.0139$, Mann-Whitney Test), and B) significantly higher levels of CD8 CD45RA$^-$ CCR7$^-$ effector T cells compared to anellovirus negative patients ($p = 0.0428$, Mann-Whitney Test) and healthy volunteers ($p = 0.0002$, Mann-Whitney Test). Median and interquartile range are shown. The lower and upper range of the healthy control T-cell populations is represented by a dotted line, as an approximation of what might be considered a 'normal' level.

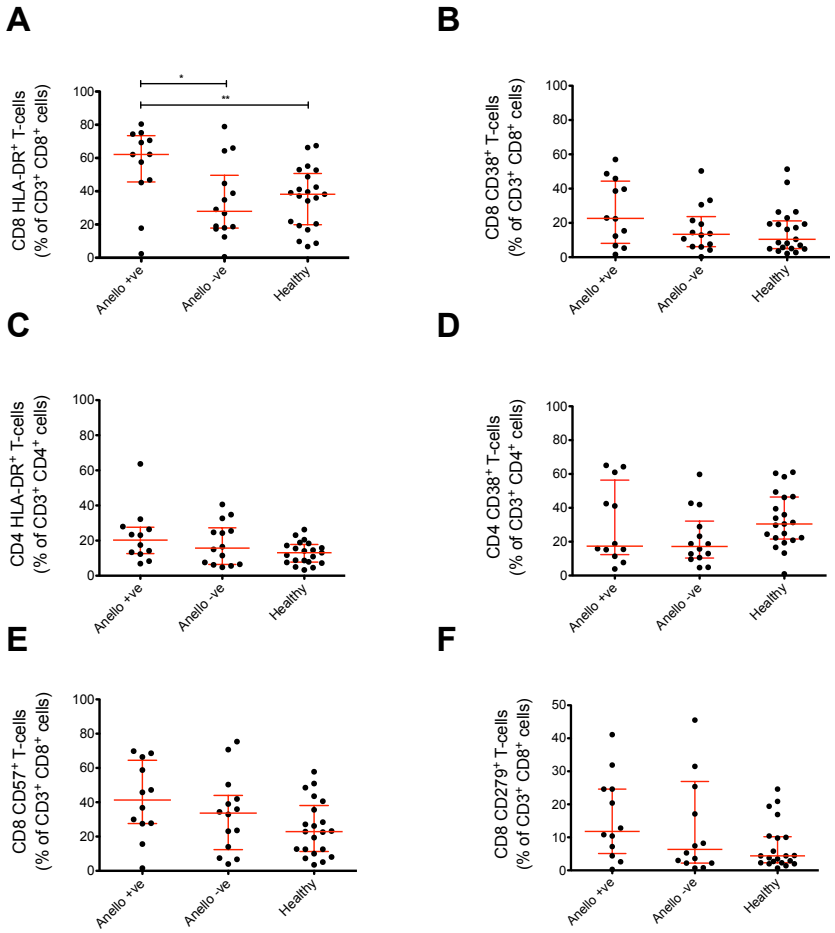


Figure 5.3. **Markers of T-cell activation and dysfunction in antibody deficient patients according to anellovirus status.**

Prospective whole blood samples taken from antibody deficient patients and healthy volunteers at a single time-point (cohort 1). A) HLA-DR expression on $CD8^+$ T-cells is increased in Anellovirus positive patients, compared to Anellovirus negative patients ($p = 0.0372$, Mann Whitney Test), and the healthy control group ($p = 0.0093$, Mann Whitney Test). Expression of other markers of T-cell activation and dysfunction did not significantly differ between groups: B) CD38 expression of $CD8^+$ T-cells. C) HLA-DR expression of $CD4^+$ T-cells. D) CD38 expression of $CD4^+$ T-cells. E) CD57 expression of $CD8^+$ T-cells. F) CD279 expression of $CD8^+$ T-cells. Median and standard deviation are shown. Kruskal Wallis Test was used to determine whether there was an overall difference between multiple groups. Mann Whitney U Test was used to compare individual groups. Only differences below the statistical significance threshold of $p < 0.05$ are shown.

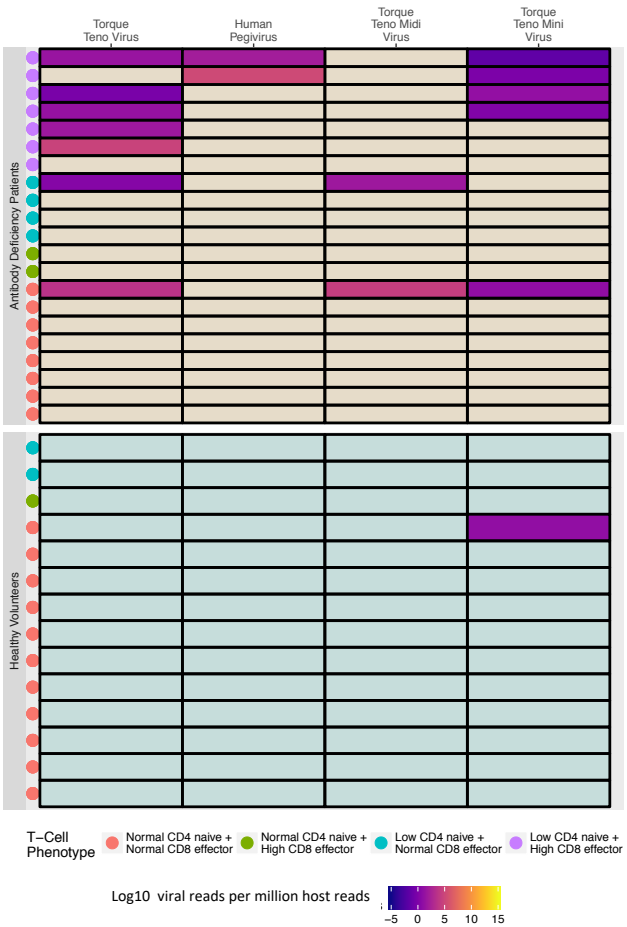


Figure 5.4. **Correlation between T-cell phenotype and the plasma virome.**

Results of metagenomic sequencing of prospective, whole blood samples taken from antibody deficient patients at a single time-point (cohort 1, n = 21) and healthy volunteers (n = 14). Metagenomic sequencing revealed a significantly greater viral prevalence in antibody deficient patients with levels of circulating CD4+ naïve cells and CD8+ effector cells outside of the normal range (defined by the interquartile range depicted in Figure 5.1) ($p < 0.01$, Fisher's Exact Test). Viral read count is shown as a proportion of host-derived reads in order to normalize between samples.

5.3.3. No correlation between anellovirus and immune-mediated diseases

A general linear model was used to test whether the presence of anellovirus was associated with past incidence of chronic inflammation or autoimmune diseases such as idiopathic thrombocytopenic purpura, or autoimmune hepatitis (see Table 5.1. for overview of patients in Cohort 1). The model revealed no significant association between TTV and chronic inflammatory diseases ($p = 0.478$) or between TTV and autoimmune diseases ($p = 0.162$). There was also no correlation when considering both immune symptoms together ($p = 0.184$).

5.3.4. The effect of antibody replacement therapy on the blood plasma virome

I next investigated whether the administration of antibody replacement therapy in antibody deficient patients leads to a decrease in viral burden. To test this, I performed metagenomic sequencing and TTV-specific qPCR on archived paired plasma samples from Cohort 2 (see methods). These samples represented time-points before and after the commencement of antibody replacement therapy.

Metagenomic sequencing revealed no significant change in the prevalence of any virus between the pre-therapy and post-therapy time-points ($p > 0.05$, McNemar's test) nor compared to the healthy control cohort ($p > 0.05$, Fisher's exact test) (Figures 5.5. and 5.6.).

5.3.5. The effect of antibody replacement therapy on anellovirus load

In order to determine whether the replacement of antibodies leads to a reduction in viral load in antibody deficient patients, anellovirus copy number was quantified by qPCR for the archived paired plasma samples where sufficient

material remained following sequencing (Figure 5.7.) (Cohort 2, see method section 5.2.1.).

The qPCR assay revealed that the average anellovirus load in CVID patients did not decrease following antibody replacement therapy ($p = 0.36$, Wilcoxon test). Virus load was also not increased compared to the healthy control group ($p = 0.20$, Mann-Whitney U test). There was also no difference in the prevalence of anellovirus in the CVID pre-therapy group compared to post-therapy ($p = 0.51$), or the control group ($p = 0.27$, Fisher's exact test). Overall, there was no trend towards increased or decreased anellovirus load following treatment. Indeed, there appears to be a roughly even split between patients whose viral load increased (highlighted in red) or decreased (blue) following treatment. This suggests that antibody replacement has absolutely no effect on anellovirus. However, it is possible that there is a difference in the underlying defect in these two groups in which antibody replacement therapy restores the component required for immune control in one group but an alternate component (e.g. recognition of bound antibodies) is missing in the second group.

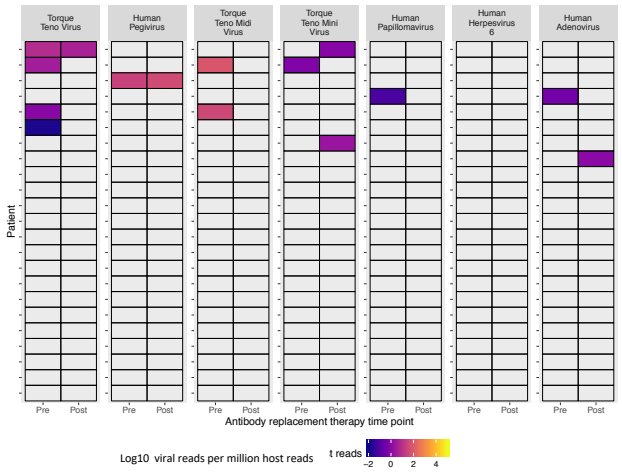


Figure 5.5. **Viruses detected by Illumina sequencing of CVID patient serum pre- and post-antibody replacement therapy**

Metagenomic sequencing of archived, paired serum samples taken from CVID patients at pre- and post-treatment time-points (cohort 2, n = 23) revealed no significant change in viral prevalence following treatment ($p > 0.05$ for all viruses, McNemar's test). Viral read count is shown as a ratio of host-derived reads in order to normalize between samples.

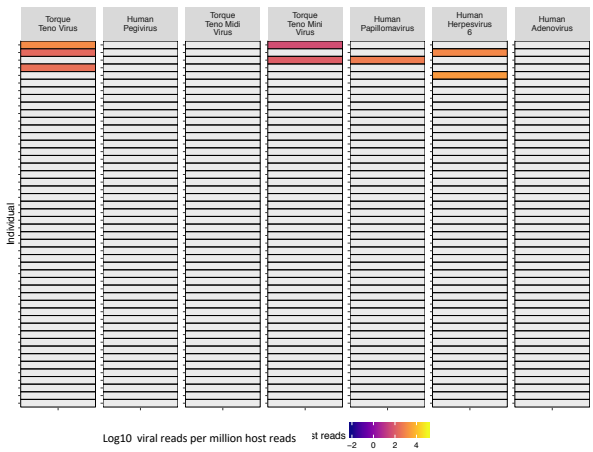


Figure 5.6. **Viruses detected by Illumina sequencing of healthy volunteer plasma.**

Archived plasma samples from healthy volunteers (n = 48) were sequenced using the metagenomic protocol. Viral reads were identified by alignment to a viral reference sequence database and confirmed by BLASTn. Viral read count is shown as a ratio of host-derived reads in order to normalize between samples.

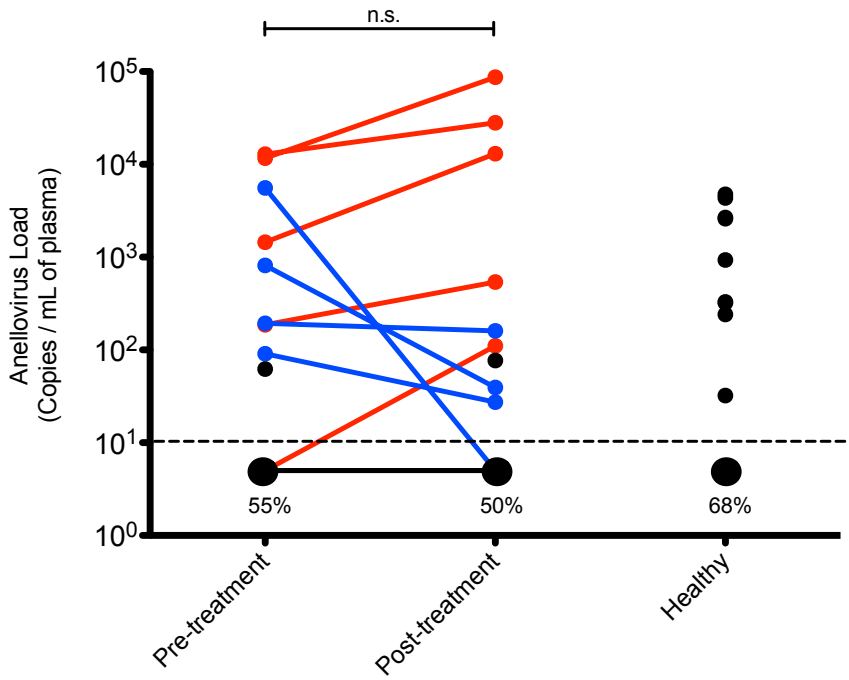


Figure 5.7. **Anellovirus load in CVID patient serum pre- and post-antibody replacement therapy.**

Cohort 2. Retrospective qPCR analysis of anellovirus load in archived, paired serum samples taken from CVID patients at pre- and post-treatment time-points (cohort 2). These samples represent pre- and post-antibody replacement time points (n = 20 and 18 respectively) and healthy volunteers (n = 28). Average anellovirus load did not alter following commencement of antibody replacement therapy (p = 0.36, Wilcoxon Test). Patients with viral loads that increased and decreased are coloured red and blue respectively, unchanged or unpaired samples are in black. Limit of detection for the assay is indicated by the dotted line and the percentage of patients with undetectable levels of virus is displayed below the line.

5.3.6. Endogenous retrovirus expression levels in antibody deficient patients

Finally, I wished to test whether a virus was correlated with another complication for which antibody deficient patients are at an increased risk: lymphoma. Here, my hypothesis was that antibody deficiency has the potential to lead to the re-activation of endogenous retroviral elements through lack of immune control at the gut-mucosal barrier, and that this reactivation can result in the development of lymphoma.

This hypothesis is based on observations from a murine model described by Young et al who demonstrated that IgM (and not IgA) deficiency results in increased gut permeability to bacteria and that the presence of bacterial products.[461] These then were shown to trigger an increase in the expression of endogenous retrovirus and the production of infectious particles. Intriguingly, increased activity of the human endogenous retrovirus K (HERV-K) has been associated with an increased incidence of lymphoma in CVID patients.[457,462]

To test this hypothesis, I used a qPCR assay to measure endogenous retrovirus RNA expression levels, an immunoglobulin M ELISA to measure IgM concentration, and an ELISA for soluble CD14 (sCD14) as a measure of bacterial translocation from the gut into the peripheral blood. sCD14 concentration is an indirect marker of bacterial translocation, which has been shown to correlate with LPS activation of monocyte / macrophages, a well-defined biomarker used to monitor disease progression in HIV/AIDS patients.[463] This was done in paired PBMC and plasma samples from the antibody deficient patients and healthy volunteers in cohort 1 (see methods section 5.2.1).

Average IgM levels were significantly reduced in untreated, antibody deficient patients ($p = 0.004$, Mann-Whitney test) and in the entire antibody deficient group irrespective of treatment status ($p = 0.01$, Mann-Whitney test) (Figure 5.8.A). However, HERV-K RNA levels were decreased compared to the healthy control group, in both treated ($p = 0.03$, Mann-Whitney test) and un-treated patients ($p = 0.02$, Mann-Whitney test) (Figure 5.8.B). Furthermore, no significant difference in soluble CD14 levels was detected between all patient and control groups (Figure 5.8.C), and reduction of IgM did not correlate with an increase in soluble CD14 ($p = 0.976$, linear regression analysis) (Figure 5.8.D) or with an increase in HERV-K expression ($p = 0.285$, linear regression analysis) (Figure 5.8.E).

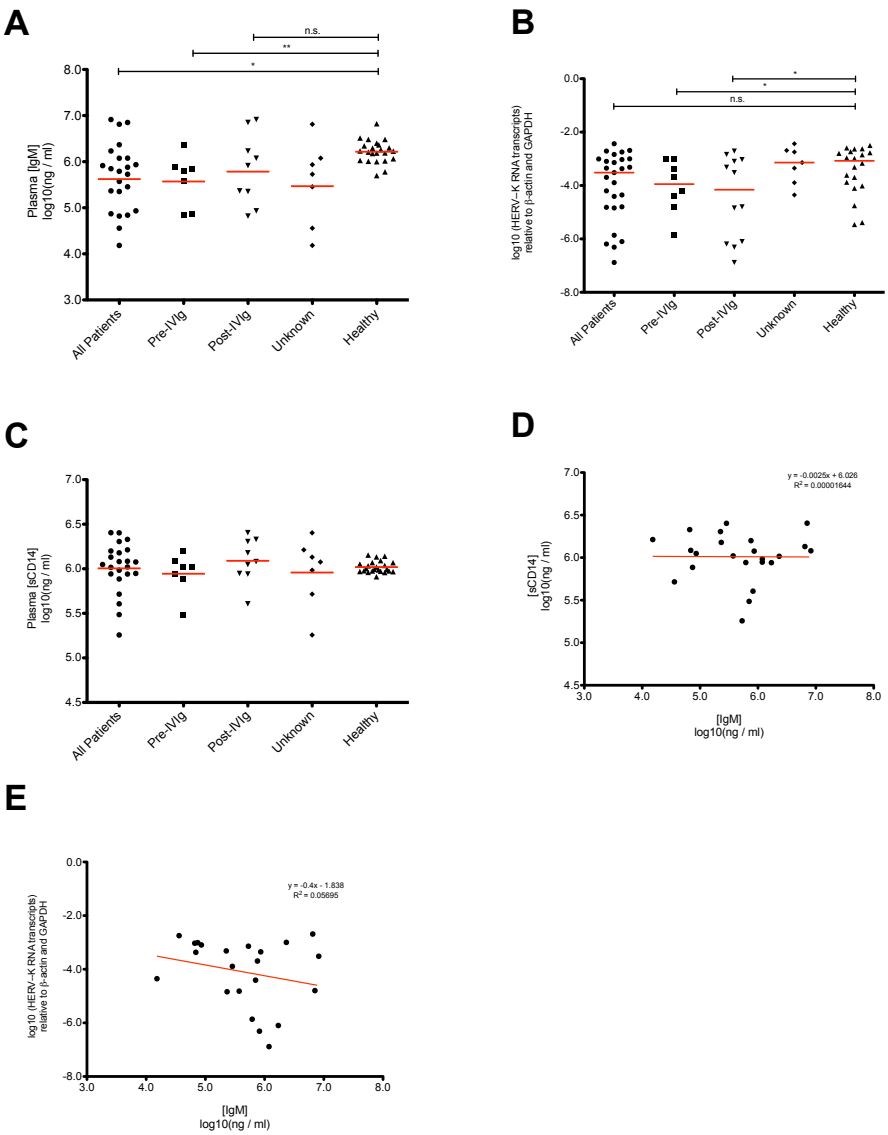


Figure 5.8. **Immunoglobulin M, soluble CD14, and human endogenous retrovirus-K levels in the blood of primary antibody deficiency patients.** Cohort 1. Prospective whole blood samples taken from antibody deficient patients at a single time-point. A) Unsurprisingly, circulating levels of IgM were decreased in untreated antibody deficient patients compared to healthy controls ($p = 0.004$, Mann-Whitney Test). However, Human Endogenous Retrovirus K RNA transcript levels in the PBMCs of both treated and untreated patients with antibody deficiency ($n = 27$) were decreased in comparison to and healthy volunteers ($n = 19$) ($p = 0.02$, and 0.03 respectively, Mann-Whitney Test) (B). Furthermore, soluble CD14 (sCD14) levels in the plasma in of the patients did not differ from that of volunteers (C). Finally, no association was observed between IgM and sCD14 (D), nor HERV-K levels (E). Patient and control samples from cohort 1 were used for this experiment where sufficient material remained following PCR and sequencing (both $n = 23$).

5.4. Discussion

5.4.1. Anellovirus prevalence in plasma is correlated with T-cell phenotype in antibody deficient patients

A combination of flow cytometry, metagenomic sequencing and qPCR data revealed a correlation between increased viral prevalence, predominantly anelloviruses, and an activated T-cell phenotype in antibody deficient patients. This activated T-cell phenotype has previously been associated with increased disease severity in antibody deficient patients, predominantly due to an increased incidence of inflammatory complications.

Statistical analysis did not reveal a significant association between the presence of anellovirus and past inflammatory complications in the antibody deficient patients. However, incomplete clinical data and a small cohort size make it unsurprising that no strong correlation could be found. Therefore, the possibility that anelloviruses are associated with inflammatory disease in these patients should not be entirely excluded. Based on previous evidence which has suggested that anelloviruses are capable of eliciting an immune response,[398,464,465] the virus may directly influence immune activation in antibody deficient patients. However, given the viruses' association with an increased susceptibility to infection demonstrated in chapter 2, presence of the virus could also be a marker, highlighting the presence of other pathogens responsible for inflammation. If the latter hypothesis proves to be true, anellovirus load may again prove a useful marker of immune dysfunction for the clinical management of these patients.

Overall, these results warrant an investigation into potential viral involvement in the development of the activated T-cell phenotype in antibody deficient patients.

The clear difference in viral prevalence between the patients with and without the activated T-cell phenotype also lends support to the recent decision to exclude the former from the classification of CVID on the basis of differences in disease severity.[466]

5.4.2. Antibody replacement therapy does not alter the blood plasma virome

Instigation of antibody replacement therapy was found to have no marked effect on either the overall virome or on anellovirus load. This suggests that the current treatment regime for antibody deficient patients has no impact on pre-existing viral infections.

With regard to the virome of antibody deficient patients, the overall picture was similar to that seen in healthy volunteers. Once again, the anelloviruses TTV, TTMDV, and TTMV were the most commonly detected viral species, in agreement with the majority of previous metagenomic studies investigating the plasma virome.[39,275,278]

HPgV was detected at a high level in a single patient pre- and post-treatment, indicating that immunoglobulin replacement therapy does not resolve this particular infection. The primary route of transmission for the virus is thought to be via contact with blood however a lack of known pathology means that despite its high prevalence in the blood supply,[467] donations are not screened for the virus.[468] Therefore it is likely that CVID patients are exposed through immunoglobulin replacement therapy rather than as a direct consequence of their immunodeficiency.

Human Papillomavirus (HPV) was detected in the plasma of both a CVID patient prior to IVIg and a healthy volunteer. Papillomaviruses are common components of the skin microbiome in healthy individuals,[469] although certain types are associated with malignancies.[470] A BLASTn search of the HPV sequences revealed gammapapillomavirus type 4 in the healthy individual and betapapillomavirus type 159 (Best match: Accession number HE963025) in the patient. These viruses are associated with the mucosa and cutaneous tissue respectively and neither has been associated with increased risk of malignancy although HPV type 4 can cause warts.[471] Detection of HPV in the blood has previously been reported at extremely low prevalence in healthy individuals[472] and in the blood of sexually naïve patients with a history of transfusion which suggests the virus may be transmissible via blood products such as IVIg.[473] Whilst uncommon in healthy individuals, both chronic and symptomatic HPV infection have been reported previously in immunocompromised populations.[369,474,475]

Human adenovirus C (HAdV-C) was detected at a single time-point in two CVID patients, one pre- and one post-IVIg therapy, and no healthy volunteers. HAdV-C is primarily associated with respiratory disease which is a common presentation for antibody deficient patients.[476] Adenoviruses are also a regular complication in immunodeficient cohorts such as stem-cell transplant recipients.[477] Intriguingly, whilst HAdV is generally cleared by healthy individuals, the virus has been found to persist in the plasma of immunodeficient patients.[478,479] A number of other, generally acute, viruses including norovirus, rotavirus, and parechovirus have been shown to chronically infect antibody deficient patients.[480–482] This highlights the possibility of the immunodeficient population acting as a reservoir for viruses, and potentially as a source of seasonal outbreaks.

Combined, the metagenomic and qPCR results suggest that, unlike bacteria, viral burden does not appear to be reduced by the administration of antibody replacement therapy. Although, it should be noted that viral burden was not significantly higher than that of healthy controls to begin with.

5.4.3. Endogenous retrovirus expression levels are not increased in antibody deficient patients

Finally, it was found that despite reduced levels of IgM, expression of HERV-K does not appear to be increased in antibody deficient patients and is therefore unlikely to be responsible for the increased incidence of lymphoma in these patients. This finding is at odds with the murine model described by Young et al.[461]

One likely explanation is that the prophylactic antibiotics given to CVID patients deplete the gut microbiota. Therefore, despite suffering from IgM deficiency, few bacteria translocate from the gut into the bloodstream, and the bacteria-driven mechanism of endogenous retrovirus activation described in IgM deficient mice is not triggered. Elevated markers of systemic bacterial translocation have previously been demonstrated in antibody deficient patients prior to the commencement of immunoglobulin replacement therapy.[448] However, other studies have been unable to detect the same result,[445] suggesting the findings in this current study are not contradictory.

5.5. Conclusion

In conclusion, the findings of this study have demonstrated that a subset of antibody deficient patients with evidence of additional T-cell dysfunction, suffer from an increased number of viral infections, particularly by anellovirus.

However, patients with antibody deficiency alone do not appear to be at a significantly increased risk of systemic viral infection, and antibody replacement therapy does not have a significant impact on viral burden. Together, this would suggest that control of chronic anellovirus infection is primarily reliant on the T-cell response, rather than the humoral response.

Antibody deficient patients with the low CD4+ naïve phenotype have previously been shown to suffer from high rates of mortality and inflammatory disease.[433] It still remains to be determined whether this phenotype is due to an underlying genetic defect, the result of environmental factors such as infection, or a combination of the two. The high proportion of activated cells observed in these patients suggests a long-term, chronic exposure. Conversely, the fact that patients do not switch between phenotypes is suggestive of a genetic rather than environmental basis. Considering this evidence in combination I would hypothesise the presence of a genetic defect in the CD4+ T-cell compartment of these patients, which predisposes to both antibody deficiency, as well as viral persistence due to the inability of CD8+ T-cells to clear infections. CD4+ T-cell deficiencies are particularly severe compared to deficiencies of the other lymphocyte subsets, as these cells are responsible for initiating both the B-cell and cytotoxic T-cell response (see Figure 1.3. of Introduction). Taking AIDS as an example, CD4+ cell depletion leads to severe immunodeficiency characterised by frequent opportunistic infections.[92]

Intriguingly, the naïve CD4+ low, effector CD8+ high phenotype detected in the antibody deficient cohort resembles that seen in AIDS patients. In AIDS, studies have shown that low CD4+ naïve cell levels are a result of viral-mediated destruction, while high effector CD8+ cell levels are driven by viral-specific activation.[483] In the antibody deficient cohorts, it appears that a congenital defect resulting in a decreased replenishment of new CD4+ T-cells has a similar effect,

which also results in CD8+ T-cell activation driven by an unknown, antigenic, stimulus.

