Advanced Sequencing Technologies Applied to Human Cytomegalovirus

By

Karman Kathy Li

Submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy (PhD), School of Infection and Immunity, College of Medical, Veterinary & Life Sciences,

University of Glasgow

September 2023

Abstract

The betaherpesvirus human cytomegalovirus (HCMV) is a ubiquitous viral pathogen. It is the most common cause of congenital infection in infants and of opportunistic infections in immunocompromised patients worldwide. The large double-stranded DNA genome of HCMV (236 kb) contains several genes that exhibit a high degree of variation among strains within an otherwise highly conserved sequence. These hypervariable genes encode immune escape, tropism or regulatory factors that may affect virulence. Variation arising from these genes and from an evolutionary history of recombination between strains has been hypothesised to be linked to disease severity. To investigate this, the HCMV genome has been scrutinised in detail over the years using a variety of molecular techniques, most looking only at one or a few of these genes at a time. The advent of high-throughput sequencing (HTS) technology 20 years ago then started to enable more in-depth whole-genome analyses. My study extends this field by using both HTS and the more recently developed long-read nanopore technology to determine HCMV genome sequences directly from clinical samples. Firstly, I used an Illumina HTS pipeline to sequence HCMV strains directly from formalin-fixed, paraffin-embedded (FFPE) tissues. FFPE samples are a valuable repository for the study of relatively rare diseases, such as congenital HCMV (cCMV). However, formalin fixation induces DNA fragmentation and cross-linking, making this a challenging sample type for DNA sequencing. I successfully sequenced five whole HCMV genomes from FFPE tissues. Next, I developed a pipeline utilising the single-molecule, long-read sequencer from Oxford Nanopore Technologies (ONT) to sequence HCMV initially from high-titre cell-cultured laboratory strains and then from clinical samples with high HCMV loads. Finally, I utilised a direct RNA sequencing protocol with the ONT sequencer to characterise novel HCMV transcripts produced during infection in cell culture, demonstrating the existence of transcript isoforms with multiple splice sites. Overall, my findings demonstrate how advanced sequencing technologies can be used to characterise the genome and transcriptome of a large DNA virus, and will facilitate future studies on HCMV prognostic factors, novel antiviral targets and vaccine development.

Table of Contents

List of Tables

List of Figures

Preface

The past decade has seen an acceleration in the technologies available for high-throughput and rapid sequencing. The cost of sequencing per base has commensurately fallen to the point where it is now within the realms of reality for sequencing, as real-time PCR had in the 1990s, to revolutionise the landscape of infectious diseases. HCMV was one of the first pathogens to have its entire genome sequenced and published just over 30 years ago, and there are now approaching 300 published genome sequences available.

I commenced this study at the end of 2017, when nanopore technology was just emerging onto the sequencing scene. In the intervening years, we experienced first-hand the utility of rapid, public health-driven sequencing during an unprecedented global pandemic. First and foremost, this emphasised the importance of close collaborations between research institutes and health protection agencies to feed rapidly into policymaking. One of legacies to emerge from this crisis has been the implementation of sequencing into many clinical diagnostic laboratories around the UK. The sequencing infrastructure, including expertise and data-sharing, will pave the way for increased sequencing of other pathogens, including that of HCMV. I am hopeful that this will translate into practical measures to ameliorate any form of HCMV disease experienced by the millions of children and immunosuppressed patients worldwide.

Acknowledgement

I am indebted to the expertise of both my supervisors in my journey as a student. Professor Andrew Davison for his extensive knowledge and expertise and for his contributions to the field of HCMV. Thank you for your eternal patience and meticulous nature – your critical input is offered with humility and humour which is much appreciated. My sincere thanks to Richard Orton, not only for explaining difficult concepts with grace, but for his endless supply of encouragement.

Practically, this thesis would not have been possible without the keen eyes of Andrew Davison, who performed the manual checking of the genome sequences and the final stages of genome annotations, and of Richard Orton who helped extensively with the bioinformatics and writing of all custom scripts. Both made the GenBank deposition of the HCMV genomes produced from this study.

I would also like to express my gratitude for my colleagues and friends throughout these years.

To my family, thank you for your love, understanding, patience, and support.

Author's Declaration

I certify that the thesis presented here for examination for a PhD degree of the University of Glasgow is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it) and that the thesis has not been edited by a third party beyond what is permitted by the University's PGR Code of Practice.

I declare that the thesis does not include work forming part of a thesis presented successfully for another degree. I declare that this thesis has been produced in accordance with the University of Glasgow's Code of Good Practice in Research.

Abbreviations

BAC	Bacterial artificial chromosome
CCMV	Chimpanzee cytomegalovirus
CF	Complement fixation
cCMV	Congenital CMV
CHX	Cycloheximide
CID	Cytomegalic inclusion disease
CPE	Cytopathic effect
Daxx	Death domain-associated protein
DE	Delayed early
DC	Dendritic cell
DNA	Deoxyribonucleic acid
dRNA-Seq	Direct RNA sequencing
DIN	DNA integrity number
dsDNA	Double-stranded DNA
ER	Endoplasmic reticulum
EGFR	Epithelial growth factor receptor
FFPE	Formalin-fixed, paraffin-embedded
GCV	Ganciclovir
gB	Glycoprotein B
gH	Glycoprotein H
gL	Glycoprotein L
gO	Glycoprotein O
GvHD	Graft-versus-host disease
G+C	Guanine plus cytosine
HPC	Haemopoietic progenitor cell
HSPG	Heparan sulphate proteoglycan
HSV1	Herpes simplex virus type 1
HSV2	Herpes simplex virus type 2
HTS	High-throughput sequencing
HDAC	Histone deacetylase
HCMV	Human cytomegalovirus
HFFF2	Human foetal foreskin fibroblast 2
HLA	Human leukocyte antigen
IE	Immediate-early
IM	Infectious mononucleosis
ISG	interferon-stimulated gene
IR_L	Internal repeat long
IR_S	Internal repeat short
kb	Kilobase/kilobase pair
LUNA	Latency-associated gene product
oriLyt	Lytic origin of DNA replication
MCP	Major capsid protein
MHC	Major histocompatibility complex

mRDA	Medaka-polished RDA
HLA-DR	MHC class II cell surface receptor
mCP	Minor capsid protein
mCP-BP	Minor capsid-binding protein
MDA	Multiple-displacement amplification
MOI	Multiplicity of infection
NIEP	Non-infectious enveloped particle
ND10	Nuclear domain 10
NFW	Nuclease-free water
ORF	Open reading frame
OLC	Overlap/layout/consensus
PFA	Phosphonoformate
PFU	Plaque forming unit
PDGFR	Platelet-derived growth factor receptor
PCR	Polymerase chain reaction
PML	Promyelocytic leukaemia protein
qPCR	Quantitative real-time PCR
RDA	Reference-dependent assembly
RIA	Reference-independent assembly
RFLP	Restriction fragment length polymorphism
Rb	Retinoblastoma protein
RNA-Seq	RNA sequencing
DSS	RNA-Seq dataset from Gatherer *et al.*
DSG	RNA-Seq dataset from Stern-Ginossar *et al.*
SNHL	Sensorineural hearing loss
ssDNA	Single-strand DNA
SPRI	Solid-phase reversible immobilisation
TR_L	Terminal repeat long
TR_S	Terminal repeat short
TES	Transcriptional end site
TSS	Transcriptional start site
UDPS	Ultra-deep pyrosequencing
U_L	Unique long
U_S	Unique short
VZV	Varicella-zoster virus
vICA	Viral inhibitor of caspase-8 activation
WGA	Whole-genome amplification

1 Introduction

1.1 Human orthoherpesviruses

The family *Orthoherpesviridae* is a large group of double-stranded DNA (dsDNA) viruses that infect mammals, birds and reptiles with high species specificity (McGeoch et al., 2006). The linear genomes of these viruses are tightly packed and enclosed in a capsid, which is then enveloped, giving the virion a characteristic “fried egg” appearance by electron microscopy. The family is divided into the three subfamilies *Alphaherpesvirinae*, *Betaherpesvirinae* and *Gammaherpesvirinae*. The characteristics of human orthoherpesviruses are summarised in **Table** 1-1. The viral genomes vary considerably in size (125-236 kb), guanine plus cytosine (G+C) content (32–75 %) and sequence organisation (Gatherer et al., 2021). Due to a long history of virus-host coevolution, primary infection by orthoherpesviruses in immunocompetent hosts is usually mild or asymptomatic, and by adulthood the majority of people will have encountered at least one, if not all, of the human orthoherpesviruses.

The diseases associated with primary infection by orthoherpesviruses are varied, with subsequent persistence of viral genomes in a latent form from which reactivation can occur. Commonly, herpes simplex virus type 1 (HSV1) and herpes simplex virus type 2 (HSV2) cause oral cold sores and genital lesions, whereas varicella-zoster virus (VZV) causes chickenpox, or vesicular eruptions in the skin and mucous mucosa, and on reactivation, shingles. Epstein-Barr virus (EBV) infection can present as infectious mononucleosis (IM) prior to resolution and establishment of latency. Similarly, primary human cytomegalovirus (HCMV) can present with IM, but even in otherwise healthy hosts, more severe disease can progress to hepatosplenomegaly, hepatitis or colitis. However, the vast burden of HCMV disease lies with congenital infections and infection of immunosuppressed patients. Following latency, as numerous HCMV strains are in existence, either reactivation or reinfection by a different strain can lead to disease. This feature of HCMV variation stems from the organisation of its genome (Sections 1.3.1 and 1.8). HCMV possesses the largest genome (236 kb) of the nine orthoherpesviruses that infect humans (Davison et al., 2003a, Dolan et al., 2004, Bradley et al., 2009) (**Table** 1-1).

1.2 HCMV discovery

The first documented description of the cytomegalic intranuclear inclusion bodies typical of infants presenting with HCMV infection was made in 1881 by Ribbert (Ribbert, 1904). Large cells containing a central nuclear body surrounded by a clear halo were found characteristically in the organs of congenitally infected infants presenting with petechial rash, hepatosplenomegaly and cerebral calcification. The causative agent was initially thought to be protozoan, but a viral origin was postulated early on, and support came when similar lesions in rodents were shown to be due to viral infection (Vonglahn and Pappenheimer, 1925). Advances in HCMV research were finally made following the development of a cell culture system adapted from Enders *et al.* for the isolation of poliovirus in 1949 (Enders et al., 1949), and Smith was able to isolate HCMV from two patients (Smith, 1956). Concurrently, Rowe and coworkers isolated a virus causing intranuclear inclusions from the adenoids of a patient diagnosed with chickenpox (Rowe et al., 1956), while Weller isolated a similar virus from a child with cytomegalic inclusion disease (CID) thought to have toxoplasmosis (Weller et al., 1957). These independently isolated viruses were noted for their similar cytopathic effect (CPE) in causing enlarged cells and christened cytomegalovirus. HCMV-associated disease was first coined “generalised cytomegalic inclusion disease” by Wyatt and colleagues (Wyatt et al., 1950). Further developments, including the isolation of HCMV antigen for use in complement-fixation assays, established both a high seroprevalence of HCMV in adults and the existence of multiple strains of HCMV (Rowe et al., 1956, Weller et al., 1960). These factors transpire to be important factors mediating the clinical features and epidemiology of HCMV (Section 1.7).

1.3 HCMV virion

Mature HCMV virions are approximately 150-200 nm in diameter. The icosahedral nucleocapsid houses the linear dsDNA genome, which is surrounded by an amorphous tegument layer and finally enveloped by a host-derived lipid membrane that is embedded with viral glycoproteins (**Figure** 1-1). In addition to mature virions, infected cells also produce non-infectious enveloped particles (NIEPs) and dense bodies, which are virions lacking the viral genome or the viral

capsid and genome, respectively. In the following sections describing the virion structure in detail, each protein is designated either by the name used in HCMV GenBank accessions (**Tables** 1-2, 1-3 and 1-4) or by the name of the gene that encodes it prefixed by “p”.

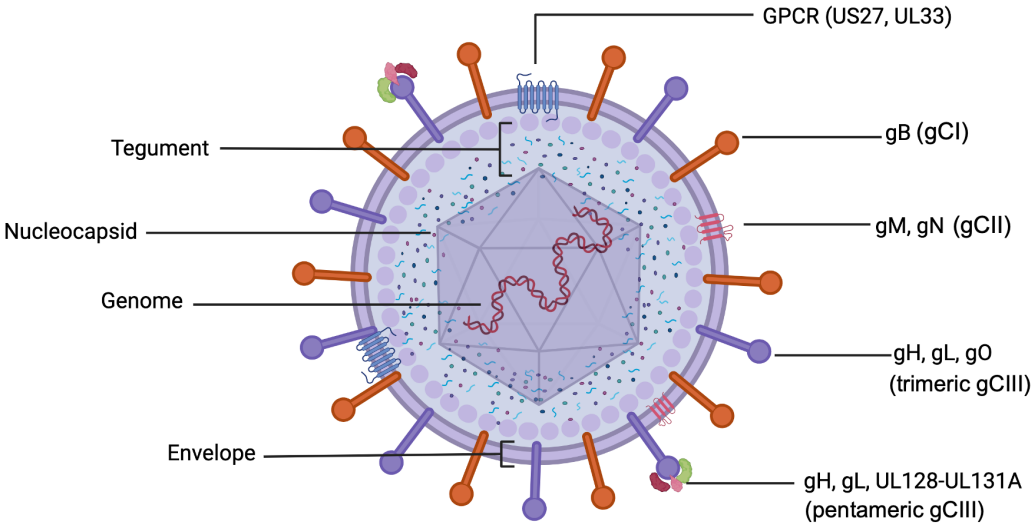


Figure 1-1. HCMV virion structure.
Mature HCMV virions are composed of three layers: the nucleocapsid enclosing the linear genome, the amorphous proteinaceous tegument and finally the envelope. The lipid bilayer envelope membrane is host-derived and associated with viral glycoproteins (prefixed “g”; the complex group is indicated in brackets). (Not to scale, created in BioRender.com.)

.

Table 1-1. Classification and characteristics of human viruses in the family *Orthoherpesviridae*.

Subfamily	Genus	Species	Virus name and abbreviation	Genome size (kb)	GenBank accession	Site of latency
Alpha-herpesvirinae	*Simplexvirus*	*Simplexvirus humanalpha1*	Herpes simplex virus type 1 (HSV1)	152	JN555585	Sensory nerve ganglia
		Simplexvirus humanalpha2	Herpes simplex virus type 2 (HSV2)	154	JN561323	Sensory nerve ganglia
	Varicellovirus	*Varicellovirus humanalpha3*	Varicella-zoster virus (VZV)	125	X04370	Sensory nerve ganglia
Beta-herpesevirinae	*Cytomegalovirus*	*Cytomegalovirus humanbeta5*	Human cytomegalovirus (HCMV)	236	AY446894	B lymphocytes
	Roseolovirus	*Roseolovirus humanbeta6a*	Human herpesvirus 6A (HHV6A)	159	X83413	T lymphocytes (CD4+), epithelial cells
		Roseolovirus humanbeta6b	Human herpesvirus 6B (HHV6B)	162	AF157706	T lymphocytes (CD4+), epithelial cells
		Roseolovirus humanbeta7	Human herpesvirus 7 (HHV7)	153	AF037218	T lymphocytes (CD4+)
Gamma-herpesvirinae	*Lymphocryptovirus*	*Lymphocryptovirus humangamma4*	Epstein-Barr virus (EBV)	172	AJ507799	Leukocytes, epithelial cells
	Rhadinovirus	*Rhadinovirus humangamma8*	Kaposi's sarcoma associated herpesvirus (KHSV)	170	AF148805	B lymphocytes, epithelial cells

1.3.1 Genome

The HCMV genome is organised in a Class E structure, with two unique sequences (unique long, U_L and unique short, U_S) that are flanked by terminal and internal inverted repeats (TR_L/IR_L and TR_S/IR_S). There is a terminal direct repeat, *a*, which is also present in reverse orientation (*a'*) in the internal repeat (**Figure** 1-2). The directly repeated *a* sequences forms a part of TR_L and TR_S, and *a'* is part of both IR_L and IR_S. Thus, the HCMV genome can be expressed as TR_L – U_L – IR_L – IR_S – U_S –TR_S or *ab* - U_L – *b'a'c'* – Us - *ca* (Davison et al., 2003b).

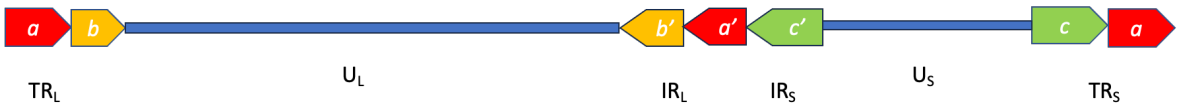


Figure 1-2. HCMV genome structure.
Unique regions U_L and U_S are represented as thin blue lines, and repeat regions are shown as block arrows, the direction of which indicates their orientation (not to scale).

Due to inversion of the U_L and U_S segments via homologous recombination between the terminal and internal repeat regions during DNA replication, viral populations exist as four isomeric forms in equimolar ratios (Chee et al., 1990, Mocarski, 1996). Early research on the genetic content of HCMV was performed on the high-passage culture-adapted strains, Towne and AD169. However, not long after the complete HCMV genome of AD169 was determined (Chee et al., 1990), it was found that high-passage strains had significant variations in genomic structure compared to clinical virus. Seminally, Cha and colleagues demonstrated the loss of 13 and 15 kb from the U_L/b' (U_L/IR_L) region of Towne and AD169, respectively, compared to the low-passage Toledo strain (Cha et al., 1996). It was thereby established that the passage of HCMV in cell culture leads to the rapid mutation of genes associated with virulence or tissue tropism, in some cases by large scale deletion and rearrangement of the U_L/b' region (Cha et al., 1996).

Determination of the genome sequence of the low-passage culture-adapted strain Merlin and comparison with the sequences of high-passage strains and clinical strains showed that Merlin accurately reflects the wild-type gene complement (Dolan et al., 2004). The predicted gene complement was adjusted in later publications (Dargan et al., 2010, Gatherer et al., 2011, Davison et al., 2013). As in the AD169 genome (Chee et al., 1990), the names of the 170 genes are prefixed by the name of the region in which each is located in the AD169 or Merlin genomes (**Figure** 1-3). The Merlin studies also confirmed that all passaged strains analysed had disabling mutations affecting one or more of UL128, UL130 and UL131A (which encode proteins forming the gCIII pentameric complex, Section 1.3.4) and various members of the RL11 glycoprotein gene family (e.g. RL5A, RL13 and UL9). Efforts to standardise a laboratory HCMV model led to the construction of a bacterial artificial chromosome (BAC) containing the Merlin genome (Stanton et al., 2010). However, even in this genome, it was known that UL128 and RL13 were both truncated by a single nucleotide substitutions causing in-frame termination codons, neither of which was present in the original Merlin clinical sample (Akter et al., 2003, Dolan et al., 2004). The Merlin BAC clone was repaired to enable study of a clinical strain generated by transfection, but RL13 mutants still emerged rapidly on growth in fibroblasts (Stanton et al., 2010).

Estimates of the number of genes encoding functional proteins in the HCMV genome range from 170 to upward of 700 (Chee et al., 1990, Davison et al., 2003a, Murphy et al., 2003, Yu et al., 2003, Stern-Ginossar et al., 2012). This inflation stems from the differing definitions and techniques by which genes are predicted from the HCMV genomic and transcriptomic data. For example, alternative splice sites can generate multiple additional exons, and these were excluded from the earlier estimates. The open reading frame (ORF) of a gene is identified by the start codon ATG encoding methionine and ends with one of the stop codons, TAA, TGA or TAG. The transcriptional start site (TSS) typically contains the ORF and starts close downstream from a promotor whilst the transcriptional end site (TES) is usually found close downstream of a polyadenylation signal, based on AATAAA. Splicing signals can also be recognised at the ends of introns. Usually, ORFs are accepted only if the sequence is over a stipulated length; for example, a threshold of 300 bp allows for a coding potential of 100 amino acids (Chee et al., 1990). Comparative genomics can then

be used to parse these sequences to distinguish ORFs likely to encode functional proteins that are recognizably conserved: first in other HCMV strains, then in other orthoherpesviruses and finally in other organisms. Also, properties of the proteins, such as membership of a family of related genes, presence of motifs and occupation of a certain region of the genome may be used as clues. For example, the original UL3 ORF was initially predicted to be protein-coding (Chee et al., 1990) but was subsequently found to have a variable TES between strains and was even truncated by an internal stop codon in an unpassaged clinical strain (Dolan et al., 2004). Also, no homologue of HCMV UL3 is present in the chimpanzee cytomegalovirus (CCMV) genome (Dolan et al., 2004). These findings indicated that UL3 is unlikely to encode a protein, and it is now absent from the HCMV genome map (**Figure** 1-3). Conversely, the CCMV genome contains three genes at the right end of U_L that are not present in HCMV strains (Dolan et al., 2004). Because of its complexity, prediction of HCMV genes is liable to miss some functional protein-coding ORFs or include non-functional ones. Additionally, assessment of features associated with protein coding would not identify genes encoding noncoding RNAs, of which there are at least five (**Figure** 1-3). Adding to the complexity, a subset of genes exhibits unusually high variation between strains. These hypervariable genes tend to encode membrane-associated proteins and proteins with immunomodulatory functions (Dolan et al., 2004, Sijmons et al., 2015). Genetic hypervariation in HCMV is discussed in Section 1.8.

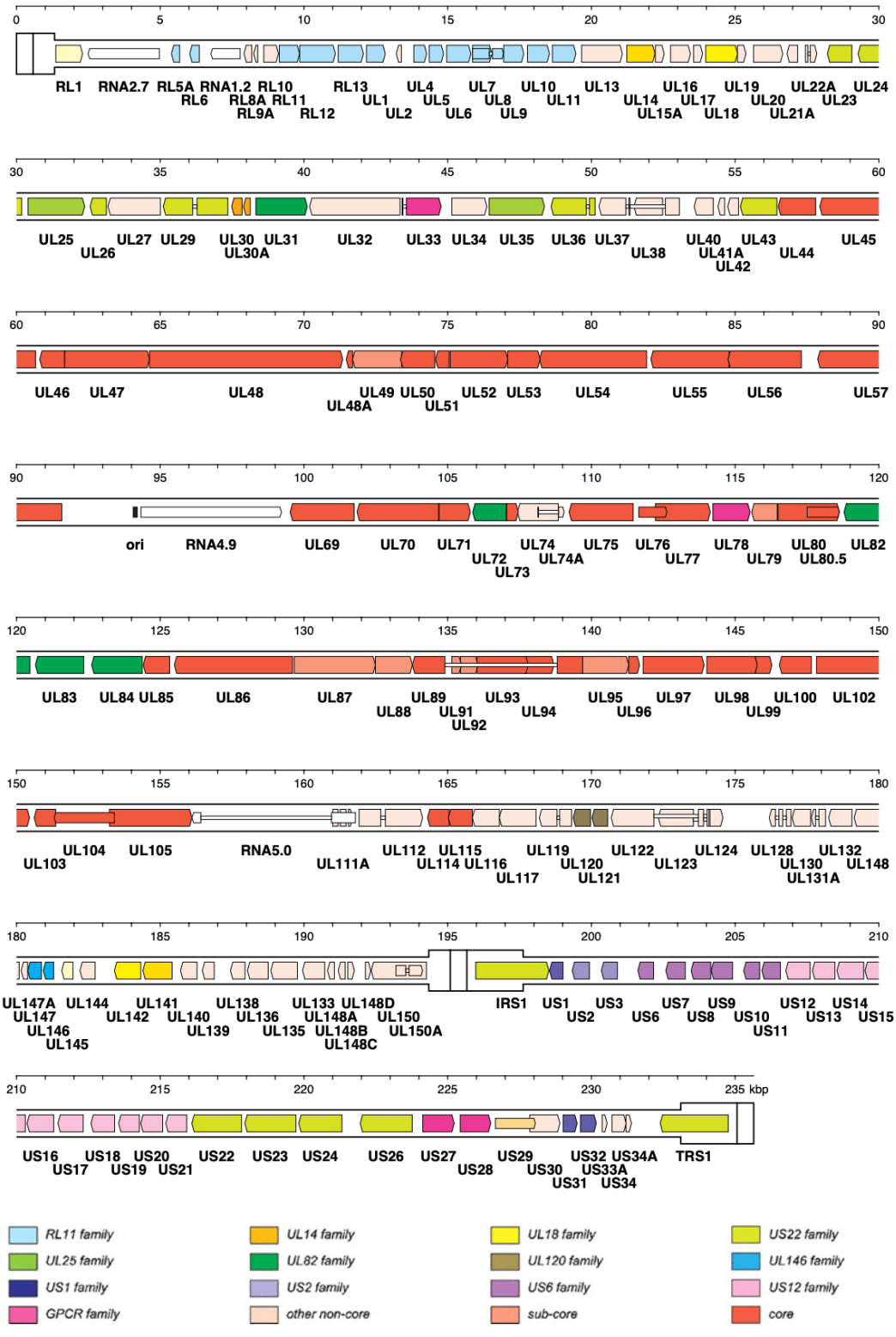


Figure 1-3. Genome map of wild-type HCMV strain Merlin.
The inverted repeats (TR_L, IR_L, IR_S and TR_S) are represented by a thicker outline than the unique regions (U_L and U_S). Predicted functional protein-coding regions are indicated by coloured arrows grouped according to the key, with gene nomenclature below. Noncoding RNAs are shown by white arrows. Introns connecting protein-coding regions (or exons in RNA5.0) are shown as narrow white bars. Colours differentiate between genes on the basis of conservation across subfamilies *Alphaherpesvirinae, Betaherpesvirinae* and *Gammaherpesvirinae* (core genes) or between subfamilies *Betaherpesvirinae* and *Gammaherpesvirinae* (sub-core genes), with subsets of the remaining non-core genes grouped into families of related genes. (Reproduced with permission from author A. Davison (Davison et al., 2013).)

1.3.2 Capsid

Encapsidation of the HCMV genome into the capsid occurs in the nucleus. The capsid is assembled from the viral proteins encoded by UL86, UL85, UL80, UL80.5, UL48A and UL46 (**Table** 1-2 and **Figure** 1-4A). Mature capsids are organised in a T=16 icosahedral lattice composed of 162 capsomeres (150 hexamers and 12 pentamers) and 320 triplexes (Chen et al., 1999) (**Figure** 1-4C). The heterotrimeric triplexes are located between the hexamers and pentamers at local threefold symmetry axes. The pentamers consist of five copies of the major capsid protein (MCP, UL86), whereas the hexamers are composed of six copies of MCP plus six copies of the small capsid protein (UL48A) (Gibson et al., 1996b). Each triplex ("Ta", "Tb", "Tc", "Td", "Te" and "Tf" in **Figure** 1-4D, E) consists of two copies of the minor capsid protein (mCP, UL85) and one copy of the mCP-binding protein (mCP-BP, UL46) (Gibson et al., 1996a). HCMV capsid reconstruction by cryoelectron microscopy has demonstrated similarities to HSV1, with differences in the capsid diameter: 125 and 130 nm for HSV1 and HCMV, respectively (Butcher et al., 1998). The greater capsid size appears insufficient for the HCMV genome, which is therefore more densely packed than in HSV1, with an average interlayer spacing of approximately 23 Å (Bhella et al., 2000). The pressure induced by high-pressure packaging has been postulated as a mechanism for the efficient delivery of the HCMV genome to the cell nucleus after infection.

Table 1-2. HCMV capsid proteins.
Summarised from the GenBank-published features of the Merlin genome (AY446894.2). Although pUL77 and pUL93 are included as capsid-associated proteins, they are putative tegument proteins that, along with pUL52, have roles in cleavage and packaging of the HCMV genome into the capsid (Borst et al., 2008, Borst et al., 2016).

Gene[a]	Family[b]	Product	Notes
UL46	Core gene	Capsid triplex subunit 1, mCP binding protein (mCP-BP)	Complexed 1:2 with capsid triplex subunit 2 to connect capsid hexons and pentons; involved in capsid morphogenesis
UL48A	Core gene	Small capsid protein pUL48.5	Located externally on capsid hexons; involved in capsid morphogenesis; possibly involved in capsid transport
UL52	Core gene	DNA packaging protein UL32	Involved in DNA encapsidation; possibly involved in capsid transport
UL77	Core gene	DNA packaging tegument protein UL25 (pUL77)	Located on capsid near vertices; possibly stabilizes the capsid and retains the genome; involved in DNA encapsidation
UL80	Core gene	Capsid maturation protease	Serine protease (N-terminal region); minor scaffold protein (remainder of protein, clipped near C terminus); involved in capsid morphogenesis
UL80.5	Core gene	Capsid scaffold protein	Clipped near C terminus; involved in capsid morphogenesis
UL85	Core gene	Capsid triplex subunit 2 (minor capsid protein, mCP)	Complexed 2:1 with capsid triplex subunit 1 to connect capsid hexons and pentons; involved in capsid morphogenesis
UL86	Core gene	Major capsid protein (MCP)	6 copies form hexons, 5 copies form pentons; involved in capsid morphogenesis
UL93	Core gene	DNA packaging tegument protein UL17	Capsid-associated; involved in DNA encapsidation; involved in capsid transport

[a] Genes in bold are discussed in detail in the text.

[b] Core genes are conserved across members of subfamilies *Alphaherpesvirinae, Betaherpesvirinae* and *Gammaherpesvirinae*, betagamma genes are conserved between subfamilies *Betaherpesvirinae* and *Gammaherpesvirinae*, and beta genes are present only within subfamily *Betaherpesvirinae*.

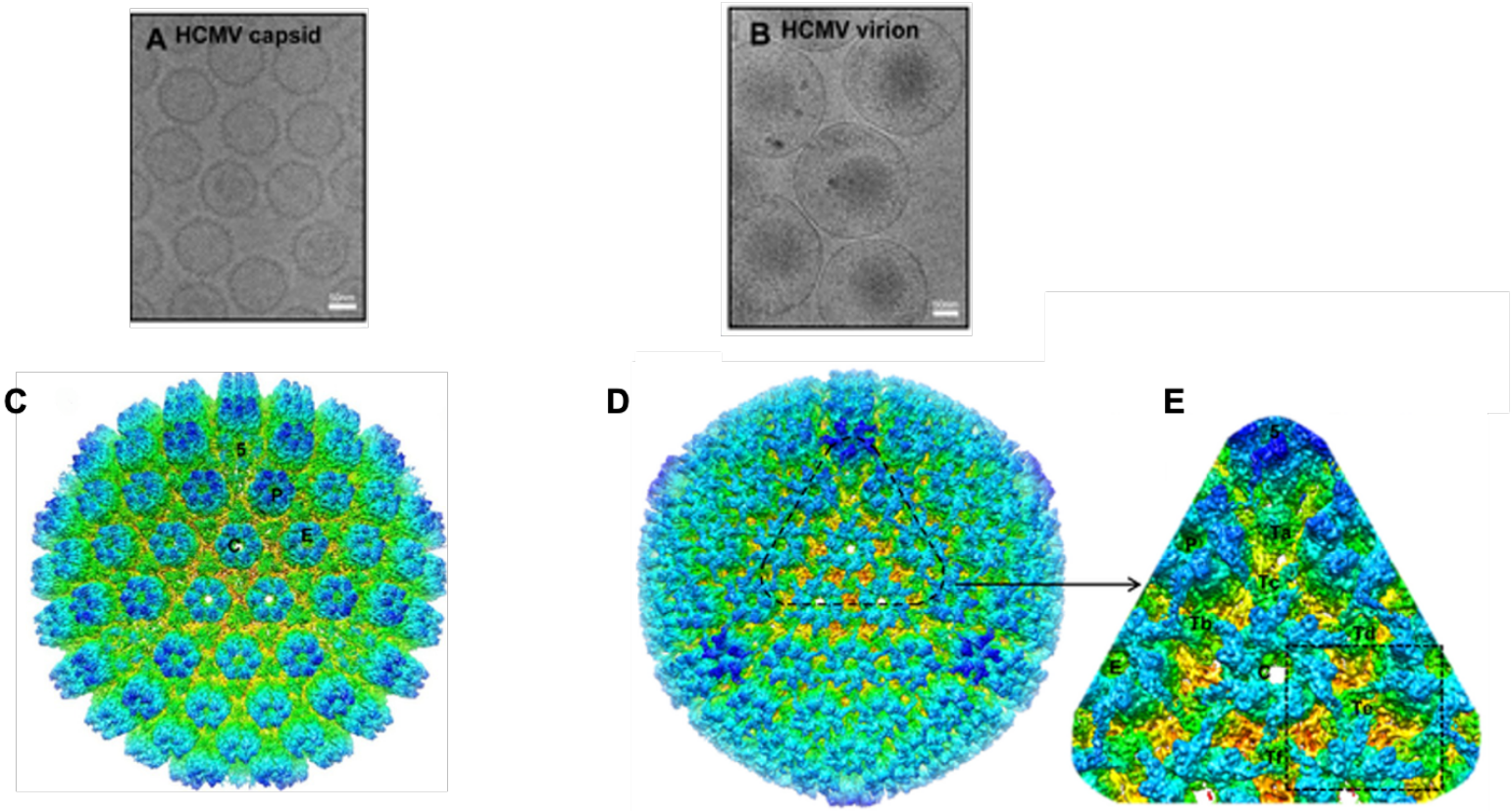


Figure 1-4. Cryoelectron microscopy images and three-dimensional reconstructions of the HCMV capsid and virion.
(A) Image of capsids. (B) Image of virions. (C) Reconstruction of a capsid showing the molecular boundaries between the 150 hexamers, 12 pentamers and 320 triplexes in the T=16 icosahedral particle. Capsomeres in an asymmetric unit, including a pentamer and three hexamers, are labelled “5”, “C”, “P” and “E”, respectively. (D) Reconstruction of an HCMV virion viewed along a threefold axis. The 9 Å reconstruction of HCMV virion shows a layer of filamentous tegument proteins bound to the capsid in an icosahedrally ordered fashion, like a net enclosing the entire capsid. (E) Zoom-in view of the area denoted in (D). Structural components in an asymmetric unit are labelled, including a pentamer (“5”), three hexamers (“C”, “P” and “E”), and six triplexes (“Ta”, “Tb”, “Tc”, “Td”, “Te” and “Tf”). The square demarcates a region encompassing Te that is segmented out for averaging with similar regions around Tb and Td. (Adapted from (Dai et al., 2013), under the terms of the Creative Commons Attribution License.)