The increased presence of anellovirus in patients with high levels of CD8+ effector T-cells suggests that these patients suffer from a lack of immune control over viruses. Therefore, it is possible that viruses are the drivers of immune activation in these patients. This remains to be proven, however if this is the case, anelloviruses may be the responsible agents. However, it is equally feasible that a yet unidentified pathogen is the cause of this phenotype. This pathogen may be a novel virus or it may be present at a level below the detection limit of the metagenomic method.

Furthermore, as antibodies are a major component of mucosal immunity, and many of the complications experienced by CVID such as gastroenteritis and granulomatous-lymphocytic interstitial lung disease are associated with mucosal surfaces, it is quite possible that in antibody deficient patients, viral infections are more prevalent at mucosal sites. Indeed, it has been shown that enteric viruses are more common in CVID patients, particularly those with low levels of IgA, compared to healthy individuals.[484] Here, the detection of the respiratory virus, Human Adenovirus C in the blood lends weight to the hypothesis that the mucosal virome goes unchecked. Therefore, it would be prudent to extend this work to 'mucosa-associated' samples such as faeces or brocho-alveolar lavages, before it is concluded that viruses have little impact in the antibody-specific deficiency setting.

In the AIDS setting, although CD4+ deficiency is the primary cause of disease, the activated CD8+ phenotype often persists, regardless of CD4+ recovery following ART.[485] Furthermore, persistence of the activated CD8+ phenotype is highly associated with non-AIDS related inflammatory complications.[485] Therefore, it is

conceivable that a similar phenomenon is also driving idiopathic inflammatory complications in the antibody deficient group. Considering this, further research is required to elucidate the stimulus that is driving CD8+ effector cell activation in these patients, if the inflammatory complications that are responsible for the majority of morbidity and mortality in antibody deficient patients are to be resolved.

Overall, further studies are needed to investigate the impact of antibody deficiency on the viral communities of the mucosal surfaces, the lungs and gut in particular, in order to elucidate the true burden of viral infections in these patients. Additionally, studies of a larger cohort may be useful in proving or disproving the role of anellovirus and determining whether it is causative of chronic inflammation or merely associated with the activated T-cell phenotype.

Chapter 6

Discussion

6.1. Summary of results

This thesis set out to investigate the prevalence and effects of viral infections in different immunocompromised populations. I began by developing a metagenomic sequencing methodology to be used for the broad detection of viruses in clinical blood samples, aiming to minimise sample loss and avoid the introduction of bias wherever possible. The optimised methodology was then applied to blood plasma samples taken from three different cohorts of immunocompromised patients; from both the primary and secondary immunodeficiency settings.

Overall, the results of this thesis have demonstrated that there are a number of clinically undiagnosed viruses infecting both healthy and immunocompromised individuals, and that immunosuppression can lead to a significant increase in the abundance and diversity of these viruses. These findings correlate well with the relatively small number of viral metagenomic studies that have previously been reported for both healthy and immunocompromised cohorts.[39,278,307,377]

6.2. Comparison of immunodeficient cohorts

A comparison of the three immunodeficient cohorts revealed interesting differences in terms of viral prevalence. The renal transplant cohort and the vasculitis cohort showed a distinct increase in viral infections following therapeutic immunosuppression, whilst the plasma virome of the antibody deficiency cohort appeared to largely resemble that of healthy individuals.

Intriguingly, antibody deficient patients with abnormalities in their T-cell compartment were an exception to this finding.

The foremost difference between the groups lies with the immune compartments that are suppressed; whilst the systemic administration of immunosuppressive drugs to vasculitis and renal transplant patients has a strong and broad effect on all lymphocytes, including NK-, B- and T-cells (see Figure 1.3), the impact of primary antibody deficiency is generally far more specific. It has previously been noted that very specific deficiencies rarely result in a marked increase in susceptibility to infections, as the immune system has evolved a high degree of redundancy between compartments.[486] Therefore, a possible reason for the disparity between the viromes of cohorts, is that other arms of the immune system effectively cover the gap in immunity created by antibody subclass deficiency, whereas the suppression of multiple compartments (e.g. B-cells and T-cells) in therapeutically immunosuppressed patients is broad enough for immune redundancy to be lost.

An alternative interpretation of the results is that T-cells are indeed the immune cells primarily responsible for efficiently resolving viral infections, and antibody deficient patients with normal T-cell compartments are at little risk. Both this theory and the theory above have the potential to explain why viral prevalence is increased only in those antibody deficient patients with additional evidence of T-cell dysfunction.

Overall, these findings are not consistent with my initial hypotheses that antibody deficient patients suffer from a number of undiagnosed viral infections. One potential caveat to this is that metagenomic sequencing was restricted to blood samples. Antibodies are known to play an important role in immunity at the mucosal surfaces, and so whilst humoral deficiency appears to not result in

an increase in blood-borne viral infections, infections at the mucosal surfaces such as the gut and respiratory tract may still be common. Indeed, the majority of CVID patients present with recurrent respiratory infections and many of the clinical complications these patients experience, such as bronchitis and gastroenteritis, are associated with either the lung or gut.[439] Therefore, further studies performing metagenomic sequencing of samples taken from the mucosal surfaces of antibody deficient patients may give a more complete picture with regards to the true prevalence of viral infections in these patients.

Furthermore, by sampling individual time points, this study is likely to have missed a number of short-lived, acute infections. Therefore, the findings are skewed towards the detection of chronic infections that are consistently present. Given that the cytotoxic T-cell response is vital for the clearance of many viral infections,[487–490] it is perhaps obvious that patients with T-cell deficiencies would consistently yield a greater abundance of viruses. Therefore, it would be prudent to repeat this study with samples taken at symptomatic time-points and with shorter periods between sampling in order to better account for acute viral infections, before fully ruling-out viral infections as a cause of morbidity in antibody deficient patients.

6.3. Anelloviruses and their association with immunosuppression and disease

The major outcome of the work of this thesis was the demonstration that infection with the highly prevalent anellovirus is associated with specific disease phenotypes and outcomes in immunocompromised cohorts. However, a specific relationship between anelloviruses and these diseases was not elucidated.

6.3.1. Anellovirus plasma load as a prognostic indicator for transplant patients

One of the findings demonstrates that anellovirus load as a prognostic indicator of future infection could prove an extremely useful tool in the management of patients under immunosuppressive regimes.

The criteria for determining whether a diagnostic test is sufficiently accurate is very much dependent on the setting; an area under the curve (AUC) of greater than 0.70 is generally regarded as a strong effect,[491] however for the diagnosis of specific pathogens a higher AUC is generally required. The qPCR assay used in this thesis generated an AUC of 0.72, which suggests that its accuracy is acceptable but there is room for improvement. However, it is also worth considering that in this scenario, where no assay currently exists, a test with any form of predictive capability may prove useful.

In order to improve the diagnostic accuracy of the prognostic test, either the sensitivity or the specificity must be increased. One possibility is that the high heterogeneity of anellovirus species in each patient, as demonstrated by the MinION sequencing experiment (see section 4.3.10.), may confound the assay results. Therefore, one possible method for improving the accuracy of the test is

by determining whether there is a single, specific anellovirus species or gene which can be used to predict future infections more accurately than the others. Sequencing the whole genomes of anelloviruses in patients using the MinION and looking for correlations could be one way of identifying such a gene or species. Here, the use of whole genome sequencing is important as it would negate the need for genome assembly, therefore ensuring that anellovirus species not included in the viral reference database are considered, as it may be one of these viruses that is found to correlate best.

Longitudinal studies have already closely examined the dynamics of anellovirus load following immunosuppression.[278,379,492] However, a deeper understanding of how the viruses are acquired and how individual species and the population as a whole reacts to immunosuppression in terms of diversity would also be beneficial when considering how to develop a test applicable to all immunosuppressed cohorts. Combined, these experiments could contribute to the development of a clinically useful test for the monitoring immunosuppressive regimes, applicable to a wide-range of cohorts who, at present, are managed based on a retrospective reaction to symptoms rather than an understanding of the underlying mechanisms which could help prevent these symptoms ever occurring.

6.3.2. Proof of causality in the metagenomic era

Somewhat conversely, anellovirus was also found to be associated with evidence of chronic T-cell activation in primary immunodeficiency patients. There are several possible explanations for this correlation, including that presence of the virus can instigate an inflammatory response. However, given the viruses abundance in individuals without these symptoms, it is clear that the mechanism would have to be more convoluted than straightforward ‘presence equals

causation'. Indeed, the discovery of the first anellovirus, Torque-Teno Virus, generated great excitement in the scientific community, as it was rapidly isolated from hepatitis patients worldwide.[493–495] However, claims implicating the virus as an agent of liver disease were abandoned just as quickly, following the revelation that it was in fact prevalent in asymptomatic, healthy blood donors.[496–498]

Historically, proof that an infectious agent is responsible for a disease has been based on the postulates established by Robert Koch,[499] which state that the putative pathogen must first be isolated and cultured from patients suffering from disease, but absent in all healthy individuals. Inoculation of the isolated, cultured pathogen into a healthy host must then cause the disease to manifest anew. Finally, the pathogen should be re-isolated from the newly diseased host. These strict guidelines were important in ensuring that the isolated pathogen was responsible for causing the disease, and not merely a contaminant, or an opportunist taking advantage of the diseased state.

However, a number of recent studies have demonstrated that whether or not disease develops can depend on the composition of the wider community of microbes present in a host, rather than the presence or absence of a single pathogen. For example, the bacteria *Clostridium difficile*, which has been previously verified by Koch's postulates as responsible for antibiotic-induced diarrhoea, was found to no longer cause disease in the presence of another bacterial species: *Clostridium scindens*.[500] A further study has reported similar results based on entire communities of bacteria, eventually identifying a combination of six species that best protected against *C. difficile* colonisation.[501] Similar findings have also been demonstrated for viruses, where the presence of murine norovirus led to germ-free mice generating a more potent immune response following infection with *Citrobacter rodentium*.[502] Overall, the results of

these studies suggest that it is perhaps time to adjust Koch's postulates, in order to better represent our understanding of disease as a consequence of multiple, inter-related factors rather than the presence of a single pathogen. Importantly, metagenomics now provides us with the tools to address this issue.

Several groups have discussed how best to adapt Koch's postulates for use in the metagenomic era.[503,504] They suggest that metagenomic investigations into agents of disease should first begin by sequencing the microbiome in both diseased and healthy individuals, followed by informatic analysis to identify common markers, such as sequencing reads or particular genes, that contribute to uniquely distinguishing between diseased and healthy subjects. Importantly, this approach should consider the possibility of multiple marker sequences, rather than sequences derived from a single pathogen. To prove causation, the identified markers can then be monitored in healthy individuals or at-risk cohorts. Alternatively, model organisms could be used through the inoculation of a healthy individual with a sample from the diseased and later by the isolated agents alone, in order to confirm that infection by the identified agents results in a diseased state. Importantly, this metagenomic approach accounts for the possibility of multiple agents combining to result in disease, and also foregoes the need for virus culture, which is a limiting and often confounding factor when testing for a number of virus species.

In this thesis, the initial two steps of these new postulates have been performed for two small cohorts of patients, renal transplant recipients and primary antibody deficient patients, both of whom experience unexplained symptoms of ill-health or disease. The resulting analysis of the two cohorts identified the presence of anellovirus as both a marker of chronic immune activation and susceptibility to infection. However, further work is required to determine how the virus is related to either of these symptoms.

6.3.3. Anellovirus and immune activation

In light of our understanding of how entire communities can impact whether infection by a single agent will or will not result in disease, I argue that the evidence implicating anellovirus in a number of diseases such as inflammation, should not be disregarded based purely on its presence in non-symptomatic individuals. The virus has been isolated from the site of inflammation in a number of diseases,[505–507] and in vitro experiments have demonstrated that the virus is capable of initiating an immune response.[465] The work of this thesis further adds to this evidence by demonstrating associations between increased presence of the virus and immune related health problems, including inflammatory disease and immune-mediated rejection episodes.

Potential parallels can be drawn between anelloviruses and the well-studied Cytomegalovirus (CMV), in terms of a highly prevalent virus, causing disease in only some individuals. Whilst the clinical impact of CMV is thought to be relatively minor in healthy individuals, the virus is a major health issue for immunosuppressed patients, in whom it is able to reactivate and cause active disease. However, active CMV infection in both immunosuppressed patients and previously healthy individuals has been reported as responsible for the induction of a range of autoimmune-like phenomena, including vasculitis and lupus,[508–510] despite no such symptoms appearing in the majority of the healthy, yet infected, population. Furthermore, a large study investigating the burden of CMV has revealed that even in apparently healthy individuals, life-long, asymptomatic infection is associated with a 3.7 year average reduction in life-span.[29] The reason for this is unknown, however one hypothesis put forward states that the reduction in life span is due to the toll taken by a lifetime of the subclinical reactivation events, engaging the immune system and gradually leading to an exhausted immune state.[511]

If anelloviruses are revealed to be responsible for generating a similar chronic immune response, it may be having the same effect, decreasing life span without necessarily generating overt symptoms except in certain cases, making it a far greater issue than currently thought. This particular hypothesis is difficult to test in humans as anellovirus is acquired essentially from birth and exists at a very high prevalence in the healthy population making recruitment of a virus-free comparison group difficult.[43,45] Furthermore, a report of murine anelloviruses at a high prevalence in laboratory-bred rats suggests that finding any mammalian model system to test anellovirus pathogenicity could be difficult.[512] However some hope remains, as a number of anelloviruses detected in wild UK mice appear to be absent from their laboratory counterparts.[513]

In order to further investigate the correlation between anellovirus and immune activation in primary antibody deficiency patients, it is important to first determine whether increased viral load and diversity is a product of immune activation or whether the virus is directly responsible (discussed in section 4.4.2). One potential method for testing this would be to isolate the cytotoxic effector T-cells from these patients and test whether the cells 1) proliferate, or 2) release pro-inflammatory cytokines, in the presence of the virus or virally infected cells. Evidence of this would suggest that they are specific for anellovirus and that the virus is indeed driving at least part of the pro-inflammatory response.

If this experiment is able to establish a strong association between the virus, activated T-cells and inflammation, a wider metagenomic study can begin in order to work out why the virus causes inflammation in some individuals and not others, based on the newly proposed update to Koch's postulates**.** The identification of potential correlates of immune activation would then

necessitate further studies requiring the inoculation of antibody-deficient mice with virus isolated from patients experiencing chronic inflammation in order to determine whether it is capable of eliciting the same T-cell activation and inflammatory profiles.

6.4. Limitations to this study

6.4.1. Limitations to the methodology

There were several limitations to the way that this metagenomic study was performed, which may have had an impact on the results. One factor to be considered is that, in order to sequence a large number of samples over a relatively short time period, a high-throughput sequencing facility was used for parts of this study. This resulted in some disparity, in terms of library preparation and sequencing, between the carefully optimised approach developed in the first chapter of this thesis, and the method used for the latter two chapters. The workflow of high-throughput sequencing facilities is primarily set-up for whole genome sequencing of samples containing abundant material from a single organism, rather than metagenomic sequencing of low abundance, clinical samples. Therefore, the samples sequenced in chapters two and three, were not necessarily prepared with the same care that they received during the validation experiments of chapter one. This is likely to account for the various cross-contamination issues described in the results section of each of the chapters, as well as the generation of fewer reads compared to those one would expect based on the results of chapter one.

The effect of varied sequencing output on the sensitivity of metagenomic sequencing is unclear. Previous studies performing rarefaction analyses of bacterial metagenomic data have suggested that the number of species detected

should increase with read number up to a point, before reaching a plateau once all possible species have been identified and additional reads provide no extra information.[514] Unfortunately, the rarefaction approach relies on the generation of a large number of target-specific reads from a diverse community of species. Therefore, it is not applicable to viral metagenomics, where generally only one or two species are detected at a time, and viral reads typically account for less than 1% of the total sequences produced. Regardless, it would be reasonable to assume that the detection of low abundance viral species is more likely in samples that generate a greater number of reads. To account for this, the average number of sequences generated from samples at each time point was compared within cohorts. No statistically significant difference was detected, suggesting that any disparity in read number between time-points did not play a confounding role in the results of this study.

The exact number of reads produced per sample cannot be controlled directly. For a multiplex sequencing run, producing an equal number of reads for each sample relies on the extremely precise quantification of DNA to enable accurate equimolar pooling. This step is made more difficult when dealing with low abundance samples such as those used in this study, as small disparities can result in a large variation in read number. There are two potential solutions to the issue: the first is to artificially control the read number by randomly selecting and analysing a given number of reads from the total. This is an excellent method for comparing fairly between samples when the read count is high. However, the approach seems wasteful if the samples with the smallest number of reads must be matched or excluded from the analysis entirely. In the interests of not reducing the size of the cohorts, I opted for a second approach: normalising the number of viral reads based on the number of reads that map to the host genome. This value was shown to correlate strongly with virus load as measured

by qPCR, suggesting that it is a valid method for normalising viral read count between samples.

Overall, it appears that the discrepancy between expected and actual read number did not prevent the effective comparison between groups in this study. However, it is possible that this study missed out on some information with regards to subclinical viral prevalence as the deficit in reads may have prevented the detection of some low level viral infections.

6.4.2. Cohort size and clinical information

The size of the primary immunodeficiency cohort is another factor which may have restricted the power of this study. However, it should be noted that any study attempting to prospectively recruit primary immunodeficiency patients is somewhat at a disadvantage owing to the rarity of these diseases. This is, in part, why it was decided to study CVID, the most common of the PIDs, despite the fact that studying this disease is somewhat hindered by its heterogeneity. Paradoxically, the divergent phenotypes of this cohort also raised interesting questions regarding the impact of viruses and the cause of chronic idiopathic inflammation. Even with this small cohort I was able to successful detect meaningful differences which shed light on answers to these questions.

Obtaining the clinical records of the renal transplant cohort made it possible to detect some interesting associations between disease and the virome. Unfortunately, the same was not possible for the primary immunodeficiency or vasculitis cohorts owing to lack of available clinical information at present. However, a similar analysis to that which was performed for the transplant cohort is an important aspect of this study that I hope to follow-up on once the information becomes available.

6.4.3. Bias in the informatics approach

The restriction of the informatic analysis to the identification of viruses of vertebrates means that, for all the work put into developing an unbiased approach, the final methodology was somewhat impartial. Identification of sequences based on any form of reference database automatically limits findings to the pathogens present in that database, i.e. those that have been previously reported to the scientific community. This method therefore precludes the detection of novel, unrelated viruses. As mentioned in the introduction to this thesis, attempts have been made to develop viral metagenomics methods capable of foregoing the reliance on alignment to databases of viral reference genomes.[311,515] However, at present, none of these methods have proven themselves capable of replacing the alignment-based method in terms of sensitivity.

Furthermore, the exclusion of sequences generated from bacteriophage, invertebrate and plant virus, bacteria, fungi, archaea and eukaryotes leaves us with a somewhat incomplete picture in terms of the overall microbial community within each patient. Variations in this community have been shown to have a dramatic impact on the health of its host, as well an ability to alter the virulence of pathogens.[500–502] A similar relationship certainly exists been the host transcriptome and genetic background, and the microbiome, as we already know that differences in how an individual's immune system reacts to a microbe can determine the microbe's pathogenic impact on its host.

The decision to exclude this information was based on two points: 1) The issue of contamination is difficult enough to control for, without considering additional microbes that are sure to be highly abundant in the environment; 2) Consideration of all of these factors would require additional sequencing to

specifically generate the required depth of information, which is expensive, and would also result in loss of statistical power if I was to attempt to consider every variable in our small cohorts. For these reasons, it was decided best to consider only viruses that are known to infect vertebrate species as they have a greater likelihood of directly infecting and causing disease in humans.

6.4.4. Consideration of novel viruses

As well as known pathogens, no consideration was made by this study for the presence of novel viruses which could be responsible for many of the manifestations of disease in these patients. Indeed, immunocompromised patients' increased susceptibility to infection makes them an ideal cohort in which to search for novel viruses capable of infecting humans. However, discovering new viruses unrelated to any known species is extremely difficult.

In particular, viral discovery methods tend to rely on the de novo assembly of contiguous sequences, which can then be interrogated in a number of ways. In this study, generally low coverage of known viruses would suggest that attempting a *de novo* assembly of unknown viruses would be highly optimistic.

One method for increasing the potential of assembling viral contigs, could be if the reads of unknown origin, which remain following mapping to the human and virus genome databases, were taken from all patients and combined. Figure 3 in chapter one of this thesis gives an example of the high proportion of reads that go unidentified. Many of these sequences are likely to represent bacteriophages and bacteria that were not tested for by this study. However, these can be subtracted through additional mapping steps and, based on my own and others experiences, I would expect many sequences to remain, which do not match any of those in the entire nucleotide database.[516,517] The assembled contigs could

then be tested for similarity to known viruses by amino acid rather than nucleotide sequence alignment, and further by searching for viral 'signatures' using the methods discussed in the introduction.[311,515]

Finally, viral discovery is limited by the fact that the majority of viral infections result in only short periods of viremia before they are cleared, making detection of these viruses very difficult without sampling from multiple time-points. In this thesis, only a single, pre-treatment time-point was taken from each cohort, providing us with a 'snap-shot' rather than a clear view of viral prevalence. Therefore, the method employed here is far more suited to detecting viruses capable of persistent infection, such as TTV and Human Pegivirus, rather than acute viral infections or those that reside in specific tissues. To improve upon the results of this study, repeated sampling over multiple time-points would be necessary to improve estimates of viral prevalence. Interestingly, Anthony et al demonstrated a method of using repeated measurements to estimate the total number of viruses that exist in a host species,[518] which may provide an additional insight into the rate of subclinical infections in immunosuppressed patients.

6.5. Towards metagenomic sequencing as a clinical tool

In this thesis I have demonstrated that metagenomic sequencing can be a useful tool for the detection of viruses in clinical samples. However, at present, it is rarely used for this purpose in the clinical setting. Metagenomics has the potential to supplement routine clinical diagnostics by being used to search for the aetiologic agent in cases of idiopathic disease; and by aiding the detection of co-infections, as well as infections that are unsuspected or sub-clinical. However, before metagenomic sequencing can be effectively utilised in the clinical setting a number of hurdles must be overcome.

The complexity of the metagenomic process means that each step, from sample collection and nucleic extraction, to the identification of viral sequences must be validated. However, as discussed in the first chapter of this thesis, there is currently no standard method for viral metagenomic sequencing protocols. To compare a number of these protocols, a panel of 25 reference viruses was developed and sent to 15 laboratories, to be extracted, prepared, sequenced and analysed.[519] Whilst the majority of techniques were extremely successful, detecting at least 20 of the 25 viruses, many yielded divergent results, highlighting the difficulty in comparing the results of studies employing different methodologies. It should however, be noted that the benefits of metagenomics were exemplified by this study, as multiple laboratories also reported the detection of additional viruses not known to have been included in the sample panel. One review of the informatics pipelines that are currently available for viral metagenomics showed that the difficulties of standardization are not limited to wet-lab protocols. The review concluded that although various tools have been made available for this purpose, their quality is highly variable and an insufficient number of benchmarking studies have been performed to test them.[520] Clearly more work is required in optimising and standardising the sequencing procedure, to meet the standards of robustness required by a clinical assay.

Contamination, such as that highlighted by this thesis, is one major issue that can have drastic consequences for diagnostics. For metagenomic sequencing to be used in the clinical setting, extreme care must be taken to avoid contamination at every stage, from sample collection onwards in order to avoid a potentially costly misdiagnosis. To minimize environmental contamination, facilities should be set-up specifically for the purpose of clinical-grade sequencing and cleaned thoroughly and regularly. Additionally, as with all diagnostic assays, careful handling of samples and a linear workflow should be used to reduce the risk of

contamination between samples. Perhaps most importantly, use of internal controls is vital in order to detect any possible contamination events and the use of confirmatory PCR assays may be warranted.

Aside from the need to validate the methodology itself, use of metagenomic sequencing for diagnosis of pathogens in the clinical setting must consider additional factors such as: in what instances it should be used; how to report only relevant findings; and what medical action should be taken based on these findings. Studies such as this thesis have shown that metagenomics can give important insights into subclinical viral infections and help to direct future diagnostics. However, the additional information generated is unnecessary for diagnosis in the majority of cases. As such, it is difficult to justify the use of sequencing when the same medically useful information can be generated by the current methods, which are generally quicker and less expensive. Therefore, if NGS is to become a useful clinical tool in infectious disease diagnostics, it will require guidelines regarding a clear clinical reason for performing the test, and empirically determined criteria regarding the medical actions to be taken based on the results.

For now, use of NGS in the clinical diagnosis of infectious disease is restricted to improving techniques which were previously reliant on Sanger sequencing, such as virus genotyping, where the output is improved but the clinical indication is unchanged. I expect future opportunities for the application of NGS to lie in areas where current assays are not sufficient. This primarily includes those that are required to target idiopathic diseases with multiple possible causes such as gastrointestinal disease, hepatitis, fever and respiratory disease. Overall, NGS has the potential to improve upon our current approach to clinical diagnosis of infectious disease, however many aspects must first be addressed and, with a

range of highly efficient assays already available, we should be careful to not attempt to progress merely for progresses sake.

6.6. Concluding remarks

Overall, this thesis has demonstrated the utility of next generation metagenomic sequencing for the detection of viruses in understudied, highly susceptible, populations of patients. The results of this work have taken an important first step towards increasing the consideration of viruses in the immunocompromised setting: generating promising early insights into the true abundance of viruses in the asymptomatic host as well as highlighting potentially important associations between the anelloviruses and disease.

The power of un-biased metagenomic sequencing has the potential to reveal much, including the answers to questions we have not yet thought to ask. Therefore, it is important to consider how best to utilise and interpret the vast amounts of information being generated. In this study alone, sequences from the host, bacteria and bacteriophages were left unanalysed, and potential novel viruses were not sought. No research group is capable of performing the work necessary to analyse all of these factors alone, and therefore making this data freely available to the research community is the best way to ensure that the full potential of next generation sequencing is reached.