1.3.3 Tegument

Surrounding the capsid is an amorphous layer of proteins, or tegument (**Figure** 1-1). The tegument is composed of approximately 60 different viral proteins, although only 31 are present at significant levels (Baldick and Shenk, 1996, Varnum et al., 2004, Gatherer et al., 2021), as reviewed by Kalejta and summarised in **Table** 1-3 (Kalejta, 2008). These proteins have roles in the following processes: i) viral entry; ii) gene expression; iii) evasion of host immune response; and iv) viral assembly and egress (**Table** 1-3).

1.3.3.1 Viral entry

During the initial stages of infection of a cell by HCMV, subsequent to membrane fusion and prior to immediate-early (IE) gene expression, tegument proteins help to deliver capsids and genomes to the nuclear pore complex. The tegument proteins pUL47, large tegument protein (UL48), DNA packaging tegument protein pUL77 and phosphoprotein 150 (pp150, UL32) are closely associated with capsids and help to direct viral genome delivery along microtubules to the nucleus (Bechtel and Shenk, 2002, AuCoin et al., 2006, Kalejta, 2008) (**Tables** 1-2 and 1-3).

1.3.3.2 Gene expression

Following genome entry into the nucleus, expression of viral IE genes during lytic infection is expedited by tegument proteins pp71 (UL82) and pUL35 (UL35). The upper matrix protein, or pp71, is a virion transactivator (Baldick et al., 1997) that suppresses cellular Daxx. Daxx normally binds to histone deacetylases (HDACs) and results in transcriptional repression of viral DNA (Saffert and Kalejta, 2006). Derepression of this intrinsic defence is modulated by the two pUL35 proteins, pUL35 and pUL35a, in conjunction with pp71 (Salsman et al., 2011) (**Figure** 1-5).

1.3.3.3 Evasion of host immune response

Suppression of all arms of the immune response, including intrinsic, innate and adaptive responses, is also facilitated by tegument proteins, including pp65 (UL83), pUL36, pUL38, pp71, pIRS1 and pTRS1. A major component of HCMV

virions, although dispensable for HCMV infection in cell culture, pp65 counters both innate and adaptive immune responses. The innate response is blunted by inhibiting natural killer cell cytotoxicity by interacting with the NKp30 activating receptor (Arnon et al., 2005) and the levels of expression of interferon-stimulated genes (ISGs) (Abate et al., 2004). The adaptive response is targeted by phosphorylation of viral IE proteins, thus preventing their presentation by major histocompatibility complex (MHC) class I molecules (Gilbert et al., 1996) and downregulating expression of HLA-DR by mediating the accumulation of HLA class II molecules in lysosomes, leading to degradation of the HLA-DR alpha-chain (Odeberg et al., 2003). Viral inhibitor of caspase-8 activation (vICA, UL36) and pUL38 both inhibit apoptosis, another part of the innate immune system (Xuan et al., 2009, McCormick et al., 2010). The IRS1/TRS1 genes, which are located in the *c* repeats flanking U_S and span into U_S, have identical sequences apart from the 3'-ends (Varnum et al., 2004). Their products are responsible for countering shutoff of host protein synthesis, thus enabling viral replication to proceed (Child et al., 2004). As infection progresses, pp71 again plays a role in downregulating MHC class I cell surface expression, thus circumventing HCMV antigen presentation to CD8+ T cells (Trgovcich et al., 2006).

1.3.3.4 Viral assembly and egress

Following the packaging of viral genomes into capsids in the nucleus, viral egress occurs via an envelopment-deenvelopment-reenvelopment process, akin to that observed in other orthoherpesviruses (Mettenleiter et al., 2006). The primary envelope is acquired by budding through the inner nuclear membrane and then lost on budding from the outer nuclear membrane, and the final envelope is acquired on budding of capsid-tegument structures into the Golgi apparatus, part of the viral assembly complex. Although the exact mechanisms by which pp150 (UL32) (Dunn et al., 2003, Yu et al., 2003) and pp28 (UL99) (Silva et al., 2003) exert their roles in this process of viral assembly and egress are still to be fully elucidated, studies have demonstrated that these tegument proteins are essential for productive viral replication and may act in a similar pathway. The protein kinase ppUL97 (UL97) is also postulated to aid incorporation of tegument proteins into virions by phosphorylation (Littler et al., 1992).

Table 1-3. HCMV tegument proteins.
Summarised from the GenBank-published features of the Merlin genome (AY446894.2) and a review paper (Kalejta, 2008).

Gene[a]	Family[b]	Product[c]	Notes
UL23	US22	Tegument protein UL23 (pUL23)	Beta gene
UL24	US22	Tegument protein UL24 (pUL24)	Beta gene
UL25	UL25	Tegument protein UL25 (pUL25)	Colocalizes with pp28
UL26	US22	Tegument protein UL26 (pUL26)	Transcriptional activator of major immediate early promoter; involved in gene regulation; increase stability of virion proteins
UL32	UL25	Tegument protein pp150	Beta gene, major tegument protein; directs capsid to site of final envelopment
UL35	Beta gene	Tegument protein UL35 (pUL35)	Activates viral gene expression
UL36	US22	Tegument protein vICA	Inhibitor of caspase-8-induced apoptosis; involved in apoptosis
UL38	Beta gene	Protein UL38 (pUL38)	Involved in apoptosis
UL43	US22	Tegument protein UL43 (pUL43)	Beta gene
UL44	Core gene	DNA polymerase processivity subunit	HCMV DNA polymerase processivity/transcription factor
UL45	Core gene	Ribonucleotide reductase subunit 1	Enzymatically inactive; tegument protein
UL47	Core gene	Tegument protein UL37 (pUL37)	Complexed with large tegument protein; involved in virion morphogenesis; release of viral DNA from capsid
UL48	Core gene	Large tegument protein	Complexed with tegument protein UL37, deubiquitinating protease
UL50	Core gene	Nuclear egress membrane protein	Type 2 membrane protein; nuclear egress of capsids
UL53	Core gene	Nuclear egress lamina protein	Nuclear egress or assembly of capsids
UL54	Core gene	DNA polymerase catalytic subunit	Viral DNA polymerase
UL57	Core gene	Single stranded DNA binding protein	Contains zinc-finger; involved in DNA replication
UL71	Core gene	Tegument protein UL51	Involved in virion morphogenesis.
UL77	Core gene	DNA packaging tegument protein UL25	Located on capsid near vertices; possibly stabilizes the capsid and retains the genome; involved in DNA encapsidation
UL82	Beta gene	Tegument protein pp71	Upper matrix protein; involved in gene regulation; transcriptional activator; targets Rb proteins for ubiquitin-independent proteosomal degradation; prevents cell surface expression of MHC
UL83	Beta gene	Tegument protein pp65	Lower matrix protein; evasion of adaptive and innate immunity
UL88	Betagamma gene	Tegument protein UL88	
UL93	Core gene	DNA packaging tegument protein UL17	Capsid-associated; involved in DNA encapsidation; involved in capsid transport.
UL94	Core gene	Tegument protein UL16 (pUL16)	Possibly involved in virion morphogenesis; putative DNA-binding protein
UL96	Core gene	Tegument protein UL14 (pUL14)	Involved in virion morphogenesis
UL97	Core gene	Tegument serine/threonine protein kinase	Involved in protein phosphorylation; mutation conveys resistance to ganciclovir; high-level resistance to maribavir
IRS1/TRS1	US22	Tegument protein IRS1 Tegument protein TRS1	Immediate early gene; transcriptional activator; blocks phosphorylation of eIF2alpha and host shutoff of protein synthesis; binds dsRNA; involved in gene regulation; involved in translational regulation
UL99	Core gene	Myristylated tegument protein pp28	Envelope-associated; involved in virion morphogenesis.
UL103	Core gene	Tegument protein UL7	involved in virion morphogenesis
US22	US22	Tegument protein US22	
US24	US22	Tegument protein US24	Activates viral gene expression

[a] Genes in bold are discussed in detail in the text.

[b] Core genes are conserved across members of subfamilies *Alphaherpesvirinae*, *Betaherpesvirinae* and *Gammaherpesvirinae*, betagamma genes are conserved between subfamilies *Betaherpesvirinae* and *Gammaherpesvirinae*, and beta genes are present only within subfamily *Betaherpesvirinae*.

[c] Suffixes: pp = phosphoprotein, p = protein

1.3.4 Envelope

The ability of HCMV to cause multisystem disease is due to its broad cellular tropism. HCMV is capable of infecting a wide range of cell types, including fibroblasts, endothelial cells, monocytes and macrophages, smooth muscle cells, stromal cells, neuronal cells, neutrophils and hepatocytes. This indicates a combination of either multiple cell specific receptors or widely expressed receptors (Myerson et al., 1984, Ibanez et al., 1991, Sinzger et al., 2000). HCMV entry into target cells is initiated by envelope glycoproteins binding to the cell surface then by specific interactions between viral glycoproteins and cellular receptors, leading to fusion between the viral envelope and the cell membrane. The essential glycoproteins of HCMV are organised in the viral envelope as three complexes, gCI, gCII and gCIII (**Figure** 1-1).

1.3.4.1 gCI

gCI is composed of two subunits gp58 and gp116, which are the furin-protease cleaved products of gB (UL55), and exists in a homotrimeric form (Vey et al., 1995). gB binds to heparan sulphate proteoglycans (HSPGs), which are sugar moieties covalently bound to cellular receptors that are found on most cell types. This initial tethering and recruitment of HCMV virions to the cell surface enables further interactions of viral glycoproteins with cell-specific receptors to mediate fusion (Boyle and Compton, 1998). Studies using neutralising antibodies against gB and complementation of UL55-defective viruses have demonstrated that gB is essential for cell entry, but not attachment, assembly or egress, thus emphasising its role in the fusion step (Boyle and Compton, 1998, Isaacson and Compton, 2009a, Pötzsch et al., 2011). The disintegrin-like domain of gB is conserved across subfamilies *Betaherpesvirinae* and *Gammaherpesvirinae* and binds to β1-integrins present in cells, promoting fusion in fibroblast, endothelial and epithelial cells (Feire et al., 2010). In nonpermissive cells, gB triggers host cell expression of myeloid cell leukaemia (Mcl)-1 protein, thus enhancing an antiapoptotic effect in myeloid cell types (Reeves et al., 2012). Four different genotypes of gB have been documented, with polymorphic sites along the most variable regions present in the N terminus, C terminus and furin cleavage sites (Chou and Dennison, 1991, Stangherlin et al., 2017). The effects of gB variation

on clinical outcomes has been investigated and is discussed in Section 1.8.1. The fact that a large proportion (40-70 %) of the total serum neutralising antibody is directed against gB also has implications for vaccine development (Britt et al., 1990) (Section 1.6.2).

1.3.4.2 gCII

The heterodimeric gCII complex is comprised of gM (UL100) and gN (UL73) and has also been shown to be essential for viral entry (Mach et al., 2000, Shimamura et al., 2006), although the distinct mechanism has yet to be fully elucidated. Alongside gCI, gCII also has HSPG-binding properties (Kari and Gehrz, 1992) and elicits a humoral response to generate neutralising antibodies (Shen et al., 2007). These antibodies appear to offer cross-protection among different strains (Shimamura et al., 2006), despite the variation of gN, which is classified into four genotypes (Pignatelli et al., 2001).

1.3.4.3 gCIII

The gCIII complex exists in two forms that are made up of distinct accessory proteins associated with gH (UL75)/gL (UL115): the trimeric gH/gL/gO complex (Huber and Compton, 1998) and the pentameric gH/gL/pUL128/pUL130/pUL131A complex (Hahn et al., 2004). The two complexes are associated with cell-type specific tropism, the gH/gL/gO complex being required for entry and dissemination among fibroblasts (Podlech et al., 2015), and the gH/gL/UL128-UL131A complex being required for entry into epithelial and endothelial cells and leukocytes (Adler et al., 2006, Wang and Shenk, 2005a). One or more of UL128, UL130 and UL131A are typically lost rapidly on passage in fibroblasts (Hahn et al., 2004, Dargan et al., 2010).

Table 1-4. HCMV envelope proteins.
Summarised from the GenBank-published features of the Merlin genome (AY446894.2).

Gene[a]	Family[b]	Product	Notes
RL10	RL11	Envelope glycoprotein RL10	Envelope glycoprotein RL10; type 1 membrane protein; possibly related to RL11 family
RL11	RL11	Membrane glycoprotein RL11	Type 1 membrane protein; binds IgG Fc; involved in immune regulation; RL11 family
UL4	RL11	Envelope glycoprotein UL4	Contains signal peptide
UL8	RL11	Membrane glycoprotein UL8	Type 1 membrane protein; RL11 family
UL7	RL11	Membrane glycoprotein UL7	Type 1 membrane protein; modulates cytokine production; involved in immune regulation
UL9	RL11	Membrane glycoprotein UL9	Type 1 membrane protein
UL11	RL11	Membrane glycoprotein UL11	Type 1 membrane protein; disrupts T cell function via inhibition of CD45; involved in immune regulation
UL16	Other, non-core gene	Membrane glycoprotein UL16	Type 1 membrane protein; inhibits NK cell cytotoxicity; sequesters MICB, ULBP1 and ULBP2; acts on activating receptor NKG2D; involved in immune regulation
UL18	UL18	Membrane glycoprotein UL18	Type 1 membrane protein; inhibits NK cell cytotoxicity; acts on inhibitory receptor LIR-1 (ILT2); involved in immune regulation; MHC family
UL22A	Other, non-core gene	Glycoprotein UL22A	Contains signal peptide; secreted protein; binds CC chemokine RANTES; involved in immune regulation
UL33	GPCR family	Envelope glycoprotein UL33	Beta gene; type 3 membrane protein; 7 transmembrane domains; putative chemokine receptor; involved in intracellular signalling;
UL37	Beta gene	Envelope glycoprotein UL37	Type 1 membrane protein; involved in apoptosis; involved in gene regulation; major isoform (mitochondrial inhibitor of apoptosis, vMIA) encoded by unspliced mRNA; IE gene
UL40	Other, non-core gene	Membrane glycoprotein UL40	Type 1 membrane protein; sequence in signal peptide inhibits NK cell cytotoxicity; upregulates cell surface expression of HLA-E; acts on inhibitory receptor CD94/NKG2A; involved in immune regulation
UL55	Core gene	Envelope glycoprotein B	Type 1 membrane protein; possible membrane fusogen; binds cell surface heparan sulphate; involved in cell entry; involved in cell-to-cell spread
UL73	Core gene	Envelope glycoprotein N	Type 1 membrane protein; complexed with envelope glycoprotein M; involved in virion morphogenesis; involved in membrane fusion
UL74	Beta gene	Envelope glycoprotein O	Contains signal peptide; associated with envelope glycoprotein H and envelope glycoprotein L; involved in virion morphogenesis
UL74A	Beta gene	Envelope glycoprotein 24	Main coding exon spliced from one or other of >45 upstream exons; type 2 membrane protein
UL75	Core gene	Envelope glycoprotein H	Type 1 membrane protein; possible membrane fusogen; complexed with envelope glycoprotein L; involved in cell entry; involved in cell-to-cell spread
UL100	Core gene	Envelope glycoprotein M	Type 3 membrane protein; 8 transmembrane domains; complexed with envelope glycoprotein N; involved in virion morphogenesis; involved in membrane fusion
UL115	Core gene	Envelope glycoprotein L	Contains signal peptide; complexed with envelope glycoprotein H; involved in cell entry; involved in cell-to-cell spread
UL119	Beta gene	Membrane glycoprotein UL119	Type 1 membrane protein; IgG Fc-binding; similar to OX-2; involved in immune regulation; possibly related to UL120 family
UL130	Other, non-core gene	Envelope glycoprotein UL130	Contains signal peptide; complexed with envelope glycoproteins H and L; essential in non-fibroblast cells; involved in cell entry

[a] Genes in bold are discussed in detail in text.

[b] Core genes are conserved across members of the subfamilies *Alphaherpesvirinae*, *Betaherpesvirinae* and *Gammaherpesvirinae* and beta genes only within the *Betaherpesvirinae* subfamily. GPCR, G-protein coupled receptor.

1.4 HCMV life cycle

HCMV can undergo lytic and latent pathways following infection (Kalejta, 2013, Jean Beltran and Cristea, 2014). Following HCMV infection, viral replication in permissive cells, such as fibroblasts, results in cell lysis. HCMV overcomes host defences to hijack cellular machinery to undergo active replication and release progeny virus in these instances (**Figure** 1-5, left). However, on infection of nonpermissive cell types, including myeloid progenitor cells (CD34+ HPCs or CD14+ monocytes), HCMV proceeds to latency, leading to a reservoir of quiescent infections ready for reactivation (**Figure** 1-5, right).

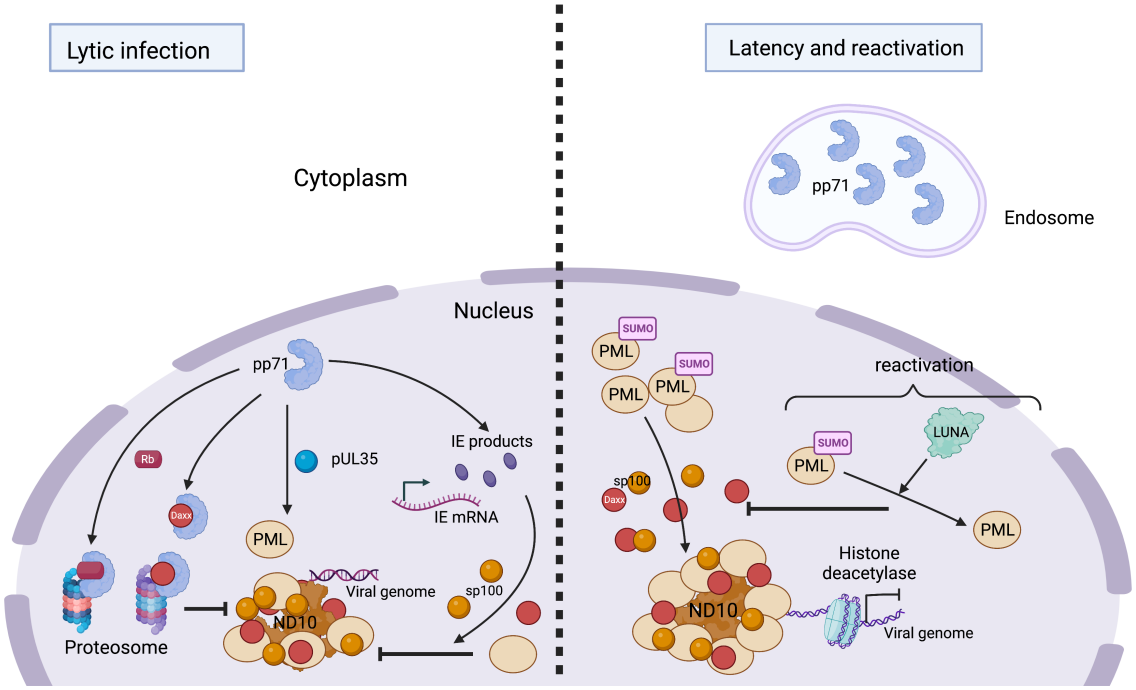


Figure 1-5. The lytic and latent cycles of HCMV infection.
These two types of infection occur in different cell types (see text). During lytic infection (left panel), nuclear pp71 induces degradation of cellular proteins, Rb and Daxx, relieving Rb-mediated cell cycle blockade and ND10 block on IE gene expression, respectively. Transactivation of IE gene expression by pp71 in turn disrupts ND10 formation. Conversely, in latent infection (right panel), pp71 is retained in the cytoplasm and ND10 (part of the innate defence system) recruits HDAC resulting in viral gene silencing. The reactivation pathway is shown within the bracket, in which isopeptidase activity of LUNA deSUMOylates PML and disrupts PML recruitment of Daxx and Sp100, thus enabling viral IE gene expression and thereby reactivation. (Created with BioRender.com.)

1.4.1 Lytic infection

Infection of permissible cells is facilitated by the interactions of envelope glycoproteins with host receptors to mediate fusion and viral entry. The three glycoprotein complexes (gCI, gCII and gCIII) on the viral envelope enable cell entry into epithelial, endothelial and myeloid cells (Hahn et al., 2004, Wang and Shenk, 2005b) (**Figure** 1-6, step 1: Binding). Initial binding is mediated through these viral receptors to ubiquitously expressed cellular proteins such as heparan sulphate (see Section 1.3.4). After sufficient viral recruitment, docking of gB with cell-specific receptors triggers conformational changes at the viral-cell membrane interface to allow for fusion (step 2: Fusion, see Section 1.3.4). One hypothesized cellular expressed receptor is epithelial growth factor receptor (EGFR). Although HCMV entry was found to be unaffected in fibroblasts, epithelial or endothelial cells when EGFR was blocked (Isaacson et al., 2007), its activation was shown to be required for viral entry into CD34+ progenitor cells (Kim et al., 2017). Other cellular receptors investigated for roles in viral entry include annexin II (Pietropaolo and Compton, 1999), CD13 (Söderberg et al., 1993) and platelet-derived growth factor receptor alpha (PDGFR-α) (Soroceanu et al., 2008). gCI mediates viral envelope and cell membrane fusion, likely by interaction with cellular integrins (see Section 1.3.4). Integrin activation leads to cytoskeletal reorganisation, and β1-integrin has been shown to be essential for the delivery of viral pp65 (UL83), which is required to initiate IE gene expression (Hashimoto et al., 2020).

After viral entry, the pp150 (UL32), pUL47), pUL48 and pUL77 tegument proteins bound to the capsid interact with the host microtubule machinery to transport capsids to the nuclear envelope and into the nucleus where viral transcription, genome replication and encapsidation occur (Ogawa-Goto et al., 2003) (**Figure** 1-6, steps 3: Microtubule transport, 4: Transcription and 5: Viral genome replication; see Section 1.3.3). Host defences to silence viral DNA by the nuclear domain 10 (ND10) nuclear bodies is overcome by pp71 (UL82). pp71 inhibits the action of ND10- and promyelocytic leukaemia protein (PML)-associated nuclear bodies by: i) inducing degradation of cellular proteins, retinoblastoma (Rb) protein and death domain-associated protein (Daxx); and ii) transactivating IE gene expression (Hofmann et al., 2002, Ishov et al., 2002) (**Figure** 1-5, left).

Daxx, which is one of the most important components of ND10, has an intrinsic defence function against HCMV and acts to restrict viral gene expression by recruiting HDAC (Cantrell and Bresnahan, 2005). Another way in which pp71 bypasses cellular defences is by overriding the effect of Rb, which interacts with the ND10 complex to suppress viral gene expression (Alcalay et al., 1998, Kalejta, 2008) (**Figure** 1-5, left). Rb and E2F complexes normally inhibit expression of E2F genes and cell cycle progression. pp71 does so by proteasome degradation of Rb proteins, but another tegument protein, ppUL97, hyperphosphorylates Rb and so reverses the effect in a PML-dependent manner (Fang et al., 2002) (Section 1.3.3.4).

Lytic infection occurs with viral genes expressed in a cascade: first IE genes, then early (E) genes, and finally late (L) genes, thus enabling virion assembly and release (Weekes et al., 2014). Lytic phase DNA replication is initiated at a *cis*-acting origin of DNA replication (oriLyt, approximately 3 kb), which is a structurally complex region containing repeat elements and transcription factor-binding sites (Anders et al., 1992). OriLyt lies between UL57 and RNA4.9, and a complex promoter region within UL57 regulates both UL57 transcription and oriLyt activation (Kiehl et al., 2003). HCMV DNA amplification occurs after E gene expression and before L gene expression and usually initiates at 24-72 h post-infection in cell culture. Tegument proteins pp71 and pUL35 initiate the temporal cascade of expression of IE1 (UL123) and IE2 (UL122), followed by that of delayed early genes (DE) genes (including UL36, UL37, UL38 and UL112), which initiate viral genome replication (Pari and Anders, 1993). L gene expression initiates capsid assembly in the nucleus and capsid egress into the cytoplasm (**Figure** 1-6, step 6: Translation). Capsids associated with tegument proteins are then trafficked to the viral assembly complex, which contain components of the endoplasmic reticulum (ER), Golgi apparatus and endosomal machinery. The capsids acquire further tegument and envelope proteins by budding into intracellular vesicles at the assembly complex (step 7: Assembly). Enveloped infectious particles are released along with noninfectious dense bodies (steps 8: Budding and 9: Release). More recently, the strictly sequential cascade of viral gene expression and protein synthesis has been challenged, suggesting that productive infection is less strictly temporal and more fluid, mediated by multiple independent modules (Rozman et al., 2022).

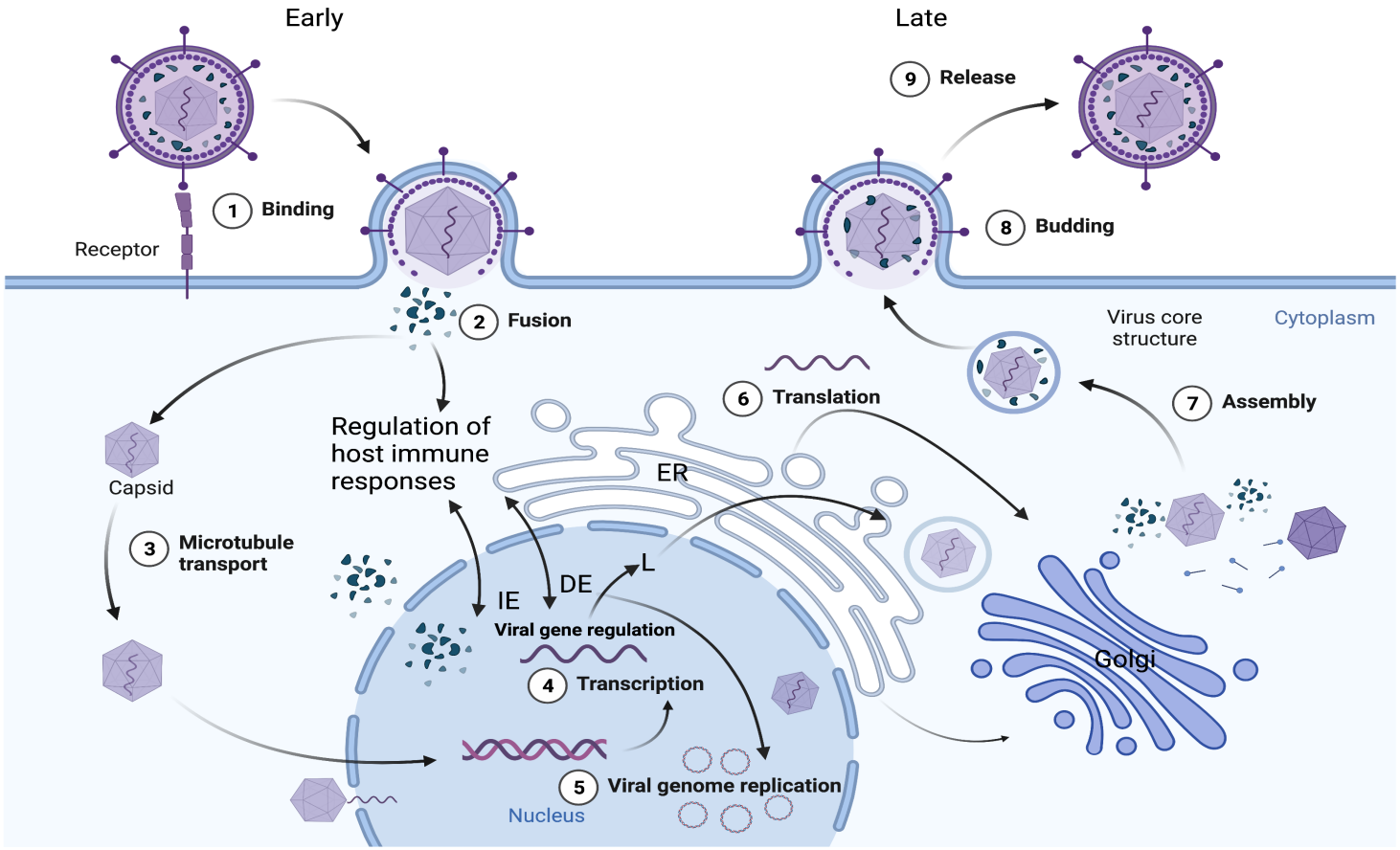


Figure 1-6. The lytic cycle of HCMV infection.
Steps 1 and 2: virions attach to the cell surface (likely via heparan sulphate) and an interaction occurs between gCI and cell-specific receptors, thus enabling fusion of the viral envelope to the cell membrane and releasing the capsid into the cytoplasm. Step 3: the capsid is transported into the nucleus where viral gene transcription and DNA replication occur (steps 4 and 5). The expression of L genes, including those encoding tegument and envelope proteins, occurs in the cytoplasm (step 6). Acquisition of viral membrane by encapsidated genomes occurs in the assembly complex (step 7), and the virion is finally released from the infected cell by budding (steps 8 and 9). (Created with BioRender.com.)

1.4.2 Latent infection

Latency and reactivation are defining features of orthoherpesviruses and occur where viral genomes persist in cells in the absence of infectious virion production. In contrast to lytic infection of fibroblasts, where pp71 (UL83) localises in the nucleus to activate its transactivation function, pp71 is retained in the cellular cytoplasm during latency in haematopoietic cells (Saffert and Kalejta, 2006) (**Figure** 1-5 right). Progenitor CD34+ cells of the myeloid lineage are the cell type capable of supporting latency (Taylor-Wiedeman et al., 1991, Kondo et al., 1994). As viral entry into CD34+ HPCs occurs via endocytosis rather than fusion, retention of pp71 in endosomes restricts its access to the cytoplasm in such cells (Lee and Kalejta, 2019). HCMV latency and reactivation is a clinically significant problem, especially following transplantation, when host immune function is iatrogenically suppressed. Peripheral blood monocytes are the major site of carriage of HCMV DNA in otherwise healthy carriers (Taylor-Wiedeman et al., 1991), and circulating dendritic cells isolated from healthy seropositive donors have been shown to be sites of HCMV reactivation (Reeves and Sinclair, 2013).

Reactivation and productive viral replication occur when pluripotent CD34+ myeloid precursor cells differentiate to macrophages or dendritic cells, in the presence of human fibroblast cytokines (interferon gamma, tumour necrosis factor alpha, interleukin-4, or granulocyte-macrophage colony-simulating factor) (Hahn et al., 1998). Daxx along with PML and Sp100 form the ND10 nuclear bodies that bind to transcriptional factors and recruit HDAC (Hofmann et al., 2002, Hollenbach et al., 2002, Maul, 2008), thus inducing repression of HCMV gene expression and therefore the establishment of latency (Saffert and Kalejta, 2006) (**Figure** 1-5, right). Additionally, SUMOylation of PML is required to recruit Daxx (Ishov et al., 2002). HCMV reactivation has been shown to be primed by the expression of a latency-associated gene product (LUNA), which has an isopeptidase activity that deSUMOylates PML and disperses ND10 bodies, thus enabling HCMV IE gene expression (Poole et al., 2018).

More recently, Rozman and colleagues (2022) were able to characterise further the dynamics of HCMV gene expression kinetics using improved gene expression

classification and transcriptome-wide measurements of an array of epigenetic inhibitors. Utilising cycloheximide (CHX) to inhibit protein synthesis and phosphonoformate (PFA) to inhibit viral DNA replication, they demonstrate that latency is defined by the repression of IE genes (Rozman et al., 2022). In contrast to previous studies advocating the maintenance of latency through total silencing of viral gene expression (Sinclair and Sissons, 2006), this and other recent investigations of HCMV latency support the low-level expression of viral genes throughout latency (Cheng et al., 2017, Shnayder et al., 2018).

In reality, the various studies performed to date to establish the mechanism behind HCMV latency have involved many cell-culture based experimental models. A variety of primary human myeloid progenitor cells have been utilised, from human foetal liver, bone marrow, umbilical cord blood, mobilised peripheral blood and myeloid progenitor cell lines to primary human CD14+ monocytes. These are a wide-ranging source of cells and more importantly, ones at different stages of cellular differentiation. As the stage of differentiation of cells has been shown to promote either latency or reactivation, caution is advisable in extrapolating the findings of each study across the HCMV latency landscape.

1.5 HCMV immune evasion

A substantial amount of genetic information encoded by the viral genome is nonessential for growth in cell culture. Thus, the virulent low-passage Toledo strain encodes a segment of approximately 13 kb that is missing from the high-passage strains AD169 and Towne (Cha et al., 1996), and site-directed mutagenesis assays on these two high-passage strains have shown that 117 and 88 genes, respectively, were individually dispensable for growth in fibroblast culture (Dunn et al., 2003, Yu et al., 2003). Many of these dispensable genes encode immune evasion functions and can broadly be categorised into those that interfere with innate and adaptive immune responses (for example by preventing the presentation of cellular antigens to T cells and NK cells) and those that interfere directly with immune effector functions (e.g. UL111A, encoding a viral IL-10 cytokine that can directly disrupt cellular signalling pathways) (Cheung et al., 2009, McSharry et al., 2012). In immunocompetent hosts, an equilibrium is

maintained between viral replication and host immune control, preventing the spread of viraemia and thus HCMV disease; establishment of HCMV latency enables reactivation from viral reservoir in the event of disruption to this balance, for example by host immunosuppression.

1.6 HCMV treatment and prophylaxis

1.6.1 Antiviral agents

Antiviral agents (antivirals) for HCMV have been available for many years for immunosuppressed patients, targeting the UL97 phosphokinase or UL54 DNA polymerase to inhibit viral replication. First-line therapy is usually ganciclovir (GCV), a guanosine analogue. GCV is first phosphorylated by the UL97 phosphokinase, then by cellular kinases to its active triphosphorylated form, which competitively binds elongating DNA chains during viral replication. However, its use is restricted due to the potential for severe myelosuppressive side-effects. Cidofovir, another guanosine analogue, requires activation only by cellular kinases and so may retain efficacy if resistance emerges in UL97. Foscarnet has a different mechanism of action, directly inhibiting the UL97 phosphokinase. These antivirals are virostatic and act only to inhibit production of new virions; HCMV clearance relies on a robust host immune response. Mutations in UL54 are less common than those in UL97 and typically occur after exposure to foscarnet or cidofovir. UL54 mutations result in complete resistance to GCV, in addition to cross-resistance to cidofovir and foscarnet in many cases. Furthermore, these antivirals are associated with toxic side-effects, most significantly, myelosuppression (GCV) and nephrotoxicity (foscarnet and cidofovir).

In recent years, two new antivirals have been added to the armamentarium against HCMV, with distinct mechanisms of action. First, letermovir, a quinazoline, targets pUL56 of the HCMV terminase complex (pUL51/pUL56/pUL89). As there is no equivalent mammalian target, side-effects are minimal. However, trials of letermovir for the treatment of HCMV disease have not been successful as there is a low barrier against the development of resistance (Goldner et al., 2014). Nonetheless, it has been licensed for

prophylactic use in post-haematopoietic stem cell transplant patients (Ligat et al., 2018). Second, maribavir is an orally bioavailable benzimidazole riboside, which acts to inhibit viral DNA replication, encapsidation and nuclear egress (Biron et al., 2002, Krosky et al., 2003, Williams et al., 2003). Mutations in UL97 (V353A, L397R, L337M, T409M, H411L, H411N, H441Y, F342 and C480F) can confer moderate- to high-level resistance, whereas mutations in UL27 (R233S, W362R, W153R, L193F, A269T, V353E, L426F, E22stop, W362stop, 218delC and 301–311del) generally confer low-level resistance. Maribavir coadministration with inducers of cellular enzyme CYP3A4 (for example rifabutin and rifampin) is not recommended due to the potential for decreased efficacy of maribavir. However, there is no need to adjust the dose if administering with CYP3A4 inhibitors (for example azole antifungals such as ketoconazole and clarithromycin, which would increase plasma maribavir levels), as there is no dose-limiting toxicity, a wide therapeutic window, and less than a threefold increase in anticipated plasma levels of maribavir. Maribavir is potentially antagonistic to GCV and should not be coadministered. Maribavir has now been approved for the treatment of resistant or refractory HCMV infection post-transplant (defined as less than a 1 $\log_{10}$ IU/mL drop in HCMV load in blood after the appropriate antiviral treatment for ≥2 weeks) (Marty et al., 2019, Papanicolaou et al., 2019).

1.6.2 Vaccines

A major obstacle in the development of a prophylactic vaccine against HCMV has been the lack of protection against reinfection and reactivation by host immunological memory. The first HCMV vaccine was developed and trialled nearly 50 years ago, using the live-attenuated high-passage strains of Towne and AD169 (Elek and Stern, 1974, Neff et al., 1979). Subsequently, a vaccine based on recombinant gB proteins promisingly demonstrated a vaccine efficacy of 50 % in prevention of incident cases of maternal and congenital HCMV infection (Pass et al., 2009). This recombinant gB vaccine adjuvanted with MF59 was further trialled in a phase-II, randomised placebo-controlled study of adults awaiting kidney or liver transplants (Griffiths et al., 2011). Again, this trial showed that vaccinated patients had a significantly higher geometric mean titre of gB antibody production than those given placebo, and that this corresponded to the

virological correlates of a lower peak or shorter viraemia. There was a corresponding decrease in the duration of viraemia, and thus the number of days receiving anti-HCMV treatment. Other strategies have used a chimaeric vaccine involving the attenuated Toledo and Towne strains, which were shown to be well-tolerated in a Phase 1 study (Adler et al., 2016). Merck recently completed a phase 2 trial of an attenuated AD169-based vaccine (Clinical Trial Registration NCT03486834, 2021; https://clinicaltrials.gov/ct2/show/NCT03486834, accessed on 21 November 2021). However, the preliminary results suggested no demonstrable efficacy in preventing HCMV infection over placebo. Moderna are progressing with phase 3 trials following successful safety and immunogenicity assessments of mRNA-1647 (John et al., 2018), a vaccine based on lipid nanoparticles encapsulating mRNA encoding gB and gCIII (Clinical Trial Registration NCT05085366, Oct 20, 2021; https://clinicaltrials.gov/ct2/show/NCT05085366?term=NCT05085366&rank=1 , accessed on 19 August 2022). mRNA vaccines now have a proven track record of eliciting both potent humoral and cell-mediated response, and it is feasible that a heterologous prime/boost vaccination regimen with gCI, gCIII and pp65 could broaden the T cell response. It will be important to determine not only immunological but also virological correlates in the optimisation of any future vaccine developments (Suárez et al., 2017).

1.7 HCMV epidemiology

Seropositivity for HCMV ranges from 30 to 90 %, being greater among lower socioeconomic groups (Zuhair et al., 2019). In the immunocompetent host, infection is usually asymptomatic or mild, although severe disease in the form of hepatitis or colitis is recognised (Rafailidis et al., 2008). After primary infection, latency is established in CD34+ HPCs, and reactivation can occur during differentiation into CD14+ monocytes (Mendelson et al., 1996, Sinclair and Sissons, 2006). Reinfection with different strains is also possible, thereby complicating the natural history of HCMV infection (Baldanti et al., 1998, Boppana et al., 2001, Gorzer et al., 2010b, Zawilinska et al., 2016).

The two patient groups in which HCMV is clinically significant include the immunosuppressed and those who are congenitally infected. In particular,

iatrogenic immunosuppression in solid organ and stem cell transplant recipients is a risk factor for HCMV disease and is associated with significant mortality. Even if non-fatal, symptoms can be severe and can include pneumonitis, encephalitis, thrombocytopenia and hepatitis, and can lead to graft loss or graft-versus-host disease (GvHD) (Ljungman et al., 2011, Kotton et al., 2013). In neonates, congenital HCMV is the largest non-genetic cause of sensorineural hearing loss (SNHL) and neurodevelopmental delay, which have a detrimental socioeconomic impact (Manicklal et al., 2013). Disease course is variable and there are no clear prognostic markers to predict patient outcome. Numerous studies have investigated the association of viral diversity with clinical outcomes in the quest for such markers (Arav-Boger et al., 2008). Hypervariable HCMV genes have been investigated individually for any association of genotype with disease, and with the advent of HTS, in various combinations or as different viral strains. Notably, due to the large number of HCMV strains in circulation and the geographically restricted nature of clinical samples collected in some of these studies, it has been difficult to extrapolate conclusions about genotype associations across the studies. Undeniably, both host immunity and virulence factors have roles in mediating the consequence of HCMV infection and disease (Boeckh et al., 2003, Lanari et al., 2008, Kotton et al., 2013). In the next section, I will review variation in the HCMV genome as a potential marker, looking at past studies on HCMV strain variation and clinical outcomes and the advances in genomic sequencing that are facilitating higher resolution analyses. Ultimately, a more detailed understanding of the HCMV genome will enable a better understanding of HCMV disease and enable targeted vaccine and antiviral approaches.

1.8 HCMV genome variation

HCMV has the largest genome of the human members of subfamily *Betaherpesvirinae*, at 236 kb and encoding at least 170 functional proteins (Davison et al., 2013). HCMV exhibits the highest level of genetic diversity amongst the human orthoherpesviruses, and this diversity is stable through multiple passages *in vivo* throughout different populations worldwide (Stanton et al., 2005, Lurain et al., 2006, Bradley et al., 2008). Many HCMV genes are nonessential for viral replication in cell culture, and these genes tend to

function in immune evasion, cell tropism and latency (for example, UL111A, encoding an IL-10 analogue (Cheung et al., 2009). Gene-disrupting mutations have been identified in clinical strains of HCMV (Cunningham et al., 2010, Sijmons et al., 2015), and HCMV evolution has also been shaped extensively by pervasive recombination (Rasmussen et al., 2003, Sijmons et al., 2015, Lassalle et al., 2016, Hage et al., 2017, Pokalyuk et al., 2017).

There are several restricted regions of hypervariability within the HCMV genome. The hypervariable genes encode surface glycoproteins (e.g. gB, gO and gN) or have immune evasion functions (e.g. truncated TNF-alpha receptor UL144 and viral CXC chemokine UL146) and have evidently been under strong selective pressure at some stage during HCMV evolution. Multiple genotypes can exist for these hypervariable genes, and each genotype is denoted conventionally by G followed by a number or a number and a letter. Thus the seven genotypes identified for gB are G1-G7 (Meyer-König et al., 1998, Haberland et al., 1999, Suárez et al., 2019b), and the 14 genotypes of UL146 are G1-G14 (Dolan et al., 2004).

UL146 is perhaps the most hypervariable HCMV gene and functions in viral virulence. The 14 genotypes have nucleotide sequence non-identity of up to 58 % between genotypes (Penfold et al., 1999, Dolan et al., 2004, Bradley et al., 2008) (**Figure** 1-7) with variation extending into the noncoding region between UL146 and the adjacent UL147 (Lurain et al., 2006). Furthermore, a common way for the virus to acquire host immune evasion genes on an evolutionary timescale is to incorporate them from the host and, in some instances, to generate paralogous gene families from them by duplication events, thus contributing to the ability of the genome to act as a genomic accordion, expanding and contracting the gene families throughout evolution (Sijmons et al., 2015). The presence of a paralogue (UL147) adjacent to UL146 is evidence of these processes having occurred (Arav-Boger et al., 2005).

Among the 12 families of paralogous genes encoded by the HCMV genome, the RL11 family is one of the most hypervariable (**Table** 1-6 and **Figure** 1-3). This family encodes a conserved domain that has also been identified within the human adenovirus E3 region and has some similarities to an immunoglobulin fold

(Chee et al., 1990, Davison et al., 2003a, Sekulin et al., 2007). The RL11 family genes have putative functions in immunomodulation and cell tropism and, like the UL146 family genes, are dispensable in cell culture (Dolan et al., 2004, Sekulin et al., 2007, Sijmons et al., 2015).

There is evidence that viral populations may exist within an individual host in varying mixtures in different body compartments (Baldanti et al., 1998, Renzette et al., 2013, Hage et al., 2017, Suárez et al., 2020). The mechanism for this phenomenon may involve the selection of a particular strain on the basis of cell tropism, bottlenecking of the viral population, or a *de novo* superinfection of a particular compartment (Renzette et al., 2013, Hage et al., 2017). Viral fitness has also been shown to be augmented by frequent coinfection of cells by trans-complementation of one HCMV strain by another (Cicin-Sain et al., 2005). *In vivo*, multiple-strain infection has been demonstrated not only in the immunosuppressed (Suárez et al., 2019b, Suárez et al., 2019a) but also in healthy, immunocompetent people (Novak et al., 2008, Gorzer et al., 2010b). Such infections can arise either by acquisition of more than one strain at primary infection (Baldanti et al., 1998) or by the accumulation of strains from multiple exposures over time (Boppana et al., 2001, Gorzer et al., 2010a). This is the likely explanation for the previous finding of the rapid evolution of HCMV on passage between compartments within a single host (Renzette et al., 2011); reinfection is common and may lead to multiple viral subpopulations, which can be erroneously interpreted as rapid divergence within an individual (Pokalyuk et al., 2017, Suárez et al., 2020).

One of the difficulties in interpreting historic strain and outcome association studies lies with the fact that the techniques used for viral isolation, amplification and genotyping have involved cell culture, PCR and Sanger sequencing, each of which can introduce technical bias. As cell culture is known to select mutants (Cha et al., 1996), studies using viruses grown in cell culture will have genomes that are not the same as those present originally in patients (Dolan et al., 2004, Stanton et al., 2005, Dargan et al., 2010, Wilkinson et al., 2015). Also, cell culture might select only certain strains present in clinical samples containing multiple strains. Selective PCR is restricted to a small region of the whole HCMV genome (Bradley et al., 2008, Görzer et al., 2010), and, like

the Sanger sequencing that usually follows it, is insensitive at detecting minor strains (Sahoo et al., 2013). Therefore, despite much research, the complexity of the HCMV genome, the limitations of the approaches used and the small size of most studies carried out to date leave the part played by viral variation in clinical outcome undetermined. Even at an early stage, Rasmussen *et al*. (2001) anticipated that an infinite number of combinations of hypervariable genes due to recombination is theoretically possible in HCMV strains, and that analysis of single genes would be insufficient to determine effects on clinical outcomes. This insight has indeed been borne out, as reflected in the following brief discussion of the studies performed to date.

1.8.1 Clinical outcomes

Several studies have attempted to correlate the genotypes of hypervariable genes in HCMV with clinical outcome. Among the genes encoding virion glycoproteins, UL55, UL73, UL74 and UL75, which specify gB, gN, gO and gH, respectively, have been analysed in greatest detail. Among the immunomodulatory genes, UL144, which specifies a truncated TNF-alpha receptor, and UL146 and UL147, which encode CXC chemokines, have been the most well-studied. Unless clinical outcome depends on one or very few of these genes, the huge number of HCMV strains circulating as a result of recombination during evolution would make seeking associations a hopeless task. Nevertheless, past studies were restricted to analyses of single genes for practical or technical reasons, and it is useful to glean findings from these studies that might be exploited by HTS of entire HCMV genomes.

UL55 encodes gB, which is an important mediator of cell receptor attachment and a key vaccine target (Section 1.3.4.1) (Isaacson and Compton, 2009b, Murthy et al., 2011). Neutralizing antibodies bind to several conformation-dependent antigenic epitopes of gB and to the linear epitopes within antigenic domains AD1 and AD2 (Spaete et al., 1988). It has also been found that 40-70 % of total serum virus-neutralising antibodies in previously infected individuals are directed against gB (Britt et al., 1990). The overall variability in the amino acid coding sequence has been estimated to be approximately 9.5 %, with the most variable

regions located at the 5’-end and the central domain encoding a proteolytic furin cleavage site (Meyer-König et al., 1998, Pignatelli et al., 2004).

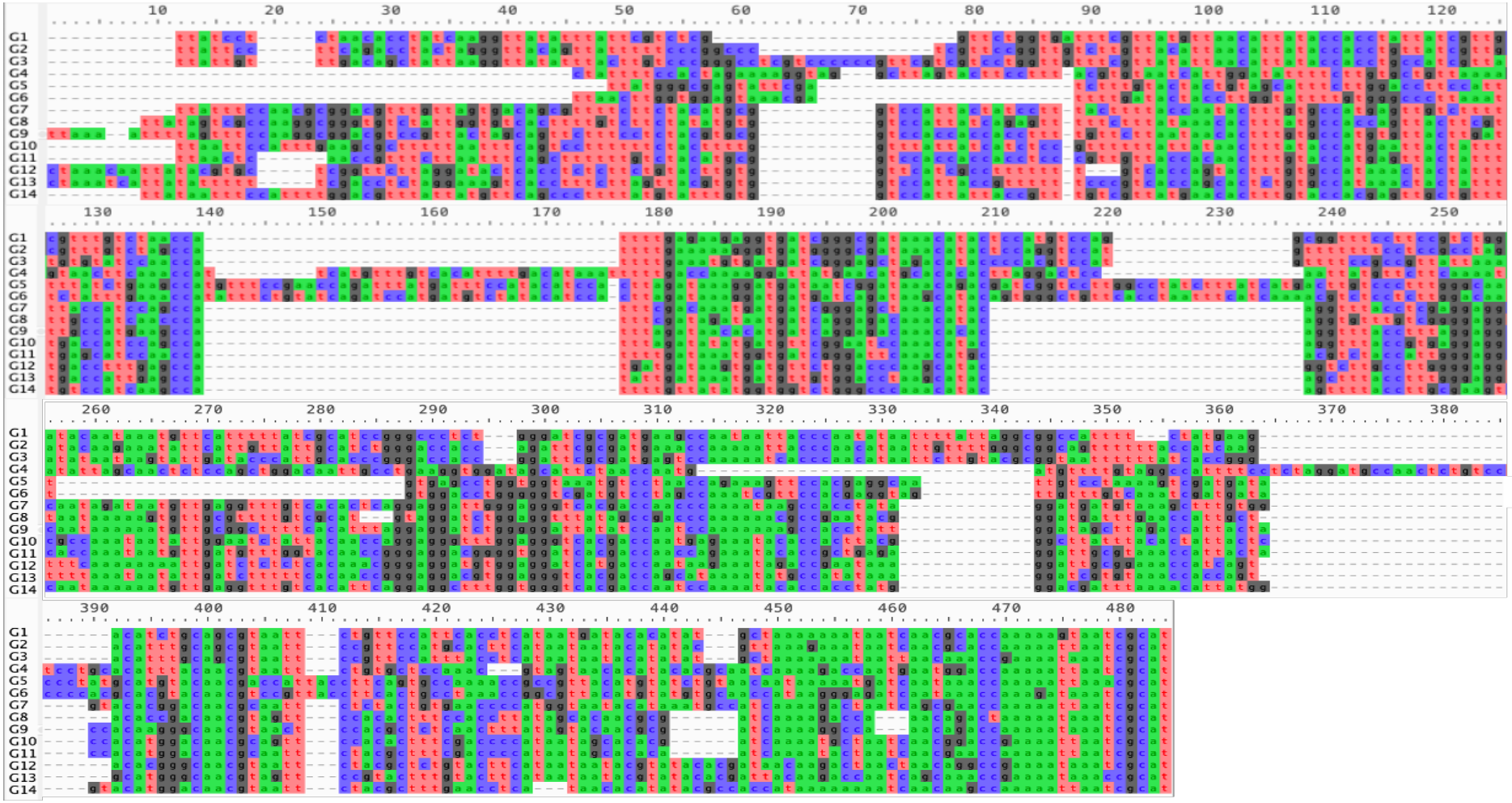


Figure 1-7. MAFFT alignment of sequences corresponding to genotypes G1 to G14 of hypervariable gene UL146.
Nucleotides labelled by colour, A, green; C, blue; G, grey; T, red. (Larsson, 2014, Suárez et al., 2019b)

Table 1-5. A list of 13 of the most hypervariable HCMV genes.
(Suárez et al., 2019b, Camiolo et al., 2021)

Gene	Family [a]	Product	Notes
RL5A	RL11	Protein RL5A	Involved in apoptosis; regulatory class TATA box and polyA signal sequence
RL6	RL11	Protein RL6	Type 1 membrane protein
RL12	RL11	Membrane protein RL12	Type 1 membrane protein
RL13	RL11	Membrane protein RL13	Type 1 membrane protein; consensus is wild type, consisting of several mutants present in different populations
UL1	RL11	Membrane protein UL1	Type 1 membrane protein
UL9	RL11	Membrane glycoprotein UL9	Type 1 membrane protein
UL11	RL11	Membrane glycoprotein UL11	Type 1 membrane protein; disrupts T cell function via inhibition of CD45; involved in immune regulation
UL20	Other, non-core	Membrane protein UL20	Type 1 membrane protein
UL73	Core	Envelope glycoprotein N	Envelope glycoprotein N; type 1 membrane protein; complexed with envelope glycoprotein M; involved in virion morphogenesis; involved in membrane fusion
UL74	Beta	Envelope glycoprotein O	Contains signal peptide; associated with envelope glycoprotein H and envelope glycoprotein L; involved in virion morphogenesis
UL120	UL120	Membrane protein UL120	Type 1 membrane protein
UL139	Other, non-core	Glycoprotein UL139	Type 1 membrane protein
UL146	CXCL	Chemokine vCXCL1	Contains signal peptide; putative secreted glycoprotein; involved in immune regulation

[a] Core genes are conserved across members of subfamilies *Alphaherpesvirinae, Betaherpesvirinae* and *Gammaherpesvirinae* and beta genes are present only within subfamily *Betaherpesvirinae*.

Homologous recombination contributes to the hypervariability of gB, both in cell culture and *in vivo* (Chou and Dennison, 1991, Haberland et al., 1999), and studies of the association of severity with some gB genotypes have led to variable results (Bale et al., 2000, Barbi et al., 2001, Arav-Boger et al., 2002, de Vries et al., 2012, Arellano-Galindo et al., 2014, Nijman et al., 2014, Paradowska et al., 2015, Rycel et al., 2015). As gB has been well-characterised as consisting of four genotypes (gB1-4) by previous restriction fragment length polymorphism (RFLP) studies (Chou and Dennison, 1991), it has been used extensively in the search for genotypic associations with virulence. Early studies in HIV patients suggested an association between gB2 and HCMV retinitis (Shepp et al., 1996), but it has transpired with larger studies that this was due to the demographic and geographic distribution of gB genotypes in this population (Chern et al., 1998, Fidouh-Houhou et al., 2001, Drew et al., 2002). One large study of liver transplant recipients showed an association of gB1 with a higher mean number of acute graft rejection episodes (p=0.027), but this was not confirmed by further studies (Rosen et al., 1998). In the transplant population, although no definitive association between specific genotype and severity of outcome has been made, the presence of multiple-strain infections has been shown to portend adverse effects, including progression to HCMV disease, as well as chronicity, higher HCMV loads, increased rate of graft rejection and poorer response to antiviral therapy (Fries et al., 1994, Vogelberg et al., 1996, Sarcinella et al., 2002, Coaquette et al., 2004, Puchhammer-Stöckl et al., 2006, Manuel et al., 2009, Zawilinska et al., 2016). It is not clear whether the association is causal, or whether the degree of immunosuppression predisposes such patients to multiple-strain infections.

Analysis of the more variable gN (UL73) has demonstrated an association of gN4 with an eightfold increase in antigenaemia in transplant patients (Rossini et al., 2005). Further studies also demonstrated the presence of multiple genotypes at a frequency of 0.1 % by ultradeep pyrosequencing (UDPS), confirming the suspicion that they may occur more frequently than are detectable by RFLP analysis, real-time PCR or PCR and Sanger sequencing (Gorzer et al., 2010a). gN has a function in cell entry and spread and is also highly immunogenic (Shimamura et al., 2006). Again, transplant recipients with multiple strain infections tended to fare worse and developed a higher level of viraemia with

longer decay kinetics and longer duration of antiviral therapy (Puchhammer-Stöckl et al., 2006, Gorzer et al., 2010a, Lisboa et al., 2012, Vinuesa et al., 2017). However, as more recent studies using HTS have detected an increasing proportions of multiple-strain infections, these results need to be interpreted with caution (Gorzer et al., 2010a, Vinuesa et al., 2017, Suárez et al., 2020).

Other studies on genotypic associations have focused on gO (UL74) and gH (UL75). Both glycoproteins are important for infection and cell-to-cell spread (Ryckman et al., 2008, Jiang et al., 2011). However, the numbers of strains analysed were small and the applicability of these individual genes as prognostic markers should be reanalysed in the context of HTS data (Fries et al., 1994, Puchhammer-Stöckl et al., 2006, Roubalova et al., 2011, de Vries et al., 2012, Pati et al., 2013, Paradowska et al., 2014b, Vinuesa et al., 2017).

To summarise, although there have been a number of individually well-designed studies seeking genotypic associations with clinical outcomes, even these were limited by the analysis of only one or a few genes, by sample sizes that were invariably small and statistically underpowered, by variations in the methods used and the fact that some studies involved cell culture and PCR, and by the complicating effects of multiple-strain infections on the understanding of HCMV diversity within an individual host. It is therefore difficult currently to draw any conclusions as to the significance of the reported findings on links (or lack of them) between viral genotype and clinical outcome. Rather, it is becoming apparent that the assessment of genome-wide variation, which is possible because of modern sequencing technology, will be essential in tackling this question.

1.9 HCMV transcriptome

The first fully annotated HCMV genome by Chee and co-workers in 1990 produced a preliminary map of 208 protein-coding ORFs. These were selected by criteria that filtered out ORFs smaller than a set threshold (300 bp) and smaller ORFs overlapped by larger ones while retaining ORFs that had homology to known genes or possessed known functional motifs (Chee et al., 1990). A similar approach has been taken to annotating HCMV genomes sequenced subsequently.

Although this approach may be relatively successful in predicting functional ORFs, it is less useful for predicting viral transcripts. Rather, it is necessary to carry out HCMV transcriptome profiling using different methods. This is a complicated and multifaceted process that involves: i) the detection of viral coding and noncoding RNAs; ii) the structural analysis of viral transcripts; iii) timepoint analysis of the differential expression of viral genes during infection; iv) the functional analysis of viral gene products; and v) detection of viral proteins. This endeavour to understand HCMV pathogenesis in the human host with regards to viral gene expression has been facilitated by powerful advances in molecular techniques.

1.9.1 HCMV RNA detection

Early HCMV transcriptome mapping was performed using conventional RNA-based methods, such as Northern blotting with reverse-transcription quantitative PCR, and DNA microarrays. The former technique relies on labelled mRNA hybridisation to a complementary single-stranded DNA probe and S1 nuclease to degrade non-hybridized regions of the probe, prior to the DNA-RNA fragments being separated by gel electrophoresis and visualized by autoradiography. This method is low-throughput and labour intensive and was superseded by transcriptome profiling using DNA microarrays. Microarrays are a powerful technique for analysing global viral gene expression, and can help identify key regulatory effects, including transcript quantitation, mapping of introns and locating the 5′- and 3′-ends of mRNAs (Chambers et al., 1999, Yang et al., 2006, Towler et al., 2012). However, this technique, like Northern blotting, is not agnostic and a prerequisite knowledge of the gene sequence is needed, with any novel or non-canonical coding regions remaining unidentified. Microarrays are also insensitive at detecting low-level expression and have a narrow dynamic range of detection due to high background and signal saturation. Sanger sequencing of cDNA to profile the HCMV transcriptome is expensive and low-throughput but can be useful when confirming novel transcripts and splice junctions. In the past decade, quantitative RNA sequencing (RNA-Seq), developed with the advent of HTS technologies, has overcome many of these limitations, superseding these traditional hybridisation-based methods, and has

made it possible to characterise the HCMV transcriptome at much higher resolution (Gatherer et al., 2011, Stern-Ginossar et al., 2012).

1.9.2 Structural analysis of viral transcripts

HCMV gene expression can produce a diverse collection of transcripts, with alternative polyadenylation, splicing, promotor use and RNA base modifications, which act to increase the coding capacity of a size-restricted viral genome (Romanowski and Shenk, 1997, Akter et al., 2003, Lurain et al., 2006, Gatherer et al., 2011, Stern-Ginossar et al., 2012). RNA-Seq studies have established that HCMV transcripts can be complex and can be expressed with different transcription start sites (TSSs) but coterminal ends, common TSSs but different transcription end sites (TESs), or nested transcripts differing at both TSSs and TESs, with partial sequence identity (Gatherer et al., 2011, Stern-Ginossar et al., 2012). Additionally, RNA splicing occurs when specific regions of the RNA transcript are cut out (introns) of primary transcripts, so that the flanking sequences (exons) encode a protein. Alternative splicing may occur; for example, high frequencies of alternative splicing have been found in RL8A, UL74, UL124 and UL150 (Stern-Ginossar et al., 2012). An example of an antisense transcript is the transcript of UL81-UL82, which encodes LUNA and has effects on latency, reactivation and lytic replication (Bego et al., 2005). The advantage of RNA-Seq in detecting noncoding RNAs, low abundance transcripts or RNAs in areas of low sequencing coverage requires a high sequencing depth, otherwise there is the risk of missing these features (Steijger et al., 2013). RNA-Seq also requires the reverse transcription of RNA into cDNA, which can lead to homology-dependent template-switching and identification of artefactual alternative transcripts (Cocquet et al., 2006). Due to the limitations of short-read sequencing, it is now known that the full transcript isoform landscape is complex and incompletely documented (Depledge et al., 2019a), and excludes spliced polyadenylated RNAs containing very short exons, those not originating via alternative splicing and with exons mapping <50 bp or >32 kb apart in the genome or those using non-canonical splicing (Gatherer et al., 2011).

1.9.3 Noncoding RNAs

Apart from encoding proteins, RNAs can also possess enzymatic functions or have regulatory roles in their own right (Mercer et al., 2009). Indeed, the majority of

polyadenylated viral RNAs are noncoding, nonoverlapping transcripts (NNTs) or long noncoding RNAs (lncRNAs) that do not substantially overlap the coding regions of other genes (Gatherer et al., 2011, Stern-Ginossar et al., 2012). At least 55% of the cDNA clones of transcriptional products in a study of HCMV during lytic infection of fibroblasts were completely or partially antisense to known or predicted HCMV genes (Zhang et al., 2007). Furthermore, lncRNAs (RNA2.7, RNA1.2, RNA4.9 and RNA5.0) form the majority (65.1 %) of polyadenylated transcripts detected in lytic infections (Gatherer et al., 2011) (**Figure** 1.3). lncRNAs are associated with regulatory roles, modulating gene expression and cellular processes. For example, a recent study reported that the most highly expressed lncRNA (RNA2.7) promotes cell movement and viral spread late in infection (Lau et al., 2021). Antisense transcripts have also been described within protein coding regions and occur throughout the HCMV genome. In addition, Stern-Ginossar and colleagues, using ribosomal profiling of lytic fibroblast infections at various timepoints, documented additional novel ORFs, derived from nested ORFs, short ORFs, antisense ORFs and short unpredicted ORFs coding for between 100-200 amino acid residues (Stern-Ginossar et al., 2012). Moreover, HCMV transcribes small, nonpolyadenylated RNAs and microRNAs (miRNAs) processed from longer transcripts in the OriLyt region and elsewhere in the genome (Meshesha et al., 2012, Stark et al., 2012); at least 16 pre-miRNAs and 26 mature miRNAs have been documented (Zhang et al., 2020).

1.9.4 Differential expression of HCMV genes

Expression of viral genes varies during the different phases of HCMV infection. Gene expression in lytic infection occurs in a cascade conventionally originating from IE, E and L genes and has been long been established from analyses of HCMV infection in permissive cell lines such as fibroblasts. Five temporal classes of viral protein expression have been defined by measuring viral protein profiles over time (Weekes et al., 2014). Latent HCMV infection in myeloid cells reactivates upon differentiation into mature DC or monocytes, suggesting that cell differentiation pathways act as determinants of reactivation (Hahn et al., 1998).

1.9.5 Functional analysis of HCMV genes

To investigate gene functions, recombinant HCMVs with targeted mutations have been introduced by recombinant engineering (recombineering) of BACs and reconstitution of mutated viruses by transfection of the mutant BACs (Ruzsics et al., 2013). Recombineering has facilitated both reverse genetics (to investigate the result of inactivation of a gene) and forward genetics (to investigate a gene essential for a certain phenotype). Although this approach makes it possible to make HCMV mutants with targeted point mutations, generation of parental BACs remains arduous, requiring insertion of the BAC vector into the HCMV genome by homologous recombination in cell culture followed by multiple rounds of clonal selection of BAC-containing bacterial colonies in order to obtain a monoclonal population of a single BAC.

1.9.6 Viral protein detection

Protein-coding transcripts seem to account for only a proportion of total transcript production. Also, 604 noncanonical HCMV ORFs have been identified by ribosomal profiling, which isolates and sequences ribosome-protected mRNA fragments (Stern-Ginossar et al., 2012). This number is far greater than the estimate of 170 canonical protein-coding genes proposed by genomic analytical methods (Gatherer et al., 2011) (Section 1.3.1 and **Figure** 1.3). In addition, quantitative temporal viromics, which employs multiplexed tandem-mass-tag-based mass spectrometry, has predicted the profiles of >80 % of HCMV canonical genes and 14 noncanonical ORFs (Weekes et al., 2014). Mass spectrometry-based techniques have also allowed host-viral protein interactions to be studied in detail (Nobre et al., 2019). However, it is not known whether any of the proteins generated from the novel ORFs highlighted by transcriptional or proteomic studies are functional or whether they are inconsequential byproducts of a loose replication system that the virus can tolerate.

1.10 Advanced sequencing technologies

1.10.1 Short-read sequencing

A review of studies on HCMV genomics provides an interesting chronology into the way that this field has developed, in particular its dependence on technology. We now know that earlier studies using cell culture for viral isolation and PCR for amplification were prone to bias, and that techniques such as RFLP analysis, PCR and Sanger sequencing (an example of first-generation sequencing), and even real-time PCR were highly limited in scope. The shortcomings of the older studies have now been overcome by the availability and applicability to clinical samples of HTS (or second-generation sequencing), which is improving at breakneck speed and carries a fraction of the cost per base of the older methods (Morey et al., 2013, Goodwin et al., 2016).

Traditionally, analysis of HCMV genomes from clinical samples required an amplification step. One method, isolation and passage in cell culture, leads rapidly to predictable and unpredictable mutations in the viral genome, rendering it an inaccurate model for clinical virus (Dolan et al., 2004, Stanton et al., 2005, Dargan et al., 2010, Wilkinson et al., 2015). Clonal amplification of HCMV genomes in BACs potentially ensures greater fidelity to clinical virus but is labour-intensive and still requires initial cell culture (Murrell et al., 2016). Progression to PCR followed by Sanger sequencing and phylogenetic analysis of the resulting sequences enabled a more accurate determination of genotypes. However, PCR is capable of introducing biases, including the selective amplification of certain genotypes and the generation of artefactual recombinants (Bradley et al., 2008, Görzer et al., 2010). Similarly, in transcriptomics, reverse transcription and amplification with DNA polymerase can introduce error and bias. Moreover, Sanger sequencing is by nature low-throughput and detects variants only if their frequency is relatively high (>20 %) (Sahoo et al., 2013). Therefore, for adequate investigations into effects of variation, recombination and multiple-strain infections on clinical outcome, these techniques fall short.

More recently, with the advent of HTS, it has become possible to analyse the HCMV genome in much greater detail, with much greater accuracy (Eckert et al., 2016, Hage et al., 2017), and at much lower cost per base (Morey et al., 2013). Similarly, RNA-Seq using HTS has enabled profiling of the viral transcriptome without *a priori* knowledge of the sequence, again at lower cost. Second-generation sequencers became available in the early 2000s and are dependent on massive, cyclic, parallel sequencing of clonally amplified and spatially separated DNA. The Illumina Genome Analyzer II was released in 2006, and was followed by a range of Illumina instruments (e.g. MiSeq, HiSeq and NextSeq) adapted for various needs in terms of throughput and turnaround times, all with error rates of <1 % per nucleotide (Goodwin et al., 2016). Illumina sequencing involves the generation of a DNA library, clonal amplification of the DNA on a flow cell via bridge amplification, and sequencing of the amplified clusters using reversible dye-terminator nucleotides (**Figure** 1-8, left). The main advantages of the Illumina platform over Sanger sequencing are speed and impressive sequencing depth. In addition to considerations of cost, the evaluation of any sequencing method focuses on throughput and read length, accuracy and depth. Illumina has a high degree of accuracy, making it suitable for applications looking into genome variation. Other second-generation sequencing platforms included 454 pyrosequencing (Roche) and Ion Torrent (Life Technologies), which relied on emulsion PCR for clonal amplification of adaptor-ligated DNA fragments on the surfaces of beads and again produced reads that were relatively shorter than, or similar in length to, those generated by Sanger sequencing (Erguner et al., 2015). However, these latter platforms had higher error rates, and this eventually resulted in Illumina instruments dominating the market, including that focusing on HCMV genomes and transcriptomes. Illumina-generated read lengths are also a few hundred bases at most, shorter than those generated by Sanger sequencing, and this presents challenges for using the data for *de novo* assembly or the detection of recombinants.