References:

1. Gardner, T. J. & Tortorella, D. Virion Glycoprotein-Mediated Immune Evasion by Human Cytomegalovirus: a Sticky Virus Makes a Slick Getaway. *Microbiol. Mol. Biol. Rev.* **80,** 663–677 (2016).
2. Summers, J., O'Connell, A. & Millman, I. Genome of hepatitis B virus: restriction enzyme cleavage and structure of DNA extracted from Dane particles. *Proc. Natl. Acad. Sci. U. S. A.* **72,** 4597–601 (1975).
3. Nguyen, M. & Haenni, A.-L. Expression strategies of ambisense viruses. *Virus Res.* **93,** 141–150 (2003).
4. McDonald, S. M., Nelson, M. I., Turner, P. E. & Patton, J. T. Reassortment in segmented RNA viruses: mechanisms and outcomes. *Nat. Rev. Microbiol.* **14,** 448–460 (2016).
5. Forterre, P. & Prangishvili, D. The origin of viruses. *Res. Microbiol.* **160,** 466–472 (2009).
6. Holmes, E. C. What Does Virus Evolution Tell Us about Virus Origins? *J. Virol.* **85,** 5247–5251 (2011).
7. Baltimore, D. Expression of Animal Virus Genomes. **35,** 235–241 (1971).
8. Carter, J. & Saunders, V. *Virology: Principles and Applications*. (John Wiley & Sons, 2013).
9. Killingley, B. & Nguyen-Van-Tam, J. Routes of influenza transmission. *Influenza Other Respi. Viruses* **7,** 42–51 (2013).
10. Ho, M. S. *et al.* Viral gastroenteritis aboard a cruise ship. *Lancet (London, England)* **2,** 961–5 (1989).
11. de Graaf, M., van Beek, J. & Koopmans, M. P. G. Human norovirus transmission and evolution in a changing world. *Nat. Rev. Microbiol.* **14,** 421–433 (2016).
12. Conry-Cantilena, C. *et al.* Routes of Infection, Viremia, and Liver Disease in Blood Donors Found to Have Hepatitis C Virus Infection. *N. Engl. J. Med.*

334, 1691–1696 (1996).

13. Conway, M. J., Colpitts, T. M. & Fikrig, E. Role of the Vector in Arbovirus Transmission. *Annu. Rev. Virol.* **1,** 71–88 (2014).
14. Fenner F. Mouse-pox; infectious ectromelia of mice; a review. *J. Immunol.* **63,** 341–73 (1949).
15. Klingen, Y., Conzelmann, K.-K. & Finke, S. Double-Labeled Rabies Virus: Live Tracking of Enveloped Virus Transport. *J. Virol.* **82,** 237–245 (2008).
16. Kristensson, K., Lycke, E., Roytta, M., Svennerholm, B. AND Vahlne, A. Neuritic Transport of Herpes Simplex Virus in Rat Sensory Neurons in vitro. Effects of Substances Interacting with Microtubular Function and Axonal Flow [Nocodazole, Taxol and Erythro-9-3-(2-hydroxynonyl)adenine]. *J. Gen. Virol.* **67,** 2023–2028 (1986).
17. Lycke, E. *et al.* Herpes simplex virus infection of the human sensory neuron. *Arch. Virol.* **101,** 87–104 (1988).
18. Grove, J. & Marsh, M. The cell biology of receptor-mediated virus entry. *J. Cell Biol.* **195,** (2011).
19. Weis, W. *et al.* Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid. *Nature* **333,** 426–431 (1988).
20. Fields, B. N., Knipe, D. M. & Howley, P. M. *Fields Virology, 6th Edition*. *Fields Virology* (Lippincott Williams & Wilkins, 2013). doi:10.1093/cid/ciu346
21. Speck, S. H. & Ganem, D. Viral latency and its regulation: lessons from the gamma-herpesviruses. *Cell Host Microbe* **8,** 100–15 (2010).
22. Looker, K. J. *et al.* Global and Regional Estimates of Prevalent and Incident Herpes Simplex Virus Type 1 Infections in 2012. *PLoS One* **10,** e0140765 (2015).
23. Mark, K. E. *et al.* Rapidly Cleared Episodes of Herpes Simplex Virus Reactivation in Immunocompetent Adults. *J. Infect. Dis.* **198,** 1141–1149 (2008).
24. Traylen, C. M. *et al.* Virus reactivation: a panoramic view in human

infections. *Future Virol.* **6,** 451–463 (2011).

25. Nagel, M. A. & Gilden, D. Complications of Varicella Zoster Virus Reactivation. *Curr. Treat. Options Neurol.* **15,** 439–453 (2013).
26. Redeker, A. G. Viral hepatitis: clinical aspects. *Am. J. Med. Sci.* **270,** 9–16 (1975).
27. Shin, H. & Wherry, E. J. CD8 T cell dysfunction during chronic viral infection. *Curr. Opin. Immunol.* **19,** 408–415 (2007).
28. Moskophidis, D., Lechner, F., Pircher, H. & Zinkernagel, R. M. Virus persistence in acutely infected immunocompetent mice by exhaustion of antiviral cytotoxic effector T cells. *Nature* **362,** 758–761 (1993).
29. Savva, G. M. *et al.* Cytomegalovirus infection is associated with increased mortality in the older population. *Aging Cell* **12,** 381–387 (2013).
30. Karst, S. M. Pathogenesis of noroviruses, emerging RNA viruses. *Viruses* **2,** 748–81 (2010).
31. Sukhrie, F. H. A., Siebenga, J. J., Beersma, M. F. C. & Koopmans, M. Chronic shedders as reservoir for nosocomial transmission of norovirus. *J. Clin. Microbiol.* **48,** 4303–5 (2010).
32. Roulston, A., Marcellus, R. C. & Branton, P. E. Viruses and Apoptosis. *Annu. Rev. Microbiol.* **53,** 577–628 (1999).
33. Bird, S. W. & Kirkegaard, K. Escape of non-enveloped virus from intact cells. *Virology* **479–480,** 444–449 (2015).
34. Alwine, J. C. Modulation of host cell stress responses by human cytomegalovirus. *Curr. Top. Microbiol. Immunol.* **325,** 263–79 (2008).
35. Kennedy, P. G. E. Viral encephalitis: causes, differential diagnosis, and management. *J. Neurol. Neurosurg. Psychiatry* **75 Suppl 1,** i10-5 (2004).
36. Takada, A. & Kawaoka, Y. The pathogenesis of Ebola hemorrhagic fever. *Trends Microbiol.* **9,** 506–511 (2001).
37. Dietzschold, B., Li, J., Faber, M. & Schnell, M. Concepts in the pathogenesis of rabies. *Future Virol.* **3,** 481–490 (2008).

38. Rascovan, N., Duraisamy, R. & Desnues, C. Metagenomics and the Human Virome in Asymptomatic Individuals. *Annu. Rev. Microbiol.* **70,** 125–141 (2016).
39. Moustafa, A. *et al.* The blood DNA virome in 8,000 humans. *PLOS Pathog.* **13,** e1006292 (2017).
40. Hayward, A. C. *et al.* Comparative community burden and severity of seasonal and pandemic influenza: results of the Flu Watch cohort study. *Lancet Respir. Med.* **2,** 445–454 (2014).
41. WHO Ebola Response Team. Ebola Virus Disease in West Africa — The First 9 Months of the Epidemic and Forward Projections. *N. Engl. J. Med.* **371,** 1481–1495 (2014).
42. Zampieri, C. A., Sullivan, N. J. & Nabel, G. J. Immunopathology of highly virulent pathogens: insights from Ebola virus. *Nat. Immunol.* **8,** 1159–1164 (2007).
43. Lin, H.-H., Kao, J.-H., Lee, P.-I. & Chen, D.-S. Early acquisition of TT virus in infants: possible minor role of maternal transmission. *J. Med. Virol.* **66,** 285–90 (2002).
44. Davidson, F. *et al.* Early Acquisition of TT Virus (TTV) in an Area Endemic for TTV Infection. *J. Infect. Dis.* **179,** 1070–1076 (1999).
45. Huang, L.-Y., Éystein Jonassen, T., Hungnes, O. & Grinde, B. High Prevalence of TT Virus-Related DNA (90%) and Diverse Viral Genotypes in Norwegian Blood Donors. *J. Med. Virol. J. Med. Virol* **64,** (2001).
46. AbuOdeh, R. *et al.* Detection and genotyping of torque teno virus (TTV) in healthy blood donors and patients infected with HBV or HCV in Qatar. *J. Med. Virol.* **87,** 1184–1191 (2015).
47. Vasilyev, E. V *et al.* Torque Teno Virus (TTV) distribution in healthy Russian population. *Virol. J.* **6,** 134 (2009).
48. Simmonds, P. *et al.* TT virus--part of the normal human flora? *J. Infect. Dis.* **180,** 1748–1750 (1999).

49. Huang, Y. *et al.* Temporal Dynamics of Host Molecular Responses Differentiate Symptomatic and Asymptomatic Influenza A Infection. *PLoS Genet.* **7,** e1002234 (2011).
50. Andersen, K. G. *et al.* Genome-wide scans provide evidence for positive selection of genes implicated in Lassa fever. *Philos. Trans. R. Soc. B Biol. Sci.* **367,** 868–877 (2012).
51. Wilkins, C. & Gale, M. Recognition of viruses by cytoplasmic sensors. *Curr. Opin. Immunol.* **22,** 41–7 (2010).
52. Pichlmair, A. *et al.* Activation of MDA5 requires higher-order RNA structures generated during virus infection. *J. Virol.* **83,** 10761–9 (2009).
53. Kato, H. *et al.* Length-dependent recognition of double-stranded ribonucleic acids by retinoic acid–inducible gene-I and melanoma differentiation–associated gene 5. *J. Exp. Med.* **205,** 1601–1610 (2008).
54. Kell, A. M. & Gale, M. RIG-I in RNA virus recognition. *Virology* **479–480,** 110–121 (2015).
55. Bauer, S. *et al.* Human TLR9 confers responsiveness to bacterial DNA via species-specific CpG motif recognition. *Proc. Natl. Acad. Sci.* **98,** 9237–9242 (2001).
56. Beachboard, D. C. & Horner, S. M. Innate immune evasion strategies of DNA and RNA viruses. *Curr. Opin. Microbiol.* **32,** 113–119 (2016).
57. Stoermer, K. A. & Morrison, T. E. Complement and viral pathogenesis. *Virology* **411,** 362–73 (2011).
58. Lachmann, P. J. & Davies, A. Complement and immunity to viruses. *Immunol. Rev.* **159,** 69–77 (1997).
59. Janeway, C. A. J., Travers, P., Walport, M. & Shlomchik, M. J. The complement system and innate immunity. in *Immunobiology: The Immune System in Health and Disease* (Garland Science, 2001).
60. McCullough, K. C., Parkinson, D. & Crowther, J. R. Opsonization-enhanced phagocytosis of foot-and-mouth disease virus. *Immunology* **65,** 187–91

(1988).

61. Ying, H. *et al.* Interaction of Mannose-Binding Lectin with HIV Type 1 Is Sufficient for Virus Opsonization But Not Neutralization. *AIDS Res. Hum. Retroviruses* **20,** 327–335 (2004).
62. Peters, M. Actions of Cytokines on the Immune Response and Viral Interactions: An Overview. *Hepatology* **23,** 909–916 (1995).
63. Isaacs, A. and Lindenmann, J. Virus interference. I. The interferon. *Proc. R. Soc. London. Ser. B, Biol. Sci.* **147,** 258–67 (1957).
64. Meraz, M. A. *et al.* Targeted Disruption of the Stat1 Gene in Mice Reveals Unexpected Physiologic Specificity in the JAK–STAT Signaling Pathway. *Cell* **84,** 431–442 (1996).
65. Slattery, E., Ghosh, N., Samanta, H. & Lengyel, P. Interferon, double-stranded RNA, and RNA degradation: activation of an endonuclease by (2'-5')An. *Proc. Natl. Acad. Sci. U. S. A.* **76,** 4778–82 (1979).
66. García, M. A., Meurs, E. F. & Esteban, M. The dsRNA protein kinase PKR: Virus and cell control. *Biochimie* **89,** 799–811 (2007).
67. Haller, O., Staeheli, P., Schwemmle, M. & Kochs, G. Mx GTPases: dynamin-like antiviral machines of innate immunity. *Trends Microbiol.* **23,** 154–163 (2015).
68. Patil, S. *et al.* Single-cell analysis shows that paracrine signaling by first responder cells shapes the interferon-β response to viral infection. *Sci. Signal.* **8,** (2015).
69. Goubau, D., Deddouche, S. & Reis e Sousa, C. Cytosolic Sensing of Viruses. *Immunity* **38,** 855–869 (2013).
70. Hewitt, E. W. The MHC class I antigen presentation pathway: strategies for viral immune evasion. *Immunology* **110,** 163–9 (2003).
71. Warren, H. S. & Smyth, M. J. NK cells and apoptosis. *Immunol. Cell Biol.* **77,** 64–75 (1999).
72. Zinkernagel, R. M. & Doherty, P. C. MHC-Restricted Cytotoxic T Cells:

Studies on the Biological Role of Polymorphic Major Transplantation Antigens Determining T-Cell Restriction-Specificity, Function, and Responsiveness. in 51–177 (1979). doi:10.1016/S0065-2776(08)60262-X

73. Rager-Zisman, B., Quan, P. C., Rosner, M., Moller, J. R. & Bloom, B. R. Role of NK cells in protection of mice against herpes simplex virus-1 infection. *J. Immunol.* **138,** (1987).
74. Colonna, M. *et al.* Plasmacytoid monocytes migrate to inflamed lymph nodes and produce largeamounts of type I interferon. *Nat. Med.* **5,** 919–923 (1999).
75. Geissmann, F. *et al.* Development of Monocytes, Macrophages, and Dendritic Cells. *Science (80-.).* **327,** (2010).
76. Fujimoto, I., Pan, J., Takizawa, T. & Nakanishi, Y. Virus clearance through apoptosis-dependent phagocytosis of influenza A virus-infected cells by macrophages. *J. Virol.* **74,** 3399–403 (2000).
77. Schneider, C. *et al.* Alveolar Macrophages Are Essential for Protection from Respiratory Failure and Associated Morbidity following Influenza Virus Infection. *PLoS Pathog.* **10,** e1004053 (2014).
78. Ni, K. & O'Neill, H. The role of dendritic cells in T cell activation. *Immunol. Cell Biol.* **75,** 223–230 (1997).
79. Janeway, C. J., Travers, P. & Walport, M. *Immunobiology: The Immune System in Health and Disease. 5th edition*. *Antiviral Research* (Garland Science, 2001). doi:10.1016/j.antiviral.2014.10.001
80. Clark, R. & Kupper, T. Old Meets New: The Interaction Between Innate and Adaptive Immunity. *J. Invest. Dermatol.* **125,** 629–637 (2005).
81. Kelsoe, G. The germinal center: a crucible for lymphocyte selection. *Semin. Immunol.* **8,** 179–184 (1996).
82. Jin, J. *et al.* Neutralizing Monoclonal Antibodies Block Chikungunya Virus Entry and Release by Targeting an Epitope Critical to Viral Pathogenesis. *Cell Rep.* **13,** 2553–2564 (2015).

83. Xu, Z., Zan, H., Pone, E. J., Mai, T. & Casali, P. Immunoglobulin class-switch DNA recombination: induction, targeting and beyond. *Nat. Rev. Immunol.* **12,** 517–531 (2012).
84. Boes, M. Role of natural and immune IgM antibodies in immune responses. *Mol. Immunol.* **37,** 1141–9 (2000).
85. Pieper, K., Grimbacher, B. & Eibel, H. B-cell biology and development. *J. Allergy Clin. Immunol.* **131,** 959–971 (2013).
86. Mäkelä, O., Rouslahti, E. & Seppälä, I. J. T. Affinity of IgM and IgG antibodies. *Immunochemistry* **7,** 917–932 (1970).
87. Bandilla, K. K., McDuffie, F. C. & Gleich, G. J. Immunoglobulin classes of antibodies produced in the primary and secondary responses in man. *Clin. Exp. Immunol.* **5,** 627–41 (1969).
88. Choudhuri, K., Kearney, A., Bakker, T. R. & van der Merwe, P. A. Immunology: How Do T Cells Recognize Antigen? *Curr. Biol.* **15,** R382–R385 (2005).
89. Russell, J. H. & Ley, T. J. Lymphocyte-Mediated Cytotoxicity. *Annu. Rev. Immunol.* **20,** 323–370 (2002).
90. Thomson, B. J. Viruses and apoptosis. *Int. J. Exp. Pathol.* **82,** 65–76 (2001).
91. Naif, H. M. Pathogenesis of HIV Infection. *Infect. Dis. Rep.* **5,** e6 (2013).
92. Sepkowitz, K. A. AIDS — The First 20 Years. *N. Engl. J. Med.* **344,** 1764–1772 (2001).
93. Marshall, N. B. & Swain, S. L. Cytotoxic CD4 T Cells in Antiviral Immunity. *J. Biomed. Biotechnol.* **2011,** 1–8 (2011).
94. Yel, L. Selective IgA Deficiency. *J. Clin. Immunol.* **30,** 10–16 (2010).
95. Mestecky, J. & Russell, M. W. Mucosal immunoglobulins and their contribution to defence mechanisms: an overview. *Biochem. Soc. Trans.* **25,** 457–62 (1997).
96. Brandtzaeg, P. *et al.* The clinical condition of IgA-deficient patients is related to the proportion of IgD- and IgM-producing cells in their nasal

mucosa. *Clin. Exp. Immunol.* **67,** 626–36 (1987).

97. Hayday, A., Theodoridis, E., Ramsburg, E. & Shires, J. Intraepithelial lymphocytes: exploring the Third Way in immunology. *Nat. Immunol.* **2,** 997–1003 (2001).
98. Hayday, A. & Gibbons, D. Brokering the peace: the origin of intestinal T cells. *Mucosal Immunol.* **1,** 172–174 (2008).
99. Hayday, A. C. γδ T Cells and the Lymphoid Stress-Surveillance Response. *Immunity* **31,** 184–196 (2009).
100. Kotenko, S. V *et al.* IFN-lambdas mediate antiviral protection through a distinct class II cytokine receptor complex. *Nat. Immunol.* **4,** 69–77 (2003).
101. Swamy, M. *et al.* Intestinal intraepithelial lymphocyte activation promotes innate antiviral resistance. *Nat. Commun.* **6,** 7090 (2015).
102. Hermant, P. & Michiels, T. Interferon-λ in the context of viral infections: production, response and therapeutic implications. *J. Innate Immun.* **6,** 563–74 (2014).
103. Palma-Ocampo, H. K. *et al.* Interferon lambda inhibits dengue virus replication in epithelial cells. *Virol. J.* **12,** 150 (2015).
104. McClellan, J. & King, M.-C. Genetic Heterogeneity in Human Disease. *Cell* **141,** 210–217 (2010).
105. Caillat-Zucman, S. *et al.* Distinct HLA class II alleles determine antibody response to vaccination with hepatitis B surface antigen. *Kidney Int.* **53,** 1626–1630 (1998).
106. McKiernan, S. M. *et al.* Distinct MHC class I and II alleles are associated with hepatitis C viral clearance, originating from a single source. *Hepatology* **40,** 108–114 (2004).
107. Spurgin, L. G. & Richardson, D. S. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc. R. Soc. London B Biol. Sci.* **277,** (2010).
108. Forterre, P. & Prangishvili, D. The Great Billion-year War between

Ribosome- and Capsid-encoding Organisms (Cells and Viruses) as the Major Source of Evolutionary Novelties. *Ann. N. Y. Acad. Sci.* **1178,** 65–77 (2009).

109. Vogel, T. U. *et al.* Major histocompatibility complex class I genes in primates: co-evolution with pathogens. *Immunol. Rev.* **167,** 327–37 (1999).
110. Van Blerkom, L. M. Role of viruses in human evolution. *Am. J. Phys. Anthropol.* **122,** 14–46 (2003).
111. Hoffmann, H.-H., Schneider, W. M. & Rice, C. M. Interferons and viruses: an evolutionary arms race of molecular interactions. *Trends Immunol.* **36,** 124–138 (2015).
112. Johnson, W. E. Rapid Adversarial Co-Evolution of Viruses and Cellular Restriction Factors. *Curr. Top. Microbiol. Immunol.* **371,** 123–151 (2013).
113. Ploegh, H. L. Viral Strategies of Immune Evasion. *Science (80-.).* **280,** (1998).
114. Novella, I. S., Presloid, J. B. & Taylor, R. T. RNA replication errors and the evolution of virus pathogenicity and virulence. *Curr. Opin. Virol.* **9,** 143–147 (2014).
115. Domingo, E. Quasispecies and the implications for virus persistence and escape. *Clin. Diagn. Virol.* **10,** 97–101 (1998).
116. Worobey, M. & Holmes, E. C. Evolutionary aspects of recombination in RNA viruses. *J. Gen. Virol.* **80,** 2535–2543 (1999).
117. Pancera, M. *et al.* Structure and immune recognition of trimeric pre-fusion HIV-1 Env. *Nature* **514,** 455–461 (2014).
118. Chan, Y. K. & Gack, M. U. Viral evasion of intracellular DNA and RNA sensing. *Nat. Rev. Microbiol.* **14,** 360–373 (2016).
119. You, Y. *et al.* The suppression of apoptosis by α-herpesvirus. *Cell Death Dis.* **8,** e2749 (2017).
120. Amara, A. & Mercer, J. Viral apoptotic mimicry. *Nat. Rev. Microbiol.* **13,** 461–469 (2015).

121. Stapleton, J. T. *et al.* A novel T cell evasion mechanism in persistent RNA virus infection. *Trans. Am. Clin. Climatol. Assoc.* **125,** 14–16 (2014).
122. Christiaansen, A., Varga, S. M. & Spencer, J. V. Viral manipulation of the host immune response. *Curr. Opin. Immunol.* **36,** 54–60 (2015).
123. Alcami, A. & Lira, S. A. Modulation of chemokine activity by viruses. *Curr. Opin. Immunol.* **22,** 482–487 (2010).
124. Slobedman, B., Barry, P. A., Spencer, J. V, Avdic, S. & Abendroth, A. Virus-encoded homologs of cellular interleukin-10 and their control of host immune function. *J. Virol.* **83,** 9618–29 (2009).
125. Ouyang, P. *et al.* IL-10 encoded by viruses: a remarkable example of independent acquisition of a cellular gene by viruses and its subsequent evolution in the viral genome. *J. Gen. Virol.* **95,** 245–262 (2014).
126. Brooks, D. G. *et al.* Interleukin-10 determines viral clearance or persistence in vivo. *Nat. Med.* **12,** 1301–1309 (2006).
127. Menzin, J., Sussman, M., Munsell, M. & Zbrozek, A. Economic impact of infections among patients with primary immunodeficiency disease receiving IVIG therapy. *Clinicoecon. Outcomes Res.* **6,** 297–302 (2014).
128. Sadeghi, B., Abolhassani, H., Naseri, A., Rezaei, N. & Aghamohammadi, A. Economic burden of common variable immunodeficiency: annual cost of disease. *Expert Rev. Clin. Immunol.* **11,** 681–688 (2015).
129. Unterman, S. *et al.* A descriptive analysis of 1251 solid organ transplant visits to the emergency department. *West. J. Emerg. Med.* **10,** 48–54 (2009).
130. Turtay, M. G. *et al.* A descriptive analysis of 188 liver transplant patient visits to an emergency department. *Eur. Rev. Med. Pharmacol. Sci.* **16 Suppl 1,** 3–7 (2012).
131. Schold, J. D. *et al.* Emergency Department Visits after Kidney Transplantation. *Clin. J. Am. Soc. Nephrol.* **11,** 674–83 (2016).
132. Maródi, L. & Notarangelo, L. D. Immunological and genetic bases of new primary immunodeficiencies. *Nat. Rev. Immunol.* **7,** 851–861 (2007).

133. Mousallem, T. *et al.* Clinical application of whole-genome sequencing in patients with primary immunodeficiency. *J. Allergy Clin. Immunol.* **136,** 476–479.e6 (2015).
134. Kelsen, J. R. *et al.* Exome Sequencing Analysis Reveals Variants in Primary Immunodeficiency Genes in Patients With Very Early Onset Inflammatory Bowel Disease. *Gastroenterology* **149,** 1415–1424 (2015).
135. Orange, J. S. *et al.* Genome-wide association identifies diverse causes of common variable immunodeficiency. *J. Allergy Clin. Immunol.* **127,** 1360–7.e6 (2011).
136. Tam, J. S. & Routes, J. M. Common variable immunodeficiency. *Am. J. Rhinol. Allergy* **27,** 260–5 (2013).
137. Warnatz, K. *et al.* B-cell activating factor receptor deficiency is associated with an adult-onset antibody deficiency syndrome in humans. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 13945–50 (2009).
138. van Zelm, M. C. *et al.* An Antibody-Deficiency Syndrome Due to Mutations in the *CD19* Gene. *N. Engl. J. Med.* **354,** 1901–1912 (2006).
139. Thiel, J. *et al.* Genetic CD21 deficiency is associated with hypogammaglobulinemia. *J. Allergy Clin. Immunol.* **129,** 801–810.e6 (2012).
140. van Zelm, M. C. *et al.* CD81 gene defect in humans disrupts CD19 complex formation and leads to antibody deficiency. *J. Clin. Invest.* **120,** 1265–74 (2010).
141. Grimbacher, B. *et al.* Homozygous loss of ICOS is associated with adult-onset common variable immunodeficiency. *Nat. Immunol.* **4,** 261–268 (2003).
142. Louis, A. G., Yel, L., Cao, J. N., Agrawal, S. & Gupta, S. Common variable immunodeficiency associated with microdeletion of chromosome 1q42.1-q42.3 and inositol 1,4,5-trisphosphate kinase B (ITPKB) deficiency. *Clin. Transl. Immunol.* **5,** e59 (2016).

143. Lopez-Herrera, G. *et al.* Deleterious mutations in LRBA are associated with a syndrome of immune deficiency and autoimmunity. *Am. J. Hum. Genet.* **90,** 986–1001 (2012).
144. Salzer, U. *et al.* Mutations in TNFRSF13B encoding TACI are associated with common variable immunodeficiency in humans. *Nat. Genet.* **37,** 820–828 (2005).
145. Oettinger, M. A., Schatz, D. G., Gorka, C. & Baltimore, D. RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. *Science* **248,** 1517–23 (1990).
146. Boyle, J. M. & Buckley, R. H. Population Prevalence of Diagnosed Primary Immunodeficiency Diseases in the United States. *J. Clin. Immunol.* **27,** 497–502 (2007).
147. Cunningham-Rundles, C., Sidi, P., Estrella, L. & Doucette, J. Identifying undiagnosed primary immunodeficiency diseases in minority subjects by using computer sorting of diagnosis codes. *J. Allergy Clin. Immunol.* **113,** 747–755 (2004).
148. Roquilly, A. *et al.* Local Modulation of Antigen-Presenting Cell Development after Resolution of Pneumonia Induces Long-Term Susceptibility to Secondary Infections. *Immunity* **47,** 135–147.e5 (2017).
149. Hughes, S. & Kelly, P. Interactions of malnutrition and immune impairment, with specific reference to immunity against parasites. *Parasite Immunol.* **28,** 577–88 (2006).
150. Powers, D. C. & Belshe, R. B. Effect of age on cytotoxic T lymphocyte memory as well as serum and local antibody responses elicited by inactivated influenza virus vaccine. *J. Infect. Dis.* **167,** 584–92 (1993).
151. Halloran, P. F. Immunosuppressive Drugs for Kidney Transplantation. *N. Engl. J. Med.* **351,** 2715–2729 (2004).
152. Fox, D. A. & McCune, W. J. Immunosuppressive drug therapy of systemic lupus erythematosus. *Rheum. Dis. Clin. North Am.* **20,** 265–99 (1994).