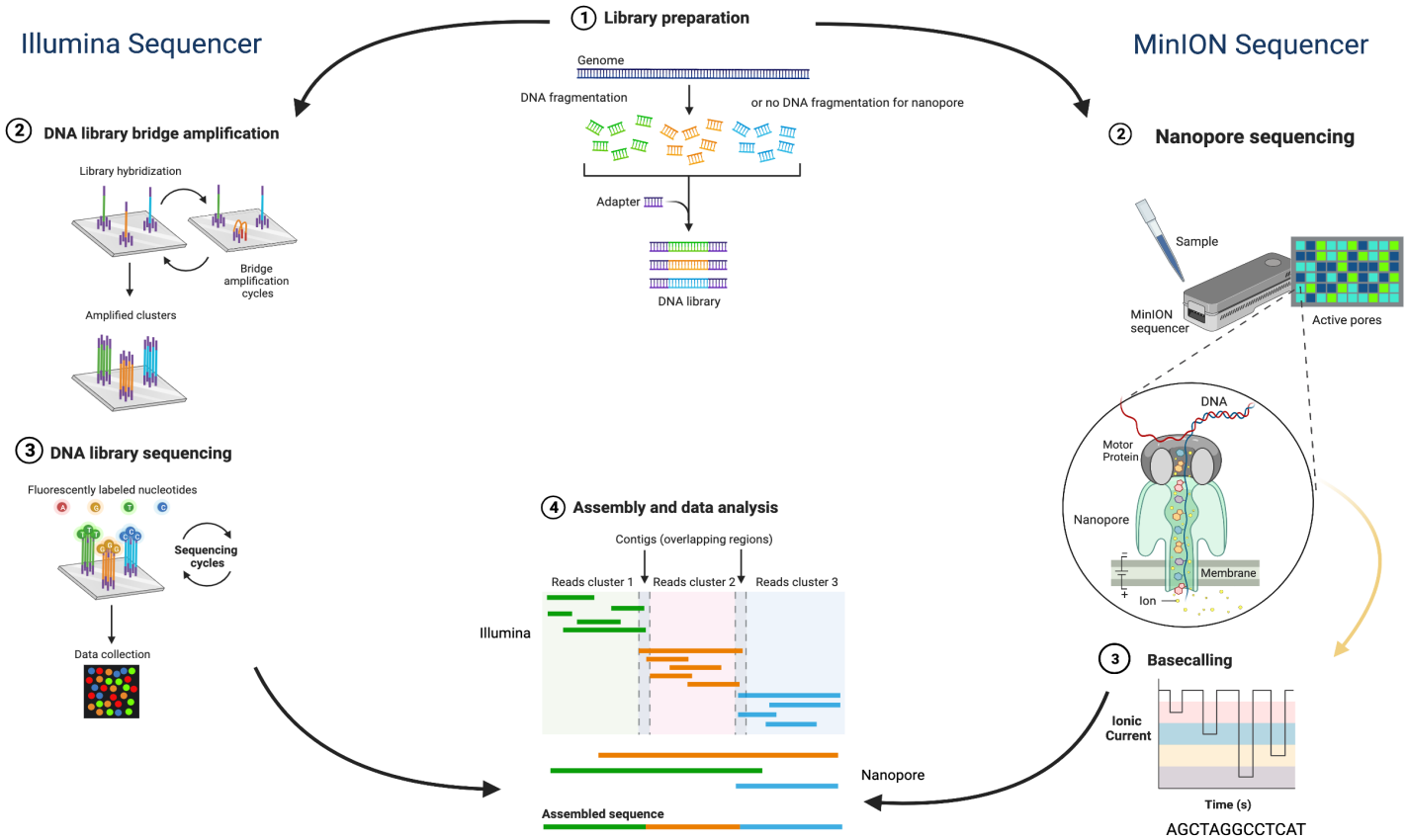


Figure 1-8. Comparison of Illumina sequencing and ONT MinION sequencing.
Illumina sequencing (left) performs massively parallel sequencing by synthesis, whereby adaptor ligated DNA hybridises to slide-fixed adaptors. Several hundred million clonal clusters are formed, generating high coverage depth. In ONT sequencing (right), a single DNA or RNA strand passes through a pore in a membrane and the detected change in current is translated into a sequence of bases (basecalling). ONT sequencing is therefore in real-time and produces much longer reads. Reads output from either platform are assembled using bioinformatic tools. (Created in BioRender.com.)

1.10.2 Long-read sequencing

Third-generation sequencing technologies offer a much longer sequencing read length. In principle, these platforms eliminate the need for clonal amplification prior to sequencing and are capable of recognising bases in an unmodified DNA or RNA strand (**Figure** 1-8, right). They rely on the use of nanotechnology and have been implemented in platforms from Pacific Biosciences (PacBio) or Oxford Nanopore Technologies (ONT). The chief benefit of these platforms is read length, which exceed 10 kb in each case and can reach >1 Mb for ONT platforms (Jain et al., 2017). In nanopore sequencing, single unmodified strands of DNA or RNA are passed through a pore in a membrane and the resulting change in electrical current is translated into a sequence in a process called basecalling. Theoretically, this can sequence whole viral genomes if intact sequences can be maintained during sequencing library preparation. For determining the relative contributions to HCMV variation afforded by recombination and multiple strain infections, the advent of ultralong reads has significant benefits over short reads. This is analogous to assembling two jigsaw puzzles that have been inadvertently mixed up – it is much easier to reassemble a puzzle if it consists of a few pieces rather than many millions (**Figure** 1-9).

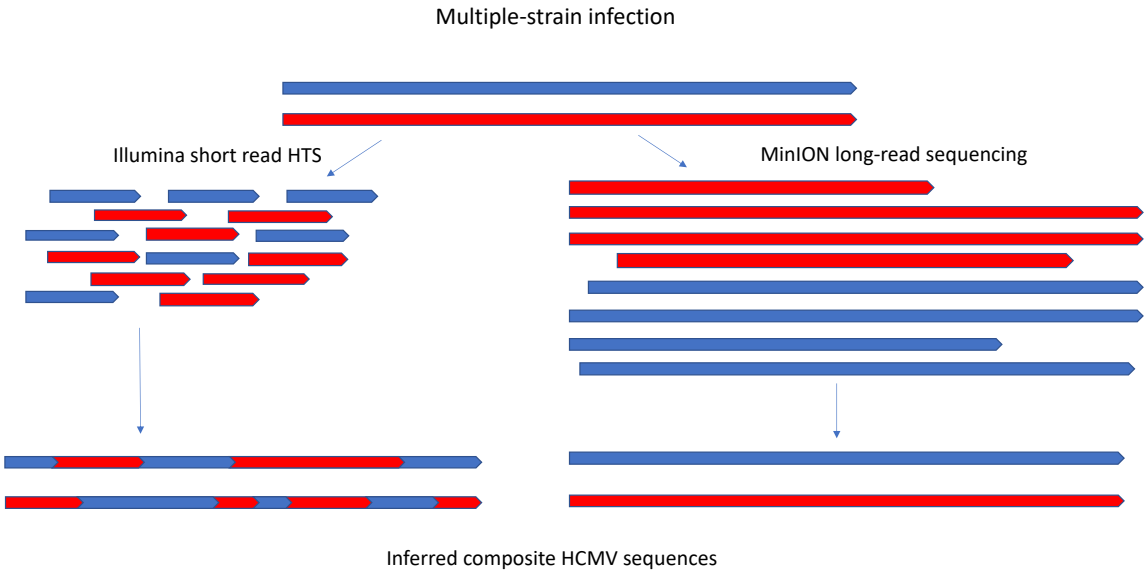


Figure 1-9. The advantage of reads provided by long-read sequencers over those provided by short-read sequencers.
Long reads can be used to assemble constituent HCMV genomes more easily than short reads.

The ONT MinION has an advantage over other third-generation sequencers by having a low capital cost (starter kits are available at $1000) and a small footprint (100 g). In particular, the high portability and USB power source of the MinION platform have facilitated its use in the field for real-time monitoring of viral outbreaks (Quick et al., 2016). On first release, basecalling had an unacceptably high error rate, which was up to 30 % for single reads. Methods to improve this have incorporated sequences generated by the more accurate Illumina platform (Madoui et al., 2015), but this undermines the benefits of the nanopore sequencing. Fortunately, improvements have been made by the release of improved pores and base-calling algorithms. The latest version of the pores uses a protein pump derived from *E. coli*, which reduces noise, and the latest analytical tools include a recurrent neural network, rather than hidden Markov model, 2D-reads (whereby both the forward and reverse strands of DNA are sequenced, to output a higher quality consensus), and better programs, such as Nanopolish and Nanocorrect, for improving the accuracy of reads (Jain et al., 2015, Jain et al., 2017, Loman et al., 2015). At the beginning of my project, the overall error rate had improved more than twofold, from 19 to 7.2 %, with miscalls at 2 %, insertions at 1.9 % and deletions at 3.3 % (Jain et al., 2017). The more current ONT updates to the flow cell (R10.4, early access release was available from Sept 2021) and basecalling have tackled homopolymer-calling accuracy to the extent that the majority of homopolymers at the consensus level are reportedly correctly resolved at lengths of <11 nt in R10.4 data, which is on a par with Illumina data (Sereika et al., 2022).

Furthermore, nanopore sequencing allows for direct sequencing without an amplification step. However, in clinical samples, the ratio of HCMV DNA may be too low relative to host DNA to generate sufficient numbers of viral reads. Enrichment protocols for viral DNA, including the use of bait hybridisation and host DNA depletion, may be a solution (Melnikov et al., 2011, Eckert et al., 2016, Brown and Christiansen, 2017), and real-time selective sequencing may be an option for minimising or avoiding host amplification (Loose et al., 2016). The development of a protocol to harness these technologies effectively would contribute greatly to sequencing HCMV, and viruses in general, in clinical samples.

Nanopore technology applied to direct RNA sequencing (dRNA-Seq) also offers the chance to characterise the structure of more complex transcript isoforms (**Figure** 1-10). With the Illumina platform, fragmentation (200-500 bp) is required, so it may not be possible to assemble complete isoform structures or characterise larger alternative transcripts (**Figure** 1-10**)**. The advent of long-read sequencing has thus enabled the initial characterisation of alternative HCMV and HSV1 transcripts (Balázs et al., 2018, Depledge et al., 2019b). Single-molecule long-read sequencing of intact transcripts offers an opportunity to capture these isoforms in greater detail.

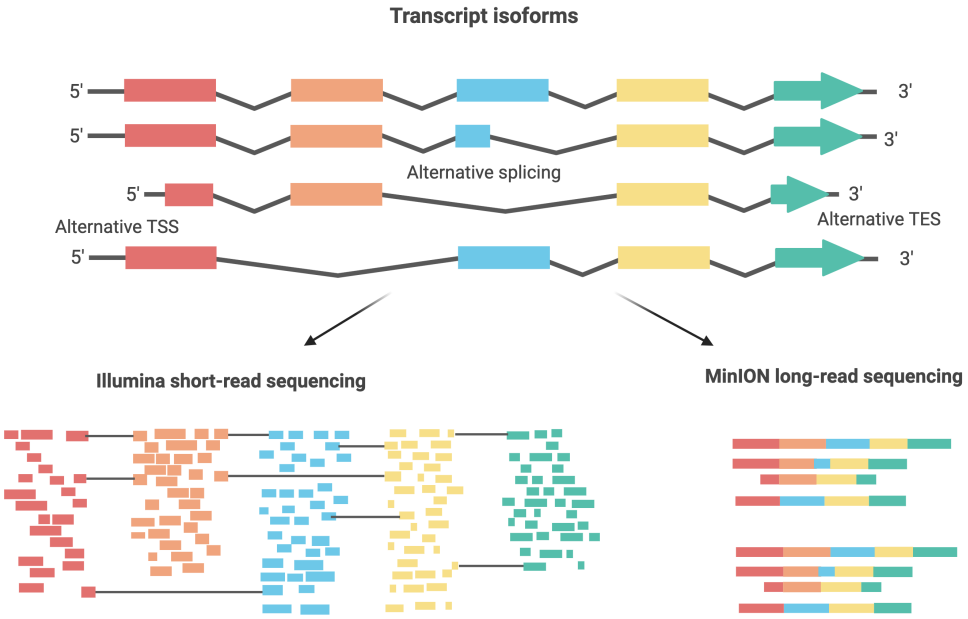


Figure 1-10. The advantage of long-read sequencing over short-read sequencing for the characterisation of transcript isoforms.
The coloured rectangles and the thin black lines connecting them represent introns and exons, respectively. TSS, transcriptional start site; TES, transcriptional end site. The same coding region is expressed as four different transcript isoforms, with alternative TSSs, splice sites or TESs, and is easily captured by long-read sequencing. Short reads may not be able to detect transcript isoforms with multiple splice sites. (Created in BioRender.com.)

Prior to the availability of long-read sequencing technology, RNA-Seq required transcripts to be fragmented and reverse transcribed to cDNA prior to sequencing, producing short reads, missing alternative TSSs and TESs and alternative splicing events. Depledge and coworkers utilised dRNA-Seq to profile the HSV1 transcriptome during productive infection of primary cells and defined

TSSs and RNA cleavage sites associated with all polyadenylated viral RNAs, identifying a novel class of chimeric HSV1 transcripts (Depledge et al., 2019b). Using the PacBio platform, Balazs and colleagues annotated an additional 291 transcript isoforms in HCMV, including eight novel antisense transcripts and a novel transcript (RS2) in TR_S, which was partially antisense to RS1, in the short repeat region (Balázs et al., 2017). They further utilised a combination of PacBio and ONT direct RNA sequencing to develop a pipeline to profile the HCMV transcriptome (Balázs et al., 2018, Kakuk et al., 2021). With the advent of a dRNA-Seq kit from ONT, the possibility exists of documenting the full repertoire of transcripts with differential structures, including those with multiple splice sites.

1.11 Project aims

HTS has helped to accelerate the study of HCMV genomics and transcriptomics. Nanopore sequencing, with its low capital cost and portability, is now also democratising sequencing as a tool for surveillance in outbreaks and for the rapid detection of resistance-associated mutations. HCMV subverts a multitude of human host immune defences, causing primary disease and secondary disease upon reactivation from latency. The latter is increasingly common in modern medicine where iatrogenic immunosuppression is often part of a patient's treatment. It is therefore important to determine the extent to which viral genome variation affects clinical outcome, and to identify relevant prognostic indicators.

In-depth profiling of the HCMV transcriptome may also help to discover novel targets for therapeutics or vaccines. Most studies to date have used technologies that have been limited in scope, with analysis performed at the level of individual genes rather than the whole genome. The advent of long-read sequencing can add higher resolution to that offered by short-read sequencing and offers the possibility for an integrated understanding of the roles of variation, recombination and multiple-strain infection in clinical outcome at the whole-genome level.

The aim of my research was to utilise the advanced sequencing technologies of Illumina and ONT to characterise HCMV from clinical samples. First, I used the

established methods of HTS with the Illumina platform to sequence HCMV from formalin-fixed paraffin embedded (FFPE) samples originating from congenitally infected infants. These are an important repository of pathology specimens, from which HCMV had not been previously sequenced, owing to the difficulty in obtaining intact DNA and the chemical changes induced by fixation. I then applied the ONT platform to sequence DNA from well-characterised high-titre cultured HCMV strains to determine the consensus error rate, and to demonstrate the ability of long-reads to detect two strains in an artificial mix and *in vitro* recombination events during coinfection. I then attempted to institute the nanopore pipeline from sequencing cultured samples to sequencing clinical samples directly. Finally, the direct-RNA sequencing application of nanopore was applied to characterise the lytic HCMV transcriptome. The goal of these individual studies was to contribute to the genomic characterisation of HCMV strains causing fatal congenital CMV disease, develop end-to-end sequencing pipelines for clinical samples using nanopore technology, and to explore the application of dRNA-Seq to the HCMV transcriptome.

2 General materials and methods

2.1 Cell culture

Human foetal foreskin fibroblast 2 (HFFF2 cells; European Collection of Authenticated Cell Cultures, Porton Down, UK; cat. no. 86031405) were cultured in 175 cm^2 culture (T175) flasks in 25 mL of growth medium (GM), which consisted of Dulbecco's modified Eagle's medium (DMEM) containing 10 % (v/v) foetal calf serum (FCS), 100 U/mL penicillin and 100 µg/mL streptomycin. When confluent (every 4-5 d), the monolayers were passaged by washing with 5 mL of versene (Gibco, ThermoFisher Scientific, Waltham, Massachussetts, USA; cat. no. 15040066) and incubating with 3 mL of trypsin-EDTA (0.5 g/L trypsin, 0.2 g/L EDTA and 0.85 g/L NaCl). The dislodged cells were resuspended in 2 mL of GM, split at a 1:4 ratio into fresh T175 flasks containing 25 mL of GM, and cultured further as described above. When confluent, the monolayers were washed, trypsinised and split into 1700 cm^2 roller bottles (Corning, New York, New York, USA; cat. no. 430852) by distributing the contents of 1.5 T175 flasks into each bottle with 100 mL of GM and 100 mL of CO_2. These and all subsequent incubations of uninfected and infected cultures were carried out in a humidified 5 % CO_2 incubator at 37 ºC.

2.2 HCMV strains

Working stocks of characterised HCMV strains were prepared (Section 2.3). Low-passage clinical strains AF1 (GenBank accession no. GU179291.1), U11 (GU179290.1) and Merlin (AY446894.2) were used (Dargan et al., 2010, Dolan et al., 2004). The RNA2.7 deletion mutant of strain Merlin, ΔRNA2.7 (Lau et al., 2021) was also used. The deletion in this mutant was at 2560–5050 nt in the Merlin genome (GU179001.1).

2.3 Preparation of working stocks of HCMV strains

Primary stocks of HCMV strains were diluted to the required volume to achieve a multiplicity of infection (MOI) of 0.15 plaque-forming units (PFU)/mL in two T175 flasks of HFFF2 cells grown to 75-80 % confluence. The volume was calculated as follows, accounting for the level of sub-confluency.

1 T175 flask contains: 1.5×10^4 cells/$cm^2 \times 175$ $cm^2 = 262.5 \times 10^4$ cells

PFU required = $(262.5 \times 10^4) \times 0.15 = 39.4 \times 10^4$ PFU

mL of stock required = 39.4×10^4/concentration of stock in PFU/mL

The GM was removed from the flasks, the inocula were added and swirled to ensure that the monolayers were covered, and the flasks were incubated for 3 h. The inocula were replaced by 25 mL of GM per flask, and the flasks were incubated overnight. The GM was replaced twice weekly until approximately 95 % cytopathic effect (CPE) was achieved. The infected flasks were washed and trypsinised (Section 2.1), and 6 mL of GM was added to each flask. For co-culture and expansion of cells, the dislodged cells from both flasks were combined and divided equally into eight 1700 cm^2 roller bottles of uninfected cells. The GM was replaced twice weekly until CPE was exhibited in most cells. Prior to the first harvest of cell-free virus in the supernatant, the amount of GM was reduced to 60 mL. The supernatant was removed from the roller bottles into 50 mL Falcon tubes (Corning; cat. no. CLS430290), and 60 mL of fresh GM was replaced into each roller bottle. The harvested supernatant was stored at -80 ºC to await virus concentration. The roller bottles were incubated for a further 2–3 d prior to the second harvest of cell-free virus in the supernatant.

The supernatant harvests were centrifuged at room temperature for 3 min at 277xg in a benchtop Megafuge 16R centrifuge (ThermoFisher Scientific, Waltham, Massachussetts, USA). The supernatants, in six 250 mL polypropylene centrifuge bottles that had been sterilised using 70 % (v/v) ethanol, were then ultracentrifuged at room temperature for 2 h 15 min at 29,994xg in an Avanti J-

25 ultracentrifuge (Beckman-Coulter, Brea, California, USA). The viral pellets were resuspended in a small volume of GM, combined, and aliquoted in 100 or 200 µL volumes into pre-labelled cryovials, which were stored at -80 ºC for use as working stocks. Viral titres were determined by plaque assay (Section 2.4).

2.4 Determination of viral titre by plaque assay

Confluent HFFF2 cells in a T175 flask were trypsinised as described above and recovered in 5 mL of GM. A 10 µL aliquot was pipetted onto a Neubauer counting chamber, and the cells were counted under a microscope. Cells were seeded at a density of 2×10^5 cells/well in nine-well flat-bottom plates (Corning; cat. no. 3516) and incubated for 1 d. Serial dilutions (10^{-3} to 10^{-8}) of virus stocks in GM were added to the cells (400 µL/well in duplicate) and incubated for 2-3 h. Semi-solid overlay was made by mixing equal volumes of GM and Avicel (FMC Corp., Pennsylvania, USA). The inocula were replaced with 2 mL of semi-solid overlay, and the plates were incubated. After 1 week, the overlay was removed and the monolayers were washed twice with 1 ml of phosphate buffered saline (PBS; Gibco, ThermoFisher Scientific) and incubated with Giemsa stain at room temperature for 1 h. The plates were washed with water and dried at room temperature overnight, and plaques were counted under a microscope. To minimise error, the dilution containing between 10 and 100 plaques was counted. Viral titre was expressed in terms of PFU/mL in the virus stock, and was calculated as an average of the replicates:

Viral titre = (average number of plaques per well/dilution factor[a]) × volume factor[b]

2.5 Testing for mycoplasma contamination

Cell cultures were tested routinely using a MycoAlert mycoplasma detection kit (Lonza, Basel, Switzerland; cat. no. LT07-318). On the rare occasions where contamination was detected, the cultures were treated every other day for 2 weeks with plasmocin (InvivoGen, San Diego, California, USA; cat. no. ant-pm-1)

[a] dilution factor = serial dilution of the well

[b] volume factor = 400 µL

and then grown in antibiotic-free media for an additional 2 weeks before being confirmed as mycoplasma-negative.

2.6 DNA extraction

2.6.1 Extraction of DNA from cultured HCMV stocks

The Qiagen genomic-tip 20/G extraction kit (QIAGEN, Hilden, Germany; cat. no. 10223) was used to extract DNA from high-titre cultured HCMV stocks. This kit preserves large DNA fragments of up to 150 kb for ONT sequencing and includes gravity-flow, anion-exchange tips that allow efficient purification of DNA from a wide range of biological samples. The protocol was modified to omit any vortexing steps. A 200 µL aliquot of a master stock of cell-free HCMV prepared as described above was resuspended in 1 mL of buffer G2 and mixed gently to lyse cellular debris and denature proteins such as nucleases, histones and viral particles. Proteinase K (25 µL; QIAGEN) was added, and the mixture was incubated at 50 ºC for 1 h to strip the DNA of bound proteins. The genomic-tip was equilibrated with buffer QBT (containing Triton X-100), and the sample was loaded onto the genomic-tip and allowed to flow through by gravity. The sample was then washed gently three times with 1 mL of buffer QC and eluted twice with 1 ml of buffer QF into a 10 mL collection tube. Isopropanol (1.4 mL) was added to the eluate to precipitate the DNA. This was mixed and centrifuged immediately at 5000x g for 15 min at 4 ºC. The centrifuged DNA pellet was washed with 1 mL of cold 70 % (v/v) ethanol and centrifuged at 5000x g for another 10 min at 4 ºC. The supernatant was removed carefully without disturbing the pellet, which was air-dried for 5 min, prior to the DNA being resuspended overnight in 50 µL of nuclease-free water (NFW) and stored at 4 ºC prior to nanopore sequencing.

2.6.2 Extraction of DNA from clinical samples

To extract DNA from residual clinical samples, which had small volumes and low DNA concentrations, the QIAamp DNA mini kit (QIAGEN; cat. no. 51304) was used. This kit can purify DNA molecules up to 50 kb in size. A 200 µL aliquot of a clinical sample and 200 µL of buffer AL were added to 20 µL of proteinase K in a 1.5 mL microcentrifuge tube, mixed and incubated in a heat block at 56 ºC for 10 min. Ethanol (200 µL) was added, and the tube was vortexed. The sample was

applied to a QIAamp mini spin column in a 2 mL collection tube and centrifuged at 6000x g for 1 min. The column was washed twice with 500 µL of buffer AW1, using a fresh 2 mL collection tube each time. The column was placed in a clean 1.5 mL Eppendorf tube, 200 µL of buffer AE was added, and the column was incubated at room temperature for 5 min and then centrifuged at 6000x g for 1 min. The eluted DNA was stored at 4 ºC prior to nanopore sequencing.

2.7 Concentration of DNA from clinical extracts

Due to the low volume and low DNA concentration of the residual clinical extracts obtained, the following steps were performed to concentrate the extracts prior to sequencing. The volume of extracted viral genomic DNA was increased to a total of 50 µL with 10 mM Tris-HCl (pH 8.5). Twice the volume of AMPure XP beads (Beckman Coulter; cat. no. A63881) was added and mixed by pipetting. The mixture was incubated at room temperature for 10 min and then placed in a magnetic rack until clear (approximately 3 min). The supernatant was discarded, and the beads were washed twice with 200 µL of freshly prepared 80 % (v/v) ethanol, waiting for at least 30 s each time and taking care not to disturb the beads. The beads were air-dried at room temperature for 3-5 min only, to avoid over-drying. The DNA was eluted in two steps, each involving adding 5.5 µL of 10 mM Tris-HCl (pH 8.5), mixing and incubating for 2 min, separating the beads in a magnetic rack, and saving 5 µL of supernatant from each elution in the same fresh 1.5 ml Eppendorf tube.

2.8 Assessment of DNA quality

2.8.1 DNA purity

This was assessed using a NanoDrop spectrophotometer (ThermoFisher Scientific). The ratio of absorbance at 260 nm to that at 280 nm (A260/A280) was used to assess DNA purity, with a ratio of approximately 1.8 indicating greater purity for DNA and a ratio of approximately 2.0 indicating the presence of RNA. A ratio appreciably lower than 1.8 indicated the presence of protein, phenol or other contaminants that absorb strongly at or near 280 nm.

2.8.2 Determination of DNA concentration

This was assessed using an Invitrogen Qubit fluorometer (ThermoFisher Scientific). The Qubit dsDNA HS assay kit (ThermoFisher Scientific; cat. no. Q32851) is accurate for initial DNA sample concentrations of 0.005-120 ng/μL, providing a detection range of 0.1-120 ng, and was used for samples with low input DNA. The Qubit dsDNA BR assay kit (ThermoFisher Scientific; cat. no. Q32850) is accurate for initial DNA sample concentrations of 0.2-2,000 ng/μL, providing a detection range of 4-2,000 ng, and was used post-amplification during library preparation for sequencing.

Prior to each assay, a fresh assay solution was prepared by diluting the stock fluorogenic dye supplied in each kit in the kit buffer in a 1:200 ratio. Kit standards 1 and 2 (10 μL) and 1-2 μL of sample were pipetted individually into clear, thin-wall 0.5 mL PCR tubes (ThermoFisher Scientific; cat. no. Q32856), and assay solution was added to a final volume of 200 μL. The tubes were vortexed for 3-5 s, incubated at room temperature for 2 min and read on the fluorometer.

2.8.3 Determination of DNA size

This was assessed using the 4200 TapeStation system (Agilent Technologies, Santa Clara, California, USA; cat. no. G2991BA) and a genomic DNA ScreenTape assay (Agilent Technologies; cat. no. 5067-5366), which is designed for analysing DNA molecules of 200 to >60,000 bp. The analysis provides automated, fast, reliable electrophoretic separation of input DNA molecules and gives a numerical measure of DNA integrity (DIN), which indicates the extent of fragmentation on a scale of 1-10, higher DINs indicating less fragmentation. The cut-off value for further processing was set at 7 for high-titre cultured samples. However, no cut-off value was set for clinical samples because they were irreplaceable.

A fresh marker ladder was prepared prior to each assay cycle by mixing 10 μL of genomic DNA sample buffer with 1 μL of genomic DNA ladder in the first tube in a PCR strip. The genomic DNA samples were prepared by mixing 10 μL of genomic DNA sample buffer with 1 μL of sample (10-100 ng/μL) in the remaining tubes. The strip was microfuged briefly, vortexed on a plate shaker for 1 min,

microfuged briefly again, and loaded into the TapeStation for electrophoresis on genomic DNA ScreenTape (Agilent Technologies; cat. no. 5067-5365).

For evaluating clinical samples with low DNA concentrations, the high sensitivity D5000 ScreenTape (Agilent Technologies; cat. no. 5067-5592) assay was used. The marker ladder was prepared by adding 2 µL of HS D5000 sample buffer to 2 µL of D5000Ladder, and the genomic DNA samples by mixing 2 µL of HS D5000 sample buffer with 2 µL of sample. The samples were then processed as described above.

2.9 Quantitative PCR of HCMV and human DNA

Quantitative real-time PCR (qPCR) was performed using the 7500 fast real-time PCR system (Applied Biosystems, Foster City, California, USA) as a means of identifying samples suitable for whole-genome amplification or direct sequencing. The samples with the highest ratio of HCMV to human DNA were chosen. The qPCR assay used primers and probes for a conserved region of HCMV gene UL97 (Slavov et al., 2016) (**Table** 2-1) and for a region of human gene *FOXP2* (Soejima et al., 2012) (**Table** 2-2). The amplification conditions were: 50 ºC for 2 min, 95 ºC for 10 min, 40 cycles of denaturation at 95 ºC for 30 s, and annealing and extension at 60 ºC for 1 min.

Table 2-1. HCMV gene UL97 primers and probe.
Genome positions refer to the HCMV strain Merlin sequence (AY446894.2). *VIC*: 2′-chloro-phenyl-1,4-dichloro-6-carboxyfluorescein. *MGB*: minor groove binder.

Name	Type	Genome Position	Sequence (5' --> 3')
UL97_F	Primer	143,285–143,303	ACCGTCTGCGCGAATGTTA
UL97_R	Primer	143,332–143,351	TCGCAGATGAGCAGCTTCTG
UL97_Probe	Probe	143,305–143,320	*VIC*-CACCCTGCTTTCCGAC-*MGB*

Table 2-2. Human gene *FOXP2* primers and probe.
FAM: 6-carboxyfluorescein. *MGB*: minor groove binder.

Name	Type	Sequence (5' --> 3')
FOXP2_FWD	Primer	TCACTACTAACAATTCCTCCTCGACTAC
FOXP2_REV	Primer	GATGAGTTATTGGTGGTGATGCTT
FOXP2_PROBE	Probe	*FAM*-TCCTCCAACACTTCC-*MGB*

2.10 Illumina short-read library preparation

Illumina libraries enriched for HCMV DNA were prepared in several stages.

2.10.1 Generation of DNA fragments

An aliquot of 50 µL of extracted DNA was sheared acoustically using an LE220 sonicator (Covaris, Woburn, MA, USA). To achieve an average fragment size of 500 bp, the following standard instrument settings were used: peak power = 175, duty factor = 10, burst cycles = 200 and time = 40 s. A Qubit HS dsDNA kit was then used to quantify DNA concentration (Section 2.8.2).

2.10.2 Generation of 5'-phosphorylated, end-repaired DNA fragments

An LTP library preparation kit (KAPA Biosystems, London, UK; cat. no. KK8232) was used to produce 5'-phosphorylated, blunt-ended DNA fragments. Sheared DNA (in 50 µL) was mixed with 6 µL of 10x end-repair buffer, 2 µL of end-repair enzyme mix and 2 µL of 10 mM Tris-HCl (pH 8.0 here and below unless stated otherwise), and incubated at 20 ºC for 30 min. The 5'-phosphorylated, end-repaired DNA fragments were purified using 0.85x (i.e. 0.85:1 bead to sample ratio) AMPure XP beads (Beckman Coulter) and eluted in 15 µL of 10 mM Tris-HCl.

2.10.3 Generation of DNA fragments with single deoxyadenosine residues at the 3'-ends

The 5'-phosphorylated, end-repaired DNA fragments were centrifuged briefly, and 2 µL of 10x A-tailing buffer, 1 µL of A-tailing enzyme and 10 µL of Tris-HCl were added. The mixture was incubated at 30 ºC for 30 min, and the A-tailed DNA fragments were purified using 1.4x AMPure beads and eluted in 15 µL of 10 mM Tris-HCl.

2.10.4 Generation of DNA fragments with adaptors at the 3'-ends

Qubit fluorometry with the high sensitivity DNA kit (Section 2.8.2) was used to calculate the concentration and volume of adaptors required for a 20:1 ratio of

adaptor to DNA. A mixture of 14 µL of A-tailed DNA (including the AMPure beads), 5 µL of 5x ligation buffer, 1 µL of T4 DNA ligase and the appropriate volumes of adaptor and NFW to bring the final volume to 25 µL was incubated at 20 ºC for 30 min. Uridine lesions were then removed by incubating with 1 µL of USER enzyme (New England Biolabs, Ipswich, Massachussetts, USA; cat. no. NEB #M5508) at 37 ºC for 15 min, and 24 µl of 10 mM Tris-HCl was added. The DNA was purified using 0.85x AMPure beads and eluted in 12 µL of 10 mM Tris-HCl.

2.10.5 Primary library amplification of DNA fragments

The library of adaptor-ligated DNA fragments was PCR-amplified with the primer-adaptors using a HiFi HotStart (high-fidelity, low-bias PCR) kit (KAPA Biosystems). To achieve this, 4 µL of the adaptor-ligated library was mixed with 7.5 µL of 2x KAPA HiFi HotStart ready mix, 0.5 µL of each primer (Illumina Universal 1.0 and 2.0 primers, 25 µM) and 2.5 µL of 10 mM Tris-HCl, and amplified under the following conditions: 95 ºC for 3 min, 12 cycles of 98 ºC for 20 s, 65 ºC for 15 s and 72 ºC for 30 s, 72 ºC for 2 min, and holding at 8 ºC. The number of cycles was limited to 12 to avoid artefacts resulting from amplification bias, high numbers of PCR duplicates and chimeric fragments. The sample was increased in volume to 30 µL by adding 10 mM Tris-HCl, purified using 0.85x AMPure XP beads, and eluted into 5 µL of 10 mM Tris-HCl. DNA concentration was measured by Qubit fluorometry, and fragment size was estimated using D1000 ScreenTape with the 4200 TapeStation (Agilent Technologies; cat. no. G2991BA).

2.10.6 Target enrichment of HCMV DNA

The PCR-amplified DNA fragments were processed using the SureSelectXT version 1.7 target enrichment system (Agilent Technologies) to enrich HCMV fragments in the final sequencing library. In clinical samples, HCMV DNA is present at a very low level relative to human DNA, and this step greatly increases the level. TruGrade oligonucleotides (Integrated DNA Technologies, Leuven, Belgium) capable of capturing sequences (including known variants) across the whole HCMV genome by hybridisation were designed for an RNA oligonucleotide bait library in 2014 by Dr Gavin Wilkie (Hage et al., 2017) on the basis of the 64 complete HCMV genome sequences available at the time. This bait library has

been used successfully in several studies sequencing complete HCMV genomes from clinical material (Govender et al., 2022, Hage et al., 2017, Suárez et al., 2017, Suárez et al., 2019a, Suárez et al., 2019b).

The protocol includes the use of blocking oligonucleotides, which prevent hybridisation of the universal adaptors ligated to the DNA fragments to each other and the consequent formation of long chimeric chains, which would lead to non-target DNA being pulled down with target HCMV DNA. For this step, the PCR-amplified DNA library in a volume of 3.4 µL was processed using the SureSelect system as follows (volumes are stated per reaction and were scaled up according to the total number of reactions). SureSelect block mastermix was prepared by mixing 2.65 µL of indexing block 1, 2.65 µL of indexing block 2 and 0.64 µL of indexing block 3. Each DNA library sample (3.4 µL) was mixed with 5.6 µL of the SureSelect block mastermix by pipetting. The sample mixes were incubated at 95 ºC for 5 min and held at 65 ºC for a minimum of 5 min.

During this incubation step, capture library hybridisation buffer mix was prepared by mixing 8.76 µL of hyb 1, 0.34 µL of hyb 2, 3.43 µL of hyb 3 and 4.64 µL of hyb4. RNase block for preventing degradation of RNA baits and tailored for capturing libraries of <3.0 Mb was also prepared (5 µL of a 1:9 dilution) and kept on ice. Then, 2 µL of bait capture library was added to the capture library hybridisation mix, having modified the capture library to suit samples with lower DNA concentrations (half reaction and quarter reactions for pre-capture DNA samples with <375 ng and <187.5 ng, respectively).

Thus, the 20 µL capture library hybridisation mix for each sample consisted of 13 µL of capture library hybridisation buffer mix, 5 µL of RNase block and the appropriate dilution of bait capture library in 2 µL. The PCR strip with the DNA libraries and SureSelect block mastermix were kept on a heat block at 65 ºC, and 20 µL of the capture library hybridisation mix was added to each sample well. The PCR strip was capped, sealed and incubated at 65 ºC overnight to complete target hybridisation.

2.10.7 Streptavidin-capture of hybridised DNA

DNA fragments that had hybridised to the RNA baits were captured using streptavidin-coated magnetic beads that had been washed and resuspended in binding buffer. All the hybridised DNA from the overnight incubation (25-29 µL) was transferred to a tube containing 200 µL of washed streptavidin-coated beads, incubated at 65 ºC for 10 min, and washed three times with wash buffer 2. Clean-up with 1x AMPure beads (Beckman Coulter) was then performed, followed by elution of the hybridised DNA into 12 µl of 10 mM Tris-HCl.

2.10.8 Indexing of enriched HCMV DNA fragments

Finally, the enriched HCMV libraries were amplified using indexing primers to enable multiplex sequencing. To avoid cross-contamination, the PCR mix for indexing was set up in a PCR cabinet with UV sterilisation and positive airflow. The mastermix was added with SureSelect ILM Indexing Post-Capture Forward PCR primer and an unique 6 bp index reverse primer per sample. The PCR plate was also set up with four standards run in triplicate.

The libraries were amplified in a thermal cycler to just below saturation to avoid the formation of heteroduplexes and other PCR artefacts. This was run on the following programme on an 7500 thermocycler (Applied Biosystems): 98 ºC for 2 min, variable cycles of 98 ºC for 30 s, 57 ºC for 30 s and 72 ºC for 1 min, final extension at 72 ºC for 10 min, and hold at 10 ºC. To optimise recovery of individual libraries, the amplification protocol was modified as follows. Real-time analysis of the multicomponent plot was performed to stop the PCR reaction once an amplification curve for a particular sample rose and crossed the curve of standard 3, taking note of the cycle threshold at which this occurs. The run was stopped manually just before the next round of PCR amplification, and the amplified sample was pipetted into a clean PCR tube and stored on ice. The PCR plate was replaced in the thermocycler, the programme was restarted at the beginning of the amplification cycle (98 ºC for 30 s), and the process was repeated. Samples reaching saturation after 18 cycles were deemed unsuccessful and not taken forward for sequencing. This process was repeated until all successfully amplified libraries were collected. The amplified libraries were

diluted to 50 µl with Tris-HCl and purified with 0.8x AMPure XP beads, washed in 80 % (v/v) ethanol and eluted in 15 µl Tris-HCl.

2.10.9 Multiplex Illumina short-read sequencing

Sequencing was carried out by the technical staff of the Genomics Facility at the MRC-University of Glasgow Centre for Virus Research (CVR). To ensure even coverage of the libraries, an equimolar pool was prepared for sequencing. Prior to loading, the concentration of DNA in the pool was quantified by Qubit fluorometry using the High Sensitivity dsDNA Assay, and the size of the fragments was determined using the TapeStation. The DNA concentration was diluted to 4 nM, and the DNA was denatured and diluted further to a final loading concentration of 1.35 pM. A 1 % spike-in of PhiX Control v3 (Illumina, San Diego, CA, USA; cat. no. FC-110-3001) was added at this point. PhiX Control v3 is an adaptor-ligated library that is used routinely as a control for Illumina sequencing runs. It provides a quality control for cluster generation, sequencing and alignment, and a calibration control for cross-talk matrix generation, phasing and pre-phasing. The reads can be rapidly aligned to estimate relevant sequencing-by-synthesis (SBS) metrics such as phasing and error rate. The pool was sequenced using a NextSeq 500/550 mid output kit v2.5 (300 cycles; IlluminaA; cat. no. 20024904). The loading amount was 1.35 pM, which achieved a density of 194,000 clusters/mm^2 and generated approximately 260 million paired-end reads of 151 nt. Since a single 6 nt index was used, a 6 nt indexing read was performed to allow the relevant reads to be assigned to each individual sample within the pool (de-multiplexing).

2.11 ONT long-read library preparation and sequencing

Genomic DNA was processed using a ligation sequencing kit 1D (ONT, Oxford, UK; cat. no. SQK-LSK109) and sequenced using R9.4.1 flow cells (ONT; cat. no. FLO-MIN106). The recommended amount of input DNA is 1 µg, and for HCMV DNA prepared from cultured samples the amount used was 2.5-7.9 µg. However, for clinical samples there was no enrichment step for HCMV DNA, and smaller amounts (100 ng) were used. The sample library was loaded onto a nanopore flow cell prior to being sequenced on the ONT MinION or the GridION platform.

2.11.1 Generation of 5'-phosphorylated, end-repaired DNA fragments

DNA repair and fragment end preparation were performed in 0.2 mL thin-walled PCR tubes by incubating each DNA sample with 3.5 µL of NEBNext FFPE DNA repair buffer, 2 µL of repair mix (New England BioLabs; cat. no. M6630), 3.5 µL of Ultra II end-prep reaction buffer, 3 µL of Ultra II end-prep enzyme mix (New England BioLabs; cat. no. E7546) and 1 µL of DNA CS (a control included in ligation sequencing kit 1D) in a final volume of 60 µL. The reactions were incubated in a thermal cycler at 20 ºC for 5 min and then 65 ºC for 5 min. The samples were processed further by adding 1.1x AMPure XP beads in LoBind tubes (Eppendorf, Hamburg, Germany; cat. no. EP0030108051), mixing by gently flicking, and incubated at room temperature for 5 min. The tubes were placed in a magnetic rack and washed twice with 200 µL of freshly prepared 70 % (v/v) ethanol, and the supernatant was removed by pipetting when clear. The pellet was dried at room temperature for only 30 s to avoiding over-drying and resuspended in 61 µL of NFW. The tubes were again placed in a magnetic rack until the supernatant was clear (approximately 1 min), and the supernatant was pipetted into a fresh LoBind tube.

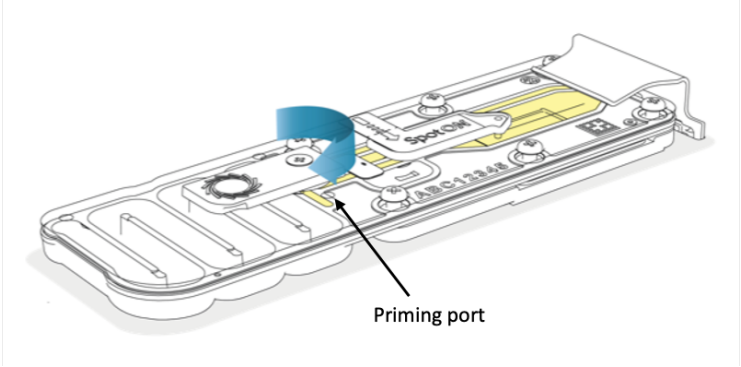
2.11.2 Generation of DNA fragments with adaptors at the 3'-ends and long fragment enrichment

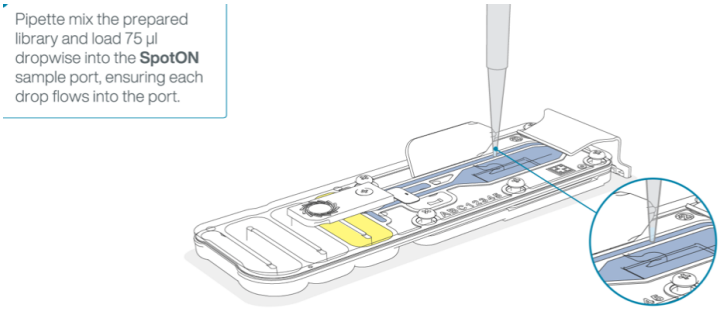
The end-repaired DNA (60 µL) was mixed in a fresh LoBind tube with 25 µL of ligation buffer, 10 µL of NEBNext Quick T4 DNA ligase (New England Biolabs; cat. no. E6056) and 5 µL of adaptor mix and the reaction was incubated at room temperature for 10 min. AMPure XP beads (40 µl) were added and mixed by pipetting, and incubation was continued for a further 5 min. The sample was microfuged briefly, placed onto a magnetic rack, and the supernatant was discarded. The beads were washed twice with 250 µL of long fragment buffer to enrich for large DNA fragments (≥3 kb), discarding the supernatant each time. The tube was again microfuged briefly and replaced on a magnet and allowed to dry for approximately 30 s. The beads were resuspended in 15 µL of elution buffer and incubated at room temperature for 10 min. After pelleting the beads in a magnetic rack, 15 µL of the eluate was transferred to a fresh 1.5 mL LoBind tube and stored on ice until ready to load into a flow cell.

2.11.3 Priming and loading of the flow cell

The flow cell was primed by removing any bubbles by drawing back a few µL from the priming port (**Figure** 2-1). The priming mix was prepared by pipetting 30 µL of the thawed and mixed flush tether into a tube of thawed and mixed flush buffer. Priming mix (800 µL) was then loaded into the flow cell via the priming port, avoiding the introduction of air bubbles, and allowed to settle for 5 min. Finally, the libraries were prepared for loading by combining the DNA library (12 µL) with loading beads (25.5 µL) and sequencing buffer (SQB, 37.5 µL). For libraries generated from high-titre cultured HCMV samples, loading beads were omitted from the final library preparation and substituted by the equivalent amount of NFW. The SpotOn sample port cover was lifted, and another 200 µL of priming mix was loaded into the flow cell via the priming port, avoiding the introduction of bubbles. The library (75 µL) was gently mixed by pipetting up and down and added dropwise to the SpotOn sample port, ensuring that each drop flowed into the port before adding the next (**Figure** 2-1). A final amount of 5-26 fmol of DNA was loaded into the flow cell, which was run for 48-72 h. Reads were acquired using ONT MinKNOW v.4.3.11 (https://community.Nanoporetech.com/downloads), and FAST5-formatted files were base-called using Guppy v.4.0 (https://community.Nanoporetech.com/downloads) in high accuracy mode with a minimum quality score of 7. Reads were assessed using MinIONQC (Lanfear et al., 2019), and adaptors were trimmed and any chimeric reads with internal adaptors discarded using PoreChop v.0.2.3 (https://github.com/rrwick/Porechop).

i)





ii)

Figure 2-1. Loading of an ONT R9.4.1 flow cell.
(i) The priming port is revealed by sliding open the cover (blue arrow). (ii) After priming, the prepared library is loaded into the flow cell via the SpotOn sample port in a dropwise fashion. Reproduced from ONT protocols.

2.12 Statistical analysis

Unless stated otherwise, comparison of the mean tests was undertaken using the unpaired, two-tailed Student's t-test with p values reported or stated as being not significant (ns). Analysis was performed using Prism v.9.4.1 (GraphPad Software, San Diego, CA, USA; https://www.graphpad.com/).

3 High-throughput sequencing of HCMV genomes in formalin-fixed paraffin-embedded tissues

3.1 Background

Congenital HCMV (cCMV) is relatively rare, affecting 1 in 100-150 live births (Dollard et al., 2007), and therefore formalin-fixed paraffin-embedded (FFPE) tissues are a valuable repository for the study of HCMV strains in cCMV disease. Formalin is used in histopathology to preserve tissues by preventing autolysis and putrefaction and maintaining the cellular architecture present in vivo by cross-linking proteins. This treatment enables tissues to be stored in a stable state for many years. Archived placental FFPE samples have demonstrated utility as an adjunct in the diagnosis of infants asymptomatic of cCMV at birth, and studies have used FFPE material to detect HCMV by immunohistochemistry or PCR amplification of short genomic fragments (Folkins et al., 2013, de la Cruz-de la Cruz et al., 2020). However, no published work has sequenced whole HCMV genomes from FFPE material. This has largely been due to the difficulty of recovering DNA of sufficient quality for sequencing from archived FFPE samples (Gilbert et al., 2007), as the characteristics that make formalin ideal for tissue preservation also adversely affect nucleic acid integrity.

There are several ways in which DNA is damaged during the formalin fixation process. Firstly, the formaldehyde component of formalin induces protein-protein, protein-DNA and inter-strand DNA crosslinks, and the cross-linking of DNA reduces the stability of dsDNA, leading to denaturation (Do and Dobrovic, 2015). Also, formaldehyde oxidises to produce formic acid, the low pH of which causes hydrolysis of purine bases in DNA chains, and thus the use of unbuffered formalin in tissue fixation results in the generation of abasic sites in DNA. Low pH also increases fragmentation of DNA, thus decreasing the amount of recoverable template. Finally, hydrolytic deamination of cytosine bases results in uracil lesions, in which artefactual C:G to T:A transitions are generated by the incorporation of adenine bases opposite the uracil lesions (Do and Dobrovic, 2012, Chen et al., 2014). Altogether, the effects of formalin can lead to sequence artefacts and false-positive mutation calls, rendering the use of FFPE repositories for HCMV strain characterisation an endeavour requiring caution.

However, recent improvements in extraction protocols specific for FFPE material have been driven by the demand for HTS investigation of somatic mutations associated with germline cancers using FFPE repositories (Carlsson et al., 2018, Bhagwate et al., 2019). As a result, there are now several commercial kits available for extracting high-quality DNA from FFPE material for downstream HTS applications (Janecka et al., 2015, McDonough et al., 2019). Most protocols involve a preliminary heat treatment step to reverse cross-linkages, and some incorporate uracil-DNA glycosylase (UNG) to remove uracil lesions, thus countering the effect of deamination from formalin fixation (Prentice et al., 2018). These kits also avoid the use of xylene, which is a toxic organic solvent that has traditionally been employed to remove paraffin from FFPE samples.

I aimed to optimise a pipeline for characterising whole HCMV genomes from archived foetal and placental FFPE blocks. For practicality, I assessed two CE-marked extraction kits that exclude xylene: one using spin-column technology (GeneRead DNA FFPE kit, QIAGEN) and the other using a paramagnetic bead-based approach (FormaPure DNA extraction and purification kit, Beckman Coulter). A target enrichment method for HCMV using a tailored oligonucleotide RNA bait library representing known HCMV sequences (Hage et al., 2017) was then used to prepare libraries for sequencing on the Illumina platform. My study was aimed at helping to develop a robust protocol for extracting, sequencing and analysing HCMV genomes from surplus FFPE samples, so that FFPE biorepositories may be confidently accessed in future for cataloguing HCMV strains associated with cCMV infections. In the broader picture, as discussed in Chapter 1, a fuller knowledge of the viral genome and its variability would enable a better understanding of the viral drivers of cCMV disease.

3.2 Objectives

1. To assess the feasibility of sequencing whole HCMV genomes from FFPE tissue samples from congenitally infected foetuses and neonates.

2. To establish a pipeline to sequence the genomes of HCMV strains causing severe congenital disease from FFPE repositories.

3.3 Materials and methods

3.3.1 Ethical approval

A total of 16 FFPE samples of foetal kidney or placenta from ten cases of cCMV in foetuses or neonates that occurred between 2008 and 2018 were identified in the Birmingham pathology laboratory archive by courtesy of Dr Phillip Cox, consultant perinatal pathologist at Birmingham Women's Hospital, UK. Delinked samples lacking associated patient-identifiable information (names, full dates of birth and hospital numbers) but including pseudonymised data (gestation at birth/death/diagnosis, histological findings, maternal infection type and antenatal findings, if available) were obtained after application for ethical approval through the Integrated Research Application System (IRAS, https://www.myresearchproject.org.uk/). Of the eight foetal demise cases, placental or foetal kidney tissue (or both) originated from intra-uterine death (n=5: cases 35, 150, 184, 239 and 473), termination of pregnancy (n=2: cases 70 and 413) and miscarriage (n=1: case 124). The remaining cases (n=2: cases 68 and 660) were infant kidney samples from neonatal deaths. Post-mortem findings from foetal examination demonstrated multi-organ effects, including encephalitis, pneumonitis, myocarditis and necrosis, with placental villitis found in all maternal tissues. The clinical details and histopathological findings of the ten cases are summarised in **Table** 3-1. Approval for sequencing HCMV from these samples was granted by the Health Research Authority Research Ethics Committee (HRA REC) (REC reference, 18/LO/1441; R&D number, 18/BW/NNU/NO17).

Table 3-1. Pseudonymised metadata from ten cases of cCMV infection used in the study.

Case no.	Age of sample (y) [a]	Tissue	Source[b]	Gestation/ Age[c]	PM findings[d]	Placenta histopathology	Maternal infection	Antenatal findings[e]
184	1	Placenta Kidney	IUD	20 w	IUGR, liver fibrosis, encephalitis/necrosis, inclusions in lung, liver, kidney, testis, thyroid, brain	Large, mild chronic villitis, abundant inclusions	Unknown	None
70	2	Placenta Kidney	TOP	38 w	Cerebral necrosis/meningoencephalitis, inclusions in lungs, pancreas, kidneys, brain	Normal size, mild plasmacytic villitis, occasional inclusions	Primary	Ventriculomegaly
150	2	Placenta	IUD	20 w	NA	Necrotising chronic villitis, plasma cells, inclusions	Unknown	SGA, echogenic bowel
413	2	Placenta Kidney	TOP	21 w	Cerebral necrosis/meningoencephalitis, polymicrogyria, vermis and corpus callosum present, splenomegaly, inclusions in lung, liver, kidney, pancreas, thyroid, adrenals	Small, severe plasmacytic villitis, occasional inclusions	Unknown	Echogenic bowel. There was also thought to be vermian agenesis, an indistinct cavum septum pellucidum raising the possibility of absence of the corpus callosum, and dilated cerebral ventricles
35	5	Placenta Kidney	IUD	22 w	Hydrops, large liver, dilated heart, pulmonary hypoplasia, scanty inclusions, normal brain	Large, hydropic, diffuse plasmacytic villitis, inclusions	HCMV IgG positive	Dilated heart, IUGR
239	5	Placenta Kidney	IUD	34 w	Microcephaly, hypoplastic corpus callosum and vermis, abnormal gyration, cholestasis, large spleen, inclusions in lung, liver, kidney, pancreas, brain	Small, plasmacytic villitis, numerous inclusions	Unknown	Growth restriction, borderline ventriculomegaly, posterior callosal deficiency, delayed sulcation with white matter volume loss, inferior vermian hypoplasia
473	6	Placenta	IUD	21 w	Micro-anencephaly, ventriculomegaly, hydrops, pulmonary hypoplasia, inclusions in lung, liver, kidney, pancreas, brain	Normal size, plasmacytic villitis, no inclusions	Unknown	HCMV DNA detected by PCR on amniocentesis
660	6	Kidney	NND	4 w	Splenomegaly, myocarditis, pneumonitis, hypoxic-ischaemic encephalopathy, inclusions in lung, pancreas	NA	Unknown	Born 35 w, IUGR, out of hospital cardiac arrest, resuscitated, ITU for 4 weeks
68	7	Kidney	NND	6 d	HCMV encephalitis and pneumonitis, inclusions in lung, kidney, ovary, adrenal, GBS pneumonia	NA	Unknown	Normal pregnancy, normal growth
124	7	Placenta Kidney	Miscarriage	19 w	Mild hydrops, chronic stress, liver necrosis, myocarditis, encephalitis, inclusions lung, liver, kidney, pancreas	Dichorionic diamniotic twin normal size. Hydropic villi, avascular villi, focal plasmacytic villitis, HCMV inclusions	Unknown	None stated

[a] The age of the sample is in years from when it was collected to when it was sequenced.

[b] IUD = intra-uterine death, TOP = termination of pregnancy, NND = neonatal death.

[c] Gestation in weeks, or age in weeks (w) or days (d).

[d] IUGR = intra-uterine growth retardation, GBS = group B streptococcus, NA = not available.

[e] SGA = small for gestational age, ITU = intensive therapy unit.

3.3.2 DNA extraction

DNA was extracted twice from each FFPE sample using a GeneRead DNA FFPE kit (QIAGEN; cat. no. 180134) and a FormaPure XL DNA reagent kit (Beckman Coulter; cat. no. C35996). From the ten individual cases, both placenta and kidney tissue were available from six, placenta only from two and kidney only from two. As outlined in **Figure** 3-1, each extraction protocol involved six major steps: deparaffinization, tissue digestion, de-crosslinking, enzyme treatment, binding and washing of DNA, and elution of DNA. Freshly cut sections of FFPE tissue (five to eight sections of 10 µm thickness) were obtained using a microtome, having discarded the first two or three air-exposed sections. The deparaffinization step involved n-hexadecane (GeneRead kit) or mineral oil (FormaPure kit), thus avoiding the use of xylene. Additionally, the GeneRead kit incorporated UNG, which counters the deamination effects of formalin. The extractions were performed using 1.5 mL Eppendorf tubes, and all heating steps were performed on a pre-set heat block.

The GeneRead extraction protocol was conducted as follows. For each sample, 160 µL of deparaffinisation solution (n-hexadecane) was vortexed with freshly cut sections and incubated at 56 ºC for 3 min. Tissue digestion was performed by adding 55 µL of NFW, 25 µL of FTB buffer and 20 µL of proteinase K, vortexing, centrifuging briefly and incubating at 56 ºC for 1 h. The sample was then de-crosslinked by incubating at 90 ºC for 1 h and centrifuged briefly to remove drops from the lid. The clear lower phase was transferred to a new tube, mixed with 115 µL of NFW, and incubated with 35 µL of UNG at 50 ºC for 1 h to remove uracil lesions. The sample was then centrifuged briefly, incubated at room temperature for 2 min with 2 µL of RNase A, vortexed with 250 µL of AL buffer, supplemented with 250 µL of 100 % (v/v) ethanol, and pulse-centrifuged (15,000x g for 10 s). The GeneRead protocol relies on spin-column technology for purifying DNA. The lysate (700 µL) was transferred to a QIAamp MinElute column (in a 2 mL collection tube) without wetting the rim and centrifuged (18,000x g for 1 min). The flow-through was discarded, and the bound DNA was washed and eluted using the same collection tube and centrifugation step. Sequential washes were performed with 500 µL of AW1 buffer, 500 µL of AW2 buffer and

250 µL of 100 % (v/v) ethanol. The spin-column was then placed on a fresh collection tube and centrifuged to remove all residual liquid. Finally, the DNA was eluted into a fresh tube by adding 45 µL of ATE buffer to the centre of the membrane, incubating at room temperature for 5 min, and centrifuging. The extracted DNA was stored at -20ºC.

The FormaPure extraction protocol was conducted as follows. For each sample, freshly cut sections were deparaffinized by immersing in 450 µL of mineral oil and incubating at 80 ºC for 5 min. The tissue was digested by adding 200 µL of LBA buffer, centrifuging at 10,000x g for 15 s, adding 20 µL of proteinase K to the lower phase, and incubating at 55 ºC for 1 h. The sample was then transferred to a heat block at 80 ºC for 1 h. The lysate in the lower phase was transferred to a new tube and incubated with 5 µL of RNase A at room temperature for 5 min. The FormaPure protocol relies on solid-phase reversible immobilisation (SPRI) beads in a binding (BBA) solution for purifying DNA. A 300 µL aliquot of SPRI beads in BBA solution was vortexed, added to the sample, mixed gently using a P1000 pipette (taking care to avoid air bubbles), and incubated at room temperature for 5 min. The tube was placed on a SPRIStand magnetic rack (Agencourt Bioscience, Beckman Coulter, Brea, California, USA) until the solution became clear (approximately 10 min), and the supernatant was discarded. The beads were washed in two steps. First, 400 µL of WBA buffer was added and mixed gently using a P1000 pipette. The sample was returned to the magnetic rack, the solution was allowed to clear, and the supernatant was discarded. Then, 750 µL of freshly prepared 80 % (v/v) ethanol was added and mixed gently using a P1000 pipette. The sample was returned to the magnetic rack, and the supernatant was again discarded, removing as much of the ethanol as possible. Finally, the DNA was eluted from the beads in 40 µL of NFW and incubated at 55 ºC for 1 min. The extracted DNA was placed into a fresh tube and stored at -20 ºC.

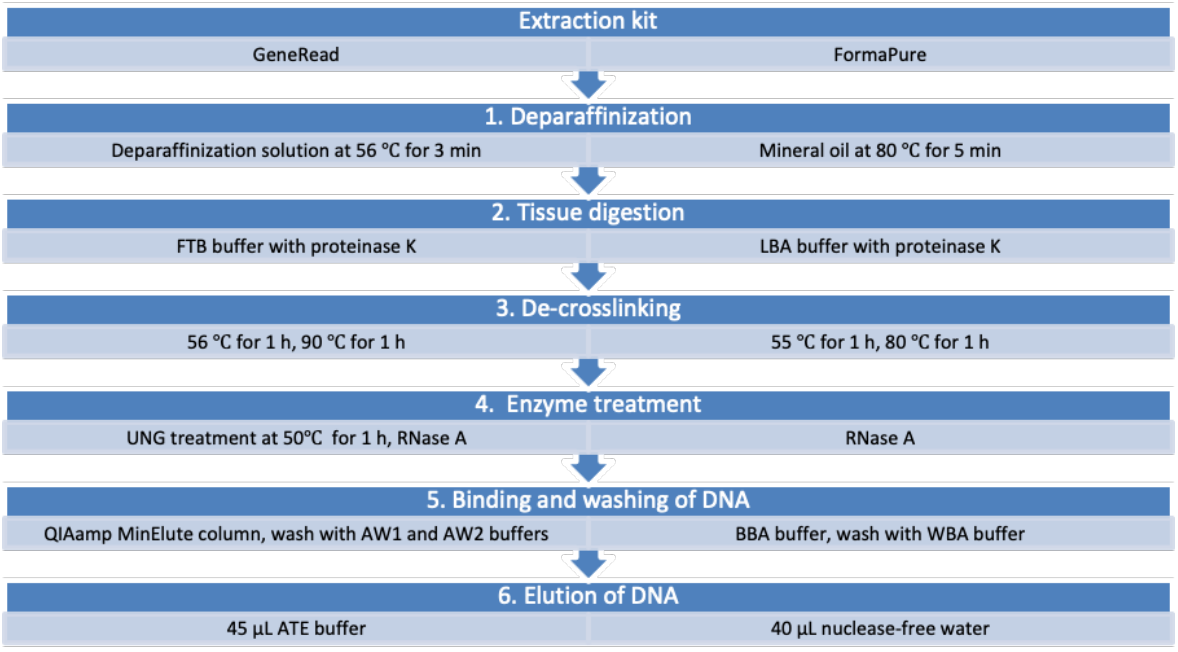


Figure 3-1. Main steps of DNA extraction using the GeneRead and FormaPure protocols.

3.3.3 DNA quantitation

As described in Section 2.8.2, the DNA concentration in the extracted samples was determined by Qubit fluorometry, and HCMV load and HCMV to human DNA ratio were determined by qPCR targeting the HCMV UL97 and human FOXP2 genes. The 16 FFPE samples received yielded 32 extracts (16 obtained by the FormaPure protocol and 16 obtained by the GeneRead protocol). The resulting datasets were labelled as the case number suffixed by a letter denoting whether the tissue was placental (P) or renal (R) and a flag denoting whether the GeneRead (_gr) or FormaPure (_fp) kit was used (Table 3-2). For example, the dataset generated from placental tissue from case number 184 extracted using the FormaPure kit was labelled 184P_fp. Quality checks were omitted in some instances, due to low the volume of available extract (NA in **Table** 3-2). Samples with a viral load of >100 IU/µL were processed for Illumina sequencing; thus, the sample from case 660 was not processed).

Table 3-2. Quality of DNA extracts as determined by DNA concentration and purity.

Case no.	Age of sample (y)[a]	Sample type	Dataset ID[b]	Qubit ng/μL	A260/A280	qPCR HCMV UL97 (IU/μL)	qPCR human FOXP2 (cp/μL)	HCMV:human DNA	HCMV viral load (log10 IU/μL)
184	1	Placenta	184P_fp	17.8	1.41	11911	672	17.72	4.08
			184P_gr	27.6	2.31	27698	960	28.85	4.44
		Kidney	184R_fp	NA	1.76	740	3861	0.19	2.87
			184R_gr	81.9	1.94	435	1073	0.41	2.64
70	2	Placenta	70P_fp	13.2	1.07	731	2061	0.35	2.86
			70P_gr	25.3	1.33	117	781	0.15	2.07
		Kidney	70R_fp	47.2	1.76	100732	43162	2.33	5.00
			70R_gr	NA	1.76	32962	3130	10.53	4.52
150	2	Placenta	150P_fp	NA	NA	NA	NA	NA	NA
			150P_gr	28.7	2.32	17726	939	18.88	4.25
413	2	Placenta	413P_fp	8.6	1.41	50974	21992	2.32	4.71
			413P_gr	177.6	1.69	6302	785	8.03	3.80
		Kidney	413R_fp	27.7	1.84	55849	45635	1.22	4.75
			413R_gr	28	2.22	6773	1116	6.07	3.83
35	5	Placenta	35P_fp	2.28	1.61	116574	13663	8.53	5.07
			35P_gr	68.3	2.16	12693	525	24.18	4.10
		Kidney	35R_fp	11.1	1.76	14647	25693	0.57	4.17
			35R_gr	37.9	2.05	745	562	1.33	2.87
239	5	Placenta	239P_fp	16.9	1.38	11204	2722	4.12	4.05
			239P_gr	25.4	2.45	5601	694	8.07	3.75
		Kidney	239R_fp	45.9	1.8	4135	498	8.30	3.62
			239R_gr	NA	NA	NA	NA	NA	NA
473	6	Placenta	473P_fp	0.629	1.38	3822	1119	3.42	3.58
			473P_gr	80.8	2.08	2974	485	6.13	3.47
660	6	Kidney	660R_fp	16.6	1.6	failed	failed	NA	NA
			660R_gr	36	2.22	86	141	0.61	1.93
68	7	Kidney	68R_fp	21.6	1.54	2198	1241	1.77	3.52
			68R_gr	49	2.11	2730	685	3.99	3.44
124	7	Placenta	124P_fp	NA	NA	NA	NA	NA	NA
			124P_gr	18.7	2.39	751	135	5.56	2.88
		Kidney	124R_fp	NA	NA	NA	NA	NA	NA
			124R_gr	92.1	2.03	125	125	1	2.10

[a] The age of the sample is in years from when it was collected to when it was sequenced.

[b] Dataset suffixes: P, placenta; R, kidney; _fp, FormaPure extraction kit; _gr, GeneRead extraction kit.

3.3.4 Illumina sequencing library preparation

Illumina libraries enriched for HCMV DNA were prepared in several stages, as described in Section 2.9. This included the following steps, with amendments specific to the processing of FFPE samples described in square brackets.

1) Generation of DNA fragments.

2) Generation of 5'-phosphorylated, end-repaired DNA fragments. [A minimum of 100 ng of DNA from FFPE samples was used when available.]

3) Generation of DNA fragments with single deoxyadenosine residues at the 3'-ends.

4) Generation of DNA fragments with adaptors at the 3'-ends.

5) Primary library amplification of DNA fragments.

Libraries were enriched further for HCMV DNA by hybridisation and capture (Sections 2.9.1.6 to 2.1.1.8) prior to indexing (addition of barcodes) by PCR in preparation for multiplex sequencing. Sequencing on NextSeq instruments (Illumina) was performed by the technical staff of the Genomics Facility at the MRC-University of Glasgow Centre for Virus Research (CVR) (Section 2.9.1.9).

3.4 Data analysis

GRACy (Genome Reconstruction and Annotation of Cytomegalovirus) is a computing toolkit for streamlining HCMV genome determination from Illumina sequence data (Camiolo et al., 2021). This user-friendly graphical user interface (GUI) application is written in python and automates various steps in the following modules: i) read filtering, ii) genotyping, iii) assembly, iv) annotation, v) single nucleotide polymorphism (SNP; a minor variant nucleotide) calling and vi) read submission (Figure 3-2). The read submission module was not used on datasets generated from sequencing the FFPE samples in this project.

Figure 3-2. The graphical user interface of GRACy.
Displaying the six modules for processing short-read Illumina sequence data obtained from clinical HCMV samples (https://github.com/salvocamiolo/GRACy/blob/273bc89a51034d8599e32d4ac8d22cbd3dcbef25/GRACy%20manual.pdf).