153. NHS Blood and Transplant. *Organ Donation and Transplantation - Activity Figures for the UK as at 7 April 2017*. (2017).
154. Cooper, G. S., Bynum, M. L. K. & Somers, E. C. Recent insights in the epidemiology of autoimmune diseases: Improved prevalence estimates and understanding of clustering of diseases. *J. Autoimmun.* **33,** 197–207 (2009).
155. Lerner, A., Jeremias, P. & Matthias, T. The World Incidence and Prevalence of Autoimmune Diseases is Increasing. *Int. J. Celiac Dis.* **3,** 151–155 (2015).
156. Montoya, J. *et al.* Infectious Complications among 620 Consecutive Heart Transplant Patients at Stanford University Medical Center. *Clin. Infect. Dis.* **33,** 629–40 (2001).
157. Kang, I. & Park, S. H. Infectious complications in SLE after immunosuppressive therapies. *Curr. Opin. Rheumatol.* **15,** 528–34 (2003).
158. The United Nations, T. D. of E. and S. A. *World Population Ageing 2015*. (2015).
159. Office for National Statistics. Ageing. (2015). Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/ageing. (Accessed: 17th May 2017)
160. Public Health England. *Seasonal influenza vaccine uptake amongst GP Patients in England 2016/2017*. (2017).
161. Weiskopf, D., Weinberger, B. & Grubeck-Loebenstein, B. The aging of the immune system. *Transpl. Int.* **22,** 1041–1050 (2009).
162. Bueno, V., Sant'Anna, O. A. & Lord, J. M. Ageing and myeloid-derived suppressor cells: possible involvement in immunosenescence and age-related disease. *Age (Dordr).* **36,** 9729 (2014).
163. Dewan, S. K., Zheng, S., Xia, S. & Bill, K. Senescent remodeling of the immune system and its contribution to the predisposition of the elderly to infections. *Chin. Med. J. (Engl).* **125,** 3325–3331 (2012).
164. Ginaldi, L. *et al.* The immune system in the elderly: I. Specific humoral

immunity. *Immunol. Res.* **20,** 101–108 (1999).

165. Ginaldi, L. *et al.* The immune system in the elderly: II. Specific cellular immunity. *Immunol. Res.* **20,** 109–115 (1999).
166. Ginaldi, L. *et al.* The immune system in the elderly: III. Innate immunity. *Immunol. Res.* **20,** 117–126 (1999).
167. Ma, Y. & Fang, M. Immunosenescence and age-related viral diseases. *Sci. China Life Sci.* **56,** 399–405 (2013).
168. Ferron, G. M. & Jusko, W. J. Species- and Gender-Related Differences in Cyclosporine/Prednisolone/Sirolimus Interactions in Whole Blood Lymphocyte Proliferation Assays. *J. Pharmacol. Exp. Ther.* **286,** (1998).
169. Hannam-Harris, A. C., Taylor, D. S. & Nowell, P. C. Cyclosporin A directly inhibits human B-cell proliferation by more than a single mechanism. *J. Leukoc. Biol.* **38,** 231–9 (1985).
170. Heidt, S. *et al.* Calcineurin inhibitors affect B cell antibody responses indirectly by interfering with T cell help. *Clin. Exp. Immunol.* **159,** 199–207 (2010).
171. Tareyeva, I. E., Shilov, E. M. & Gordovskaya, N. B. The effects of azathioprine and prednisolone on T- and B-lymphocytes in patients with lupus nephritis and chronic glomerulonephritis. *Clin. Nephrol.* **14,** 233–7 (1980).
172. Eickenberg, S. *et al.* Mycophenolic acid counteracts B cell proliferation and plasmablast formation in patients with systemic lupus erythematosus. *Arthritis Res. Ther.* **14,** R110 (2012).
173. Meehan, A. C. *et al.* Impact of commonly used transplant immunosuppressive drugs on human NK cell function is dependent upon stimulation condition. *PLoS One* **8,** e60144 (2013).
174. Morteau, O. *et al.* Renal Transplant Immunosuppression Impairs Natural Killer Cell Function In Vitro and In Vivo. *PLoS One* **5,** e13294 (2010).
175. Cseuz, R. & Panayi, G. S. The inhibition of NK cell function by azathioprine

during the treatment of patients with rheumatoid arthritis. *Br. J. Rheumatol.* **29,** 358–62 (1990).

176. Systemic Anti-Cancer Therapy. *Chemotherapy Dataset Completeness Report 2016/17*. (2017).
177. Hazenberg, M. D., Hamann, D., Schuitemaker, H. & Miedema, F. T cell depletion in HIV-1 infection: how CD4+ T cells go out of stock. *Nat. Immunol.* **1,** 285–289 (2000).
178. NHS Choices. HIV and AIDS. (2014). Available at: http://www.nhs.uk/conditions/hiv/pages/introduction.aspx. (Accessed: 17th May 2017)
179. Cotton, M. F. & Rabie, H. Impact of earlier combination antiretroviral therapy on outcomes in children. *Curr. Opin. HIV AIDS* **10,** 12–17 (2015).
180. Phan, T. X., Jaruga, B., Pingle, S. C., Bandyopadhyay, B. C. & Ahern, G. P. Intrinsic Photosensitivity Enhances Motility of T Lymphocytes. *Sci. Rep.* **6,** 39479 (2016).
181. Edgar, R. S. *et al.* Cell autonomous regulation of herpes and influenza virus infection by the circadian clock. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 10085–90 (2016).
182. Úbeda, F., Jansen, V. A. A., Roberts, M. G., Blattner, W. & Day, T. The evolution of sex-specific virulence in infectious diseases. *Nat. Commun.* **7,** 13849 (2016).
183. Dropulic, L. K. & Cohen, J. I. Severe viral infections and primary immunodeficiencies. *Clin. Infect. Dis.* **53,** 897–909 (2011).
184. Zhang, S.-Y. *et al.* TLR3 Deficiency in Patients with Herpes Simplex Encephalitis. *Science (80-.).* **317,** 1522–1527 (2007).
185. Guo, Y. *et al.* Herpes simplex virus encephalitis in a patient with complete TLR3 deficiency: TLR3 is otherwise redundant in protective immunity. *J. Exp. Med.* **208,** (2011).
186. Conley, M. E., Mathias, D., Treadaway, J., Minegishi, Y. & Rohrer, J.

Mutations in btk in patients with presumed X-linked agammaglobulinemia. *Am. J. Hum. Genet.* **62,** 1034–43 (1998).

187. Winkelstein, J. A. *et al.* X-Linked Agammaglobulinemia. *Medicine (Baltimore).* **85,** 193–202 (2006).
188. Faulkner, G. C. *et al.* X-Linked agammaglobulinemia patients are not infected with Epstein-Barr virus: implications for the biology of the virus. *J. Virol.* **73,** 1555–64 (1999).
189. Fischer, A. Severe combined immunodeficiencies (SCID). *Clin. Exp. Immunol.* **122,** 143–9 (2000).
190. Taylor, C. E. *et al.* Parainfluenza virus and respiratory syncytial virus infection in infants undergoing bone marrow transplantation for severe combined immunodeficiency. *Commun. Dis. public Heal.* **1,** 202–3 (1998).
191. Crooks, B. N. A. *et al.* Respiratory viral infections in primary immune deficiencies: significance and relevance to clinical outcome in a single BMT unit. *Bone Marrow Transplant.* **26,** 1097–1102 (2000).
192. Monforte-Muñoz, H., Kapoor, N. & Saavedra, J. A. Epstein-Barr virus-associated leiomyomatosis and posttransplant lymphoproliferative disorder in a child with severe combined immunodeficiency: case report and review of the literature. *Pediatr. Dev. Pathol.* **6,** 449–57
193. Patel, N. C. *et al.* Vaccine-Acquired Rotavirus in Infants with Severe Combined Immunodeficiency. *N. Engl. J. Med.* **362,** 314–319 (2010).
194. Wunderli, W. *et al.* Astrovirus Infection in Hospitalized Infants with Severe Combined Immunodeficiency after Allogeneic Hematopoietic Stem Cell Transplantation. *PLoS One* **6,** e27483 (2011).
195. Waggoner, J. J., Soda, E. A. & Deresinski, S. Rare and Emerging Viral Infections in Transplant Recipients. *Clin. Infect. Dis.* **57,** 1182–1188 (2013).
196. Waghmare, A., Englund, J. A. & Boeckh, M. How I treat respiratory viral infections in the setting of intensive chemotherapy or hematopoietic cell transplantation. *Blood* **127,** (2016).

197. Talbot, H. K. & Falsey, A. R. The Diagnosis of Viral Respiratory Disease in Older Adults. *Clin. Infect. Dis.* **73,** 100201102709029–0 (2010).
198. Ellis, S. E., Coffey, C. S., Mitchel, E. F., Dittus, R. S. & Griffin, M. R. Influenza- and respiratory syncytial virus-associated morbidity and mortality in the nursing home population. *J. Am. Geriatr. Soc.* **51,** 761–7 (2003).
199. Mertz, D. *et al.* Populations at risk for severe or complicated influenza illness: systematic review and meta-analysis. *BMJ* **347,** (2013).
200. Koenig, S. *et al.* Detection of AIDS Virus in Macrophages in Brain Tissue from AIDS Patients with Encephalopathy. *Science* **233,** 1089–1093
201. Casper, C. *et al.* Frequent and Asymptomatic Oropharyngeal Shedding of Human Herpesvirus 8 among Immunocompetent Men. *J. Infect. Dis.* **195,** 30–36 (2007).
202. Jenkins, F. J., Rowe, D. T., Rinaldo, C. R. & Jr. Herpesvirus infections in organ transplant recipients. *Clin. Diagn. Lab. Immunol.* **10,** 1–7 (2003).
203. Stowe, R. *et al.* Chronic herpesvirus reactivation occurs in aging. *Exp. Gerontol.* **42,** 563–570 (2007).
204. Lok, A. S. F. *et al.* Reactivation of hepatitis B virus replication in patients receiving cytotoxic therapy. *Gastroenterology* **100,** 182–188 (1991).
205. Levitsky, J., Doucette, K. & Levitsky, J. Viral Hepatitis in Solid Organ Transplantation the AST Infectious Diseases Community of Practice. *Am. J. Transplant.* **13,** 147–168 (2013).
206. Loustaud-Ratti, V., Jacques, J., Debette-Gratien, M. & Carrier, P. Hepatitis B and elders: An underestimated issue. *Hepatol. Res.* **46,** 22–28 (2016).
207. Hayes, C. N. & Chayama, K. HBV culture and infectious systems. *Hepatol. Int.* **10,** 559–566 (2016).
208. Lohmann, V. & Bartenschlager, R. On the History of Hepatitis C Virus Cell Culture Systems. *J. Med. Chem.* **57,** 1627–1642 (2014).
209. Block, T. M., Alter, H. J., London, W. T. & Bray, M. A historical perspective on the discovery and elucidation of the hepatitis B virus. *Antiviral Res.* **131,**

109–123 (2016).

210. Engvall, E. & Perlmann, P. Enzyme-linked immunosorbent assay (ELISA) quantitative assay of immunoglobulin G. *Immunochemistry* **8,** 871–874 (1971).
211. Skinner, G. R. Transformation of primary hamster embryo fibroblasts by type 2 simplex virus: evidence for a "hit and run" mechanism. *Br. J. Exp. Pathol.* **57,** 361–76 (1976).
212. Saiki, R. *et al.* Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science (80-.).* **239,** (1988).
213. Gerard, G. F., Fox, D. K., Nathan, M. & D'Alessio, J. M. Reverse transcriptase. The use of cloned Moloney murine leukemia virus reverse transcriptase to synthesize DNA from RNA. *Mol. Biotechnol.* **8,** 61–77 (1997).
214. Wang, S., Xu, F. & Demirci, U. Advances in developing HIV-1 viral load assays for resource-limited settings. *Biotechnol. Adv.* **28,** 770–81 (2010).
215. Zipper, H., Brunner, H., Bernhagen, J. & Vitzthum, F. Investigations on DNA intercalation and surface binding by SYBR Green I, its structure determination and methodological implications. *Nucleic Acids Res.* **32,** e103 (2004).
216. Holland, P. M., Abramson, R. D., Watson, R. & Gelfand, D. H. Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of Thermus aquaticus DNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 7276–80 (1991).
217. Ririe, K. M., Rasmussen, R. P. & Wittwer, C. T. Product Differentiation by Analysis of DNA Melting Curves during the Polymerase Chain Reaction. *Anal. Biochem.* **245,** 154–160 (1997).
218. Sedlak, R. H. & Jerome, K. R. Viral diagnostics in the era of digital polymerase chain reaction. *Diagn. Microbiol. Infect. Dis.* **75,** 1–4 (2013).
219. Waggoner, J., Ho, D. Y., Libiran, P. & Pinsky, B. A. Clinical Significance of Low Cytomegalovirus DNA Levels in Human Plasma. *J. Clin. Microbiol.* **50,**

2378–2383 (2012).

220. Park, Y. *et al.* A Novel Multiplex Real-Time PCR Assay for the Concurrent Detection of Hepatitis A, B and C Viruses in Patients with Acute Hepatitis. *PLoS One* **7,** e49106 (2012).
221. Wang, W. *et al.* design of multiplexed detection assays for identification of avian influenza a virus subtypes pathogenic to humans by SmartCycler real-time reverse transcription-PCR. *J. Clin. Microbiol.* **47,** 86–92 (2009).
222. Dieffenbach, C. W., Lowe, T. M. J. & Dveksler, G. S. General concepts for PCR primer design. *Genome Res.* **3,** 30–37 (1993).
223. Chen, E. C., Miller, S. A., DeRisi, J. L. & Chiu, C. Y. Using a pan-viral microarray assay (Virochip) to screen clinical samples for viral pathogens. *J. Vis. Exp.* (2011). doi:10.3791/2536
224. Abruzzo, L. V *et al.* Validation of oligonucleotide microarray data using microfluidic low-density arrays: a new statistical method to normalize real-time RT-PCR data. *Biotechniques* **38,** 785–92 (2005).
225. Snijders, A. M., Meijer, G. A., Brakenhoff, R. H., van den Brule, A. J. & van Diest, P. J. Microarray techniques in pathology: tool or toy? *Mol. Pathol.* **53,** 289–94 (2000).
226. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
227. Datta, S. *et al.* Next-generation sequencing in clinical virology: Discovery of new viruses. *World J. Virol.* **4,** 265–76 (2015).
228. Ganova-Raeva, L. *et al.* Cryptic Hepatitis B and E in Patients With Acute Hepatitis of Unknown Etiology. *J. Infect. Dis.* **212,** 1962–1969 (2015).
229. Sardi, S. *et al.* Co-Infections from Zika and Chikungunya Virus in Bahia, Brazil Identified by Metagenomic Next-Generation Sequencing. *J. Clin. Microbiol.* **54,** 2348-2353 (2016).
230. Quick, J. *et al.* Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.*

12, 1261–1276 (2017).

231. Lisitsyn, N., Lisitsyn, N. & Wigler, M. Cloning the differences between two complex genomes. *Science* **259,** 946–51 (1993).
232. Memar, O. M., Rady, P. L. & Tyring, S. K. Human herpesvirus-8: detection of novel herpesvirus-like DNA sequences in Kaposi's sarcoma and other lesions. *J. Mol. Med.* **73,** 603–609 (1995).
233. van der Hoek, L. *et al.* Identification of a new human coronavirus. *Nat. Med.* **10,** 368–373 (2004).
234. van Leeuwen, M. *et al.* Human Picobirnaviruses Identified by Molecular Screening of Diarrhea Samples. *J. Clin. Microbiol.* **48,** 1787–1794 (2010).
235. Daly, G. M. *et al.* A viral discovery methodology for clinical biopsy samples utilising massively parallel next generation sequencing. *PLoS One* **6,** e28879 (2011).
236. He, S. *et al.* Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat. Methods* **7,** 807–812 (2010).
237. Rosseel, T., Ozhelvaci, O., Freimanis, G. & Van Borm, S. Evaluation of convenient pretreatment protocols for RNA virus metagenomics in serum and tissue samples. *J. Virol. Methods* **222,** 72–80 (2015).
238. Hall, R. J. *et al.* Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *J. Virol. Methods* **195,** 194–204 (2014).
239. Li, L. *et al.* Comparing viral metagenomics methods using a highly multiplexed human viral pathogens reagent. *J. Virol. Methods* **213,** (2015).
240. Kohl, C. *et al.* Protocol for metagenomic virus detection in clinical specimens. *Emerg. Infect. Dis.* **21,** 48–57 (2015).
241. Briese, T. *et al.* Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis. *MBio* **6,** e01491--15 (2015).
242. Naccache, S. N. *et al.* The perils of pathogen discovery: origin of a novel

parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *J. Virol.* **87,** 11966–77 (2013).

243. Liu, H. *et al.* Complete genome sequence of a bovine viral diarrhea virus 2 from commercial fetal bovine serum. *J. Virol.* **86,** 10233 (2012).
244. Smith, R. A. Contamination of clinical specimens with MLV-encoding nucleic acids: implications for XMRV and other candidate human retroviruses. *Retrovirology* **7,** 112 (2010).
245. Popgeorgiev, N. *et al.* Marseillevirus-like virus recovered from blood donated by asymptomatic humans. *J. Infect. Dis.* **208,** 1042–1050 (2013).
246. Popgeorgiev, N. *et al.* Marseillevirus prevalence in multitransfused patients suggests blood transmission. *J. Clin. Virol.* **58,** 722–725 (2013).
247. Sauvage, V. *et al.* No Evidence of Marseillevirus-like Virus Presence in Blood Donors and Recipients of Multiple Blood Transfusions. *J. Infect. Dis.* **210,** 2017–2018 (2014).
248. Popgeorgiev, N., Michel, G., Lepidi, H., Raoult, D. & Desnues, C. Marseillevirus Adenitis in an 11-Month-Old Child. *J. Clin. Microbiol.* **51,** 4102–4105 (2013).
249. Aherfi, S. *et al.* Marseillevirus in lymphoma: a giant in the lymph node. *Lancet Infect. Dis.* **16,** e225–e234 (2016).
250. Halary, S., Temmam, S., Raoult, D. & Desnues, C. Viral metagenomics: are we missing the giants? *Curr. Opin. Microbiol.* **31,** 34–43 (2016).
251. Sehn, J. K. *et al.* Occult Specimen Contamination in Routine Clinical Next-Generation Sequencing Testing. *Am. J. Clin. Pathol.* **144,** 667–674 (2015).
252. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12,** 87 (2014).
253. Chen, Y. *et al.* VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics* **29,** 266–267 (2013).
254. Zhou, Y. *et al.* Metagenomic Approach for Identification of the Pathogens

Associated with Diarrhea in Stool Specimens. *J. Clin. Microbiol.* **54,** 368–75 (2016).

255. Del Fabbro, C., Scalabrin, S., Morgante, M. & Giorgi, F. M. An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. *PLoS One* **8,** e85024 (2013).
256. Pruitt, K. D. *et al.* The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* **19,** 1316–23 (2009).
257. Daly, G. M. *et al.* Host subtraction, filtering and assembly validations for novel viral discovery using next generation sequencing data. *PLoS One* **10,** (2015).
258. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
259. R Development Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing Vienna Austria* **0,** {ISBN} 3-900051-07-0 (2016).
260. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17,** 333–351 (2016).
261. Lewandowska, D. W. *et al.* Metagenomic sequencing complements routine diagnostics in identifying viral pathogens in lung transplant recipients with unknown etiology of respiratory infection. *PLoS One* **12,** e0177340 (2017).
262. Towner, J. S. *et al.* Newly Discovered Ebola Virus Associated with Hemorrhagic Fever Outbreak in Uganda. *PLoS Pathog.* **4,** e1000212 (2008).
263. Lakis, N. S. *et al.* Novel Poxvirus Infection in an Immune Suppressed Patient. *Clin. Infect. Dis.* **61,** 1543–8 (2015).
264. Drosten, C. *et al.* Identification of a Novel Coronavirus in Patients with

Severe Acute Respiratory Syndrome. *N. Engl. J. Med.* **348,** 1967–1976 (2003).

265. Yu, G. *et al.* Discovery of a novel polyomavirus in acute diarrheal samples from children. *PLoS One* **7,** e49449 (2012).
266. Chiu, C. Y. *et al.* A novel adenovirus species associated with an acute respiratory outbreak in a baboon colony and evidence of coincident human infection. *MBio* **4,** (2013).
267. Victoria, J. G. *et al.* Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus. *J. Virol.* **84,** 6033–40 (2010).
268. Norman, J. M. *et al.* Disease-Specific Alterations in the Enteric Virome in Inflammatory Bowel Disease. *Cell* **160,** 447–460 (2015).
269. Handley, S. A. *et al.* Pathogenic simian immunodeficiency virus infection is associated with expansion of the enteric virome. *Cell* **151,** 253–266 (2012).
270. Li, L. *et al.* AIDS alters the commensal plasma virome. *J. Virol.* **87,** 10912–5 (2013).
271. Virgin, H. W. The virome in mammalian physiology and disease. *Cell* **157,** 142–50 (2014).
272. Savard, M. *et al.* Infection of primary human monocytes by Epstein-Barr virus. *J. Virol.* **74,** 2612–9 (2000).
273. Chun, T.-W. *et al.* Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature* **387,** 183–188 (1997).
274. De Vlaminck, I. *et al.* Temporal response of the human virome to immunosuppression and antiviral therapy. *Cell* **155,** 1178–1187 (2013).
275. Li, L. *et al.* AIDS alters the commensal plasma virome. *J. Virol.* **87,** 10912–10915 (2013).
276. Yin, L. *et al.* High-resolution deep sequencing reveals biodiversity, population structure, and persistence of HIV-1 quasispecies within host

ecosystems. *Retrovirology* **9,** 108 (2012).

277. Gyarmati, P. *et al.* Metagenomic analysis of bloodstream infections in patients with acute leukemia and therapy-induced neutropenia. *Sci. Rep.* **6,** 23532 (2016).
278. De Vlaminck, I. *et al.* Temporal Response of the Human Virome to Immunosuppression and Antiviral Therapy. *Cell* **155,** 1178–1187 (2013).
279. Kapoor, A. *et al.* Virome Analysis of Transfusion Recipients Reveals a Novel Human Virus That Shares Genomic Features with Hepaciviruses and Pegiviruses. *MBio* **6,** e01466--15 (2015).
280. Sauvage, V. *et al.* Viral metagenomics applied to blood donors and recipients at high risk for blood-borne infections. *Blood Transfus.* 1–8 (2016). doi:10.2450/2016.0160-15
281. Read, S. J. Recovery efficiences on nucleic acid extraction kits as measured by quantitative LightCycler PCR. *Mol. Pathol.* **54,** 86–90 (2001).
282. Victoria, J. G. *et al.* Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *J. Virol.* **83,** 4642–51 (2009).
283. Matranga, C. B. *et al.* Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome Biol.* **15,** 519 (2014).
284. Bent, Z. W. *et al.* Enriching pathogen transcripts from infected samples: A capture-based approach to enhanced host-pathogen RNA sequencing. *Anal. Biochem.* **438,** 90–96 (2013).
285. van Dijk, E. L., Jaszczyszyn, Y. & Thermes, C. Library preparation methods for next-generation sequencing: Tone down the bias. *Exp. Cell Res.* **322,** 12–20 (2014).
286. Conceição-Neto, N. *et al.* Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci. Rep.* **5,** 16532 (2015).
287. Lewandowska, D. W. *et al.* Optimization and validation of sample

preparation for metagenomic sequencing of viruses in clinical samples. *Microbiome* **5,** 94 (2017).

288. Kim, K.-H. & Bae, J.-W. Amplification Methods Bias Metagenomic Libraries of Uncultured Single-Stranded and Double-Stranded DNA Viruses. *Appl. Environ. Microbiol.* **77,** 7663–7668 (2011).
289. Reyes, G. R. & Kim, J. P. Sequence-independent, single-primer amplification (SISPA) of complex DNA populations. *Mol. Cell. Probes* **5,** 473–481 (1991).
290. Djikeng, A. *et al.* Viral genome sequencing by random priming methods. *BMC Genomics* **9,** 5 (2008).
291. Tan, L. Van *et al.* Random PCR and ultracentrifugation increases sensitivity and throughput of VIDISCA for screening of pathogens in clinical specimens. *J. Infect. Dev. Ctries.* **5,** 142–8 (2011).
292. Blomström, A.-L., Widén, F., Hammer, A.-S., Belák, S. & Berg, M. Detection of a novel astrovirus in brain tissue of mink suffering from shaking mink syndrome by use of viral metagenomics. *J. Clin. Microbiol.* **48,** 4392–6 (2010).
293. Rosseel, T. *et al.* The Origin of Biased Sequence Depth in Sequence-Independent Nucleic Acid Amplification and Optimization for Efficient Massive Parallel Sequencing. *PLoS One* **8,** (2013).
294. Karlsson, O. E., Belák, S. & Granberg, F. The effect of preprocessing by sequence-independent, single-primer amplification (SISPA) on metagenomic detection of viruses. *Biosecur. Bioterror.* **11 Suppl 1,** (2013).
295. Aigrain, L., Gu, Y. & Quail, M. A. Quantitation of next generation sequencing library preparation protocol efficiencies using droplet digital PCR assays - a systematic comparison of DNA library preparation kits for Illumina sequencing. *BMC Genomics* **17,** 458 (2016).
296. Quail, M. A. *et al.* Optimal enzymes for amplifying sequencing libraries. *Nat. Methods* **9,** 10–11 (2012).
297. Dabney, J. & Meyer, M. Length and GC-biases during sequencing library

amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques* **52,** 87–94 (2012).

298. Solonenko, S. A. *et al.* Sequencing platform and library preparation choices impact viral metagenomes. *BMC Genomics* **14,** 320 (2013).
299. Frey, K. G. *et al.* Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood. *BMC Genomics* **15,** (2014).
300. Qiu, Y. *et al.* Detection of viromes of RNA viruses using the next generation sequencing libraries prepared by three methods. *Virus Res.* **237,** 22–26 (2017).
301. GLENN, T. C. Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* **11,** 759–769 (2011).
302. Fox, E. J., Reid-Bayliss, K. S., Emond, M. J. & Loeb, L. A. Accuracy of Next Generation Sequencing Platforms. *Next Gener. Seq. Appl.* **1,** (2014).
303. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20,** 265–72 (2010).
304. Illumina. HiSeq 2500 Specifications. Available at: https://www.illumina.com/systems/sequencing-platforms/hiseq-2500/specifications.html. (Accessed: 10th July 2017)
305. Simmonds, P. Methods for virus classification and the challenge of incorporating metagenomic sequence data. *J. Gen. Virol.* **96,** 1193–1206 (2015).
306. Young, J. C. *et al.* Viral metagenomics reveal blooms of anelloviruses in the respiratory tract of lung transplant recipients. *Am. J. Transplant* **15,** 200–9 (2015).
307. Rani, A. *et al.* A diverse virome in kidney transplant patients contains multiple viral subtypes with distinct polymorphisms. *Sci. Rep.* **6,** 33327 (2016).