3.4.1 Read filtering

The read-filtering module can be used to trim sequencing adaptors, remove reads containing human sequences (which is useful for clinical samples), and deduplicate reads that are produced by the extensive use of PCR amplification during library preparation. Even when target enrichment for HCMV is used, datasets derived from clinical samples may contain a large proportion of human reads. Also, adaptors incorporated during library preparation and high levels of read clonality resulting from PCR amplification may lead to aberrant de novo assembly. Thus, filtering of human reads, adaptors and clonal reads is required prior to assembly. First, datasets were depleted of human reads using Bowtie 2 v.2.3 (Langmead and Salzberg, 2012), using the -local option with the human reference genome (Hg38; http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml), trimmed of default or user-specified adaptors and low-quality nucleotides using Trim Galore v.0.6.4 (https://github.com/FelixKrueger/TrimGalore), and deduplicated of clonal reads using FastUniq v.1.1 (Xu et al., 2012), which operates on the basis of paired reads sharing the same sequences.

3.4.2 Genotyping

The genotyping module utilises a motif-based approach to genotype 13 of the most hypervariable HCMV genes (RL5A, RL6, RL12, RL13, UL1, UL9, UL11, UL20, UL73, UL74, UL120, UL146 and UL139) (Camiolo et al., 2021) (**Table** 1-5). The positions of these genes within the UL region of the genome are shown in **Figure** 3-3. For each dataset, the module enumerates the occurrences of sequence motifs (kmers; sequences of k nt; in GRACy, k=17) that are specific to each genotype of each hypervariable gene (Suárez et al., 2019b, Camiolo et al., 2022). The number of kmers per genotype ranges from 2 to 1007 (Camiolo et al., 2021). Thus, the module can identify the HCMV strain (or strains) as a constellation of 13 genotypes (or multiple genotypes; each denoted as G followed by a number or a number and a letter) directly from the reads, without (and prior to) genome construction (Section 1.8). This is particularly useful when the HCMV data are limited or when there is incomplete genome coverage. GRACy only reports genotypes that are confirmed by an appropriate number of reads. This threshold is set as the number of reads identified for the genotype being ≥2 % of the average coverage of the whole Merlin reference genome. A folder is output with details of the genotypes detected for each of the hypervariable genes. Mixed genotypes may be detected if reads are detected with kmers specific to more than one genotype for a gene. The proportion of reads matching each genotype detected is given, as is the total number of reads for the gene.

The genotyping results are also plotted graphically as concentric annular rings representing one or more samples (**Figure** 3-4). This allows the easy identification of multiple genotypes present in a sample, with the caveat that the significance of each needs to be confirmed from the report. As a genotype is registered only when the number of reads with genotype-specific motifs is accounted for by >2 % of the total number of reads matching the gene, cross-contamination during sample processing can be largely filtered out. The number of strains in the sample is then estimated as the maximum number of genotypes detected in at least two genes (Suárez et al., 2019b).

Separately, each genotype designation is assigned a one-letter code (**Table** 3-3) so that the genotype constellation in a library of published 244 HCMV genomes can be represented as a 13-string code for comparison purposes (Suárez et al.,

2019b, Camiolo et al., 2022) (https://github.com/salvocamiolo/minion_Genotyper/blob/master/depositedSequences_codes.txt). For example, Merlin is represented by AANNNHACIPYHD (**Table** 3-4). This step facilitates the rapid identification of genotype constellations matching published HCMV genome sequences and the identification of novel HCMV strains.

3.4.3 Assembly

The assembly module was used to construct draft HCMV genome sequences from the datasets. It includes components dedicated to quality filtering, normalisation, sub-sampling, contig assembly, scaffold building, gap resolution and major variant and error correction (**Figure** 3-5). Several subsampled datasets containing 20-100 % of the reads were selected randomly and assembled into contigs using SPAdes v.3.12 (Bankevich et al., 2012). The contig set with the highest N50 value (that is, 50 % of the genome is contained in contigs longer than this length) was aligned to the Merlin reference genome to determine the position and orientation of each contig, and the contigs were joined to create a scaffold using Scaffold_builder v.2.0 (Silva et al., 2013) and Ragout v.2.2 (Kolmogorov et al., 2014). Gap resolution was achieved by locating the flanking 100 nt regions in a database of published HCMV genomes using BLASTn v.2.9 (Altschul et al., 1990). Alignment of reads to the genome sequence with the closest similarity to both flanking regions enabled the construction of a consensus, which was incorporated into the scaffold to fill the gap. Any remaining gaps after this iterative process were filled using GapFiller v.1.11 (Boetzer and Pirovano, 2012).

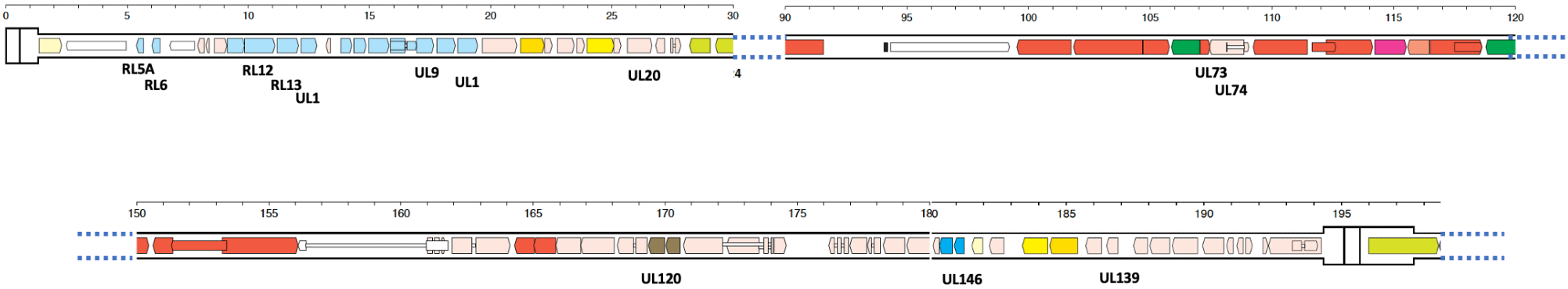


Figure 3-3. Locations of hypervariable genes within the UL region of the HCMV genome used by the genotyping module of GRACy.
The thinner outline represents the UL region, and the thicker outline the flanking inverted repeats (US starts at the right end). Blue hashed lines represent a section break in the genome, which is not displayed for clarity. (Modified from Figure 1-3.)

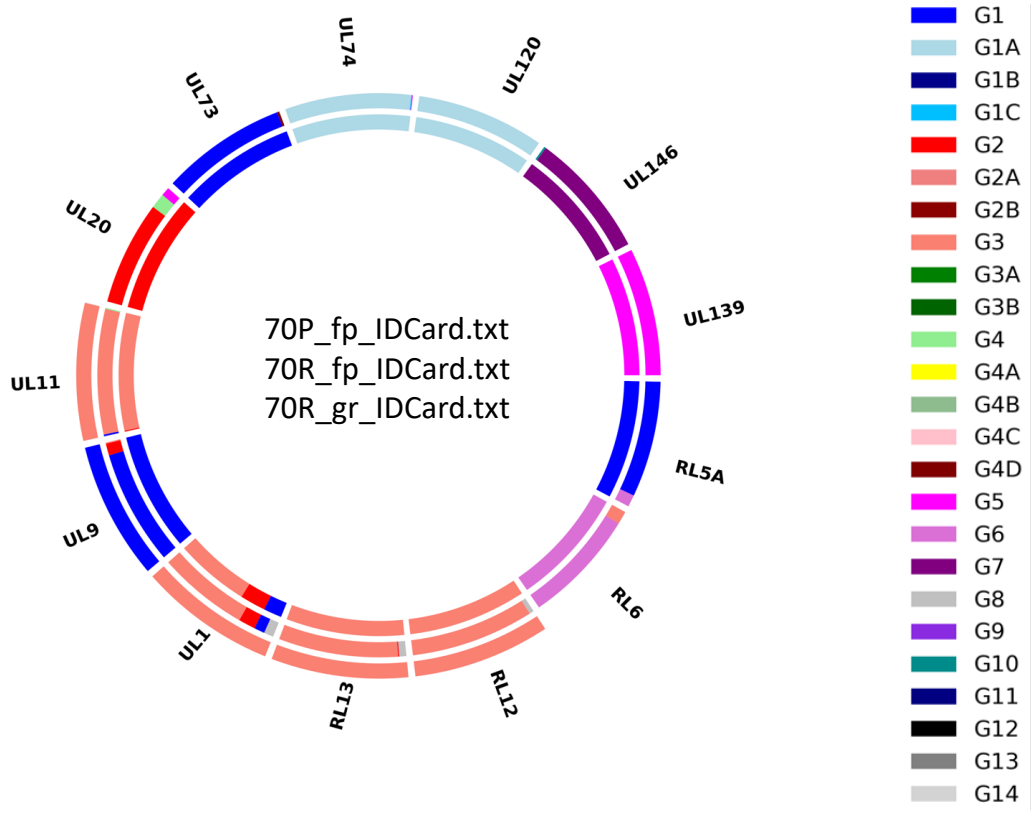


Figure 3-4. Example of the graphical representation of the genotypes detected in datasets processed by the genotyping module of GRACy.
Each ring denotes a dataset separated into the 13 most hypervariable HCMV genes as segments of the ring and labelled at the perimeter with the gene name. The genotypes present are represented by the colour codes displayed in the panel on the right. An absent segment block indicates that there were insufficient reads matching the kmer combinations for genotyping that gene (e.g. for 70P_fp, represented by the outer ring, no motifs specific to UL20, UL73, UL74, UL120, UL146, UL139, RL5A or RL6 were detected). The presence of more than one genotype in a sample is represented by multiple colours in a segment (as exemplified by genes UL20, UL11, UL9, UL1, RL13, RL12, RL6, RL5A and UL74 from dataset 70R_fp in the middle ring).

Table 3-3. List of known genotypes of the 13 hypervariable genes and the one-letter code assigned by GRACy.
The one-letter code remains the same regardless of the gene to which it refers. The genotype constellation of an HCMV genome can thus be expressed as a 13-letter code, with each letter corresponding to the genotype of a hypervariable gene.

Genotype	One letter code
G1	A
G1A	D
G1B	F
G1C	G
G2	H
G2A	K
G2B	L
G3	Q
G3A	W
G3B	E
G4	R
G4A	T
G4B	Y
G4D	I
G4C	O
G5	P
G6	C
G7	V
G8	N
G9	M
G10	S
G11	J
G12	Z
G13	X
G14	B

Table 3-4. Example of how the Merlin genome is expressed as a 13-string strain code.
Each hypervariable gene and its corresponding genotype name represented by its position in the string and the one-letter genotype code, respectively. Strain Merlin is therefore represented by the 13-string code AANNNHACIPYHD.

	Merlin	
Gene	Genotype	One-letter code
RL5A	G1	A
RL6	G1	A
RL12	G8	N
RL13	G8	N
UL1	G8	N
UL9	G2	H
UL11	G1	A
UL20	G6	C
UL73	G4D	I
UL74	G5	P
UL120	G4B	Y
UL146	G2	H
UL139	G1A	D

However, a scarcity of reads in some regions of the genome tended to leave some gaps unresolved (e.g. the G+C-rich, repetitive sequences in the centre of the genome between UL57 and RNA4.9 containing the origin of DNA replication, in addition to the b' sequence and some parts of the flanking *a'* and *c'* sequences). To address the *b'a'c'* region, the reads that aligned to a sequence of approximately 10,000 nt of the Merlin reference genome containing this region were extracted and assembled de novo using SPADes. The resulting contigs were then scaffolded to the Merlin reference genome (using Scaffold_builder), which resulted in an improved consensus for the *b'a'c'* sequence.

Coverage depth was calculated at each position by aligning all reads to the draft genome (using Bowtie 2). Misassembly errors were then identified as having low or no coverage and were removed from the assembly. Gaps were filled by an internal algorithm using an overlap/layout/consensus (OLC) method, which employs read overlap to extend flanking regions. Having produced an improved genome consensus, the terminal *ab* and *ca* inverted repeats were formed by reverse complementation of the internal *b'a'* and *a'c'* sequences and added to the ends of the sequence, and, if necessary, filling any gaps between these sequences and the rest of the genome by the OLC method. Major variants, including substitutions, indels and variations in homopolymer tract length were identified using Varscan 2 v.2.4 (Koboldt et al., 2012). The most frequent variants were incorporated, and any residual errors were corrected.

Finally, the GRACy draft genome was curated manually by aligning the reads to this draft genome (using Bowtie 2), inspecting the alignment in Tablet v.1.19.09.03 (Milne et al., 2016), and instituting final corrections and improvements to obtain the final genome.

3.4.4 SNP calling

The GRACy SNP (variant) calling module trims the reads in a dataset (if this has not already been done in the trimming module) using Trim Galore and then quality-filters the trimmed reads using PRINSEQ v.0.20 (Schmieder and Edwards, 2011) with stringent parameters (*-min_qual_mean 25 -trim_qual_right 30 -trim_qual_window 5 -trim_qual_step 1 -min_len 80 -trim_ns_right 20)*. The

trimmed, quality-filtered reads were then aligned to the genome using Bowtie 2 v.2.4.2 (Langmead and Salzberg, 2012), and the aligned reads were deduplicated using Picard v.2.21 (https://github.com/broadinstitute/picard/releases). Variants were called using LoFreq v.2.1.4 (Wilm et al., 2012), which considers only nucleotides with a Phred quality score of ≥30. The output listed the position and frequency of each SNP in relation to the genome, the affected gene, the modified codon and the amino acid substitution if the mutation was non-synonymous.

3.5 Results

3.5.1 HCMV genome sequences

In total, five complete HCMV genomes (cases 35, 150, 184, 239 and 413) were obtained using the GRACy pipeline from 16 samples originating from nine unique cCMV cases (**Table** 3-4). Of the ten original cases, a single case of neonatal death at 4 weeks of age (case 660) could not be sequenced (**Table** 3-1) This case had a very low level of HCMV DNA (144 IU/µL; **Table** 3-2), possibly reflecting postnatal treatment with ganciclovir for cCMV; however, ascertainment of treatment history was not possible with the retrospective metadata available.

Both the GeneRead and FormaPure kits successfully extracted DNA from FFPE samples for sequencing. Complete genomes were obtained from FFPE samples that had been stored for no longer than 5 years from time of fixation to sequencing and had a minimum coverage depth of 1,904 reads/nt (**Table** 3-5). However, only four samples were stored for >5 years, and these had a very low HCMV loads (<3,802 IU/µL). A minimum viral load of 11,204 IU/µL (average 42,872 IU/µL) was required to obtain complete HCMV genomes from FFPE samples (**Table** 3-5), in contrast to the estimated 1000 IU/µL threshold established for successful sequencing from non-FFPE clinical samples (Hage et al., 2017). GRACy assembled incomplete draft genomes for cases 68, 70 and 473, with the presence of many gaps (summing to 29,464, 6,360 and 9,918 unresolved nucleotides, respectively) that could not be closed.

Table 3-5. Coverage statistics for the FFPE sequence datasets.
(Using Bowtie 2 with the ***--local*** option and the appropriate final genome as reference.)

Dataset name	Age of sample (y) [a]	Original reads (no.)	Trimmed reads (no.) [b]	Target reads (no.)	Target reads (%)	Coverage (reads/nt) [c]	Coverage all (reads/nt) [d]	Deduplicated reads (no.) [e]	Duplication factor [e]	GenBank accession no. [f]	Mutations
184P_fp	1	16,364,538	16,348,704	13,919,532	85	8,724	19,276	5,416,142	3.02	OR546128	RL5A, UL1, UL150
184P_gr	1	20,182,582	20,162,458	16,704,168	83	10,552		9,989,220	2.02		
70R_fp	2	9,762,406	9,644,848	1,194,788	12	616	1,750	3,611,788	2.67	NA	
70R_gr	2	13,762,840	13,732,790	1,881,980	14	1,134		6,603,148	2.08		
150P_fp	2	10,207,696	10,184,744	7,389,855	73	4,566	7,410	5,754,016	1.77	OR546127	None
150P_gr	2	17,007,560	16,986,544	4,529,233	27	2,844		13,950,366	1.22		
413P_fp	2	12,841,786	12,814,432	1,197,392	9	739	7,060	6,971,286	1.84	OR546130	None
413R_fp	2	16,105,508	16,079,892	775,709	5	477		10,969,092	1.47		
413R_gr	2	18,376,158	18,348,900	2,472,291	13	1,558		12,879,022	1.42		
413P_gr	2	17,447,048	17,377,286	6,982,420	40	4,286		5,689,606	3.05		
35P_fp	5	12,891,102	12,865,982	3,060,824	24	1,893	7,332	8,099,780	1.59	OR546126	RL5A, UL150
35R_gr	5	13,331,384	13,282,588	239,218	2	142		7,112,434	1.87		
35P_gr	5	21,232,492	21,189,510	8,540,357	40	5,297		12,783,670	1.66		
239P_fp	5	18,438,992	18,378,390	604,116	3	365	1,904	15,245,000	1.21	OR546129	None
239P_gr	5	20,297,308	20,274,658	2,459,534	12	1,539		15,443,438	1.31		
473P_fp	6	18,865,854	18,847,210	430,527	2	324	324	8,680,490	2.17	NA	
68R_fp	7	19,176,338	19,129,082	647,447	3	422	422	7,763,736	2.46	NA	
124R_fp	7	2,087,288	2,085,358	3466	0.2	4	188	1,625,396	1.28	NA	
124P_fp	7	17,347,294	17,308,248	270,675	2	184		6,051,918	2.86		

[a] The age of the sample is in years from when it was collected to when it was sequenced.

[b] The read filtering module of GRACy was used to obtain input reads that were trimmed of sequencing adaptors, depleted of human reads and deduplicated of clonal reads (see Section 3.4.2).

[c] Average coverage depth by unique reads, the depth of coverage from combined datasets from each case.

[d] Average coverage depth by all reads, the depth of coverage from combined datasets from each case.

[e] Deduplicated read numbers were calculated using FastUniq for determining the duplication rate (clonality) of each dataset. The duplication factor is the ratio of the deduplicated reads to the original reads.

[f] NA, not available.

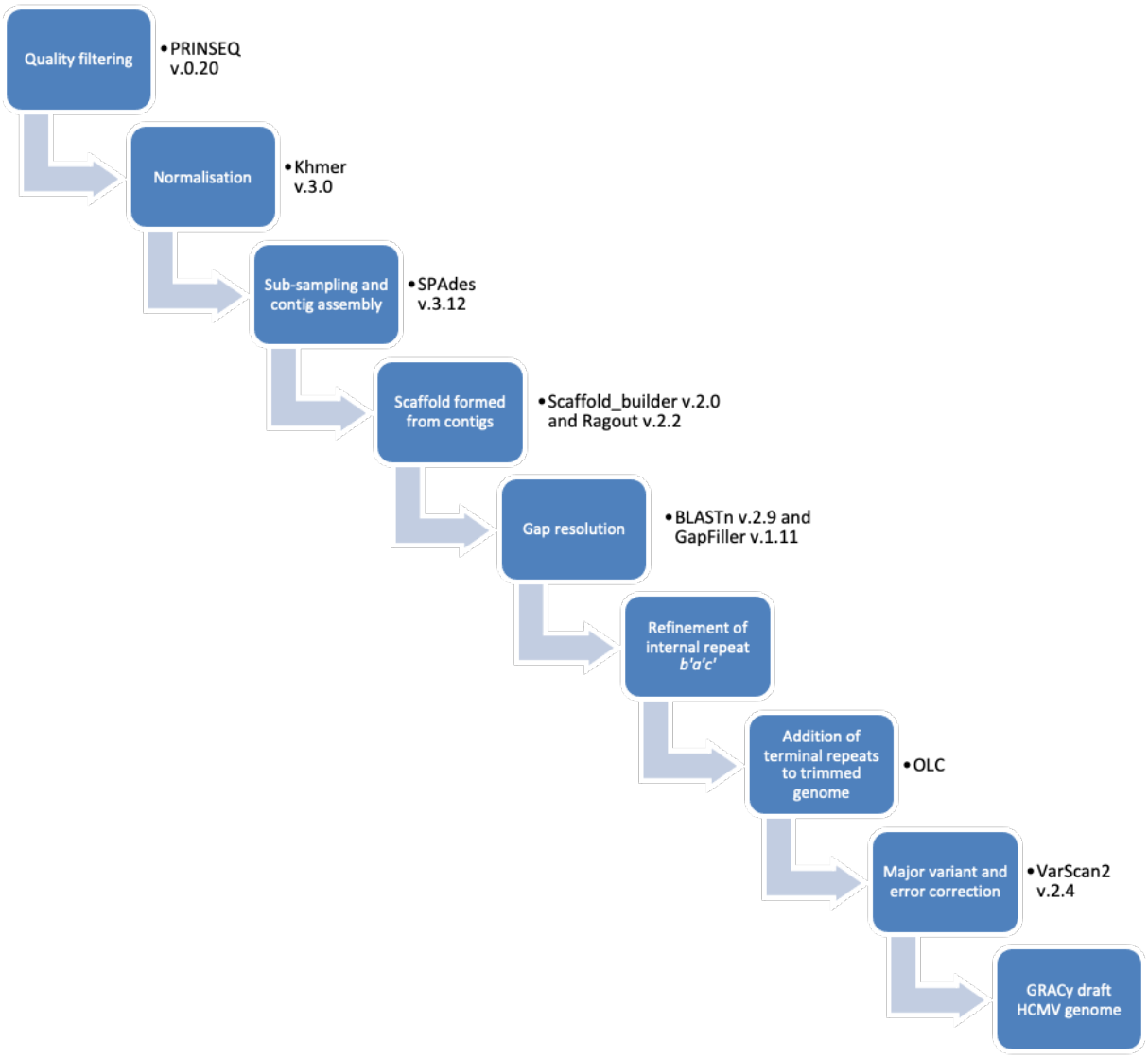


Figure 3-5. Outline of the steps and programs used for HCMV genome assembly by GRACy. The GRACy draft HCMV genome requires final manual inspection and curation in an alignment viewer (see text for details).

Two of the five completely sequenced genomes (cases 413 and 239) exhibited unusual characteristics. On manual curation of the GRACy assembled draft genome from case 413, a heterogenous region in the inverted repeat (*c/c'*) flanking U_S was detected (**Figure** 3-6). First, the terminal repeat regions from the GRACy draft genome were removed so that a single copy of the repeats (*b'a'c'*) was present in the alignment reference. Then the reads from the four datasets for this case (413P_fp, 413P_gr, 413R_fp and 413R_gr) were mapped to this reference. This demonstrated a region of heterogeneity in the *c'* region.

Two versions of this region were present in each dataset in approximately equal proportions and were thus unlikely to have been due to a fixation-induced artefact. The two alternative sequences were curated into the final genome with one version (288 nt) in *c'* and the other (318 nt) in *c* (**Figure** 3-5). However, given the limitations of short-read data, they could exist in a single genome population with the versions exchanged or even in two separate genome populations with identical copies in each population. To resolve this experimentally, PCR and sequencing could be carried out using primers within *c*/*c'* and the two ends of U_S, or nanopore sequencing could be conducted. However, the fragmented nature of DNA preserved in FFPE may preclude success with these approaches. Using BLASTn (Zhang et al., 2000, Morgulis et al., 2008) to compute a pairwise alignment between the unique *c'* and *c* sequences and the GenBank nucleotide database, the *c'* version was found to be 100 % identical to that in three previously sequenced HCMV strains (GenBank accession numbers KY490086.1, KJ361964.1 and KP745639.1) (**Figure** 3-6), and the *c* version was found to be novel.

Similarly, in case 239, the *a* sequence at the left genome end was different from the *a'* sequence region internally, the latter consisting of two fused, dissimilar *a'* sequences, and the former being identical to one of these except for 8 nt at one end (Figures 3-7 and 3-8). Again, this feature was present in both datasets (239P_fp and 239P_gr).

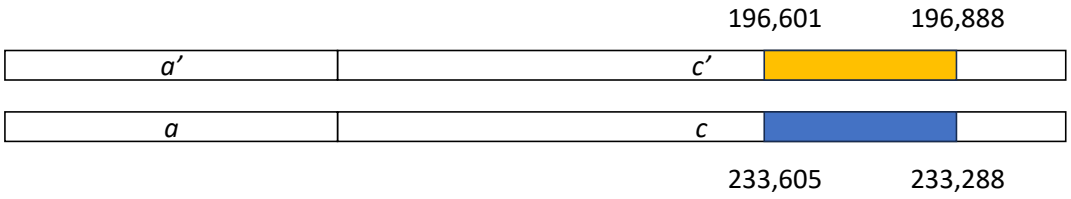


Figure 3-6. Diagrammatic representation of an alignment of the reverse complement of the *c'* sequence to the *c* sequence from case 413.
The heterogenous regions are shown as yellow in *c'* and blue in *c*, marked with the genome coordinates.

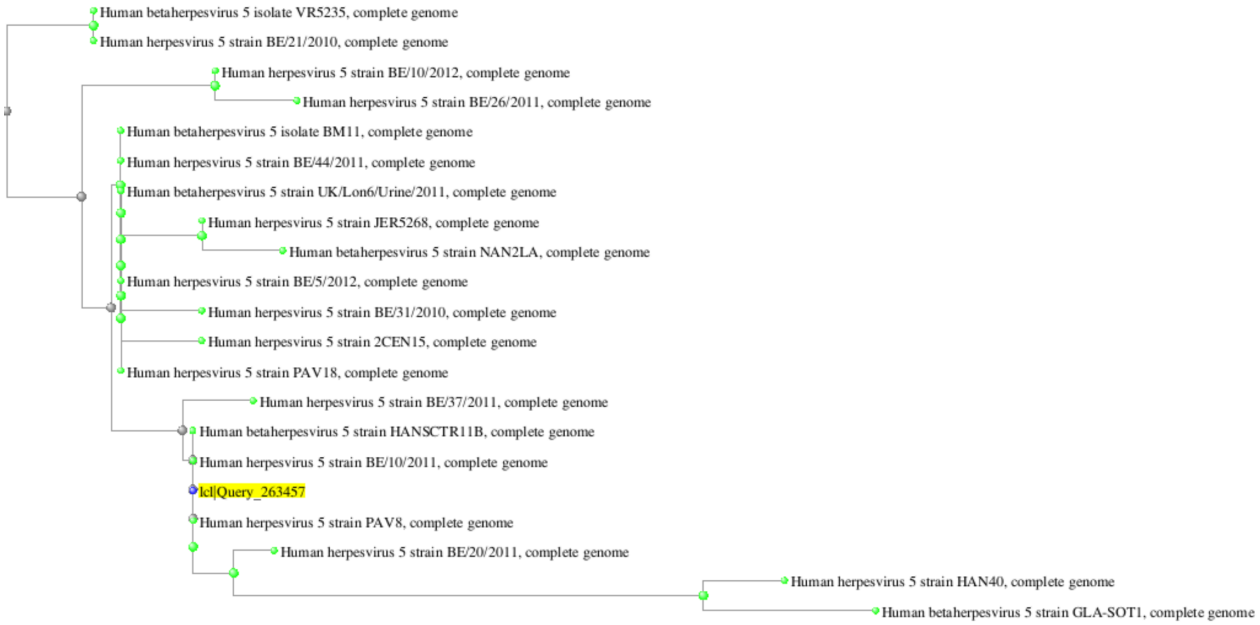


Figure 3-7. BLASTn tree view of alignment of the heterogenous *c'* sequence from case 413 to the GenBank nucleotide database.
The *c'*-sequence (lcl|Query_263457, in yellow highlight) had 100 % sequence identity to published strains HANSCTR11B, PAV8 and BE/10/2011 (GenBank accession no. KY490086.1, KJ361964.1 and KP745639.1, respectively) (Saitou and Nei, 1987).

```
a       1 CCATTCCGGGCCGCGTGGTGGGTCCCCGAGGGGCGGGGGGGGTGTTTTTAG     50
          ||||||||||||||||||||||||||||||||||||||||||||||||||||
a'      1 CCATTCCGGGCCGCGTGGTGGGTCCCCGAGGGGCGGGGGGGTGTTTTTAG     50

a      51 CGGGGGGGTGAAAATTGGAGTCTTGGAGTCGCGTGTGCTGTGGAGGACGG    100
          ||||||||||||||||||||||||||||||||||||||||||||||||||||
a'     51 CGGGGGGGTGAAAATTGGAGTCTTGGAGTCGCGTGTGCTGTGGAGGACGG    100

a     101 TGACGGTGCTGCGGTGCGGTTGGGACGGCGGCGGCGAATAAAAGCGAAAA    150
          ||||||||||||||||||||||||||||||||||||||||||||||||||||
a'    101 TGACGGTGCTGCGGTGCGGTTGGGACGGCGGCGGCGAATAAAAGCGAAAA    150

a     151 GCAGACGCGTGTCTGTGTCTGTTTGACGACGGCAGCGCGGGTCCCTGGGG    200
          ||||||||||||||||||||||||||||||||||||||||||||||||||||
a'    151 GCAGACGCGTGTCTGTGTCTGTTTGACGACGGCAGCGCGGGTCCCTGGGG    200

a     201 ACACACGAAGAACGCCTCCCAGGGGCCAGCGACGCGGAAACGACGCGGAA    250
          ||||||||||||||||||||||||||||||||||||||||||||||||||||
a'    201 ACACACGAAGAACGCCTCCCAGGGGCCAGCGACGCGGAAACGACGCGGAA    250

a     251 AAAAGGAAGTCCCCAGGGGGACGGCCGCGCGCGGAAAAGGGGAAGCGCTC    300
          ||||||||||||||||||||||||||||||||||||||||||||||||||||
a'    251 AAAAGGAAGTCCCCAGGGGGACGGCCGCGCGCGGAAAAGGGGAAGCGCTC    300

a     301 GGCGCGACTCGTCCCTAGGGGACCGCGGGGGAAGTAACGGCCGCCCAGGG    350
          ||||||||||||||||||||||||||||||||||||||||||||||||||||
a'    301 GGCGCGACTCGTCCCTAGGGGACCGCGGGGGAAGTAACGGCCGCCCAGGG    350

a     351 GAGGGGGGGGGGCTCGCGGGCCCCGGGCCACACACACGCACCCCGAAGCC    400
          ||||||||||||||||||||||||||||||||||||||||||||||||||||
a'    351 GAGGGGGGGGGGCTCGCGGGCCCCGGGCCACACACACGCACCCCGAAGCC    400

a     401 ACGCACACCGCGGGGGAAACAACAAAGTACGGCACAACCCGCGTCGAGCA    450
          ||||||||||||||||||||||||||||||||||||||||||||||||||||
a'    401 ACGCACACCGCGGGGGAAACAACAAAGTACGGCACAACCCGCGTCGAGCA    450

a     451 CACACGCAGAAGAAGCGTCCGGCCGAGGGGAGGGGGGGGGGCTCGCGGGC    500
          ||||||||||||||||||||||||||||||||||||||||||||||||||||
a'    451 CACACGCAGAAGAAGCGTCCGGCCGAGGGGAGGGGGGGGGGCTCGCGGGC    500

a     501 CCCGGGGCACACTGCTTCCATCCACCCGCGCACACCCCGCCACACACCCC    550
          ||||||||||||||||||||||||||||||||||||||||||||||||||||
a'    501 CCCGGGGCACACTGCTTCCATCCACCCGCGCACACCCCGCCACACACCCC    550

a     551 TGACACACCCGGCACACGCCCGCGACACACCCGACCGACACACCCGCGAC    600
```

```
                ||||||||||||||||||||||||||||||||||||||||||||||||||
a'    551 TGACACACCCGGCACACGCCCGCGACACACCCGACCGACACACCCGCGAC    600

a     601 ACACCCGCGACACACCCGGCACACGCCCGCGACACACCCGGCACACACCC    650
          ||||||||||||||||||||||||||||||||||||||||||||||||||
a'    601 ACACCCGCGACACACCCGGCACACGCCCGCGACACACCCGGCACACACCC    650

a     651 ACCCAGCCGCGCCCCGACACACCCAACCCGACGCCGGTGCGG--------    692
          ||||||||||||||||||||||||||||||||||||||||
a'    651 ACCCAGCCGCGCCCCGACACACCCAACCCGACGCCGGTGCGGCGACGGCG    700

a     693 --------------------------------------------------    692

a'    701 AATAAAAGCGACGTGCGGCGCGCACGTCGAAAGCCAGACGCGCGTCTGTC    750

a     693 --------------------------------------------------    692

a'    751 TTGTGTGTGTTTGCATCCCCGGGGAAAAAAGAGGAAGAAGTCCCAAAGGG    800

a     693 --------------------------------------------------    692

a'    801 ACGGCAGCGCGAGTCTCTGGGGACGCGACGATCAACTCCCGAGGGGTGAG    850

a     693 --------------------------------------------------    692

a'    851 GGAAAAGACGCGGGGACGGCCGCAGGCGGAAAAGAAGAGGAAGCCGCGGC    900

a     693 --------------------------------------------------    692

a'    901 GCACAACCCGCGTCGAGGACACACGCAGAAACGCCCTCGCGCGGGGAAGG    950

a     693 --------------------------------------------------    692

a'    951 GGGGGGGGAGTCGCGGGCCCCGGGGCACACTGCTGCCAGCCAGCCGCGCA   1000

a     693 --------------------------------------------------    692

a'   1001 CACCCGCGCCAAACCCCTGACACACCCCGCCCGGCACGCGCCCGCGGCAC   1050

a     693 --------------------------------------------------    692

a'   1051 ACCCGGCACACGCCCGTGACACACCCCGCCGACACACCCGCGACACACCC   1100

a     693 --------------------------------------------------    692

a'   1101 GCCACACGCCCGCCACACGCCCGCGACACACTCCCCTGACACACCCAGCC   1150

a     693 --------------------------------------------------    692

a'   1151 AACACACCCCCGACACACCCGGCACACACCCACCCAGCCGCGCCCCAGAC   1200

a     693 ------------------------GACGGGGG    700
                                  |||.|||.
a'   1201 ACACCCCGACCGCCGTCGGTGCGGGACCGGGC   1232
```

Figure 3-8. Sequence alignment of the *a* sequence and the reverse complement of the *a'* sequence in the case 239 final genome.
The first 692 nt from both sequences are identical, but the *a'* sequence is a composite of this sequence and an additional *a* sequence of 534 nt.

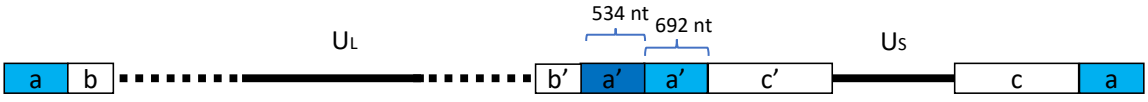


Figure 3-9. Schematic representation of the dissimilar fused *a'* sequences in the case 239 final genome.
The *a* sequence at the left genome end is 692 nt, whereas the fused *a'* sequences located internally are 534 nt longer. The situation with the *a* sequence at the right genome end was not determined, but is shown to correspond to that at the left genome end.

3.5.2 SNP analysis

Deduplication and stringent quality filtering (Phred score ≥30) of the datasets followed by application of the variant-calling module of GRACy, which incorporates LoFreq, identified 23, 21, 42, 11 and 35 SNPs in the datasets for cases 35, 150, 184, 239 and 413, respectively (**Appendix Table** A1). However, application of the thresholds used in human somatic allelic calling (cut-off frequency of 5 % and coverage of ≥50; (Bhagwate et al., 2019, Mathieson and Thomas, 2020) left no SNPs detected in case 35, fewer than five SNPs for each of cases 150, 184, 239, and eight SNPs in case 413 (**Table** 3-6). Furthermore, all bar one of the SNPs identified were present in only a single dataset, at a frequency of ≤8.2 %, with the majority being C:G to T:A mutations, which are known to occur in FFPE samples in association with the formalin-mediated deamination. Notably, seven of the ten C:G to T:A substitutions were detected in samples extracted using the FormaPure kit, which, unlike the GeneRead kit, does not incorporate UNG into the extraction process. Setting aside these as potential false discovery mutations, only one variant stood out, having been detected at high frequency (≥36 %): case 239, in gene UL147, by an A to G mutation at position 180,263, leading to a non-synonymous mutation (amino acid residue change from Y to C). This is likely the only credible variant as it was present in both sequencing datasets (sequenced from extracts obtained using the GeneRead and FormaPure kits) and is not a C:G to T:A substitution (**Table** 3-6).

3.5.3 Strain enumeration

The genotypes of 13 hypervariable HCMV genes were determined using the genotyping module of GRACy (**Figure** 3-10 and **Table** 3-7). Previous studies utilising the SureSelect enrichment protocol for Illumina sequencing of HCMV from clinical samples employed a threshold for calling a genotype when it was represented by >25 reads at >5 % of the total number of reads detected for all genotypes of that gene (Suárez et al., 2019b, Suárez et al., 2019a). This threshold effectively filters out contaminating reads. Here, a higher threshold of 100 reads was used to filter FFPE mutational artefacts (Bhagwate et al., 2019). The number of strains present was then scored as being the greatest number of genotypes present for at least two genes. Doughnut plots of concentric annuli (rings), with each annulus representing one dataset showing the 13

hypervariable genotypes as colour-coded segments (genotypes are represented proportionately by the size of the colour block) are shown for the nine cases for which sufficient HCMV reads were obtained (**Figure** 3-10 and Section 3.4.2).

The genotyping module identifies genotypes based on kmers (as 17 nt sequence motifs) specific to the individual genotypes of a gene (Section 3.4.2); if no reads with these specific motifs are found, the segment is left blank. Examples of blank segments are included in **Figure** 3-10: dataset 35R_gr, with insufficient reads for UL20 and UL139; dataset 70P_fp, which lacked sufficient reads for eight genes; and dataset 124R_fp, for which only RL12 was genotyped. Blank segments may be due to lack of relevant reads but may also be caused by the gene featuring relatively few genotype-specific kmers that were unmatched due to substitutions, or poor dataset quality due to low HCMV load or sub-optimal enrichment of HCMV DNA. As the genotyping module outputs a doughnut plot regardless of the quality of the dataset, confirmation was obtained from the accompanying output text file detailing the proportion of all reads mapping to the gene and the total number of reads matching the genotype. By this means, datasets 124R_fp, 35R_gr and 70R_fp were found to be poor quality, with the cut-off threshold unmet, a read coverage of <150 reads/nt, inconsistent genotypes with those found in counterpart datasets, or having insufficient reads to genotype a gene; these datasets were not analysed further. As an example, the case 35 kidney dataset displayed in the outer annulus (35R_gr) appears to have different genotypes in UL1, RL13, UL9 and UL74 compared to those in the placental dataset (35P_gr and 35P_fp), suggesting the transmission of a different HCMV strain in maternal placenta compared to that found in foetal kidney. Further assessment demonstrated that this was an artefact of a poor-quality dataset. Three cases (35, 70 and 150) had evidence of multiple genotypes present in at least one gene. For case 35, gene UL1 was detected as genotypes G1 and G2, both at levels over the significant threshold. Reads with G2-specific kmers were present in two datasets at greater than 12 % and were present in more than 550 reads (**Table** 3-7). Similarly, case 70 had a genotypic profile suggestive of multiple-strain infection, with the presence of two genotypes in genes RL13 (G3/G8), UL9 (G1/G2) and UL1 (G3/G2); the lattermost supported by two datasets (70R_fp and 70R_gr). The existence of gene RL12 as genotypes G3 and G1A in case 150 was supported by two datasets (150P_gr and 150P_fp). In

the remaining cases (68, 184, 239, 413 and 473), genotyping provided evidence for single-strain infections only. By assigning the 13-string code (**Table** 3-3), the genotyped HCMV strains (samples 35, 68, 70, 124, 150, 184, 239 and 413) were found to be unique when compared to a database of 244 published genomes (**Table** 3-8 and https://github.com/salvocamiolo/minion_Genotyper/blob/master/depositedSequences_codes.txt). Of note, cases 124 and 184 had the same 13-string code, suggesting infection with the same HCMV strain (**Table** 3-8). As these samples were obtained from a regional pathology laboratory, it is possible that this strain was prevalent in the local region. However, the case 124 datasets were of poor quality and incomplete, with only 2 % of reads (equating to 353,010 reads) remaining after removal of human reads. In addition, no alignment file was generated and a consensus was not obtained. Also, alignment of the reads from case 124 datasets to the 184 genome yielded no alignments due to the poor quality of these datasets.

Table 3-6 Summary of SNPs detected by the variant-calling module of GRACy at a frequency of >5 % and at a coverage of ≥50 reads.
The nucleotide changes and resultant codon and amino acid changes are listed, as are the C:G to T:A mutations, which may be a result of deamination during formalin fixation.

Dataset	Gene	Position in genome	Position in protein	Frequency (%)	Coverage	Nucleotide change	Codon change	Amino acid change	C:G to T:A
150P_fp	UL52	74897	268	6.06	66	G→ A	GCC→ ACC	Ala→ Thr	+
150P_fp	UL57	89641	1481	7.14	84	G→ A	CGC→ CAC	Arg→ His	+
184P_fp	UL85	124864	105	5.13	117	G→ A	CCG→ CCA	synonymous	+
239P_fp	UL45	58950	1403	5.26	76	C → T	GCT→ GTT	Ala→ Val	+
239P_fp	UL147	180263	272	36.36	99	A → G	TAT→ TGT	Tyr→ Cys	-
239P_gr	UL147	180263	272	38.46	143	A → G	TAT→ TGT	Tyr→ Cys	-
413P_fp	RL13	11231	483	5	80	C→ T	TGC→ TGT	synonymous	+
413P_fp	UL16	22931	612	6.06	66	C→ T	GCC→ GCT	synonymous	+
413P_fp	UL54	78547	2957	6.35	63	C→ T	ACG→ ATG	Thr→ Met	+
413P_gr	UL74	107530	845	7.41	54	C→ T	ACA→ ATA	Thr→ Ile	+
413P_gr	UL123	172012	1046	7.02	57	A→ G	AAG→ AGG	Lys→ Arg	-
413P_gr	UL128	175882	34	8.2	61	C→ T	GCG→ TCG	Ala→ Ser	+
413P_gr	US24	220840	12	8	50	G→ A	CCG→ CCA	synonymous	+
413P_gr	US28	225284	327	8	50	C→ T	GCC→ GCT	synonymous	+

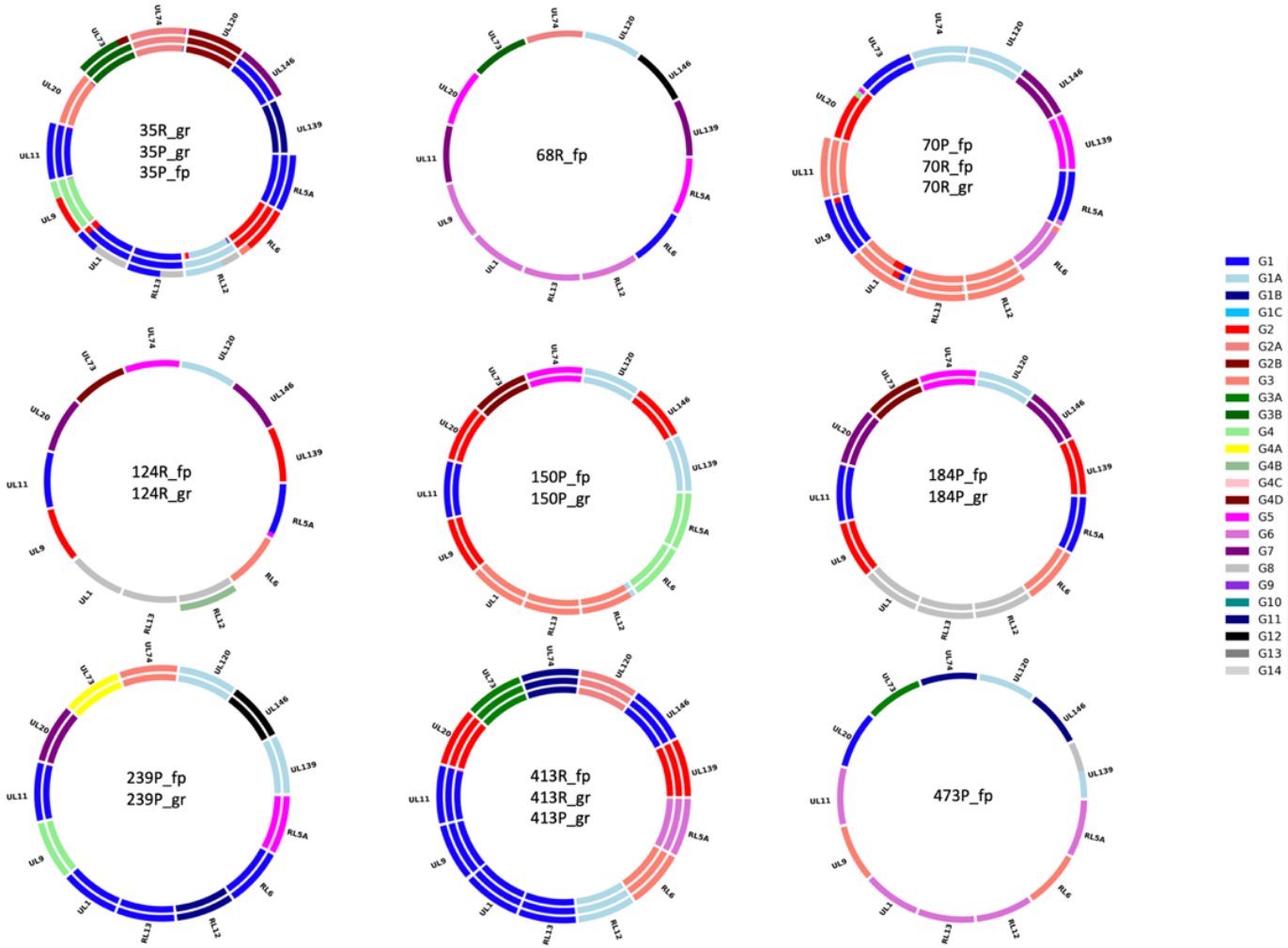


Figure 3-10. Doughnut plots reporting the results of genotyping the samples from nine cases.
Each annulus represents an individual dataset, divided into sections representing the 13 hypervariable genes analysed. Datasets are listed from the outer annulus inwards. The size of the coloured bars corresponds to the proportion of genotypes detected for each gene, as coded in the panel on the right.

Table 3-7. Genotypes of 13 hypervariable genes for each case and dataset.
Where multiple genotypes were identified for a gene, the percentage and number of reads matching the minor genotype(s) are reported in parentheses. Only the top two genotypes detected in >5 % of reads mapping to the gene and numbering ≥100 reads are reported.

Case	Dataset[a]	Gene					
		RL5A	RL6	RL12	RL13	UL1	UL9
35	35P_fp	G1	G2	G1A	G1	**G1**(82.5%, 3052), G2 (14.9%, 550)	G4
	35P_gr	G1	G2	G1A	G1	**G1**(88%, 10043), G2 (12%, 1366)	G4
	35R_gr	G1	G2	G1A	G1	IR	G2
68	68R_fp	G5	G1	G6	G6	G6	G6
70	70R_fp	G1	G6	G3	**G3**(93.6%, 2748), G8 (5.7%, 166)	**G3** (70.0%, 958), G2(13.3%, 180),	**G1**(89.9%, 1454), G2 (9.3%, 150)
	70R_gr	G1	G6	G3	G3	**G3** (62.9%, 307), G2(21.7%, 106)	G1
	70P_fp	NR	NR	G3	G3	G3	G1
124	124P_fp	G1	G3	G8	G8	G8	G2
150	150P_gr	G4	G4	**G3** (90.7%, 11854), G1A (9.3%, 1218)	G3	G3	G2
	150P_fp	G4	G4	**G3** (91.5%, 30012), G1A (8.5%, 2795)	G3	G3	G2
184	184P_fp	G1	G3	G8	G8	G8	G2
	184P_gr	G1	G3	G8	G8	G8	G2
239	239P_fp	G5	G1	G1B	G1	G1	G4
	239P_gr	G5	G1	G1B	G1	G1	G4
413	413R_fp	G6	G3	G1A	G1	G1	G1
	413R_gr	G6	G3	G1A	G1	G1	G1
	413P_gr	G6	G3	G1A	G1	G1	G1
473	473P_fp	G6	G3	G6	G6	G6	G3

Table 3-7 *(continued)*

Case	Dataset[a]	Gene						
		UL11	UL20	UL73	UL74	UL120	UL139	UL146
35	35P_fp	G1	G3	G3B	G2A	G2B	G1B	G1
	35P_gr	G1	G3	G3B	G2A	G2B	G1B	G1
	35R_gr	G1	NR	G3B	G2B	G2B	NR	IR
68	68R_fp	G7	G5	G3B	G2A	G1A	G7	G12
70	70R_fp	G3	G2	G1	G1A	G1A	G5	G7
	70R_gr	G3	G2	G1	G1A	G1A	G5	G7
	70P_fp	G3	NR	NR	NR	NR	NR	NR
124	124P_fp	G1	G7	G4D	G5	G1A	G2	G7
150	150P_gr	G2	G2	G4D	G5	G1A	G1A	G2
	150P_fp	G1	G2	G4D	G5	G1A	G1A	G2
184	184P_fp	G1	G7	G4D	G5	G1A	G2	G7
	184P_gr	G1	G7	G4D	G5	G1A	G2	G7
239	239P_fp	G1	G7	G4A	G3	G1A	G1A	G12
	239P_gr	G1	G7	G4A	G3	G1A	G1A	G12
413	413R_fp	G1	G2	G3A	G1B	G2A	G2	G1
	413R_gr	G1	G2	G3A	G1B	G2A	G2	G1
	413P_gr	G1	G2	G3A	G1B	G2A	G2	G1
473	473P_fp	G6	G1	G3A	G1B	G1A	IR	G11

[a] Dataset suffixes: P, placenta; R, kidney; _fp, FormaPure extraction kit, _gr, GeneRead extraction kit. Single datasets (35R_gr and 70R_fp) reporting multiple genotypes or having no reads (NR) or insufficient reads (IR) for genotyping were discarded as unreliable.

Table 3-8. Strains represented as a 13-string code, replacing the genotype with the corresponding one-letter code (Table 3-3).

Case	RL5a	RL6	RL12	RL13	UL1	UL9	UL11	UL20	UL73	UL74	UL120	UL139	UL146	13-letter code
35	A	H	D	A	A/H	R	A	Q	E	K	L	F	A	AHDAARAQEKLFA AHDAHRAQEKLFA
68	P	A	C	C	C	C	V	P	E	K	D	V	Z	PACCCCVPEKDVZ
70	A	C	Q	Q/N	Q/H	A/H	Q	H	A	D	D	P	V	ACQQQAQHADDPV ACQNHHQHADDPV
124	A	Q	N	N	N	H	A	V	I	P	D	H	V	AQNNNHAVIPDHV
150	R	R	Q/D	Q	Q	H	A	H	I	P	D	D	H	RRQQQHAHIPDDH RRDQQHAHIPDDH
184	A	Q	N	N	N	H	A	V	I	P	D	H	V	AQNNNHAVIPDHV
239	P	A	F	A	A	R	A	V	T	Q	D	D	Z	PAFAARAVTQDDZ
413	C	Q	D	A	A	A	A	H	W	F	K	H	A	CQDAAAAHWFKHA

3.5.4 Comparison of FFPE extraction kits

Qubit fluorometric quantitation, using the high sensitivity dsDNA assay, demonstrated a higher mean DNA concentration in samples extracted using the GeneRead kit (55.52 vs. 19.15 ng/µL; p=0.01; **Table** 3-2 and **Figure** 3-11a). Nanodrop analysis indicated that the GeneRead kit produced purer nucleic acid (mean A260/A280 of 2.06 vs. 1.56; p=0.0004) (**Figure** 3-11b). However, this could have been a confounding effect of the quantity of DNA extracted by the FormaPure kit because, although low A260/A280 ratios usually indicate sample contamination by protein or residual guanidine from the extraction protocol, they may also be secondary to low concentrations (<20 ng/µL) of nucleic acids (Koetsier, 2019). Additionally, shifts in the pH of the solution can also cause A260/A280 to vary, with a basic solution over-representing the ratio by 0.2-0.3 (Wilfinger et al., 1997). Thus, the slightly basic elution buffer (ATE buffer, pH 8) used in the GeneRead kit may have contributed to higher A260/A280 ratios, compared to FormaPure, which was eluted in NFW (pH 7). Quantitation of HCMV load from the samples demonstrated insignificant differences (within 3.16 IU/µL) between the samples extracted using either kit (**Table** 3-2), with the overall quality of the DNA libraries generated being reliant on the quality of the DNA in the FFPE sample (**Figure** 3-11c).

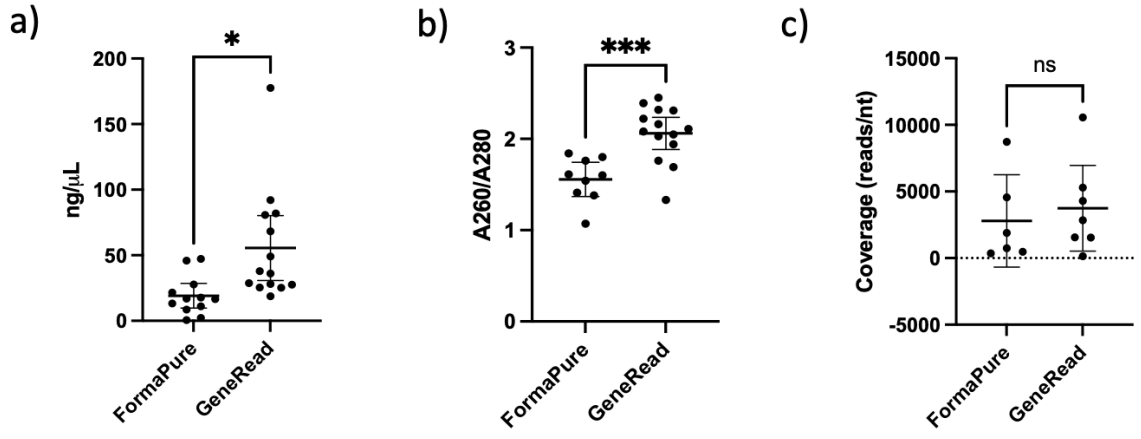


Figure 3-11. Jitter plots comparing DNA purity and concentration in DNA extracts.
a) DNA concentration (Qubit), b) A260/A280 (Nanodrop) for DNA extracted using the FormaPure kit or the GeneRead kit. c) Quality of DNA library as measured by average coverage depth of the Merlin reference genome. Error bars demonstrate mean and 95 % CI (* p<0.05, *** p<0.001; two-tailed Student's t-test with p values calculated and plotted using GraphPad Prism v.9.4.1).

3.6 Discussion

This study was aimed at demonstrating the feasibility of using the Illumina platform to generate HCMV sequence data from cCMV FFPE material that was of sufficient quality to generate complete HCMV genomes. I achieved this by employing the latest available FFPE-specific DNA extraction kits and by utilising the recently developed toolkit GRACy for downstream bioinformatic processing of the datasets. In total, five complete HCMV genome sequences were obtained from FFPE samples originating from nine cases of cCMV. The most important factor for successful sequencing was a high starting HCMV load. DNA of sufficient quality for sequencing was recovered from FFPE tissues that had been stored for up to 5 years, but not from older samples. This is consistent with the results of previous studies showing that sequenceable libraries could be obtained in 97 % of FFPE samples stored for up to 3 years, with the success rate decreasing to 50 % in samples stored for longer periods (14-21 years) (Bolognesi et al., 2016). However, that study used the older QIAamp DNA FFPE tissue kit (QIAGEN), and the failure to obtain genomes in my study from the older samples (cases 473, 68 and 124) was more likely due to low starting HCMV load and subsequent poor genome coverage (<500 reads/nt). Evaluation of two commercial FFPE extraction kits demonstrated that the GeneRead kit performed better than the FormaPure kit in obtaining higher DNA concentrations and by reducing deamination artefacts by UNG treatment.

Eight novel genotypic combinations of the 13 most hypervariable genes were identified by using the genotyping module alone, without the necessity of obtaining genome sequences. This is consistent with previous observations that vast numbers of combinations of the genotypes of hypervariable gene are possible in the otherwise well-conserved HCMV genome (Rasmussen et al., 2003, Lassalle et al., 2016, Suárez et al., 2019b). Of the genotypes identified for UL146 from the FFPE samples, each of G1, G7 and G12 were present in at least two individual cases (35 and 413, 10 and 184, and 68 and 239, respectively). In one study conducted in Poland, UL146 G1 accounted for around a third of cCMV cases (Paradowska et al., 2014a). In another study conducted in HIV-infected mothers' breastmilk in Zambia, G9 was the most commonly identified genotype (present in 55 % of cases) (Suárez et al., 2019a). A more geographically inclusive study involving samples from Asia, Europe, Africa and North America found a

more even distribution of UL146 genotypes (the top genotypes being G7 at 16 %, G1 at 9.7 % and G2 at 7.14 %), although this study included a range of clinical samples that were not specifically from cCMV cases (Bradley et al., 2008). These findings signify the need to consider the geographical prevalence of HCMV genotypes when analysing inference studies of the effects on outcomes. This is supported by a recent study finding a similar proportion of UL146 genotypes in cCMV cases compared to that in the surveyed population (Berg et al., 2021).

All the cCMV cases included in my study had fatal outcomes, but of those for which complete HCMV genomes were obtained (cases 184, 150, 35 and 239), only one (case 150) was suggestive of multiple-strain infection (gene RL12, with genotypes G3 and G1A present in a 9:1 ratio; **Table** 3-7). However, in order to call a multiple-strain infection reliably, multiple genotypes should be present in at least two of the hypervariable genes. Including the genotyped cases from which it was not possible to obtain full genomes, and for which the data were less reliable due to lower quality or sparse genome coverage, only case 70 fulfilled this criterion (multiple genotypes detected in three genes: RL13, UL1 and UL9). This is in contrast to some studies reporting that multiple-strain infections are associated with severe clinical disease in infants with cCMV (Coaquette et al., 2004, Arav-Boger, 2015), but in keeping with other findings that most congenital or postnatal infections involve single strains (Görzer et al., 2015). A survey of HCMV variation in breastmilk found multiple-strain infections to be common in HIV-infected mothers (Suárez et al., 2019a), with the implication that only certain strains cross the placenta or are transmitted post-partum. This finding was echoed in a smaller survey of five HIV-infected mothers of cCMV infants, where it was found that single-strain transmission is typical and associated with tropism or immunomodulatory genes exhibiting significant similarities (Pang et al., 2020). Again, the generalisability of such surveys is difficult, as these studies were conducted in Zambia and Kenya, respectively, whereas most samples that have been analysed (including mine) have been from high-income countries. The study of cCMV has thus been restricted thus far and has been constrained by the availability of sequence data from geographically inclusive sites as well as the paucity of cCMV samples sequenced. However, my study was not designed to investigate the effect of multiple-strain infections on outcomes, as the number of cases analysed was too small to permit any

inference on the relationship between disease outcome and the number of strains or the combination of genotypes involved. Rather, I have demonstrated the feasibility of using HTS to determine complete genomes from FFPE material, a difficult but abundant sample type from which larger, multi-centre studies spanning different continents may begin to answer such questions.

A limitation of the variant analysis was the anticipated higher error rate resulting from the degradation of nucleic acid in FFPE samples. To control for this, more stringent quality thresholds for filtering reads and calling genotypes were set, in line with standards set from human somatic variant calling from FFPE tissues (Bhagwate et al., 2019). Using these thresholds, omitting C:G to T:A substitutions as deamination artefacts of formalin fixation, and acknowledging only the variants represented in datasets from both FormaPure and GeneRead extractions, there was but a single variant (UL147 of case 239, present in 37 % of reads). This mutation encodes a nonsynonymous substitution that would change a tyrosine to a cysteine residue in the encoded homologue of the chemokine CXCL-2 that may be involved in immune evasion (Lüttichau, 2010). Case 239 presented with intra-uterine death at 34 w gestation with gross pathology evident in both the placental and foetal tissues, suggestive of an extended infection from early pregnancy and exposure to immune-driven selective pressure, making this a biologically plausible mutation. This general lack of mutability of HCMV genomes during *in vivo* infections is in keeping with previous studies (Stanton et al., 2005, Lurain et al., 2006).

Further work would need to be done to compare fresh tissues alongside FFPE material using the protocol developed here to investigate the extent to which formalin fixation generates artefactual lesions in HCMV DNA and to establish more robust thresholds to rule out false discovery. Much of the DNA in FFPE material exists as single-stranded DNA (ssDNA) due to the formalin fixation process. Strand-split artefacts are hypothesized to originate from ssDNA during sequencing library construction: T4 ligase repair of ragged overhangs on the end of dsDNA molecules can extend annealed ssDNA repetitive sequences and produce chimaeric reads derived from non-contiguous portions of the genome (Haile et al., 2019). Since these artefacts can be removed enzymatically by treating FFPE samples with S1 nuclease to degrade ssDNA (Haile et al., 2019),

this would be a useful extra step to test in future studies. Alternatively, in clinical sequencing of tumours, a bioinformatic filtering pipeline has been developed to detect these chimeric reads in human breast tissue samples (Ikegami et al., 2021), and a similar approach could be tested to eliminate chimaeric HCMV reads from FFPE datasets.

In summary, I have demonstrated the feasibility of obtaining whole HCMV genomes from FFPE tissues using extraction kits designed for this material, an established bait-based target enrichment protocol, and a powerful bioinformatics approach (GRACy). I obtained complete genomes from five of ten cCMV cases, with both extraction kits performing well, GeneRead giving slightly better-quality DNA than FormaPure. Genotyping of the sequence datasets demonstrated the diversity of HCMV strains seen in previous studies. Using thresholds established in human somatic variant calling, mutations were found to be rare. This demonstrates the importance of evaluating and setting thresholds to avoid false discoveries, with future studies investigating HCMV from FFPE samples achieving at least the standards set recently in the HCMV context (Camiolo et al., 2022). As cCMV is a relatively rare disease, most studies to date involve only a small number of samples that are incapable of providing the statistical power to investigate the association between multiple-strain infection or of genotypes with severe disease. FFPE repositories provide an invaluable resource for characterising both HCMV strains and genotypes. The ability to retrieve and analyse HCMV genomes from these samples, from biorepositories worldwide with associated metadata, will facilitate more comprehensive future studies.

4 Nanopore sequencing of high-titre laboratory-cultured HCMV strains

4.1 Background

HCMV infection can cause a spectrum of diseases ranging from mild to severe. Investigation of viral genotypes aimed at identifying associations with disease severity has been hindered by the inability to assess entire HCMV genomes simultaneously, with individual studies focusing on one or a few genes at a time (Section 1.8). The natural history of HCMV infection also means that patients can be infected with multiple strains, often in highly asymmetric proportions, with consequent confounding of strain composition by lack of recognition of minor strains or by recombination. The ONT platform is capable of long-read sequencing and can potentially resolve these issues, with the additional benefit of low capital cost. However, the high error-rate of early iterations of ONT pores and basecallers, and the requirement for a substantial input of relatively pure target DNA, posed major challenges to sequencing HCMV genomes directly from clinical samples. Nanopore sequencing of cultured HCMV samples performed to date has utilised enrichment using custom RNA baits or transposase-based library preparation, with maximum HCMV read-lengths of 5.6 kb and 56 kb achieved, respectively (Eckert et al., 2016, Karamitros et al., 2016).

To assess the potential of current iterations of the ONT hardware and basecallers to overcome these limitations, and to establish the impact of the read error-rate on consensus calling, I sequenced three well-characterised high-titre laboratory-cultured HCMV strains (Merlin, U11 and AF1) (Dolan et al., 2004, Dargan et al., 2010). These strains were established at the time of sequencing to consist of single strains and had been sequenced using Sanger (capillary electrophoresis) technology, considered as gold-standard in terms of accuracy for clinical research (Dolan et al., 2004, Gatherer et al., 2011). Additionally, Merlin has also been sequenced independently on the Illumina platform. Although the GenBank sequences are referenced for the purpose of comparison of the nanopore generated consensus, these references nonetheless may not be entirely error-free. Sanger sequencing of AF1 and U11 involved PCR amplification of the whole genome (Dargan et al., 2010), and PCR is known to incur slippage in homopolymeric tracts of approximately 8 bp (Shinde et al., 2003). Nevertheless,

these strains are relatively low-passage strains of known provenance and are more akin to the clinical HCMV than the highly used high-passage laboratory-adapted strains. They were therefore chosen to assess the accuracy of nanopore consensus sequences.

Two approaches were used to derive HCMV genomes from nanopore datasets. First, a *r*eference-*d*ependent *a*ssembled (RDA) genome sequence was created by mapping reads to the cognate reference sequence. Second, a *r*eference-*i*ndependent *a*ssembled (RIA) genome sequence was created using a long-read *de novo* assembler that does not rely on a reference sequence.

After establishing an appropriate workflow, I tested the ability of nanopore sequencing to discriminate two HCMV strains in a simulated multiple-strain infection, by combining DNA extracted separately from two strains in a 50:50 ratio. HCMV genomes in this equimolar ratio are challenging to tease apart using short-read technology due to the large regions of sequence similarity connecting the relatively rare islands of hypervariability. Furthermore, to test the ability of long-read sequencing to detect recombinants, I used my pipeline to search for recombinants arising by co-infection with two distinct HCMV strains in cell culture.

4.2 Objectives

1. To obtain whole HCMV genome sequences using nanopore technology from high-titre laboratory-cultured HCMVs without enrichment, and concurrently to assess the error rate of nanopore-generated RDA sequences in these circumstances.

2. To exploit the long-read sequencing capability of nanopore technology to differentiate individual HCMV genomes present in an equimolar mixture of DNA extracted from two strains.

3. To explore the long-read sequencing capability of nanopore technology to detect recombination between two HCMV strains in cell culture.

4.3 Materials and Methods

4.3.1 Experiments 1-3: single-strain infection

To achieve the first objective and assess the consensus sequence error-rate, cell culture and propagation of individual HCMV strains Merlin, U11 and AF1 (GenBank accession numbers AY446894.2, GU179290.1 and GU179291.1, respectively) were performed as described (Sections 2.1 to 2.3). These strains were sequenced individually in Experiments 1 to 3 (**Table** 4-1). Briefly, viral stocks were used to infect HFFF2 cells at an MOI of 0.15 PFU/cell, and cell-free virus was harvested by collection of the supernatant. Each virus was pelleted from the supernatant and resuspended in 1 mL of medium per 3400 cm^2 infected cells and stored at -80ºC. Infectious titres were determined by plaque assay (Section 2.4). DNA was extracted using Genomic-tip 20/g extraction kits (QIAGEN; Section 2.5). The integrity and average fragment size of the extracted DNA was measured using the 4200 TapeStation system (Agilent; cat. no. G2991BA; Section 2.6)(**Figure** 4-1). The DNA was stored at 4 ºC and then sequenced on the MinION.

4.3.2 Experiment 4: simulation of multiple-strain infection

To achieve the second objective of separating equimolar mix of DNA from two strains, I mixed DNA extracts from cultured AF1 and U11. The Qubit dsDNA high sensitivity assay kit was used to determine DNA concentration in the AF1 and U11 extracts, and qPCR with the WHO Merlin international standard was used to measure HCMV loads (Section 2.7). Genomic fragment size and integrity were determined using genomic DNA ScreenTape on the 4200 TapeStation system (QIAGEN). Aliquots of the AF1 and U11 extracts were combined in a 1:1 ratio (3 µg DNA in total) prior to library preparation for sequencing, and a final library concentration of 39 ng/µL was sequenced on the MinION.