308. Zhao, G. *et al.* VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology* **503,** 21–30 (2017).
309. Kostic, A. D. *et al.* PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat. Biotechnol.* **29,** 393–396 (2011).
310. Gish, W. & States, D. J. Identification of protein coding regions by database similarity search. *Nat. Genet.* **3,** 266–272 (1993).
311. Trifonov, V. & Rabadan, R. Frequency analysis techniques for identification of viral genetic data. *MBio* **1,** e00156-10 (2010).
312. Minot, S., Wu, G. D., Lewis, J. D., Bushman, F. D. & Ideker, T. Conservation of Gene Cassettes among Diverse Viruses of the Human Gut. *PLoS One* **7,** e42342 (2012).
313. Forterre, P., Krupovic, M. & Prangishvili, D. Cellular domains and viral lineages. *Trends Microbiol.* **22,** 554–558 (2014).
314. Harish, A., Abroi, A., Gough, J. & Kurland, C. Did Viruses Evolve As a Distinct Supergroup from Common Ancestors of Cells? *Genome Biol. Evol.* **8,** 2474–2481 (2016).
315. Kalpoe, J. S. *et al.* Validation of clinical application of cytomegalovirus plasma DNA load measurement and definition of treatment criteria by analysis of correlation to antigen detection. *J. Clin. Microbiol.* **42,** 1498–504 (2004).
316. Dumoulin, A. & Hirsch, H. H. Reevaluating and optimizing polyomavirus BK and JC real-time PCR assays to detect rare sequence polymorphisms. *J. Clin. Microbiol.* **49,** 1382–8 (2011).
317. Candotti, D., Temple, J., Owusu-Ofori, S. & Allain, J.-P. Multiplex real-time quantitative RT-PCR assay for hepatitis B virus, hepatitis C virus, and human immunodeficiency virus type 1. *J. Virol. Methods* **118,** 39–47 (2004).
318. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).

319. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–9 (2012).
320. Genome Research Ltd. SMALT.
321. Qiagen. CLC Genomics Workbench.
322. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
323. Luft, L. M., Gill, M. J. & Church, D. L. HIV-1 viral diversity and its implications for viral load testing: review of current platforms. *Int. J. Infect. Dis.* **15,** e661–e670 (2011).
324. Pawlotsky, J.-M. *et al.* Standardization of Hepatitis C Virus RNA Quantification. *Hepatology* **32,** 654–659 (2000).
325. Païssé, S. *et al.* Comprehensive description of blood microbiome from healthy donors assessed by 16S targeted metagenomic sequencing. *Transfusion* **56,** 1138–1147 (2016).
326. Mariscal, L. F. *et al.* TT virus replicates in stimulated but not in nonstimulated peripheral blood mononuclear cells. *Virology* **301,** 121–129 (2002).
327. Madsen, C. D. *et al.* TTV viral load as a marker for immune reconstitution after initiation of HAART in HIV-infected patients. *HIV Clin. Trials* **3,** 287–295 (2002).
328. Thom, K. & Petrik, J. Progression towards AIDS leads to increased Torque teno virus and Torque teno minivirus titers in tissues of HIV infected individuals. *J. Med. Virol.* **79,** 1–7 (2007).
329. Cooling, L. L., Koerner, T. A. & Naides, S. J. Multiple glycosphingolipids determine the tissue tropism of parvovirus B19. *J. Infect. Dis.* **172,** 1198–205 (1995).
330. Cassinotti, P. & Siegl, G. Quantitative Evidence for Persistence of Human Parvovirus B19 DNA in an Immunocompetent Individual. *Eur J Clin Microbiol Infect Dis* **19,** 886–895 (2000).

331. Stapleton, J. T. *et al.* A novel T cell evasion mechanism in persistent RNA virus infection. *Trans. Am. Clin. Climatol. Assoc.* **125,** 14-24–6 (2014).
332. Bourlet, T. *et al.* Detection of GB virus C/hepatitis G virus in semen and saliva of HIV type-1 infected men. *Clin. Microbiol. Infect.* **8,** 352–357 (2002).
333. Mellor, J., Haydon, G., Blair, C., Livingstone, W. & Simmonds, P. Low level or absent in vivo replication of hepatitis C virus and hepatitis G virus/GB virus C in peripheral blood mononuclear cells. *J. Gen. Virol.* **79 (Pt 4),** 705–714 (1998).
334. Pontisso, P. *et al.* Hepatitis C Virus RNA Profiles in Chronically Infected Individuals: Do They Relate to Disease Activity?
335. Reploeg, M. D., Storch, G. A. & Clifford, D. B. BK Virus: A Clinical Review. *Clin. Infect. Dis.* **33,** 191–202 (2001).
336. Jeffers, L. K., Madden, V. & Webster-Cyriaque, J. BK virus has tropism for human salivary gland cells in vitro: Implications for transmission. *Virology* **394,** 183–193 (2009).
337. Burger-Calderon, R. *et al.* Replication of Oral BK Virus in Human Salivary Gland Cells. *J. Virol.* **88,** 559–573 (2014).
338. Zimber, U. *et al.* Geographical prevalence of two types of Epstein-Barr virus. *Virology* **154,** 56–66 (1986).
339. Scott, M. K. *et al.* Human Adenovirus Associated with Severe Respiratory Infection, Oregon, USA, 2013–2014. *Emerg. Infect. Dis.* **22,** 1044–1051 (2016).
340. Segerman, A., Lindman, K., Mei, Y.-F., Allard, A. & Wadell, G. Adenovirus types 11p and 35 attach to and infect primary lymphocytes and monocytes, but hexon expression in T-cells requires prior activation. *Virology* **349,** 96–111 (2006).
341. van der Veen, J. & Lambriex, M. Relationship of adenovirus to lymphocytes in naturally infected human tonsils and adenoids. *Infect. Immun.* **7,** 604–

609 (1973).

342. Flomenberg, P., Gutierrez, E., Piaskowski, V. & Casper, J. T. Detection of adenovirus DNA in peripheral blood mononuclear cells by polymerase chain reaction assay. *J. Med. Virol.* **51,** 182–8 (1997).
343. Echavarría, M. Adenoviruses in immunocompromised hosts. *Clin. Microbiol. Rev.* **21,** 704–15 (2008).
344. Lukasik, J., Scott, T. M., Andryshak, D. & Farrah, S. R. Influence of salts on virus adsorption to microporous filters. *Appl. Environ. Microbiol.* **66,** 2914–20 (2000).
345. Dugan, A. S. *et al.* Human alpha-defensins inhibit BK virus infection by aggregating virions and blocking binding to host cells. *J. Biol. Chem.* **283,** 31125–32 (2008).
346. Kumar, A., Murthy, S. & Kapoor, A. Evolution of selective-sequencing approaches for virus discovery and virome analysis. *Virus Res.* (2017). doi:10.1016/j.virusres.2017.06.005
347. Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. Viral mutation rates. *J. Virol.* **84,** 9733–48 (2010).
348. Duffy, S., Shackelton, L. A. & Holmes, E. C. Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* **9,** 267–276 (2008).
349. Zhao, G. *et al.* VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology* **503,** 21–30 (2017).
350. Naeem, R., Rashid, M. & Pain, A. READSCAN: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation. *Bioinformatics* **29,** 391–392 (2013).
351. Lee, W.-P. *et al.* MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping. *PLoS One* **9,** e90581 (2014).
352. Frey, K. G. *et al.* Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood. *BMC Genomics* **15,** 96 (2014).

353. Manso, C. F., Bibby, D. F. & Mbisa, J. L. Efficient and unbiased metagenomic recovery of RNA virus genomes from human plasma samples. *Sci. Rep.* **7,** 4173 (2017).
354. Rosseel, T., Pardon, B., De Clercq, K., Ozhelvaci, O. & Van Borm, S. False-positive results in metagenomic virus discovery: A strong case for follow-up diagnosis. *Transbound. Emerg. Dis.* **61,** (2014).
355. Laurence, M., Hatzis, C., Brash, D. E., Lei, H. & Li, T. Common Contaminants in Next-Generation Sequencing That Hinder Discovery of Low-Abundance Microbes. *PLoS One* **9,** e97876 (2014).
356. Naccache, S. N. *et al.* The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-Like Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns. *J. Virol.* **87,** (2013).
357. Bolin, S. R., Matthews, P. J. & Ridpath, J. F. Methods for detection and frequency of contamination of fetal calf serum with bovine viral diarrhea virus and antibodies against bovine viral diarrhea virus. *J Vet Diagn Invest* **3,** 199–203
358. Flickinger, M. *et al.* Correcting for Sample Contamination in Genotype Calling of DNA Sequence Data. *Am. J. Hum. Genet.* **97,** 284–290 (2015).
359. Meier-Kriesche, H.-U., Schold, J. D., Srinivas, T. R. & Kaplan, B. Lack of improvement in renal allograft survival despite a marked decrease in acute rejection rates over the most recent era. *Am. J. Transplant* **4,** 378–83 (2004).
360. Lodhi, S. A., Lamb, K. E. & Meier-Kriesche, H. U. Solid Organ Allograft Survival Improvement in the United States: The Long-Term Does Not Mirror the Dramatic Short-Term Success. *Am. J. Transplant.* **11,** 1226–1235 (2011).
361. Lechler, R. I., Sykes, M., Thomson, A. W. & Turka, L. A. Organ transplantation--how much of the promise has been realized? *Nat. Med.* **11,** 605–613 (2005).

362. Fishman, J. A. Opportunistic infections--coming to the limits of immunosuppression? *Cold Spring Harb. Perspect. Med.* **3,** a015669 (2013).
363. Unterman, S. *et al.* A descriptive analysis of 1251 solid organ transplant visits to the emergency department. *West. J. Emerg. Med.* **10,** 48–54 (2009).
364. Pinson, C. W. *et al.* Health-related quality of life after different types of solid organ transplantation. *Ann. Surg.* **232,** 597–607 (2000).
365. Martin-Gandul, C. *et al.* Preventive Strategies Against Cytomegalovirus and Incidence of α-Herpesvirus Infections in Solid Organ Transplant Recipients: A Nationwide Cohort Study. *Am. J. Transplant.* (2017). doi:10.1111/ajt.14192
366. Le, J., Durand, C. M., Agha, I. & Brennan, D. C. Epstein–Barr virus and renal transplantation. *Transplant. Rev.* **31,** 55–60 (2017).
367. Lamarche, C. *et al.* BK Polyomavirus and the Transplanted Kidney: Immunopathology and Therapeutic Approaches. *Transplantation* **100,** 2276–2287 (2016).
368. Funk, G. A., Gosert, R. & Hirsch, H. H. Viral dynamics in transplant patients: implications for disease. *Lancet Infect. Dis.* **7,** 460–472 (2007).
369. Tan, H.-H. & Goh, C.-L. Viral infections affecting the skin in organ transplant recipients: epidemiology and current management strategies. *Am. J. Clin. Dermatol.* **7,** 13–29 (2006).
370. Shahani, L., Ariza-Heredia, E. J. & Chemaly, R. F. Antiviral therapy for respiratory viral infections in immunocompromised patients. *Expert Rev. Anti. Infect. Ther.* **15,** 401–415 (2017).
371. Angarone, M. & Ison, M. G. Diarrhea in solid organ transplant recipients. *Curr. Opin. Infect. Dis.* **28,** 308–316 (2015).
372. Green, K. Y. Norovirus infection in immunocompromised hosts. *Clin. Microbiol. Infect.* **20,** 717–723 (2014).
373. Engelmann, I. *et al.* In Vivo Persistence of Human Rhinoviruses in Immunosuppressed Patients. *PLoS One* **12,** e0170774 (2017).

374. Lakis, N. S. *et al.* Novel Poxvirus Infection in an Immune Suppressed Patient. *Clin. Infect. Dis.* **61,** 1543–1548 (2015).
375. Roossinck, M. J. The good viruses: viral mutualistic symbioses. *Nat. Publ. Gr.* **9,** (2011).
376. Kernbauer, E., Ding, Y. & Cadwell, K. An enteric virus can replace the beneficial function of commensal bacteria. *Nature* **516,** 94–8 (2014).
377. Young, J. C. *et al.* Viral metagenomics reveal blooms of anelloviruses in the respiratory tract of lung transplant recipients. *Am. J. Transplant* **15,** 200–209 (2015).
378. Focosi, D., Macera, L., Pistello, M. & Maggi, F. Torque Teno virus viremia correlates with intensity of maintenance immunosuppression in adult orthotopic liver transplant. *J. Infect. Dis.* **210,** 667–668 (2014).
379. Focosi, D. *et al.* Torquetenovirus viremia kinetics after autologous stem cell transplantation are predictable and may serve as a surrogate marker of functional immune reconstitution. *J. Clin. Virol.* **47,** 189–192 (2010).
380. Görzer, I., Haloschan, M., Jaksch, P., Klepetko, W. & Puchhammer-Stöckl, E. Plasma DNA levels of Torque teno virus and immunosuppression after lung transplantation. *J. Heart Lung Transplant.* **33,** 320–323 (2014).
381. Li, S.-K. *et al.* Detection and identification of plasma bacterial and viral elements in HIV/AIDS patients in comparison to healthy adults. *Clin. Microbiol. Infect.* **18,** 1126–1133 (2012).
382. Cadwell, K. The virome in host health and disease. *Immunity* **42,** 805–13 (2015).
383. Li, L. *et al.* AIDS Alters the Commensal Plasma Virome. *J. Virol.* **87,** 10912–10915 (2013).
384. Parker, W. B. Enzymology of purine and pyrimidine antimetabolites used in the treatment of cancer. *Chem. Rev.* **109,** 2880–93 (2009).
385. Pillans, P. Experimental and Clinical Pharmacology: Immunosuppressants - mechanisms of action and monitoring. *Aust. Prescr.* **29,** 99–101 (2006).

386. Meier, C. A. Mechanisms of immunosuppression by glucocorticoids. *Eur. J. Endocrinol.* **134,** 50 (1996).
387. Harper, L. & Savage, C. O. S. Pathogenesis of ANCA-associated systemic vasculitis. *J. Pathol.* **190,** 349–359 (2000).
388. Berden, A. *et al.* Diagnosis and management of ANCA associated vasculitis. *BMJ* **344,** e26 (2012).
389. Hamour, S., Salama, A. D. & Pusey, C. D. Management of ANCA-associated vasculitis: Current trends and future prospects. *Ther. Clin. Risk Manag.* **6,** 253–64 (2010).
390. Sović, I. *et al.* Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat. Commun.* **7,** 11307 (2016).
391. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–9 (2009).
392. Wickham, H. *Ggplot2 : elegrant graphics for data analysis*. (Springer-Verlag, 2009).
393. Naimi, A. I., Moodie, E. E. M., Auger, N. & Kaufman, J. S. Constructing Inverse Probability Weights for Continuous Exposures. *Epidemiology* **25,** 292–299 (2014).
394. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*. *World* **53,** (Springer International Publishing, 2002).
395. Lumley, A. T. survey: analysis of complex survey samples. *R documentation* 128 (2016).
396. Thomas Lumley. Analysis of Complex Survey Samples. *J. Stat. Softw.* **Vol. 9,** (2004).
397. Hajian-Tilaki, K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Casp. J. Intern. Med.* **4,** 627–35 (2013).
398. Tsuda, F. *et al.* IgM-class antibodies to TT virus (TTV) in patients with acute TTV infection. *Hepatol. Res.* **19,** 1–11 (2001).

399. Itoh, Y. *et al.* Visualization of TT virus particles recovered from the sera and feces of infected humans. *Biochem. Biophys. Res. Commun.* **279,** 718–724 (2000).
400. Sospedra, M. *et al.* Recognition of conserved amino acid motifs of common viruses and its role in autoimmunity. *PLoS Pathog.* **1,** e41 (2005).
401. Maggi, F. *et al.* Blood levels of TT virus following immune stimulation with influenza or hepatitis B vaccine. *J. Med. Virol.* **75,** 358–365 (2005).
402. Alter, H. J. Emerging, re-emerging and submerging infectious threats to the blood supply. *Vox Sang.* **87 Suppl 2,** 56–61 (2004).
403. Williams, C. F. *et al.* Persistent GB virus C infection and survival in HIV-infected men. *N. Engl. J. Med.* **350,** 981–990 (2004).
404. Lanteri, M. C. *et al.* Downregulation of Cytokines and Chemokines by GB Virus C After Transmission Via Blood Transfusion in HIV-Positive Blood Recipients. *J. Infect. Dis.* **211,** 1585–1596 (2015).
405. Toupance, O. *et al.* Cytomegalovirus-related disease and risk of acute rejection in renal transplant recipients: a cohort study with case-control analyses. *Transpl. Int.* **13,** 413–9 (2000).
406. Hermans, P. E., Diaz-Buxo, J. A. & Stobo, J. D. Idiopathic late-onset immunoglobulin deficiency. *Am. J. Med.* **61,** 221–237 (1976).
407. Lischner, H. W. & Huang, N. N. Respiratory complications of primary hypogammaglobulinemia. *Pediatr. Ann.* **6,** 514–25 (1977).
408. Goldman, A. S. & Goldblum, R. M. Primary deficiencies in humoral immunity. *Pediatr. Clin. North Am.* **24,** 277–91 (1977).
409. Sanna, P. P. & Burton, D. R. Role of antibodies in controlling viral disease: lessons from experiments of nature and gene knockouts. *J. Virol.* **74,** 9813–9817 (2000).
410. Chapel, H. *et al.* Common variable immunodeficiency disorders: division into distinct clinical phenotypes. *Blood* **112,** 277–286 (2008).
411. van Kessel, D. A. *et al.* Long-term Clinical Outcome of Antibody

Replacement Therapy in Humoral Immunodeficient Adults With Respiratory Tract Infections. *EBioMedicine* **18,** 254–260 (2017).

412. Weck, K. E., Kim, S. S., Virgin HW, I. V, Speck, S. H. & Speck, S. H. B cells regulate murine gammaherpesvirus 68 latency. *J. Virol.* **73,** 4651–61 (1999).
413. Graham, M. B. & Braciale, T. J. Resistance to and Recovery from Lethal Influenza Virus Infection in B Lymphocyte–deficient Mice. *J. Exp. Med.* **186,** (1997).
414. Woodward, J. M. *et al.* The Role Of Chronic Norovirus Infection In The Enteropathy Associated With Common Variable Immunodeficiency. *Am. J. Gastroenterol.* **110,** 320–327 (2015).
415. Kris, R. M., Yetter, R. A., Cogliano, R., Ramphal, R. & Small, P. A. Passive serum antibody causes temporary recovery from influenza virus infection of the nose, trachea and lung of nude mice. *Immunology* **63,** 349–53 (1988).
416. Primary immunodeficiency diseases. Report of an IUIS Scientific Committee. International Union of Immunological Societies. *Clin. Exp. Immunol.* 1–28 (1999). doi:10.1046/j.1365-2249.1999.00109.x
417. Abolhassani, H. *et al.* A hypomorphic recombination-activating gene 1 (RAG1) mutation resulting in a phenotype resembling common variable immunodeficiency. *J. Allergy Clin. Immunol.* **134,** 1375–80 (2014).
418. van de Ven, A. A. J. M., Compeer, E. B., van Montfrans, J. M. & Boes, M. B-cell defects in common variable immunodeficiency: BCR signaling, protein clustering and hardwired gene mutations. *Crit. Rev. Immunol.* **31,** 85–98 (2011).
419. Kruger, G. *et al.* Interleukin-2 correction of defective in vitro T-cell mitogenesis in patients with common varied immunodeficiency. *J. Clin. Immunol.* **4,** 295–303 (1984).
420. Mayer, L., Fu, S. M., Cunningham-Rundles, C. & Kunkel, H. G. Polyclonal

immunoglobulin secretion in patients with common variable immunodeficiency using monoclonal B cell differentiation factors. *J. Clin. Invest.* **74,** 2115–20 (1984).

421. Farrington, M. *et al.* CD40 ligand expression is defective in a subset of patients with common variable immunodeficiency. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 1099–103 (1994).
422. Nonoyama, S., Farrington, M. L. & Ochs, H. D. Effect of IL-2 on immunoglobulin production by anti-CD40-activated human B cells: synergistic effect with IL-10 and antagonistic effect with IL-4. *Clin. Immunol. Immunopathol.* **72,** 373–9 (1994).
423. Cunningham-Rundles, C. & Ponda, P. P. Molecular defects in T- and B-cell primary immunodeficiency diseases. *Nat. Rev. Immunol.* **5,** 880–892 (2005).
424. Cunningham-Rundles, C. & Bodian, C. Common Variable Immunodeficiency: Clinical and Immunological Features of 248 Patients. *Clin. Immunol.* **92,** 34–48 (1999).
425. Di Renzo, M., Zhou, Z., George, I., Becker, K. & Cunningham-Rundles, C. Enhanced apoptosis of T cells in common variable immunodeficiency (CVID): role of defective CD28 co-stimulation. *Clin. Exp. Immunol.* **120,** 503–11 (2000).
426. Giovannetti, A. *et al.* Unravelling the complexity of T cell abnormalities in common variable immunodeficiency. *J. Immunol.* **178,** 3932–43 (2007).
427. Mouillot, G. *et al.* B-Cell and T-Cell Phenotypes in CVID Patients Correlate with the Clinical Phenotype of the Disease. *J. Clin. Immunol.* **30,** 746–755 (2010).
428. Wong, G. K. & Huissoon, A. P. T-cell abnormalities in common variable immunodeficiency: the hidden defect. *J. Clin. Pathol.* **69,** 672–676 (2016).
429. Fiedler, W. *et al.* T-cell activation defect in common variable immunodeficiency: restoration by phorbol myristate acetate (PMA) or

allogeneic macrophages. *Clin. Immunol. Immunopathol.* **44,** 206–18 (1987).

430. Gregersen, S. *et al.* Lung disease, T-cells and inflammation in common variable immunodeficiency disorders. *Scand. J. Clin. Lab. Invest.* **73,** 514–522 (2013).
431. Mullighan, C. G., Fanning, G. C., Chapel, H. M. & Welsh, K. I. TNF and lymphotoxin-alpha polymorphisms associated with common variable immunodeficiency: role in the pathogenesis of granulomatous disease. *J. Immunol.* **159,** (1997).
432. Mannon, P. J. *et al.* Excess IL-12 but not IL-23 Accompanies the Inflammatory Bowel Disease Associated With Common Variable Immunodeficiency. *Gastroenterology* **131,** 748–756 (2006).
433. Bertinchamp, R. *et al.* Exclusion of Patients with a Severe T-Cell Defect Improves the Definition of Common Variable Immunodeficiency. *J. Allergy Clin. Immunol. Pract.* **4,** 1147–1157 (2016).
434. Lanio, N., Sarmiento, E., Gallego, A. & Carbone, J. Immunophenotypic profile of T cells in common variable immunodeficiency: is there an association with different clinical findings? *Allergol. Immunopathol. (Madr).* **37,** 14–20
435. Viallard, J.-F. *et al.* CD8+HLA-DR+ T lymphocytes are increased in common variable immunodeficiency patients with impaired memory B-cell differentiation. *Clin. Immunol.* **119,** 51–58 (2006).
436. Carbone, J., Sarmiento, E., Micheloud, D., Rodríguez-Molina, J. & Fernández-Cruz, E. Elevated levels of activated CD4 T cells in common variable immunodeficiency: association with clinical findings. *Allergol. Immunopathol. (Madr).* **34,** 131–5
437. Wehr, C. *et al.* The EUROclass trial: defining subgroups in common variable immunodeficiency. *Blood* **111,** 77–85 (2008).
438. Aghamohammadi, A., Abolhassani, H., Moazzami, K., Parvaneh, N. & Rezaei, N. Correlation between common variable immunodeficiency clinical

phenotypes and parental consanguinity in children and adults. *J. Investig. Allergol. Clin. Immunol.* **20,** 372–9 (2010).

439. Cunningham-Rundles, C. The many faces of common variable immunodeficiency. *Hematology Am. Soc. Hematol. Educ. Program* **2012,** 301–305 (2012).
440. Cunningham-Rundles, C. How I treat common variable immune deficiency. *Blood* **116,** 7–15 (2010).
441. López-Pérez, P. *et al.* [Study of quality of life in adults with common variable immunodeficiency by using the Questionnaire SF-36]. *Rev. Alerg. Mex.* **61,** 52–58 (2014).
442. Chapel, H. & Cunningham-Rundles, C. Update in understanding common variable immunodeficiency disorders (CVIDs) and the management of patients with these conditions. *Br. J. Haematol.* **145,** 709–27 (2009).
443. Resnick, E. S., Moshier, E. L., Godbold, J. H. & Cunningham-Rundles, C. Morbidity and mortality in common variable immune deficiency over 4 decades. *Blood* **119,** 1650–1657 (2012).
444. Park, J. *et al.* Interferon Signature in the Blood in Inflammatory Common Variable Immune Deficiency. *PLoS One* **8,** e74893 (2013).
445. Paquin-Proulx, D. & Sandberg, J. K. Persistent Immune Activation in CVID and the Role of IVIg in Its Suppression. *Front. Immunol.* **5,** 637 (2014).
446. Paquin-Proulx, D. *et al.* IVIg Immune Reconstitution Treatment Alleviates the State of Persistent Immune Activation and Suppressed CD4 T Cell Counts in CVID. *PLoS One* **8,** e75199 (2013).
447. Bateman, E. A. L. *et al.* T cell phenotypes in patients with common variable immunodeficiency disorders: associations with clinical phenotypes in comparison with other groups with recurrent infections. *Clin. Exp. Immunol.* **170,** 202–211 (2012).
448. Perreau, M. *et al.* Exhaustion of bacteria-specific CD4 T cells and microbial translocation in common variable immunodeficiency disorders. *J. Exp. Med.*

211, 2033–2045 (2014).

449. Vlkova, M. *et al.* Age dependency and mutual relations in T and B lymphocyte abnormalities in common variable immunodeficiency patients. *Clin. Exp. Immunol.* **143,** 373–379 (2006).
450. Marashi, S. M. *et al.* Inflammation in common variable immunodeficiency is associated with a distinct CD8+ response to cytomegalovirus. *J. Allergy Clin. Immunol.* **127,** 1385–1393.e4 (2011).
451. Raeiszadeh, M. *et al.* The T cell response to persistent herpes virus infections in common variable immunodeficiency. *Clin. Exp. Immunol.* **146,** 234–42 (2006).
452. Yap, P. L. *et al.* Hepatitis C virus transmission by intravenous immunoglobulin. *J. Hepatol.* **21,** 455–460 (1994).
453. Wheat, W. H. *et al.* Possible role of human herpesvirus 8 in the lymphoproliferative disorders in common variable immunodeficiency. *J. Exp. Med.* **202,** 479–84 (2005).
454. Sallusto, F., Lenig, D., Förster, R., Lipp, M. & Lanzavecchia, A. Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. *J. Immunol.* **402,** 708–712 (1999).
455. Liu, Z. *et al.* Elevated CD38 antigen expression on CD8+ T cells is a stronger marker for the risk of chronic HIV disease progression to AIDS and death in the Multicenter AIDS Cohort Study than CD4+ cell count, soluble immune activation markers, or combinations of HLA-DR and CD38 expression. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.* **16,** 83–92 (1997).
456. Hulspas, R., O'Gorman, M. R. G., Wood, B. L., Gratama, J. W. & Sutherland, D. R. Considerations for the control of background fluorescence in clinical flow cytometry. *Cytom. Part B Clin. Cytom.* **76B,** 355–364 (2009).
457. Depil, S., Roche, C., Dussart, P. & Prin, L. Expression of a human endogenous retrovirus, HERV-K, in the blood cells of leukemia patients.

Leukemia **16,** 254–259 (2002).

458. Andersen, C. L., Ledet Jensen, J. & Ørntoft, T. F. Normalization of Real-Time Quantitative Reverse Transcription-PCR Data: A Model-Based Variance Estimation Approach to Identify Genes Suited for Normalization, Applied to Bladder and Colon Cancer Data Sets. *CANCER Res.* **64,** 5245–5250 (2004).
459. Chege, D. *et al.* Evaluation of a Quantitative Real-Time PCR Assay to Measure HIV-Specific Mucosal CD8+ T Cell Responses in the Cervix. *PLoS One* **5,** (2010).
460. Burgmann, H. *et al.* Increased serum concentration of soluble CD14 is a prognostic marker in gram-positive sepsis. *Clin. Immunol. Immunopathol.* **80,** 307–10 (1996).
461. Young, G. R. *et al.* Resurrection of endogenous retroviruses in antibody-deficient mice. *Nature* **491,** 774–778 (2012).
462. Januszkiewicz-Lewandowska, D. *et al.* Env gene expression of human endogenous retrovirus-k and human endogenous retrovirus-w in childhood acute leukemia cells. *Acta Haematol.* **129,** 232–237 (2013).
463. Marchetti, G., Tincati, C. & Silvestri, G. Microbial translocation in the pathogenesis of HIV infection and AIDS. *Clin. Microbiol. Rev.* **26,** 2–18 (2013).
464. Maggi, F. & Bendinelli, M. Immunobiology of the Torque teno viruses and other anelloviruses. *Curr. Top. Microbiol. Immunol.* **331,** 65–90 (2009).
465. Rocchi, J. *et al.* Torquetenovirus DNA drives proinflammatory cytokines production and secretion by immune cells via toll-like receptor 9. *Virology* **394,** 235–242 (2009).
466. Ameratunga, R. *et al.* Comparison of diagnostic criteria for common variable immunodeficiency disorder. *Front. Immunol.* **5,** 415 (2014).
467. Lau, P. *et al.* Metagenomics analysis of red blood cell and fresh-frozen plasma units. *Transfusion* (2017). doi:10.1111/trf.14148

468. Sauvage, V. & Eloit, M. Viral metagenomics and blood safety. *Transfus. Clin. Biol. J. la Soci{é}t{é} fran{ç}aise Transfus. Sang.* **23,** 28–38 (2016).
469. Wylie, K. M. *et al.* Metagenomic analysis of double-stranded DNA viruses in healthy adults. *BMC Biol.* **12,** 71 (2014).
470. Braaten, K. P. & Laufer, M. R. Human Papillomavirus (HPV), HPV-Related Disease, and the HPV Vaccine. *Rev. Obstet. Gynecol.* **1,** 2–10 (2008).
471. Gross, G., Pfister, H., Hagedorn, M. & Gissmann, L. Correlation between human papillomavirus (HPV) type and histology of warts. *J. Invest. Dermatol.* **78,** 160–4 (1982).
472. Chen, A. C.-H. *et al.* Human papillomavirus DNA detected in peripheral blood samples from healthy Australian male blood donors. *J. Med. Virol.* **81,** 1792–1796 (2009).
473. Bodaghi, S. *et al.* Could human papillomaviruses be spread through blood? *J. Clin. Microbiol.* **43,** 5428–5434 (2005).
474. Maglennon, G. A., McIntosh, P. B. & Doorbar, J. Immunosuppression facilitates the reactivation of latent papillomavirus infections. *J. Virol.* **88,** 710–716 (2014).
475. Peyton, C. L. *et al.* A novel human papillomavirus sequence from an international cervical cancer study. *J. Infect. Dis.* **170,** 1093–1095 (1994).
476. Jesenak, M., Banovcin, P., Jesenakova, B. & Babusikova, E. Pulmonary manifestations of primary immunodeficiency disorders in children. *Front. Pediatr.* **2,** 77 (2014).
477. Watanabe, A. *et al.* Human adenovirus detection among immunocompetent and immunocompromised patients presenting acute respiratory infection. *Rev. Soc. Bras. Med. Trop.* **46,** 161–165 (2013).
478. Matthes-Martin, S., Boztug, H. & Lion, T. Diagnosis and treatment of adenovirus infection in immunocompromised patients. *Expert Rev. Anti. Infect. Ther.* **11,** 1017–1028 (2013).
479. Echavarria, M. Adenoviruses in immunocompromised hosts. *Clin.*

Microbiol. Rev. **21,** 704–715 (2008).

480. Woodward, J., Gkrania-Klotsas, E. & Kumararatne, D. Chronic norovirus infection and common variable immunodeficiency. *Clin. Exp. Immunol.* (2016). doi:10.1111/cei.12884
481. van de Ven, A. A. J. M. *et al.* Pleconaril-resistant chronic parechovirus-associated enteropathy in agammaglobulinaemia. *Antivir. Ther.* **16,** 611–614 (2011).
482. Saulsbury, F. T., Winkelstein, J. A. & Yolken, R. H. Chronic rotavirus infection in immunodeficiency. *J. Pediatr.* **97,** 61–65 (1980).
483. Margolick, J. B. *et al.* Failure of T-cell homeostasis preceding AIDS in HIV-1 infection. The Multicenter AIDS Cohort Study. *Nat. Med.* **1,** 674–80 (1995).
484. van de Ven, A. A. J. M. *et al.* Increased Prevalence of Gastrointestinal Viruses and Diminished Secretory Immunoglobulin a Levels in Antibody Deficiencies. *J. Clin. Immunol.* **34,** 962–970 (2014).
485. Cao, W., Mehraj, V., Kaufmann, D. E., Li, T. & Routy, J.-P. Elevation and persistence of CD8 T-cells in HIV infection: the Achilles heel in the ART era. *J. Int. AIDS Soc.* **19,** 20697 (2016).
486. Fischer, A. & Rausell, A. Primary immunodeficiencies suggest redundancy within the human immune system. *Sci. Immunol.* **1,** eaah5861 (2016).
487. Fung-Leung, W. P., Kündig, T. M., Zinkernagel, R. M. & Mak, T. W. Immune response against lymphocytic choriomeningitis virus infection in mice without CD8 expression. *J. Exp. Med.* **174,** 1425–9 (1991).
488. Thimme, R. *et al.* CD8(+) T cells mediate viral clearance and disease pathogenesis during acute hepatitis B virus infection. *J. Virol.* **77,** 68–76 (2003).
489. Thimme, R. *et al.* Determinants of Viral Clearance and Persistence during Acute Hepatitis C Virus Infection. *J. Exp. Med.* **194,** (2001).
490. Simmons, A. & Tscharke, D. C. Anti-CD8 impairs clearance of herpes simplex virus from the nervous system: implications for the fate of virally

infected neurons. *J. Exp. Med.* **175,** (1992).

491. Rice, M. E. & Harris, G. T. Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law Hum. Behav.* **29,** 615–620 (2005).
492. Görzer, I. *et al.* Pre-transplant plasma Torque Teno virus load and increase dynamics after lung transplantation. *PLoS One* **10,** e0122975 (2015).
493. Ikeda, H. *et al.* Infection with an unenveloped DNA virus (TTV) in patients with acute or chronic liver disease of unknown etiology and in those positive for hepatitis C virus RNA. *J. Hepatol.* **30,** 205–212 (1999).
494. Höhne, M., Berg, T., Müller, A. R. & Schreier, E. Detection of sequences of TT virus, a novel DNA virus, in German patients. *J. Gen. Virol.* **79 (Pt 11),** 2761–4 (1998).
495. Tanaka, H. *et al.* Infection with an unenveloped DNA virus (TTV) associated with posttransfusion non-A to G hepatitis in hepatitis patients and healthy blood donors in Thailand. *J. Med. Virol.* **56,** 234–8 (1998).
496. Naoumov, N. V, Petrova, E. P., Thomas, M. G. & Williams, R. Presence of a newly described human DNA virus (TTV) in patients with liver disease. *Lancet* **352,** 195–197 (1998).
497. Takahashi, K. Very high prevalence of TT virus (TTV) infection in general population of Japan revealed by a new set of PCR primers. *Hepatol. Res.* **12,** 233–239 (1998).
498. Viazov, S. *et al.* Lack of evidence for an association between TTV infection and severe liver disease. *J. Clin. Virol.* **11,** 183–187 (1998).
499. Koch R. Die Ätiologie der Tuberkulose. *Berliner Klin. Wochenschrift* **19,** (1882).
500. Buffie, C. G. *et al.* Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. *Nature* **517,** 205–208 (2014).
501. Lawley, T. D. *et al.* Targeted Restoration of the Intestinal Microbiota with a Simple, Defined Bacteriotherapy Resolves Relapsing Clostridium difficile Disease in Mice. *PLoS Pathog.* **8,** e1002995 (2012).

502. Kernbauer, E., Ding, Y. & Cadwell, K. An enteric virus can replace the beneficial function of commensal bacteria. *Nature* **516,** 94–8 (2014).
503. Byrd, A. L. & Segre, J. A. Adapting Koch's postulates. *Science (80-.).* **351,** 224–226 (2016).
504. Mokili, J. L., Rohwer, F. & Dutilh, B. E. Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* **2,** 63–77 (2012).
505. Mancuso, R. *et al.* Torque teno virus (TTV) in multiple sclerosis patients with different patterns of disease. *J. Med. Virol.* **85,** 2176–2183 (2013).
506. Gergely, P. *et al.* Increased prevalence of transfusion-transmitted virus and cross-reactivity with immunodominant epitopes of the HRES-1/p28 endogenous retroviral autoantigen in patients with systemic lupus erythematosus. *Clin. Immunol.* **116,** 124–134 (2005).
507. Gergely, P., Blazsek, A., Dankó, K., Ponyi, A. & Poór, G. Detection of TT virus in patients with idiopathic inflammatory myopathies. *Ann. N. Y. Acad. Sci.* **1050,** 304–313 (2005).
508. Varani, S. & Landini, M. P. Cytomegalovirus-induced immunopathology and its clinical consequences. *Herpesviridae* **2,** 6 (2011).
509. Meyer, M. F., Hellmich, B., Kotterba, S. & Schatz, H. Cytomegalovirus infection in systemic necrotizing vasculitis: causative agent or opportunistic infection? *Rheumatol. Int.* **20,** 35–8 (2000).
510. Pérez-Mercado, A. E. & Vilá-Pérez, S. Cytomegalovirus as a Trigger for Systemic Lupus Erythematosus. *JCR J. Clin. Rheumatol.* **16,** 335–337 (2010).
511. Weinberger, B. *et al.* Healthy Aging and Latent Infection with CMV Lead to Distinct Changes in CD8+ and CD4+ T-Cell Subsets in the Elderly. *Hum. Immunol.* **68,** 86–90 (2007).
512. Nishiyama, S., Dutia, B. M. & Sharp, C. P. Complete genome sequences of novel anelloviruses from laboratory rats. *Genome Announc.* **3,** (2015).
513. Nishiyama, S. *et al.* Identification of novel anelloviruses with broad

diversity in UK rodents. *J. Gen. Virol.* **95,** 1544–53 (2014).

514. Ugland, K. I., Gray, J. S. & Ellingsen, K. E. The species-accumulation curve and estimation of species richness. *J. Anim. Ecol.* **72,** 888–897 (2003).
515. Willner, D., Thurber, R. V. & Rohwer, F. Metagenomic signatures of 86 microbial and viral metagenomes. *Environ. Microbiol.* **11,** 1752–1766 (2009).
516. Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5,** 4498 (2014).
517. Huson, D. H., Richter, D. C., Mitra, S., Auch, A. F. & Schuster, S. C. Methods for comparative metagenomics. *BMC Bioinformatics* **10,** S12 (2009).
518. Anthony, S. J. *et al.* A strategy to estimate unknown viral diversity in mammals. *MBio* **4,** e00598-13 (2013).
519. Mee, E. T., Preston, M. D., Minor, P. D., Schepelmann, S. & CS533 Study Participants, C. S. Development of a candidate reference material for adventitious virus detection in vaccine and biologicals manufacturing by deep sequencing. *Vaccine* **34,** 2035–43 (2016).
520. Rose, R., Constantinides, B., Tapinos, A., Robertson, D. L. & Prosperi, M. Challenges in the analysis of viral metagenomes. *Virus Evol.* **2,** vew022 (2016).

Appendices

Appendix 1. Database of vertebrate virus genomes used for the identification of viral reads by sequence alignment

Accession Number	Virus Name
NC_001499.1	Abelson murine leukemia virus
NC_002077.1	Adeno-associated virus 1
NC_001401.2	Adeno-associated virus 2
NC_001729.1	Adeno-associated virus 3
NC_001829.1	Adeno-associated virus 4
NC_006152.1	Adeno-associated virus 5
NC_006260.1	Adeno-associated virus 7
NC_006261.1	Adeno-associated virus 8
NC_007548.1	Adult diarrheal rotavirus strain J19
NC_007549.1	Adult diarrheal rotavirus strain J19
NC_007550.1	Adult diarrheal rotavirus strain J19
NC_007551.1	Adult diarrheal rotavirus strain J19
NC_007552.1	Adult diarrheal rotavirus strain J19
NC_007553.1	Adult diarrheal rotavirus strain J19
NC_007554.1	Adult diarrheal rotavirus strain J19
NC_007555.1	Adult diarrheal rotavirus strain J19
NC_007556.1	Adult diarrheal rotavirus strain J19
NC_007557.1	Adult diarrheal rotavirus strain J19
NC_007558.1	Adult diarrheal rotavirus strain J19
NC_004763.2	African green monkey polyomavirus
NC_006021.1	African horsesickness virus segment 1
NC_005996.1	African horsesickness virus segment 2
NC_006017.1	African horsesickness virus segment 3
NC_006012.1	African horsesickness virus segment 4
NC_006020.1	African horsesickness virus segment 5
NC_006018.1	African horsesickness virus segment 6
NC_006011.1	African horsesickness virus segment 7
NC_006016.1	African horsesickness virus segment 8
NC_006019.1	African horsesickness virus segment 9
NC_006009.1	African horsesickness virus segment 10
NC_001659.1	African swine fever virus
NC_001918.1	Aichi virus
NC_018465.1	Aino virus segment L
NC_018459.1	Aino virus segment M
NC_018460.1	Aino virus segment S
NC_009894.1	Akabane virus segment L
NC_009895.1	Akabane virus segment M
NC_009896.1	Akabane virus segment S
NC_001662.1	Aleutian mink disease virus
NC_004355.1	Alkhurma virus
NC_010249.1	Allpahuayo virus segment L
NC_010253.1	Allpahuayo virus segment S
NC_010251.1	Amapari virus segment L
NC_010247.1	Amapari virus segment S
NC_005832.1	Ambystoma tigrinum virus
NC_003468.2	Andes virus segment L
NC_003467.2	Andes virus segment M
NC_003466.1	Andes virus segment S
NC_003676.1	Apoi virus
NC_007582.1	Aquareovirus A segment 1
NC_007583.1	Aquareovirus A segment 2
NC_007584.1	Aquareovirus A segment 3
NC_007585.1	Aquareovirus A segment 4
NC_007592.1	Aquareovirus A segment 6
NC_007588.1	Aquareovirus A segment 8
NC_007589.1	Aquareovirus A segment 9
NC_007590.1	Aquareovirus A segment 10
NC_007591.1	Aquareovirus A segment 11
NC_020808.1	Aravan virus
NC_016752.1	Artibeus jamaicensis parvovirus 1
NC_014320.1	Astrovirus MLB1 HK05
NC_011400.1	Astrovirus MLB1
NC_016155.1	Astrovirus MLB2
NC_019028.1	Astrovirus MLB3
NC_013060.1	Astrovirus VA1
NC_018669.1	Astrovirus VA2
NC_019026.1	Astrovirus VA3
NC_019027.1	Astrovirus VA4
NC_016896.1	Astrovirus wild boar
NC_019853.1	Ateles paniscus polyomavirus 1
NC_007654.1	Atlantic salmon swim bladder sarcoma virus
NC_003900.1	Aura virus
NC_003243.1	Australian bat lyssavirus
NC_004828.1	Avian adeno-associated virus
NC_006263.1	Avian adeno-associated virus
NC_001402.1	Avian carcinoma virus
NC_003990.1	Avian encephalomyelitis virus
NC_005947.1	Avian endogenous retrovirus
NC_015396.1	Avian gyrovirus 2
NC_001451.1	Avian infectious bronchitis virus
NC_015116.1	Avian leukemia virus
NC_001408.1	Avian leukosis virus
NC_007652.1	Avian metapneumovirus
NC_001866.1	Avian myelocytomatosis virus
NC_015126.1	Avian orthoreovirus segment L1
NC_015127.1	Avian orthoreovirus segment L2
NC_015128.1	Avian orthoreovirus segment L3
NC_015129.1	Avian orthoreovirus segment M1
NC_015130.1	Avian orthoreovirus segment M2
NC_015131.1	Avian orthoreovirus segment M3
NC_015132.1	Avian orthoreovirus segment S1
NC_015133.1	Avian orthoreovirus segment S2
NC_015134.1	Avian orthoreovirus segment S3
NC_015135.1	Avian orthoreovirus segment S4
NC_019531.1	Avian paramyxovirus 4
NC_003043.1	Avian paramyxovirus 6
NC_006553.1	Avian sapelovirus
NC_015877.1	Baboon orthoreovirus segment L1
NC_015878.1	Baboon orthoreovirus segment L2
NC_015879.1	Baboon orthoreovirus segment L3
NC_015880.1	Baboon orthoreovirus segment M1
NC_015881.1	Baboon orthoreovirus segment M2
NC_015882.1	Baboon orthoreovirus segment M3
NC_015883.1	Baboon orthoreovirus segment S1
NC_015884.1	Baboon orthoreovirus segment S2
NC_015886.1	Baboon orthoreovirus segment S3
NC_015885.1	Baboon orthoreovirus segment S4
NC_012534.1	Bagaza virus
NC_010107.1	Bandicoot papillomatosis carcinomatosis virus type 1
NC_010817.1	Bandicoot papillomatosis carcinomatosis virus type 2
NC_015399.1	Barbel circovirus
NC_013458.1	Barfin flounder nervous necrosis virus RNA 1
NC_013459.1	Barfin flounder nervous necrosis virus RNA 2
NC_011063.1	Barfin flounder virus RNA1
NC_011064.1	Barfin flounder virus RNA2
NC_001786.1	Barmah Forest virus
NC_014468.1	Bat adeno-associated virus
NC_015932.1	Bat adenovirus 2
NC_016895.1	Bat adenovirus TJM
NC_021206.1	Bat circovirus isolate XOR7
NC_008315.1	Bat coronavirus (BtCoV/133/2005)
NC_010437.1	Bat coronavirus 1A
NC_010436.1	Bat coronavirus 1B
NC_014470.1	Bat coronavirus BM48-31/BGR/2008
NC_022103.1	Bat coronavirus CDPHE15/USA/2006
NC_009988.1	Bat coronavirus HKU2
NC_009019.1	Bat coronavirus HKU4-1
NC_009020.1	Bat coronavirus HKU5-1
NC_010438.1	Bat coronavirus HKU8
NC_009021.1	Bat coronavirus HKU9-1
NC_020881.1	Bat hepatitis virus
NC_018382.1	Bat hepevirus
NC_015940.1	Bat picornavirus 1
NC_015941.1	Bat picornavirus 2
NC_015934.1	Bat picornavirus 3
NC_017936.1	Bat sapovirus
NC_001944.1	Beak and feather disease virus
NC_010255.1	Bear Canyon virus segment L
NC_010256.1	Bear Canyon virus segment S
NC_016962.1	Bebaru virus
NC_007803.1	Beilong virus
NC_010646.1	Beluga Whale coronavirus
NC_019843.2	Betacoronavirus England 1
NC_014143.1	Bettongia penicillata papillomavirus 1
NC_001538.1	BK polyomavirus
NC_005982.1	Blotched snakehead virus
NC_005983.1	Blotched snakehead virus
NC_018506.1	Bluegill picornavirus
NC_006023.1	Bluetongue virus segment 1
NC_006013.1	Bluetongue virus segment 2
NC_006014.1	Bluetongue virus segment 3
NC_006024.2	Bluetongue virus segment 4
NC_006025.1	Bluetongue virus segment 5
NC_006010.1	Bluetongue virus segment 6
NC_006022.1	Bluetongue virus segment 7
NC_006007.1	Bluetongue virus segment 8

Accession Number	Virus Name
NC_006008.1	Bluetongue virus segment 9
NC_006015.1	Bluetongue virus segment 10
NC_014358.1	Bocavirus gorilla
NC_003679.1	Border disease virus
NC_001607.1	Borna disease virus
NC_005889.1	Bovine adeno-associated virus
NC_020074.1	Bovine adenovirus 6
NC_006324.1	Bovine adenovirus A
NC_001876.1	Bovine adenovirus B
AC_000002.1	Bovine adenovirus B
NC_002685.2	Bovine adenovirus D
NC_003045.1	Bovine coronavirus
NC_001859.1	Bovine enterovirus
NC_002526.1	Bovine ephemeral fever virus
NC_001831.1	Bovine foamy virus
NC_018668.1	Bovine hungarovirus
NC_001413.1	Bovine immunodeficiency virus
NC_004421.1	Bovine kobuvirus
NC_001414.1	Bovine leukemia virus
NC_001522.1	Bovine papillomavirus 1
NC_004197.1	Bovine papillomavirus 3
NC_004195.1	Bovine papillomavirus 5
NC_007612.1	Bovine papillomavirus 7
NC_005337.1	Bovine papular stomatitis virus
NC_002161.1	Bovine parainfluenza virus 3
NC_001540.1	Bovine parvovirus
NC_006259.1	Bovine parvovirus 2
NC_001442.1	Bovine polyomavirus
NC_012948.1	Bovine respiratory coronavirus
NC_012949.1	Bovine respiratory coronavirus
NC_001989.1	Bovine respiratory syncytial virus
NC_010354.1	Bovine rhinitis B virus
NC_001461.1	Bovine viral diarrhea virus 1
NC_002032.1	Bovine viral diarrhea virus 2
NC_012812.1	Bovine viral diarrhea virus 3
NC_007447.1	Breda virus
NC_014236.1	Broome virus segment L1
NC_014237.1	Broome virus segment L2
NC_014238.1	Broome virus segment L3
NC_014239.1	Broome virus segment M1
NC_014240.1	Broome virus segment M2
NC_014241.1	Broome virus segment M3
NC_014242.1	Broome virus segment S1
NC_014243.1	Broome virus segment S2
NC_014244.1	Broome virus segment S3
NC_014245.1	Broome virus segment S4
NC_004764.2	Budgerigar fledgling disease virus 1
NC_014373.1	Bundibugyo ebolavirus
NC_001925.1	Bunyamwera virus L segment
NC_001926.1	Bunyamwera virus M segment
NC_001927.1	Bunyamwera virus segment S
NC_009026.2	Bussuquara virus
NC_006875.1	Calicivirus isolate TCG
NC_012699.1	Calicivirus pig/AB90/CAN
NC_004064.1	Calicivirus strain NB
NC_012126.1	California sea lion anellovirus
NC_013796.1	California sea lion polyomavirus 1
NC_003391.1	Camelpox virus
NC_015267.1	Camelus dromedarius papillomavirus type 1
NC_015268.1	Camelus dromedarius papillomavirus type 2
NC_003410.1	Canary circovirus
NC_017085.1	Canary polyomavirus
NC_005309.1	Canarypox virus
NC_015374.1	Candiru virus segment L
NC_015373.1	Candiru virus segment M
NC_015375.1	Candiru virus segment S
AC_000003.1	Canine adenovirus 1
AC_000020.1	Canine adenovirus type 2
NC_001734.1	Canine adenovirus
NC_020499.1	Canine bocavirus
NC_004542.1	Canine calicivirus
NC_020904.1	Canine circovirus
NC_001921.1	Canine distemper virus
NC_004442.1	Canine minute virus
NC_001619.1	Canine oral papillomavirus
NC_006564.1	Canine papillomavirus 2
NC_008297.1	Canine papillomavirus 3
NC_010226.1	Canine papillomavirus 4
NC_013237.1	Canine papillomavirus 6
NC_016014.1	Canine papillomavirus 8
NC_016074.1	Canine papillomavirus 9
NC_016075.1	Canine papillomavirus 10
NC_019852.1	Canine papillomavirus 14
NC_001539.1	Canine parvovirus
NC_021178.1	Canine picodicistrovirus
NC_016964.1	Canine picornavirus
NC_008032.1	Capra hircus papillomavirus type 1
NC_011051.1	Capreolus capreolus papillomavirus 1
NC_001463.1	Caprine arthritis encephalitis virus
NC_020067.1	Cardioderma polyomavirus
NC_011530.1	Caretta caretta papillomavirus 1
NC_018484.1	CAS virus segment L
NC_018481.1	CAS virus segment S
NC_019854.1	Cebus albifrons polyomavirus 1
NC_001564.1	Cell fusing agent virus
NC_020065.1	Chaerephon polyomavirus 1
NC_020805.1	Chandipura virus
NC_010563.1	Chapare virus segment L
NC_010562.1	Chapare virus segment S
NC_001427.1	Chicken anemia virus
NC_003790.1	Chicken astrovirus
NC_004162.2	Chikungunya virus
NC_017825.1	Chimpanzee adenovirus Y25
NC_014743.1	Chimpanzee polyomavirus
NC_007586.1	Chum salmon reovirus CS segment 5
NC_007587.1	Chum salmon reovirus CS segment 7
NC_005990.1	Chuzan virus segment 1
NC_005995.1	Chuzan virus segment 10
NC_005986.1	Chuzan virus segment 2
NC_005989.1	Chuzan virus segment 3
NC_005991.1	Chuzan virus segment 4
NC_005993.1	Chuzan virus segment 5
NC_005987.1	Chuzan virus segment 6
NC_005988.1	Chuzan virus segment 7
NC_005994.1	Chuzan virus segment 8
NC_005992.1	Chuzan virus segment 9
NC_013020.1	Circovirus-like genome BBC-A
NC_013028.1	Circovirus-like genome CB-A
NC_013029.1	Circovirus-like genome CB-B
NC_013023.1	Circovirus-like genome RW-A
NC_013024.1	Circovirus-like genome RW-B
NC_013025.1	Circovirus-like genome RW-C
NC_013026.1	Circovirus-like genome RW-D
NC_013027.1	Circovirus-like genome RW-E
NC_013030.2	Circovirus-like genome SAR-A
NC_013018.1	Circovirus-like genome SAR-B
NC_002657.1	Classical swine fever virus
NC_015692.1	Colobus guereza papillomavirus type 2
NC_004181.1	Colorado tick fever virus segment 1
NC_004182.1	Colorado tick fever virus segment 2
NC_004183.1	Colorado tick fever virus segment 3
NC_004184.1	Colorado tick fever virus segment 4
NC_004185.1	Colorado tick fever virus segment 5
NC_004186.1	Colorado tick fever virus segment 6
NC_004187.1	Colorado tick fever virus segment 7
NC_004188.1	Colorado tick fever virus segment 8
NC_004180.1	Colorado tick fever virus segment 9
NC_004189.1	Colorado tick fever virus segment 10
NC_004191.1	Colorado tick fever virus segment 11
NC_004190.1	Colorado tick fever virus segment 12
NC_002361.1	Columbid circovirus
NC_016996.1	Common-moorhen coronavirus HKU21
NC_012800.1	Cosavirus A strain HCoSV-A1 polyprotein gene
NC_014372.1	Cote d'Ivoire ebolavirus
NC_016924.1	Cotia virus
NC_001541.1	Cottontail rabbit papillomavirus
NC_003663.2	Cowpox virus
NC_005301.3	Crimean-Congo hemorrhagic fever virus segment L
NC_005300.2	Crimean-Congo hemorrhagic fever virus segment M
NC_005302.1	Crimean-Congo hemorrhagic fever virus segment S
NC_018575.1	Crocuta crocuta papillomavirus 1
NC_007922.1	Crow polyomavirus
NC_008604.2	Culex flavivirus
NC_010252.1	Cupixi virus segment L
NC_010254.1	Cupixi virus segment S
NC_015521.1	Cutthroat trout virus
NC_014929.1	Cyclovirus bat/USA/2009
NC_014930.1	Cyclovirus NGchicken15/NGA/2009

Accession Number	Virus Name
NC_014928.1	Cyclovirus PKgoat11/PAK/2009
NC_014927.1	Cyclovirus PKgoat21/PAK/2009
NC_021707.1	Cyclovirus VN isolate ds3 capsid protein gene
NC_016154.1	Cynomolgus macaque cytomegalovirus strain Ottawa
NC_001523.1	Deer papillomavirus
NC_006967.1	Deerpox virus W
NC_006966.1	Deerpox virus W
NC_001477.1	Dengue virus 1
NC_001474.2	Dengue virus 2
NC_001475.2	Dengue virus 3
NC_002640.1	Dengue virus 4
NC_005234.1	Dobrava virus segment M
NC_005233.1	Dobrava virus segment S
NC_005235.1	Dobrava-Belgrade virus
NC_005283.1	Dolphin morbillivirus
NC_001813.1	Duck adenovirus A
AC_000004.1	Duck adenovirus A
NC_012437.1	Duck astrovirus C-NGB
NC_007220.1	Duck circovirus
NC_008250.2	Duck hepatitis A virus
NC_001344.1	Duck hepatitis B virus
NC_004159.1	Dugbe virus segment L
NC_004158.1	Dugbe virus segment M
NC_004157.1	Dugbe virus segment S
NC_020810.1	Duvenhage virus
NC_003899.1	Eastern equine encephalitis virus
NC_002549.1	Ebola virus --- Mayinga
NC_004105.1	Ectromelia virus
NC_016744.1	Eidolon helvum parvovirus 1
NC_020068.1	Eidolon polyomavirus 1
NC_001479.1	Encephalomyocarditis virus
NC_008718.1	Entebbe bat virus
NC_021220.1	Enterovirus F strain BEV---261 polyprotein gene
NC_010415.1	Enterovirus J strain 1631
NC_013695.1	Enterovirus J strain N203
NC_004994.2	Enzootic nasal tumour virus of goats
NC_004137.1	Epinephelus tauvina nervous necrosis virus RNA 1
NC_004136.1	Epinephelus tauvina nervous necrosis virus RNA 2
NC_013396.1	Epizootic hemorrhagic disease virus serotype 1 segment 1
NC_013397.1	Epizootic hemorrhagic disease virus serotype 1 segment 2
NC_013398.1	Epizootic hemorrhagic disease virus serotype 1 segment 3
NC_013399.1	Epizootic hemorrhagic disease virus serotype 1 segment 4
NC_013400.1	Epizootic hemorrhagic disease virus serotype 1 segment 5
NC_013401.1	Epizootic hemorrhagic disease virus serotype 1 segment 6
NC_013402.1	Epizootic hemorrhagic disease virus serotype 1 segment 7
NC_013403.1	Epizootic hemorrhagic disease virus serotype 1 segment 8
NC_013404.1	Epizootic hemorrhagic disease virus serotype 1 segment 9
NC_013405.1	Epizootic hemorrhagic disease virus serotype 1 segment 10
NC_002532.2	Equine arteritis virus
NC_010327.1	Equine coronavirus
NC_002201.1	Equine foamy virus
NC_001450.1	Equine infectious anemia virus
NC_012123.1	Equine papillomavirus 2
NC_017862.1	Equine papillomavirus 3
NC_020500.1	Equine papillomavirus type 6
NC_020902.1	Equine Pegivirus 1 isolate C0035
NC_017982.1	Equine polyomavirus
NC_003982.1	Equine rhinitis A virus
NC_003983.1	Equine rhinitis B virus 1
NC_003748.1	Equus caballus papillomavirus --- 1
NC_020085.1	Equus ferus caballus papillomavirus type 4
NC_020084.1	Equus ferus caballus papillomavirus type 5
NC_020501.1	Equus ferus caballus papillomavirus type 7
NC_006951.1	Erethizon dorsatum papillomavirus type 1
NC_011765.1	Erinaceus europaeus papillomavirus
NC_009527.1	European bat lyssavirus 1
NC_009528.1	European bat lyssavirus 2
NC_002615.1	European brown hare syndrome virus
NC_001524.1	European elk papillomavirus
NC_017940.1	European sheatfish virus
NC_003696.1	Eyach virus segment 1
NC_003697.1	Eyach virus segment 2
NC_003698.1	Eyach virus segment 3
NC_003699.1	Eyach virus segment 4
NC_003700.1	Eyach virus segment 5
NC_003701.1	Eyach virus segment 6
NC_003702.1	Eyach virus segment 7
NC_003703.1	Eyach virus segment 8
NC_003704.1	Eyach virus segment 9
NC_003705.1	Eyach virus segment 10
NC_003706.1	Eyach virus segment 11
NC_003707.1	Eyach virus segment 12
NC_022249.1	Feline astrovirus 2 strain 1637F
NC_017823.1	Feline bocavirus
NC_001481.2	Feline calicivirus
NC_001871.1	Feline foamy virus
NC_001482.1	Feline immunodeficiency virus
NC_002306.3	Feline infectious peritonitis virus
NC_001940.1	Feline leukemia virus
NC_016156.1	Feline picornavirus
NC_021472.1	Felis catus papillomavirus 3
NC_004765.1	Felis domesticus papillomavirus type 1
NC_005084.2	Fer-de-lance virus
NC_022253.1	Ferret papillomavirus isolate MpPV1
NC_008522.1	Finch circovirus
NC_007923.1	Finch polyomavirus
NC_010759.1	Flexal virus segment L
NC_010757.1	Flexal virus segment S
NC_004004.1	Foot-and-mouth disease virus
NC_013528.1	Fort Morgan virus
NC_021221.1	Fowl adenovirus 5 strain 340
NC_001720.1	Fowl adenovirus A
AC_000014.1	Fowl adenovirus A
NC_015323.1	Fowl adenovirus C
NC_000899.1	Fowl adenovirus D
AC_000013.1	Fowl adenovirus D
NC_014969.1	Fowl adenovirus E
NC_002188.1	Fowlpox virus
NC_013117.1	Francolinus leucoscepus papillomavirus 1
NC_001362.1	Friend murine leukemia virus
NC_004068.1	Fringilla coelebs papillomavirus
NC_002501.1	Frog adenovirus 1
NC_005946.1	Frog virus 3
NC_001403.1	Fujinami sarcoma virus
NC_014469.1	Gammapapillomavirus HPV127
NC_001710.1	GB virus C/Hepatitis G virus
NC_006558.1	Getah virus
NC_001885.2	Gibbon ape leukemia virus
NC_010306.1	Gill---associated virus
NC_004003.1	Goatpox virus Pellor
NC_018482.1	Golden Gate virus segment L
NC_018483.1	Golden Gate virus segment S
NC_005166.1	Golden shiner reovirus segment 1
NC_005167.1	Golden shiner reovirus segment 2
NC_005168.1	Golden shiner reovirus segment 3
NC_005169.1	Golden shiner reovirus segment 4
NC_005170.1	Golden shiner reovirus segment 5
NC_005171.1	Golden shiner reovirus segment 6
NC_005172.1	Golden shiner reovirus segment 7
NC_005173.1	Golden shiner reovirus segment 8
NC_005174.1	Golden shiner reovirus segment 9
NC_005175.1	Golden shiner reovirus segment 10
NC_005176.1	Golden shiner reovirus segment 11
NC_017979.1	Goose adenovirus 4
NC_003054.1	Goose circovirus
NC_004800.1	Goose hemorrhagic polyomavirus
NC_005036.1	Goose paramyxovirus SF02
NC_001701.1	Goose parvovirus
NC_014522.1	Great Island virus segment 1
NC_014523.1	Great Island virus segment 2
NC_014524.1	Great Island virus segment 3
NC_014525.1	Great Island virus segment 4
NC_014526.1	Great Island virus segment 5
NC_014527.1	Great Island virus segment 6
NC_014528.1	Great Island virus segment 7
NC_014529.1	Great Island virus segment 8
NC_014530.1	Great Island virus segment 9
NC_014531.1	Great Island virus segment 10
NC_001484.1	Ground squirrel hepatitis virus
NC_005082.1	Guanarito virus segment L
NC_005077.1	Guanarito virus segment S
NC_008521.1	Gull circovirus
NC_018401.1	Gyrovirus 4
NC_017091.1	Gyrovirus GyV3
NC_016418.1	Halastavi arva RNA virus
NC_001663.1	Hamster polyomavirus
NC_005222.1	Hantaan virus segment L
NC_005218.1	Hantaan virus

Accession Number	Virus Name
NC_005219.1	Hantaan virus
NC_006435.1	Hantavirus Z10 chromosome L
NC_006433.1	Hantavirus Z10 chromosome S segment
NC_006437.1	Hantavirus Z10 segment M
NC_001906.3	Hendra virus
NC_001489.1	Hepatitis A virus
NC_003977.1	Hepatitis B virus
NC_004102.1	Hepatitis C virus genotype 1
NC_009823.1	Hepatitis C virus genotype 2
NC_009824.1	Hepatitis C virus genotype 3
NC_009825.1	Hepatitis C virus genotype 4
NC_009826.1	Hepatitis C virus genotype 5
NC_009827.1	Hepatitis C virus genotype 6
NC_001653.2	Hepatitis delta virus
NC_001434.1	Hepatitis E virus
NC_001837.1	Hepatitis GB virus A
NC_001655.1	Hepatitis GB virus B
NC_001486.1	Heron hepatitis B virus
NC_012561.1	Highlands J virus
NC_005093.1	Hirame rhabdovirus
NC_013443.1	HMO Astrovirus A
AC_000007.1	Human adenovirus 2
AC_000008.1	Human adenovirus 5
NC_012959.1	Human adenovirus 54
AC_000005.1	Human adenovirus A
NC_001460.1	Human adenovirus A
NC_011203.1	Human adenovirus B1
NC_011202.1	Human adenovirus B2
NC_001405.1	Human adenovirus C
NC_010956.1	Human adenovirus D
AC_000006.1	Human adenovirus D
NC_003266.2	Human adenovirus E
NC_001454.1	Human adenovirus F
AC_000017.1	Human adenovirus type 1
AC_000019.1	Human adenovirus type 35
AC_000018.1	Human adenovirus type 7
NC_001943.1	Human astrovirus
NC_012042.1	Human bocavirus 2
NC_012564.1	Human bocavirus 3
NC_012729.2	Human bocavirus 4
NC_007455.1	Human bocavirus
NC_002645.1	Human coronavirus 229E
NC_006577.2	Human coronavirus HKU1
NC_005831.2	Human coronavirus NL63
NC_005147.1	Human coronavirus OC43
NC_012801.1	Human cosavirus B1
NC_012802.1	Human cosavirus D1
NC_012798.1	Human cosavirus E1
NC_021568.1	Human cyclovirus VS5700009
NC_012950.1	Human enteric coronavirus strain 4408
NC_001612.1	Human enterovirus A
NC_001472.1	Human enterovirus B
NC_001430.1	Human enterovirus D
NC_004295.1	Human erythrovirus V9
NC_015630.1	Human gyrovirus type 1
NC_001806.1	Human herpesvirus 1
NC_001798.1	Human herpesvirus 2
NC_001348.1	Human herpesvirus 3
NC_007605.1	Human herpesvirus 4 type 1
NC_009334.1	Human herpesvirus 4
NC_006273.2	Human herpesvirus 5
NC_001664.2	Human herpesvirus 6A
NC_000898.1	Human herpesvirus 6B
NC_001716.2	Human herpesvirus 7
NC_009333.1	Human herpesvirus 8
NC_001802.1	Human immunodeficiency virus 1
NC_001722.1	Human immunodeficiency virus 2
NC_004148.2	Human metapneumovirus
NC_022095.1	Human papillomavirus complete genome
NC_001356.1	Human papillomavirus type 1
NC_001352.1	Human papillomavirus type 2
NC_001457.1	Human papillomavirus type 4
NC_001531.1	Human papillomavirus type 5
NC_001355.1	Human papillomavirus type 6b
NC_001595.1	Human papillomavirus type 7
NC_001596.1	Human papillomavirus type 9
NC_001576.1	Human papillomavirus type 10
NC_001526.2	Human papillomavirus type 16
NC_001357.1	Human papillomavirus type 18
NC_001583.1	Human papillomavirus type 26
NC_001586.1	Human papillomavirus type 32
NC_001587.1	Human papillomavirus type 34
NC_001354.1	Human papillomavirus type 41
NC_001690.1	Human papillomavirus type 48
NC_001591.1	Human papillomavirus type 49
NC_001691.1	Human papillomavirus type 50
NC_001593.1	Human papillomavirus type 53
NC_001676.1	Human papillomavirus type 54
NC_001694.1	Human papillomavirus type 61
NC_001693.1	Human papillomavirus type 60
NC_001458.1	Human papillomavirus type 63
NC_010329.1	Human papillomavirus type 88
NC_004104.1	Human papillomavirus type 90
NC_004500.1	Human papillomavirus type 92
NC_005134.2	Human papillomavirus type 96
NC_008189.1	Human papillomavirus type 101
NC_008188.1	Human papillomavirus type 103
NC_012213.1	Human papillomavirus type 108
NC_012485.1	Human papillomavirus type 109
NC_012486.1	Human papillomavirus type 112
NC_013035.1	Human papillomavirus type 116
NC_014185.1	Human papillomavirus type 121
NC_016157.1	Human papillomavirus type 126
NC_014952.1	Human papillomavirus type 128
NC_014953.1	Human papillomavirus type 129
NC_014954.1	Human papillomavirus type 131
NC_014955.1	Human papillomavirus type 132
NC_014956.1	Human papillomavirus type 134
NC_017993.1	Human papillomavirus type 135
NC_017994.1	Human papillomavirus type 136
NC_017995.1	Human papillomavirus type 137
NC_017996.1	Human papillomavirus type 140
NC_017997.1	Human papillomavirus type 144
NC_021483.1	Human papillomavirus type 154
NC_019023.1	Human papillomavirus type 166
NC_003461.1	Human parainfluenza virus 1
NC_003443.1	Human parainfluenza virus 2
NC_001796.2	Human parainfluenza virus 3
NC_021928.1	Human parainfluenza virus 4
NC_001897.1	Human parechovirus
NC_007018.1	Human parvovirus 4
NC_000883.2	Human parvovirus B19
NC_007026.1	Human picobirnavirus RNA segment 1
NC_007027.1	Human picobirnavirus RNA segment 2
NC_020890.1	Human polyomavirus 12
NC_015150.1	Human polyomavirus 9
NC_001781.1	Human respiratory syncytial virus
NC_001490.1	Human rhinovirus 14
NC_001617.1	Human rhinovirus 89
NC_009996.1	Human rhinovirus C
NC_021545.1	Human rotavirus B strain Bang373 VP2 gene
NC_021544.1	Human rotavirus B strain Bang373 VP6 gene
NC_021548.1	Human rotavirus B strain Bang373 NSP2 gene
NC_021547.1	Human rotavirus B strain Bang373 NSP3 gene
NC_021550.1	Human rotavirus B strain Bang373 NSP4 gene
NC_021549.1	Human rotavirus B strain Bang373 NSP5 gene
NC_021546.1	Human rotavirus B strain Bang373 NSP1 gene
NC_021543.1	Human rotavirus B strain Bang373 VP4 gene
NC_021542.1	Human rotavirus B strain Bang373 VP7 gene
NC_021541.1	Human rotavirus B strain Bang373 VP1 mRNA
NC_021551.1	Human rotavirus B strain Bang373 VP3 mRNA
NC_001436.1	Human T-lymphotropic virus 1
NC_001488.1	Human T-lymphotropic virus 2
NC_011800.1	Human T-lymphotropic virus 4
NC_010810.1	Human TMEV-like cardiovirus
NC_018629.1	Ikoma lyssavirus
NC_004178.1	Infectious bursal disease virus segment A
NC_004179.1	Infectious bursal disease virus segment B
NC_001652.1	Infectious hematopoietic necrosis virus
NC_002190.2	Infectious hypodermal and hematopoietic necrosis virus
NC_001915.1	Infectious pancreatic necrosis virus segment A
NC_001916.1	Infectious pancreatic necrosis virus segment B
NC_003494.1	Infectious spleen and kidney necrosis virus
NC_007358.1	Influenza A virus (Goose/Guangdong/96(H5N1)) segment 2
NC_007362.1	Influenza A virus (Goose/Guangdong/96(H5N1)) segment 4
NC_007363.1	Influenza A virus (Goose/Guangdong/96(H5N1)) segment 7
NC_007361.1	Influenza A virus (Goose/Guangdong/96(H5N1)) segment 6
NC_007364.1	Influenza A virus (Goose/Guangdong/96(H5N1)) segment 8

Accession Number	Virus Name
NC_007357.1	Influenza A virus (Goose/Guangdong/96(H5N1)) segment 1
NC_007359.1	Influenza A virus (Goose/Guangdong/96(H5N1)) segment 3
NC_007360.1	Influenza A virus (Goose/Guangdong/96(H5N1)) segment 5
NC_004910.1	Influenza A virus (Hong Kong/1073/99(H9N2)) segment 1
NC_004911.1	Influenza A virus (Hong Kong/1073/99(H9N2)) segment 2
NC_004912.1	Influenza A virus (Hong Kong/1073/99(H9N2)) segment 3
NC_004908.1	Influenza A virus (Hong Kong/1073/99(H9N2)) segment 4
NC_004909.1	Influenza A virus (Hong Kong/1073/99(H9N2)) segment 6
NC_004907.1	Influenza A virus (Hong Kong/1073/99(H9N2)) segment 7
NC_004906.1	Influenza A virus (Hong Kong/1073/99(H9N2)) segment 8
NC_004905.2	Influenza A virus (Hong Kong/1073/99(H9N2)) segment 5
NC_007378.1	Influenza A virus (Korea/426/68(H2N2)) segment 1
NC_007375.1	Influenza A virus (Korea/426/68(H2N2)) segment 2
NC_007376.1	Influenza A virus (Korea/426/68(H2N2)) segment 3
NC_007374.1	Influenza A virus (Korea/426/68(H2N2)) segment 4
NC_007381.1	Influenza A virus (Korea/426/68(H2N2)) segment 5
NC_007382.1	Influenza A virus (Korea/426/68(H2N2)) segment 6
NC_007377.1	Influenza A virus (Korea/426/68(H2N2)) segment 7
NC_007380.1	Influenza A virus (Korea/426/68(H2N2)) segment 8
NC_007373.1	Influenza A virus (New York/392/04(H3N2)) segment 1
NC_007372.1	Influenza A virus (New York/392/04(H3N2)) segment 2
NC_007366.1	Influenza A virus (New York/392/04(H3N2)) segment 4
NC_007369.1	Influenza A virus (New York/392/04(H3N2)) segment 5
NC_007368.1	Influenza A virus (New York/392/04(H3N2)) segment 6
NC_007367.1	Influenza A virus (New York/392/04(H3N2)) segment 7
NC_007370.1	Influenza A virus (New York/392/04(H3N2)) segment 8
NC_007371.1	Influenza A virus (New York/392/04(H3N2)) segment 3
NC_002023.1	Influenza A virus (A/Puerto Rico/8/34(H1N1)) segment 1
NC_002021.1	Influenza A virus (A/Puerto Rico/8/34(H1N1)) segment 2
NC_002022.1	Influenza A virus (A/Puerto Rico/8/34(H1N1)) segment 3
NC_002017.1	Influenza A virus (A/Puerto Rico/8/34(H1N1)) segment 4
NC_002019.1	Influenza A virus (A/Puerto Rico/8/34(H1N1)) segment 5
NC_002018.1	Influenza A virus (A/Puerto Rico/8/34(H1N1)) segment 6
NC_002016.1	Influenza A virus (A/Puerto Rico/8/34(H1N1)) segment 7
NC_002020.1	Influenza A virus (A/Puerto Rico/8/34(H1N1)) segment 8
NC_002204.1	Influenza B virus RNA 1
NC_002207.1	Influenza B virus RNA 4
NC_002208.1	Influenza B virus RNA 5
NC_002209.1	Influenza B virus RNA 6
NC_002210.1	Influenza B virus RNA 7
NC_002211.1	Influenza B virus RNA 8
NC_002205.1	Influenza B virus RNA-2
NC_002206.1	Influenza B virus RNA-3
NC_006307.1	Influenza C virus (C/Ann Arbor/1/50) segment 1
NC_006308.1	Influenza C virus (C/Ann Arbor/1/50) segment 2
NC_006309.1	Influenza C virus (C/Ann Arbor/1/50) segment 3
NC_006310.1	Influenza C virus (C/Ann Arbor/1/50) segment 4
NC_006311.1	Influenza C virus (C/Ann Arbor/1/50) segment 5
NC_006312.1	Influenza C virus (C/Ann Arbor/1/50) segment 6
NC_006306.2	Influenza C virus (C/Ann Arbor/1/50) segment 7
NC_007906.1	Ippy virus segment L
NC_007905.1	Ippy virus segment S
NC_020809.1	Irkut virus
NC_020806.1	Isfahan virus N gene
NC_007454.1	J-virus
NC_001494.1	Jaagsiekte sheep retrovirus
NC_015123.1	Japanese eel endothelial cells-infecting virus
NC_001437.1	Japanese encephalitis virus
NC_001699.1	JC polyomavirus
NC_005080.1	Junin virus segment L
NC_005081.1	Junin virus segment S
NC_004210.1	Kadipiro virus segment 1
NC_004212.1	Kadipiro virus segment 2
NC_004213.1	Kadipiro virus segment 3
NC_004214.1	Kadipiro virus segment 4
NC_004215.1	Kadipiro virus segment 5
NC_004216.1	Kadipiro virus segment 6
NC_004209.1	Kadipiro virus segment 7
NC_004208.1	Kadipiro virus segment 8
NC_004207.1	Kadipiro virus segment 9
NC_004206.1	Kadipiro virus segment 10
NC_004205.1	Kadipiro virus segment 11
NC_004199.1	Kadipiro virus segment 12
NC_005064.1	Kamiti River virus
NC_006947.1	Karshi virus
NC_012533.1	Kedougou virus
NC_009238.1	KI polyomavirus
NC_021704.1	Koala retrovirus
NC_009029.2	Kokobera virus
NC_017714.1	Kotonkan virus
NC_004108.1	La Crosse virus segment L
NC_004109.1	La Crosse virus segment M
NC_004110.1	La Crosse virus segment S
NC_001639.1	Lactate dehydrogenase-elevating virus
NC_020807.1	Lagos bat virus
NC_001608.3	Lake Victoria marburgvirus
NC_003690.1	Langat virus
NC_004297.1	Lassa virus segment L
NC_004296.1	Lassa virus segment S
NC_010760.1	Latino virus segment L
NC_010758.1	Latino virus segment S
NC_007736.1	Liao ning virus segment 1
NC_007737.1	Liao ning virus segment 2
NC_007738.1	Liao ning virus segment 3
NC_007739.1	Liao ning virus segment 4
NC_007740.1	Liao ning virus segment 5
NC_007741.1	Liao ning virus segment 6
NC_007742.1	Liao ning virus segment 7
NC_007743.1	Liao ning virus segment 8
NC_007744.1	Liao ning virus segment 9
NC_007745.1	Liao ning virus segment 10
NC_007746.1	Liao ning virus segment 11
NC_007747.1	Liao ning virus segment 12
NC_003976.2	Ljungan virus
NC_016144.1	Lloviu virus
NC_021242.1	Lone Star virus isolate TMA 1381 segment L
NC_021243.1	Lone Star virus isolate TMA 1381 segment M
NC_021244.1	Lone Star virus isolate TMA 1381 segment S
NC_001809.1	Louping ill virus
NC_004713.1	LuIII virus
NC_012777.1	Lujo virus segment L
NC_012776.1	Lujo virus segment S
NC_003027.1	Lumpy skin disease virus NI-2490
NC_016153.1	Luna virus segment L
NC_016152.1	Luna virus segment S
NC_018711.1	Lunk virus NKS-1 segment L
NC_018710.1	Lunk virus NKS-1 segment S
NC_005902.1	Lymphocystis disease virus
NC_001824.1	Lymphocystis disease virus 1
NC_004291.1	Lymphocytic choriomeningitis virus segment L
NC_004294.1	Lymphocytic choriomeningitis virus segment S
NC_015691.1	Macaca fascicularis papillomavirus type 2
NC_019851.1	Macaca fascicularis polyomavirus 1
NC_010819.1	Macaque simian foamy virus
NC_005079.1	Machupo virus segment L
NC_005078.1	Machupo virus segment S
NC_016993.1	Magpie---robin coronavirus
NC_013225.1	Mammalian orthoreovirus 3 segment L1
NC_004282.4	Mammalian orthoreovirus 3 segment L1
NC_013226.1	Mammalian orthoreovirus 3 segment L2
NC_004275.1	Mammalian orthoreovirus 3 segment L2
NC_013229.1	Mammalian orthoreovirus 3 segment L3
NC_004274.1	Mammalian orthoreovirus 3 segment L3
NC_013227.1	Mammalian orthoreovirus 3 segment M1
NC_004280.1	Mammalian orthoreovirus 3 segment M1
NC_013228.1	Mammalian orthoreovirus 3 segment M2
NC_004278.1	Mammalian orthoreovirus 3 segment M2
NC_013230.1	Mammalian orthoreovirus 3 segment M3
NC_004281.1	Mammalian orthoreovirus 3 segment M3
NC_013231.1	Mammalian orthoreovirus 3 segment S1
NC_004277.1	Mammalian orthoreovirus 3 segment S1
NC_013232.1	Mammalian orthoreovirus 3 segment S2
NC_004279.1	Mammalian orthoreovirus 3 segment S2
NC_013233.1	Mammalian orthoreovirus 3 segment S3
NC_004283.1	Mammalian orthoreovirus 3 segment S3
NC_013234.1	Mammalian orthoreovirus 3 segment S4
NC_004276.1	Mammalian orthoreovirus 3 segment S4
NC_009489.1	Mapuera virus
NC_001550.1	Mason---Pfizer monkey virus
NC_008519.1	Mastomys coucha papillomavirus 2
NC_001605.1	Mastomys natalensis papillomavirus
NC_003417.1	Mayaro virus
NC_001498.1	Measles virus
NC_020439.1	Melaka orthoreovirus segment L1
NC_020447.1	Melaka orthoreovirus segment L2
NC_020440.1	Melaka orthoreovirus segment L3
NC_020441.1	Melaka orthoreovirus segment M1
NC_020442.1	Melaka orthoreovirus segment M2

Accession Number	Virus Name
NC_020443.1	Melaka orthoreovirus segment M3
NC_020448.1	Melaka orthoreovirus segment S1
NC_020444.1	Melaka orthoreovirus segment S2
NC_020445.1	Melaka orthoreovirus segment S3
NC_020446.1	Melaka orthoreovirus segment S4
NC_007620.1	Menangle virus
NC_020900.1	Meno virus
NC_010277.1	Merkel cell polyomavirus
NC_012702.1	Midway virus
NC_020069.1	Miniopterus polyomavirus isolate KY369
NC_004579.1	Mink astrovirus
NC_019712.1	Mink calicivirus
NC_007904.1	Mobala virus segment L
NC_007903.1	Mobala virus segment S
NC_003635.1	Modoc virus
NC_006429.1	Mokola virus
NC_001731.1	Molluscum contagiosum virus subtype 1
NC_001501.1	Moloney murine leukemia virus
NC_001502.1	Moloney murine sarcoma virus
NC_003310.1	Monkeypox virus Zaire-96-I-16
NC_004119.1	Montana myotis leukoencephalitis virus
NC_006572.1	Mopeia Lassa reassortant 29 segment L
NC_006573.1	Mopeia Lassa reassortant 29 segment S
NC_006574.1	Mopeia virus AN20410 segment L
NC_006575.1	Mopeia virus AN20410 segment S
NC_016013.1	Morelia spilota papillomavirus 1
NC_013058.1	Morogoro virus segment L
NC_013057.1	Morogoro virus segment S
NC_021069.1	Mosquito flavivirus isolate LSFlaviV-A20-09
NC_005339.1	Mossman virus
NC_015935.1	Mouse astrovirus M-52/USA/2008
NC_015936.1	Mouse kobuvirus M-5/USA/2010
NC_001503.1	Mouse mammary tumor virus
NC_001630.1	Mouse parvovirus 1
NC_008186.1	Mouse parvovirus 2
NC_008185.1	Mouse parvovirus 3
NC_011619.1	Mouse parvovirus 4
NC_011618.1	Mouse parvovirus 5
NC_005053.1	Mulard duck circovirus
NC_002200.1	Mumps virus
NC_011550.1	Munia coronavirus HKU13-3514
NC_014899.1	Murine adenovirus 2
NC_012584.1	Murine adenovirus 3
NC_000942.1	Murine adenovirus A
AC_000012.1	Murine adenovirus A
NC_018702.1	Murine astrovirus
NC_001846.1	Murine hepatitis virus strain A59
NC_006852.1	Murine hepatitis virus strain JHM
NC_008311.1	Murine norovirus 1
NC_001506.1	Murine osteosarcoma virus
NC_001505.2	Murine pneumotropic virus
NC_001515.1	Murine polyomavirus
NC_001702.1	Murine type C retrovirus
NC_000943.1	Murray Valley encephalitis virus
NC_014326.1	Mus musculus papillomavirus type 1
NC_006561.1	Muscovy duck circovirus
NC_006147.2	Muscovy duck parvovirus
NC_018102.1	MW polyomavirus
NC_011310.1	Myotis polyomavirus VM---2008
NC_001132.2	Myxoma virus
NC_017937.1	Nariva virus
NC_016963.1	New World begomovirus associated satellite DNA
NC_007916.1	Newbury agent 1 virus
NC_002617.1	Newcastle disease virus B1
NC_013955.1	Ngaingan virus
NC_016994.1	Night-heron coronavirus HKU19
NC_008030.1	Nile crocodilepox virus
NC_002728.1	Nipah virus
NC_001959.2	Norwalk virus
NC_018705.1	Ntaya virus isolate IPDIA
NC_012703.1	Nyamanini virus
NC_001512.1	O'nyong-nyong virus
NC_008582.1	Old World harvest mouse papillomavirus
NC_010250.1	Oliveros virus segment L
NC_010248.1	Oliveros virus segment S
NC_005062.1	Omsk hemorrhagic fever virus
NC_013439.1	Orangutan polyomavirus
NC_005336.1	Orf virus
NC_005776.1	Oropouche virus segment L
NC_005775.1	Oropouche virus segment M
NC_005777.1	Oropouche virus segment S
NC_020071.1	Otomops polyomavirus 1 isolate KY157
NC_020066.1	Otomops polyomavirus 2 isolate KY156
NC_002513.1	Ovine adenovirus A
AC_000001.1	Ovine adenovirus A
NC_004037.2	Ovine adenovirus D
NC_002469.1	Ovine astrovirus
NC_007015.1	Ovine enzootic nasal tumour virus
NC_001511.1	Ovine lentivirus
NC_001789.1	Ovine papillomavirus 1
NC_019858.1	Pan troglodytes schweinfurthii polyomavirus 2
NC_019855.1	Pan troglodytes verus polyomavirus 3
NC_019856.1	Pan troglodytes verus polyomavirus 4
NC_019857.1	Pan troglodytes verus polyomavirus 5
NC_017716.1	Papio hamadryas papillomavirus type 1
NC_006430.1	Parainfluenza virus 5
NC_010761.1	Parana virus segment L
NC_010756.1	Parana virus segment S (small)
NC_016561.1	Parrot hepatitis B virus
NC_001358.1	Parvovirus H1
NC_022089.1	Parvovirus NIH-CQV putative 15kDa protein
NC_020803.1	Perch rhabdovirus isolate PRV nucleocapsid (N)
NC_007748.1	Peruvian horse sickness virus segment 1
NC_007749.1	Peruvian horse sickness virus segment 2
NC_007750.1	Peruvian horse sickness virus segment 3
NC_007751.1	Peruvian horse sickness virus segment 4
NC_007755.1	Peruvian horse sickness virus segment 5
NC_007752.1	Peruvian horse sickness virus segment 6
NC_007756.1	Peruvian horse sickness virus segment 7
NC_007754.1	Peruvian horse sickness virus segment 8
NC_007753.1	Peruvian horse sickness virus segment 9
NC_007757.1	Peruvian horse sickness virus segment 10
NC_006383.2	Peste-des-petits-ruminants virus
NC_003678.1	Pestivirus Giraffe-1
NC_018713.1	Pestivirus strain Aydin/04-TR
NC_018074.1	Phocoena phocoena papillomavirus 1
NC_018075.1	Phocoena phocoena papillomavirus 2
NC_018076.1	Phocoena phocoena papillomavirus 4
NC_003348.1	Phocoena spinipinnis papillomavirus
NC_006439.1	Pichinde virus L RNA
NC_006447.1	Pichinde virus
NC_017916.1	Pig stool associated circular ssDNA virus GER2011
NC_015626.1	Pigeon picornavirus B
NC_019850.1	Piliocolobus rufomitratus polyomavirus 1
NC_005897.1	Pirital virus segment L
NC_005894.1	Pirital virus segment S
NC_015639.1	Piscine myocarditis virus AL V---708
NC_006579.1	Pneumonia virus of mice J3666
NC_002058.3	Poliovirus
NC_014406.1	Polyomavirus HPyV6
NC_014407.1	Polyomavirus HPyV7
NC_005869.1	Porcine adenovirus A
NC_002702.1	Porcine adenovirus C
AC_000009.1	Porcine adenovirus C
NC_018617.1	Porcine associated stool circular virus
NC_019494.1	Porcine astrovirus 3 isolate US-MO123
NC_016031.1	Porcine bocavirus 3
NC_016032.1	Porcine bocavirus 4-1
NC_016647.1	Porcine bocavirus 5/JS677
NC_001792.2	Porcine circovirus 1
NC_005148.1	Porcine circovirus 2
NC_013774.1	Porcine circovirus type 1/2a
NC_016990.1	Porcine coronavirus HKU15
NC_022233.1	Porcine cytomegalovirus strain BJ09
NC_003059.1	Porcine endogenous retrovirus E
NC_000940.1	Porcine enteric sapovirus
NC_004441.1	Porcine enterovirus B
NC_003436.1	Porcine epidemic diarrhea virus
NC_007732.1	Porcine hemagglutinating encephalomyelitis virus
NC_016769.1	Porcine kobuvirus SH-W-CHN/2010/China
NC_011829.1	Porcine kobuvirus S-1-HUN/2007/Hungary
NC_022104.1	Porcine partetravirus strain FMV10---1437266
NC_014665.1	Porcine parvovirus 4
NC_001718.1	Porcine parvovirus
NC_001961.1	Porcine respiratory and reproductive syndrome virus
NC_009640.1	Porcine rubulavirus
NC_003987.1	Porcine sapelovirus 1
NC_021203.1	Porcine stool-associated circular virus 2

Accession Number	Virus Name
NC_021204.1	Porcine stool-associated circular virus 3
NC_003985.1	Porcine teschovirus 1
NC_008714.1	Possum enterovirus W1
NC_008715.1	Possum enterovirus W6
NC_003687.1	Powassan virus
NC_007150.1	Procyon lotor papillomavirus 1
NC_013804.1	Pseudocowpox virus
NC_003973.1	Psittacus erithacus timneh papillomavirus
NC_020070.1	Pteronotus polyomavirus
NC_005225.1	Puumala virus segment L
NC_005223.1	Puumala virus segment M
NC_005224.1	Puumala virus segment S
NC_016403.1	Quail picornavirus
NC_011704.1	Rabbit calicivirus Australia 1 MIC-07
NC_017083.1	Rabbit coronavirus HKU14
NC_001266.1	Rabbit fibroma virus
NC_001543.1	Rabbit hemorrhagic disease virus-FRG
NC_002232.1	Rabbit oral papillomavirus
NC_008580.1	Rabbit vesivirus
NC_001542.1	Rabies virus
NC_021930.1	Rangifer tarandus papillomavirus 2
NC_015455.1	Raptor adenovirus A
NC_012936.1	Rat coronavirus Parker
NC_001819.1	Rauscher murine leukemia virus
NC_008375.1	Raven circovirus
NC_001803.1	Respiratory syncytial virus
NC_004161.1	Reston ebolavirus
NC_006934.1	Reticuloendotheliosis virus
NC_001678.1	Rhesus monkey papillomavirus
NC_014397.1	Rift Valley fever virus segment L
NC_014396.1	Rift Valley fever virus segment M
NC_014395.1	Rift Valley fever virus segment S
NC_006296.2	Rinderpest virus (strain Kabete O)
NC_003675.1	Rio Bravo virus
NC_021153.1	Rodent hepacivirus isolate RHV---339
NC_021154.1	Rodent pegivirus isolate CC61
NC_001544.1	Ross River virus
NC_005888.1	Ross's goose hepatitis B virus
NC_011507.2	Rotavirus A segment 1
NC_011506.2	Rotavirus A segment 2
NC_011508.2	Rotavirus A segment 3
NC_011510.2	Rotavirus A segment 4
NC_011500.2	Rotavirus A segment 5
NC_011509.2	Rotavirus A segment 6
NC_011501.2	Rotavirus A segment 7
NC_011502.2	Rotavirus A segment 8
NC_011503.2	Rotavirus A segment 9
NC_011504.2	Rotavirus A segment 10
NC_011505.2	Rotavirus A segment 11
NC_007547.1	Rotavirus C segment 1
NC_007546.1	Rotavirus C segment 2
NC_007572.1	Rotavirus C segment 3
NC_007574.1	Rotavirus C segment 4
NC_007570.1	Rotavirus C segment 5
NC_007543.1	Rotavirus C segment 6
NC_007544.1	Rotavirus C segment 7
NC_007571.1	Rotavirus C segment 8
NC_007545.1	Rotavirus C segment 9
NC_007569.1	Rotavirus C segment 10
NC_007573.1	Rotavirus C segment 11
NC_014511.1	Rotavirus D chicken/05V0049/DEU/2005 segment 1
NC_014512.1	Rotavirus D chicken/05V0049/DEU/2005 segment 2
NC_014513.1	Rotavirus D chicken/05V0049/DEU/2005 segment 3
NC_014514.1	Rotavirus D chicken/05V0049/DEU/2005 segment 4
NC_014515.1	Rotavirus D chicken/05V0049/DEU/2005 segment 5
NC_014516.1	Rotavirus D chicken/05V0049/DEU/2005 segment 6
NC_014517.1	Rotavirus D chicken/05V0049/DEU/2005 segment 7
NC_014518.1	Rotavirus D chicken/05V0049/DEU/2005 segment 8
NC_014519.1	Rotavirus D chicken/05V0049/DEU/2005 segment 9
NC_014520.1	Rotavirus D chicken/05V0049/DEU/2005 segment 10
NC_014521.1	Rotavirus D chicken/05V0049/DEU/2005 segment 11
NC_021625.1	Rotavirus F chicken/03V0568/DEU/2003 segment 1
NC_021626.1	Rotavirus F chicken/03V0568/DEU/2003 segment 2
NC_021631.1	Rotavirus F chicken/03V0568/DEU/2003 segment 3
NC_021630.1	Rotavirus F chicken/03V0568/DEU/2003 segment 4
NC_021632.1	Rotavirus F chicken/03V0568/DEU/2003 segment 5
NC_021635.1	Rotavirus F chicken/03V0568/DEU/2003 segment 6
NC_021633.1	Rotavirus F chicken/03V0568/DEU/2003 segment 7
NC_021628.1	Rotavirus F chicken/03V0568/DEU/2003 segment 8
NC_021627.1	Rotavirus F chicken/03V0568/DEU/2003 segment 9
NC_021629.1	Rotavirus F chicken/03V0568/DEU/2003 segment 10
NC_021634.1	Rotavirus F chicken/03V0568/DEU/2003 segment 11
NC_021590.1	Rotavirus G chicken/03V0567/DEU/2003 segment 1
NC_021580.1	Rotavirus G chicken/03V0567/DEU/2003 segment 2
NC_021581.1	Rotavirus G chicken/03V0567/DEU/2003 segment 3
NC_021589.1	Rotavirus G chicken/03V0567/DEU/2003 segment 4
NC_021583.1	Rotavirus G chicken/03V0567/DEU/2003 segment 5
NC_021588.1	Rotavirus G chicken/03V0567/DEU/2003 segment 6
NC_021585.1	Rotavirus G chicken/03V0567/DEU/2003 segment 7
NC_021584.1	Rotavirus G chicken/03V0567/DEU/2003 segment 8
NC_021582.1	Rotavirus G chicken/03V0567/DEU/2003 segment 9
NC_021586.1	Rotavirus G chicken/03V0567/DEU/2003 segment 10
NC_021587.1	Rotavirus G chicken/03V0567/DEU/2003 segment 11
NC_001407.1	Rous sarcoma virus
NC_008298.1	Rousettus aegyptiacus papillomavirus type 1
NC_018871.1	Rousettus bat coronavirus HKU10
NC_001545.2	Rubella virus
NC_006313.1	Sabia virus segment L
NC_006317.1	Sabia virus
NC_009448.2	Saffold virus
NC_012986.1	Salivirus A isolate 02394-01
NC_012957.1	Salivirus NG-J1
NC_003930.1	Salmon pancreas disease virus
NC_015412.1	Sandfly Sicilian Turkey virus segment L
NC_015411.1	Sandfly Sicilian Turkey virus segment M
NC_015413.1	Sandfly Sicilian Turkey virus segment S
NC_004718.3	SARS coronavirus
NC_018461.1	Sathuperi virus segment L
NC_018466.1	Sathuperi virus segment M
NC_018462.1	Sathuperi virus segment S
NC_011659.1	Schlumbergera virus X
NC_009657.1	Scotophilus bat coronavirus 512
NC_012094.1	Sea turtle tornovirus 1
NC_015212.1	Seal anellovirus TFFN/USA/2006
NC_009891.1	Seal picornavirus type 1
NC_021482.1	Sebokele virus 1 ORF1 gene for polyprotein
NC_003215.1	Semliki forest virus
NC_001552.1	Sendai virus
NC_011349.1	Seneca valley virus
NC_005237.1	Seoul virus segment M
NC_005236.1	Seoul virus strain 80-39 segment S
NC_005238.1	Seoul virus strain Seoul 80-39 clone 1
NC_008719.1	Sepik virus
NC_018136.1	SFTS virus HB29 segment L
NC_018138.1	SFTS virus HB29 segment M
NC_018137.1	SFTS virus HB29 segment S
NC_018463.1	Shamonda virus segment L
NC_018467.1	Shamonda virus segment M
NC_018464.1	Shamonda virus segment S
NC_004002.1	Sheeppox virus 17077-99
NC_005890.1	Sheldgoose hepatitis B virus
NC_018476.1	Simbu virus segment L
NC_018478.1	Simbu virus segment M
NC_018477.1	Simbu virus segment S
NC_006879.1	Simian adenovirus 1
NC_022266.1	Simian adenovirus 18
NC_020485.1	Simian adenovirus 20 strain ATCC VR-541
AC_000010.1	Simian adenovirus 21
AC_000011.1	Simian adenovirus 25
NC_006144.1	Simian adenovirus 3
NC_015225.1	Simian adenovirus 49
NC_021168.1	Simian adenovirus C isolate BaAdV-2
NC_003988.1	Simian enterovirus A
NC_010820.1	Simian foamy virus 3
NC_001364.1	Simian foamy virus
NC_003092.1	Simian hemorrhagic fever virus
NC_004455.1	Simian immunodeficiency virus SIV-mnd 2
NC_001549.1	Simian immunodeficiency virus
NC_014474.1	Simian retrovirus 4
NC_004451.1	Simian sapelovirus 1
NC_011546.1	Simian T-cell lymphotropic virus 6
NC_000858.1	Simian T-lymphotropic virus 1
NC_001815.1	Simian T-lymphotropic virus 2
NC_003323.1	Simian T-lymphotropic virus 3
NC_007611.1	Simian virus 12
NC_001669.1	Simian virus 40
NC_006428.1	Simian virus 41
NC_005217.1	Sin Nombre virus segment L

Accession Number	Virus Name
NC_005215.1	Sin Nombre virus segment M
NC_005216.1	Sin Nombre virus segment S
NC_001547.1	Sindbis virus
NC_006549.1	Singapore grouper iridovirus
NC_008514.1	Siniperca chuatsi rhabdovirus
NC_003433.1	Sleeping disease virus
NC_007013.1	Small anellovirus 1
NC_007014.1	Small anellovirus 2
NC_009989.1	Snake adenovirus
NC_006148.1	Snake parvovirus 1
NC_001724.1	Snakehead retrovirus
NC_000903.1	Snakehead rhabdovirus
NC_005950.1	Snow goose hepatitis B virus
NC_016437.1	South polar skua adenovirus-1
NC_016960.1	Southern elephant seal virus
NC_016992.1	Sparrow coronavirus HKU17
NC_001500.1	Spleen focus-forming virus
NC_002803.1	Spring viraemia of carp virus
NC_009951.1	Squirrel monkey polyomavirus
NC_001514.1	Squirrel monkey retrovirus - HLB
NC_007580.2	St. Louis encephalitis virus
NC_008033.1	Starling circovirus
NC_011050.1	Steller sea lion vesivirus
NC_020106.1	STL polyomavirus strain MA138
NC_006432.1	Sudan ebolavirus
NC_011280.1	Sus scrofa papillomavirus type 1
NC_018226.1	Swine pasivirus 1
NC_003389.1	Swinepox virus
NC_004292.1	Tacaribe virus segment L
NC_004293.1	Tacaribe virus segment S
NC_003996.1	Tamana bat virus
NC_010702.1	Tamiami virus segment L
NC_010701.1	Tamiami virus segment S (small)
NC_009888.1	Tanapox virus
NC_008291.1	Taterapox virus
NC_015843.2	Tembusu virus strain JS804
NC_001366.1	Theilovirus
NC_006495.1	Thogoto virus
NC_006496.1	Thogoto virus
NC_006508.1	Thogoto virus segment 1
NC_006507.1	Thogoto virus segment 5
NC_006504.1	Thogoto virus segment 6
NC_006506.1	Thogoto virus
NC_010707.1	Thottapalayam virus segment L
NC_010708.1	Thottapalayam virus segment M
NC_010704.1	Thottapalayam virus segment S
NC_011549.1	Thrush coronavirus HKU12-600
NC_020804.1	Tibrogargan virus strain CS132
NC_001672.1	Tick---borne encephalitis virus
NC_013460.1	Tiger puffer nervous necrosis virus RNA 1
NC_013461.1	Tiger puffer nervous necrosis virus RNA2
NC_004074.1	Tioman virus
NC_020487.1	Titi monkey adenovirus ECC-2011
NC_014071.1	Torque teno canis virus
NC_014087.1	Torque teno douroucouli virus
NC_014072.1	Torque teno felis virus
NC_009225.1	Torque teno midi virus 1
NC_014093.1	Torque teno midi virus 2
NC_014097.1	Torque teno mini virus 1
NC_014086.1	Torque teno mini virus 2
NC_014088.1	Torque teno mini virus 3
NC_014090.1	Torque teno mini virus 4
NC_014089.1	Torque teno mini virus 5
NC_014095.1	Torque teno mini virus 6
NC_014082.1	Torque teno mini virus 7
NC_014068.1	Torque teno mini virus 8
NC_002195.1	Torque teno mini virus 9
NC_014070.1	Torque teno sus virus 1
NC_014092.2	Torque teno sus virus k2 isolate 2p
NC_014085.1	Torque teno tamarin virus
NC_002076.2	Torque teno virus 1
NC_014076.1	Torque teno virus 10
NC_014075.1	Torque teno virus 12
NC_014077.1	Torque teno virus 14
NC_014096.1	Torque teno virus 15
NC_014091.1	Torque teno virus 16
NC_014078.1	Torque teno virus 19
NC_014480.2	Torque teno virus 2
NC_014083.1	Torque teno virus 25
NC_014079.1	Torque teno virus 26
NC_014074.1	Torque teno virus 27
NC_014073.1	Torque teno virus 28
NC_014081.1	Torque teno virus 3
NC_014069.1	Torque teno virus 4
NC_014094.1	Torque teno virus 6
NC_014080.1	Torque teno virus 7
NC_014084.1	Torque teno virus 8
NC_015783.1	Torque teno virus
NC_006319.1	Toscana virus segment L
NC_006320.1	Toscana virus segment M
NC_006318.1	Toscana virus segment S
NC_004453.1	Tree shrew adenovirus
NC_006563.1	Trichechus manatus latirostris papillomavirus 1
NC_016898.1	Trichechus manatus latirostris papillomavirus 2
NC_014361.1	Trichodysplasia spinulosa-associated polyomavirus
NC_020498.1	TTV-like mini virus isolate TTMV_LY1
NC_005226.1	Tula virus segment L
NC_005228.1	Tula virus segment M
NC_005227.2	Tula virus segment S
NC_002199.1	Tupaia paramyxovirus
NC_007020.1	Tupaia virus
NC_014411.1	Turdivirus 1
NC_014412.1	Turdivirus 2
NC_014413.1	Turdivirus 3
NC_014564.2	Turkey adenovirus 1
NC_001958.1	Turkey adenovirus A
AC_000016.1	Turkey adenovirus A
NC_005790.1	Turkey astrovirus 2
NC_002470.1	Turkey astrovirus
NC_010800.1	Turkey coronavirus
NC_018400.1	Turkey gallivirus
NC_021201.1	Turkey hepatitis virus 2993D polyprotein gene
NC_008184.1	Tursiops truncatus papillomavirus 2
NC_011109.1	Tursiops truncatus papillomavirus type 1
NC_001618.1	UR2 sarcoma virus
NC_010739.1	Ursus maritimus papillomavirus 1
NC_006551.1	Usutu virus
NC_006998.1	Vaccinia virus
NC_001611.1	Variola virus
NC_001449.1	Venezuelan equine encephalitis virus
NC_019844.1	Vervet monkey polyomavirus 1 DNA
NC_002551.1	Vesicular exanthema of swine virus
NC_001560.1	Vesicular stomatitis Indiana virus
NC_000855.1	Viral hemorrhagic septicemia virus Fil3
NC_001452.1	Visna/Maedi virus
NC_001867.1	Walleye dermal sarcoma virus
NC_004541.1	Walrus calicivirus
NC_012735.1	Wesselsbron virus
NC_009942.1	West Nile virus
NC_001563.2	West Nile virus
NC_003908.1	Western equine encephalomyelitis virus
NC_016961.1	Whataroa virus
NC_016991.1	White-eye coronavirus HKU16
NC_010703.1	Whitewater Arroyo virus segment L
NC_010700.1	Whitewater Arroyo virus segment S
NC_016995.1	Wigeon coronavirus HKU20
NC_011639.1	Wongabel virus
NC_004107.1	Woodchuck hepatitis virus
NC_009424.4	Woolly monkey sarcoma virus
NC_009539.1	WU Polyomavirus
NC_008094.1	Y73 sarcoma virus
NC_005179.1	Yaba monkey tumor virus
NC_002642.1	Yaba-like disease virus
NC_002031.1	Yellow fever virus
NC_004168.1	Yellowtail ascites virus segment A
NC_004176.1	Yellowtail ascites virus segment B
NC_015960.1	Yoka poxvirus
NC_005039.1	Yokose virus
NC_007656.1	Yunnan orbivirus segment 1
NC_007665.1	Yunnan orbivirus segment 10
NC_007657.1	Yunnan orbivirus segment 2
NC_007658.1	Yunnan orbivirus segment 3
NC_007659.1	Yunnan orbivirus segment 4
NC_007660.1	Yunnan orbivirus segment 5
NC_007661.1	Yunnan orbivirus segment 6
NC_007662.1	Yunnan orbivirus segment 7
NC_007663.1	Yunnan orbivirus segment 8
NC_007664.1	Yunnan orbivirus segment 9
NC_015325.1	Zalophus californianus papillomavirus 1
NC_012532.1	Zika virus

Appendix 2. Josh Quick's low input native barcoding library preparation for MinION sequencing protocol

- Dilute 100 ng pooled amplicons into 30 ul for each sample
- Add 4.2 ul Ultra II buffer and 1.8 ul Ultra II enzyme mix (0.6x reaction size)
- Incubate at 20°C for 5 minutes and 65°C for 5 minutes
- Perform SPRI clean-up and elute in 22.5 ul
- Continue without quantification
- Perform native barcode ligation as per protocol
- Heat denature ligase at 65°C for 10 minutes
- Combine all samples giving a total of 600 ul
- Perform 1x SPRI clean-up and elute in 58 ul
- Complete adapter ligation and pull-down as per protocol
- Quantify library add 80-100ng total into MinION flowcell

Available at: http://www.zibraproject.org/blog/protocol-low-input-native-barcoding-protocol/