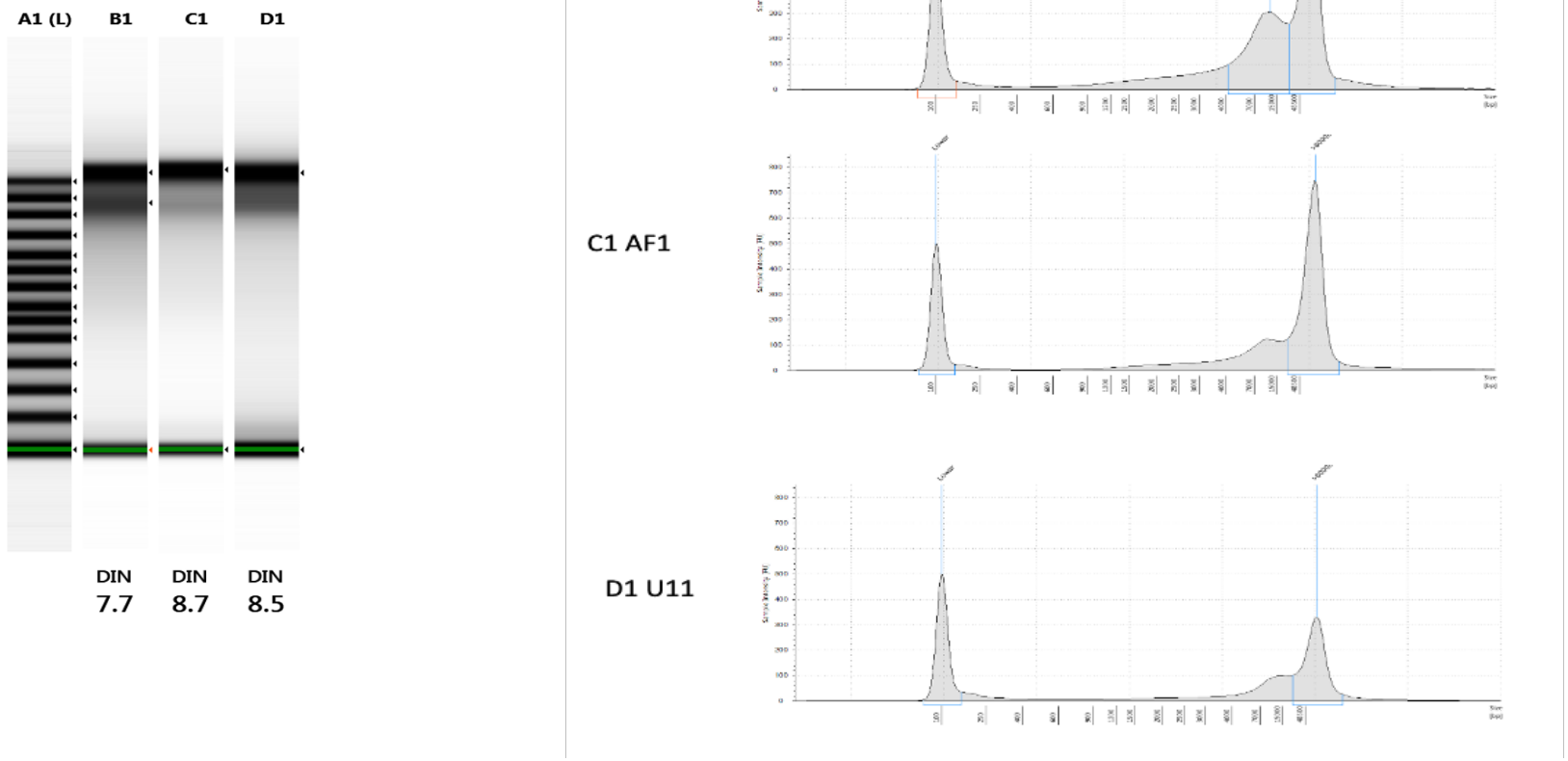


Figure 4-1. TapeStation gel image and corresponding electropherograms of DNA extracted from HCMV strains grown in cell culture.
A1 shows the marker ladder, and B1-D1 show the results for HCMV strains Merlin, AF1 and U11, respectively. Larger molecules are represented by a band near the top of the gel image and as a peak towards the right of the electropherograms. The DINs are shown below the gel image.

4.3.3 Experiments 5-6: recombination between strains

In order to stimulate recombination in cell culture, two HCMV strains were grown together in the following two experiments.

4.3.3.1 Experiment 5: Rec AF1/U11

The stock titres of U11 and AF1 were $6.8x10^5$ and $1.6x10^6$ PFU/mL, respectively. A high MOI of 3 PFU/cell of each strain was used to infect six-well plates of HFFF2 cells containing $2x10^5$ cells/well to promote co-infection with both strains and to achieve infection of 95 % of cells. Equal titres of U11 and AF1 were used to infect the cells (equivalent to 900 µL of U11 stock and 387 µL of AF1 stock). The inoculum was removed after 5 h, and the cells were washed and replenished with GM, then incubated for a total of 72 h. Virus was propagated and concentrated as described in Section 2.2. A Genomic-tip 20/g extraction kit (QIAGEN) was used to obtain DNA from the viral concentrate (Section 2.5), and 3.8 µg of DNA was subjected to library preparation using the SQK-LSK-109 protocol (Section 2.9). A final library DNA concentration of 85 ng/µL was sequenced on the MinION.

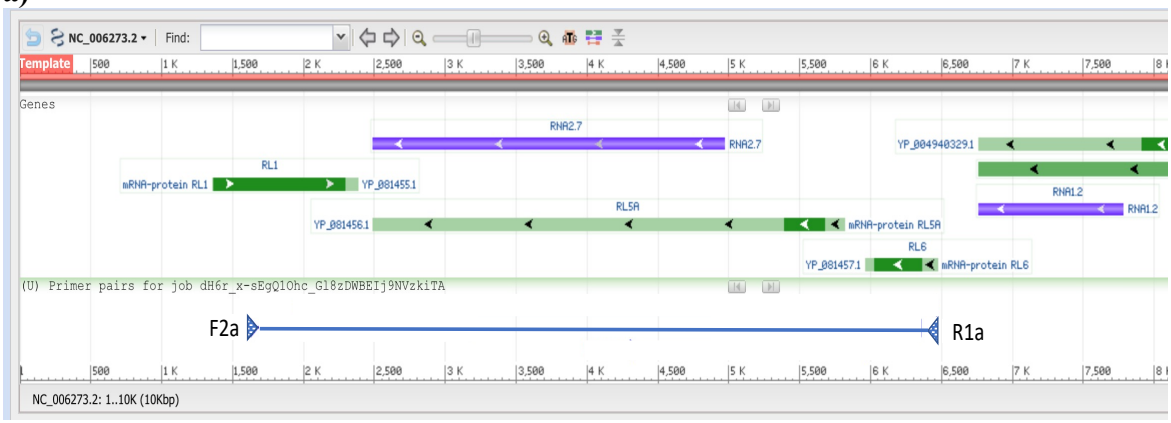
4.3.3.2 Experiment 6: Rec AF1/ΔRNA2.7

The RCMV1111 mutant (ΔRNA2.7) of strain Merlin (Lau et al., 2021), which lacks gene RNA2.7 (coordinates 2560 - 5050 relative to the Merlin genome) (Stanton et al., 2010) and AF1 were co-cultured in a further recombination experiment. The stock titres of AF1 and ΔRNA2.7 were $1.6x10^6$ and $1.18x10^7$ PFU/mL, respectively. Infection was carried out as described above. AF1 and ΔRNA2.7 were propagated as mixtures in different ratios to increase the probability of coinfection by both strains and to accommodate competition by either strain. In Experiments 6A, 6B and 6C, the AF1:ΔRNA2.7 ratio was 1:3, 1:1 and 3:1, respectively. The infections were propagated and DNA was extracted as described above. To demonstrate the presence of both strains in the concentrated DNA extracts, PCR was performed, and the products were visualised on an agarose gel, using primers designed to amplify the region encompassing RNA2.7 (**Figure** 4-2a). The thermal cycling conditions were as follows: denature at 95 ºC for 2 min; ten cycles at 95 ºC for 30 s, 55 ºC for 30 s

and 68 °C for 4.5 min; 25 cycles at 95 °C for 30 s, 55 °C for 30 s and 68 °C for 4.5 min, increasing the third step by 20 s at each successive cycle; extend at 68 °C for 15 min; and hold at 4 °C.

The expected PCR product sizes were 4778 bp for AF1 and 2078 bp for ΔRNA2.7. Experiment 6A presented strong bands for both AF1 and ΔRNA2.7 and was selected to proceed to library preparation for nanopore sequencing (RecR1 AF1/ΔRNA2.7) (**Figure** 4-2b). Experiments 6B and 6C were discarded, as the weak bands at 2078 bp indicated a low level of ΔRNA2.7. An aliquot of 2.5 µg of extracted DNA was used for library preparation, and the library was sequenced on the MinION.

a)



b)

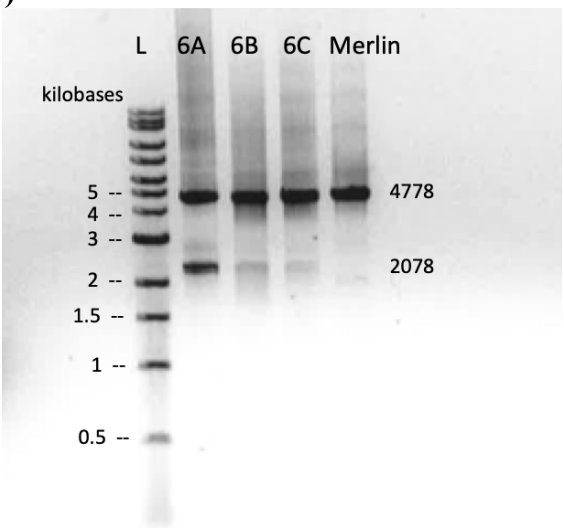


Figure 4-2. Detection of input strains AF1 and ΔRNA2.7 after co-culture in Experiment 6. a) Graphical view of primers encompassing the RNA2.7 gene, which is deleted in ΔRNA2.7. The blue hatched arrows label primers F2a and R1a). b) Gel electrophoresis demonstrating semi-quantitatively the presence of AF1 (band at 4778 bp) and ΔRNA2.7 (band at 2078 bp) in Experiments 6A-C. L = Quick-Load 1 kb extend DNA ladder.

4.3.4 Nanopore sequencing

The ONT genomic DNA by ligation protocol (SQK-LSK109) was used to prepare the libraries in each experiment, which were sequenced using ONT MinION R9.4.1 flow cells. Extracted DNA concentrations were 67-690 ng/µL (**Table** 4-1). DNA end-repair and adaptor ligation were performed as detailed by the SQK-LSK109 protocol (Section 2.11). The flow cell was primed according to the protocol, apart from the substitution of loading beads by NFW. Final libraries with input genomic DNA of 2.5-7.9 µg were loaded onto the flow cell for sequencing. Sequencing was stopped after 24 h, except for Experiment 4, which was stopped after 48 h.

4.3.5 Bioinformatic analysis

FAST5 files from the completed runs were compressed using gzip and transferred to a computer cluster. The single strain data (for Merlin, U11 and AF1) were basecalled locally on the computer cluster into fastq format using Albacore v.2.3.3 and subsequent runs by the Guppy basecaller (https://community.nanoporetech.com/protocols/Guppy-protocol/v/gpb_2003_v1_revt_14dec2018). Quality checks were performed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and MinIONQC v.1.3.5 (Lanfear et al., 2019) to check the run statistics and the quality scores of the reads. Diamond (Buchfink et al., 2015), which is a sequence aligner for protein and translated DNA, confirmed successful sequencing of HCMV and the absence of any bacterial or fungal contaminant from the cultures. Adaptor trimming was performed using Porechop v.0.2.3 (Wick et al., 2017) with the option to discard reads with internally located adaptors to enable downstream processing with the signal-level consensus callers Nanopolish (https://github.com/nanoporetech/nanopolish) and Medaka (https://github.com/nanoporetech/medaka). Run statistics were summarised using MinIONQC v.1.3.5 (https://doi.org/10.1093/bioinformatics/bty654) (**Table** 4-2).

Table 4-1. Summary of the high-titre laboratory cultured HCMV DNA samples sequenced in Experiments 1- 6A.
Experiments 1-3 involved DNA from single strains. Experiment 4 sequenced combined DNA extracts from two strains (U11 and AF1 in an approximate 50:50 ratio to simulate a multiple-strain infection). Experiments 5-6A involved DNA from co-infections by two strains to investigate inter-strain recombination in culture.

Experiment	Strain(s) present	DNA extract concentration (ng/µL)	DNA purity (A260/A280)	HCMV UL97 qPCR (IU/µL)[b]	Genomic DNA size (bp)	DNA library concentration (ng/µL)	Input DNA for sequencing (µg)
1	Merlin	690	1.9	ND	>60,000	29.2	7.9
2	U11	243	1.9	51,286,138	18,222	20	4.8
3	AF1	103	1.9	141,253,754	>60,000	17	3.7
4	Mixed AF1/U11	--	--	--	--	39	3.0
5	Rec AF1/U11[a]	102	1.87	ND	>60,000	85	3.8
6A	Rec AF1/ΔRNA2.7[a]	67	1.87	ND	50,893	52	2.5

[a] Rec, Recombination experiments.

[b] ND, not done.

Table 4-2. Summary statistics for sequencing runs.

	Single strain			Simulated multiple strain	Recombination	
Experiment	**1**	**2**	**3**	**4**	**5**	**6A**
Strain(s)	Merlin	U11	AF1	AF1/U11 50:50 mix	Rec AF1/U11	Rec AF1/ΔRNA2.7
Total run time (h)	24	24	24	48	24	24
Total gigabases	6.16	5.38	5.99	11.08	9.92	8.00
Total reads (no.)	2,466,578	1,122,159	1,599,910	4,190,064	3,688,404	2,340,881
N50 length (bp)	3,990	6,813	7,208	4,634	4,830	19,113
Mean length (bp)	2,498	4,798	3,742	2,645	2,689	10,577
Median length (bp)	1,531	3,231	1,805	1,459	1,410	5,710
Max length (bp)	234,877	234,848	298,823	299,468	225,306	500,622
Mean q-score	9.6	9.3	10.6	10.4	10.3	32.0
Median q-score	9.7	9.5	10.9	10.6	10.5	33.0
Read length (no.)						
>10 kb	60,108	98,201	104,914	136,726	139,673	145,372
>20 kb	10,255	22,586	38,920	32,638	41,635	46,996
>50 kb	1,736	4,036	8,725	7,449	6,404	4,690
>100 kb	241	643	885	1,579	364	80
>200 kb	3	30	11	161	2	0

4.3.6 Determination of HCMV consensus genomes

4.3.6.1 RDA

To determine the accuracy of the nanopore-generated consensus genomes, Minimap2 v.2.17 (Li, 2018) was used to align the reads generated in Experiments 1 to 3 (single-strain runs) to the respective reference genomes for Merlin, U11 and AF1. Samtools v.1.9 (Li et al., 2009) was used to generate BAM and mpileup files. A java program, vsensus.jar (https://github.com/rjorton/VSensus), parsed these files using appropriate thresholds (minimum coverage 20 and minimum Q-score 0), and outputted the RDA sequence and a file containing details of errors. Error correction of the RDA sequence was attempted, using two nanopore consensus refinement tools, Nanopolish v.0.11.2 and Medaka v.1.3.3, independently. The baseline error rate and indels were calculated from the polished RDA sequence, using an in-house script written by R.J. Orton, AlignMuts.jar (https://github.com/rjorton), which outputs a text file recording the differences between the polished sequence and the published reference.

4.3.6.2 RIA

A long-read assembler, Canu v.1.9, was used to perform *de novo* assembly with the nanopore datasets for Experiments 1 – 3 (Koren et al., 2017). Canu is specifically designed for noisy single-molecule sequences and can cope with large repeats and closely related haplotypes by using an adaptive OLC strategy based on a sparse assembly graph construction that avoids collapsing diverged repeats and haplotypes. Canu assembly includes three stages: correction, trimming and assembly. The correction stage selects the best overlaps to use for correction, estimates corrected read lengths, and generates corrected reads. The trimming stage identifies unsupported regions in the input and trims or splits reads to their longest supported range. The assembly stage makes a final pass to identify sequencing errors, constructs the best overlap graph and outputs contigs (consensus), an assembly graph and summary statistics. Assembly was facilitated by limiting the readSamplingCoverage parameter to 200x (by random sampling), as the coverage exceeded 6000x in some instances (in repetitive regions) and led to excessive computing times. The following line command was used to run Canu.

```
$ canu -p HCMV_strain -d HCMV_strain_folder readSamplingCoverage=200x
batMemory=64g genomeSize=250k maxMemory=100g maxThreads=12 -nanopore-
raw HCMV_strain.fasta
```

Canu outputs the assembled contigs as a fasta file, and includes the sequence length. Manual curation of the contig was necessary to produce the HCMV genome sequence structure as *ab* - U_L - *b'a'c'* - U_S - *ca*, the final RIA sequence.

Mummerplot v3.5 (Marçais et al., 2018) was used to visualise the RIA sequence.

```
$ mummer -mum -n -b -c -l 30 ref.fasta read.fasta > ref_read.mums
$ mummerplot -postscript -p ref_read ref_read.mums
```

4.3.7 Identification of strains based on read data alone

4.3.7.1 By read genotyping

In the first approach, the software minion_Genotyper (https://github.com/salvocamiolo/minion_Genotyper), which was developed by Salvo Camiolo, was used to parse HCMV sequence reads for specific kmers within 13 hypervariable genes (RL5A, RL6, RL12, RL13, UL1, UL9, UL11, UL20, UL73, UL74, UL120, UL146 and UL139) (**Table** 1.6 and **Figure** 3.3). It does so in the same fashion as GRACy for Illumina data (Section 3.4.2). The program outputs a 13-string code corresponding to the genotypes of these genes as they are located along the genome (**Table** 3.3). A file suffixed by *_statistics.txt* is the main output and contains the genotype code for each gene and the number of occurrences of reads mapping to each genotype. The reconstructed strain (or strains in the case of multiple-strain infections) is denoted as a 13-string strain code. If the strain code has been previously observed in any of a collection of 244 HCMV genomes deposited in GenBank (https://github.com/salvocamiolo/minion_Genotyper/blob/master/depositedSequences_codes.txt), the accession number is also reported.

In Experiment 4, which simulated a multiple-strain infection, the presence of the two strains was identified when multiple strain codes were detected in

separate reads. The number of strains present and their ratio was derived from the number of reads reported to match each strain.

In Experiments 5 and 6A, recombination was identified when single reads contained a genotypic combination originating from both parental strains. For example, AF1 is annotated using the 13-string strain code as "AHHHHAQVHLDRN", and ΔRNA2.7 (which has a 13-string code identical to that of Merlin) is "AANNNHACIPYHD". Reads were identified as recombinant when the genotype combinations of both strains were detected in a single read. For example, a read containing the first eight of the 13 hypervariable genes typed as "AAHHHAQV-----" would have originated from AF1, a read coded "AANNNHAC----" would have originated from Merlin, and a read containing "AANNNAQV-----" would have originated from both ΔRNA2.7 and AF1 and would be suggestive of recombination. The reads identified as recombinant by minion_Genotyper were then individually analysed by aligning to an alignment of the AF1 and ΔRNA2.7 references, using the *-add* and *-keeplength* options in MAFFT v.7.475 (Katoh et al., 2019). The original alignment length was therefore retained when aligning the recombinant sequence, maintaining the co-ordinates across all the single-read alignments. A three-sequence alignment was created for each read identified as recombinant by minion_Genotyper: the two references AF1 and ΔRNA2.7 and the recombinant read. The number of base matches, including gaps, between the read and each of the reference sequences was then calculated within adjacent 1000 nt windows, and the similarity of the read to each reference was reported as a proportion for each window. Recombination could be confirmed if the read had a higher similarity to one reference for one part of the read and switched to have a higher similarity to the other reference in another part of the read. This was visualised using similarity plots of the alignments (Excel v.16.77).

4.3.7.2 By read alignment to published HCMV sequences

In a second read-based method for identifying HCMV strains, reads were aligned to 244 published HCMV genomes using minimap2 v.2.17 (Li, 2018). Secondary alignments were disallowed, as HCMV strains are generally highly similar in sequence. Thus, the results of primary alignment alone were output in order to visualise the most highly matched strains. Summary alignment statistics,

including the total number of reads and the percentage of total reads aligned to the matched strains, were obtained using samtools idxstats (Li et al., 2009). This strategy was performed using all reads, and then using reads of ≥10 kb, ≥25 kb and ≥100 kb. To establish an optimal read-length threshold for identifying the input strain, this process was performed using data from Experiments 1 to 3 (the single-strain infections) prior to analysing the composite strains in Experiment 4 (the simulated multiple-strain infection).

4.4 Results

4.4.1 Detection of HCMV genome isomers from long single reads

HCMV sequences were identified in 52, 27 and 29 % of reads from Experiments 1 to 3, for Merlin, U11 and AF1, respectively, using DIAMOND (Buchfink et al., 2015). The 50 longest reads from each run were aligned to the respective reference genome, and the alignments were visualised using the online version of MAFFT v.7 (https://mafft.cbrc.jp/alignment/server/), which displays a LAST dot plot consisting of a graph of the alignment of one sequence against another (Katoh et al., 2019). Experiments 2 and 3 contained long reads that captured most of the HCMV genome and readily demonstrated the existence of the four genome isomers that result from inversion of U_L and U_S by recombination between the flanking repeats in concatemeric DNA during viral DNA replication. Representative reads demonstrating this phenomenon from Experiment 2 are shown in **Figure** 4-3, with a diagrammatic representation of the genome structure and arrows indicating the orientation of U_L and U_S relative to the reference U11 genome below each corresponding dot plot. In Experiment 3, the longest read was 298,823 nt, but on alignment with the reference AF1 genome, it was apparent that this was chimaeric, with half of the read mapping in the forward orientation relative to U_L and the other half to the same sequence in the reverse orientation. This is a recognised feature of nanopore sequencing and occurs either physically by ligation of two distinct molecules during library preparation or *in silico* when a single DNA strand is sequenced through a nanopore and followed in quick succession by the other DNA strand (White et al., 2017).

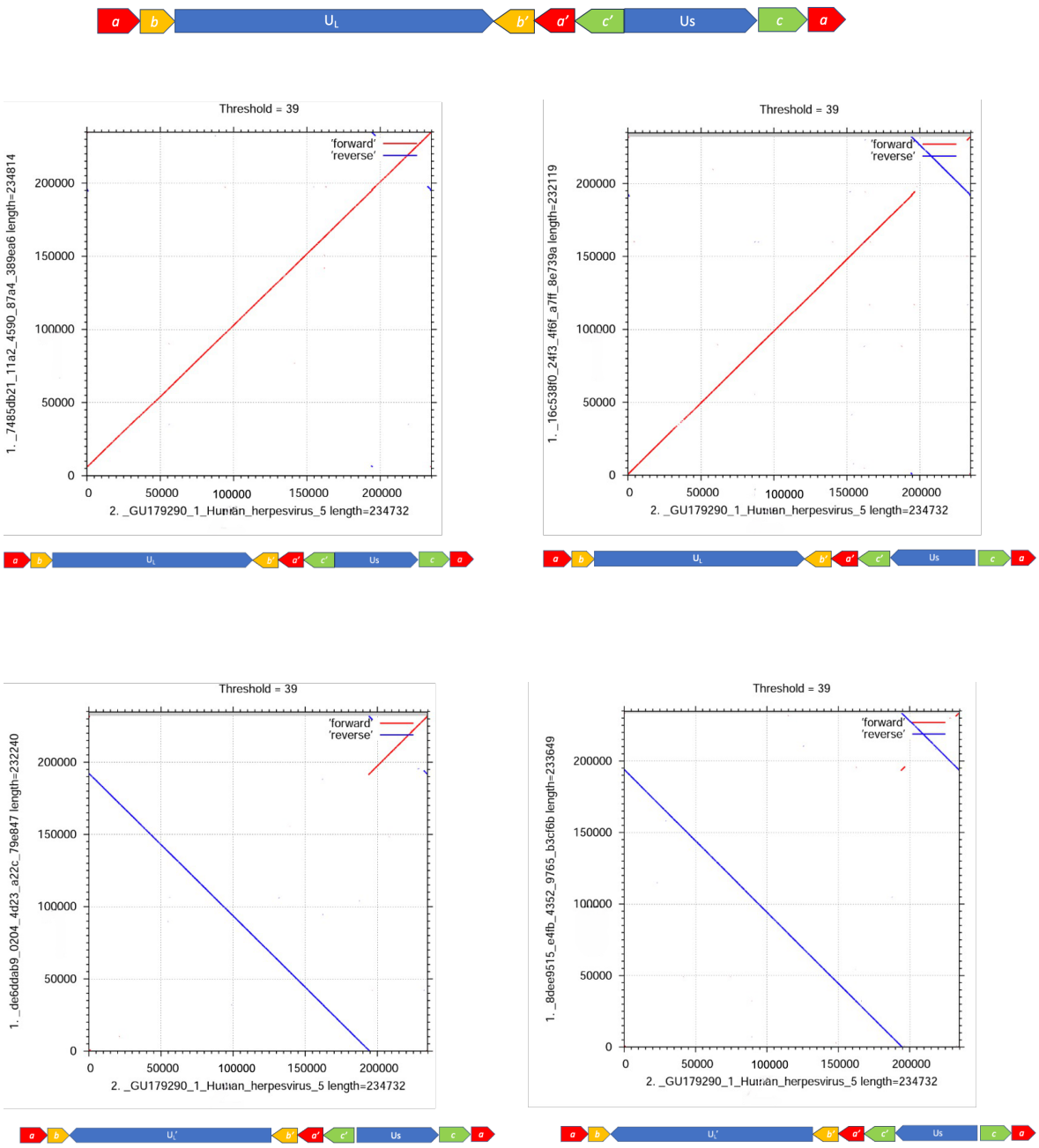


Figure 4-3. LAST dot plots of pairwise sequence similarity of four representative single HCMV reads from Experiment 2 (single strain infection with U11) against the reference U11 genome.
The read is plotted on the y-axis and the U11 genome is plotted on the *x*-axis, with sequence length explicit. Below each plot is a diagrammatic representation of the relevant orientations of U_L and U_S relative to the terminal and internal repeat regions (*ab*, *b'a'c'* and *ca*; not to scale). The standard genome structure is displayed above the plots.

4.4.2 RDA sequence error rate

To ascertain the RDA sequence error rate (rather than the single read error rate), the polished RDA sequences from Experiments 1 – 3 (the single-strain infections) were compared to the respective reference genomes. A number of tools are available for polishing consensus sequences assembled using nanopore data. I utilised one of the earlier tools (Nanopolish) and a more recently developed tool (Medaka) independently to polish the mapped genomes (Section 4.3.6.1). Nanopolish, which corrects draft sequences using signal-level data, did not alter the consensus. In contrast, Medaka, which uses neural networks applied to the pileup files of sequencing reads against the final assembly, amended many of the homopolymer tracts in the mapped consensus. It also corrected ambiguity codes used by Minimap2, where it was unable to call the base. The total number of deletion, insertion, substitution and ambiguity calls of the RDA sequence and the Medaka-polished RDA (mRDA) sequence from the cognate published references are shown in **Table** 4-4. Although Medaka improved the final consensus, it increased the overall percentage accuracy by only 0.1 - 0.2 %. Close inspection of the sequence alignments of the RDA and mRDA sequences and the Merlin reference showed that the differences occurred largely in homopolymer tracts, as expected from nanopore single-read sequence data. As an example, in an alignment from Experiment 1 in **Figure** 4-5, a part of the mRDA consensus and Merlin reference are shown. The differences at positions 7,518 and 7,519 are noted as substitutions, whereas they are likely to be due to an insertion error at position 7,518 and a deletion error at position 7,519. These errors, as well as the deletion at position 7,531, are in keeping with nanopore basecalling errors associated with homopolymer tracts. Such alignments may thus lead to an overcalling of substitutions and an underestimate of deletions and insertions.

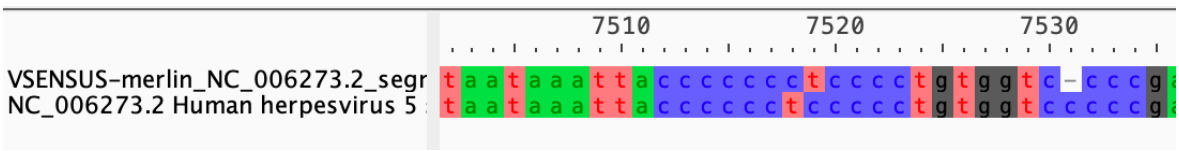


Figure 4-4. Alignment of a short sequence from the final Medaka-polished genome and reference strain Merlin. The reference sequence is the second one shown.

Table 4-3. Comparison of RDA and mRDA sequences from Experiments 1-3 to the respective published references.

Experiment	1		2		3	
Strain	Merlin		U11		AF1	
Genome size (bp)	235,646		234,732		235,937	
Assembled genome[a]	RDA	mRDA	RDA	mRDA	RDA	mRDA
Deletion (no.)	773	328	787	567	992	498
Insertion (no.)	0	22	0	43	0	17
Mutation (no.)	36	9	36	138	39	21
Ambiguity (no.)	3	0	7	0	11	0
Total no. mutations	813	359	823	748	1042	536
Percentage accuracy (%)	99.7	99.8	99.6	99.7	99.6	99.8

[a] RDA, reference-dependent assembly of nanopore data; mRDA, Medaka-polished RDA sequence.

4.4.3 RIA of HCMV genomes

As read length progressively increased in successive runs of the single-strain high-titre cultured HCMV strains in Experiments 1–3 to encompass nearly the entire HCMV genome (Section 4.4.1), *de novo* assembly of reads can seem superfluous, as the use of alignment tools (BLASTn) with a database of published HCMV genomes was sufficient to identify the parental strain. Nevertheless, I used Canu with the datasets from Experiments 1-3 to establish a pipeline for *de novo* assembly of HCMV sequences. This was in anticipation of the fact that clinical samples have lower HCMV DNA concentrations and more fragmented viral DNA (Chapter 5) and that the HCMV genome sequences were very unlikely to be identical to any already published. Canu performed *de novo* assembly of the datasets from Experiments 1-3 with input reads of ≥1000 nt. The Merlin, U11 and AF1 data assembled into HCMV contigs of 373,469, 310,284 and 273,789 nt, respectively, each exceeding the genome size by 20-60 %. In each case, the excess sequence represented an artefactual extension of the HCMV genome due to the presence of the repeat regions. Manual curation of each contig, by trimming the excess sequences and orienting the U_L and U_S regions, enabled the construction of a complete HCMV genome (RIA). A LAST dotplot showing the alignment of the RIA sequence from Experiment 3 against the AF1 reference genome is shown in **Figure** 4-6 (Katoh et al., 2019). Alignment of the RIA sequence to the published HCMV database using BLASTn also confirmed the strain to be AF1. Therefore, construction of the RIA sequence can be used independently to identify the most appropriate HCMV strain for reference-based assembly, where the constituent strain is unknown, as is the case for clinical samples. This initial study using nanopore data from previously sequenced laboratory strains thus provided a pipeline for the agnostic assembly of clinical HCMV genomes and is tested further in Chapter 5.

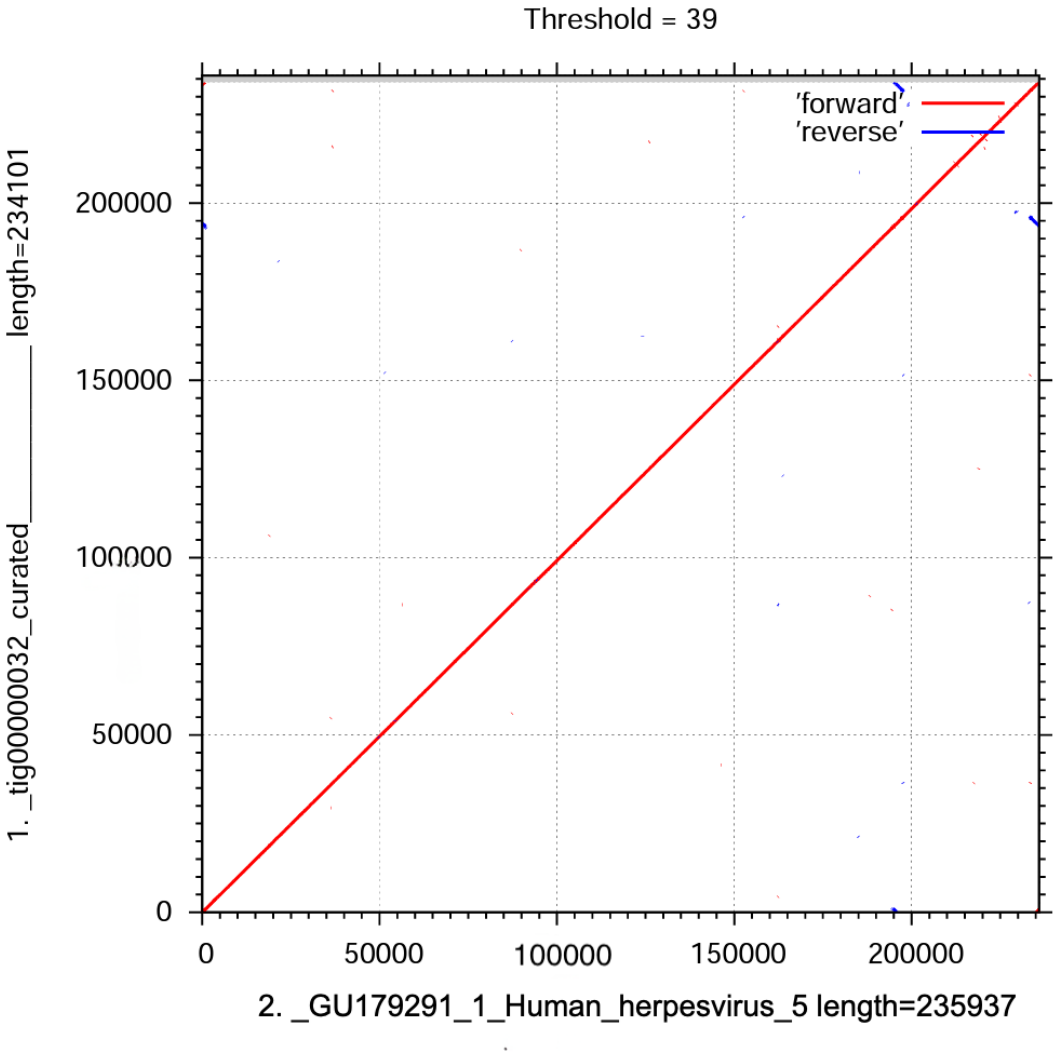


Figure 4-5. Genome-wide similarity comparison of the RIA sequence from Experiment 3 sequence dataset (*y*-axis) to the reference AF1 genome (*x*-axis).

4.4.4 HCMV strain identification from a simulated multiple-strain infection

In a first investigation, Minimap2 was used to align the reads from Experiment 4 to the collection of 244 published HCMV genomes (Section 3.3.7). The optimal read length threshold was established using the single strain data from Experiments 1-3, using reads of ≥10,000, ≥25,000 and ≥100,000 nt. Each of these thresholds allowed the correct identification of the input strain (**Table** 4-5). For the purposes of setting a reliable threshold, a read length of ≥25,000 nt was chosen for a high signal-to-noise ratio that filtered out shorter non-strain-specific sequences without compromising the total number of reads available, as would have been the case if the higher threshold of ≥100,000 nt had been used.

Using the threshold of ≥25,000 nt, U11 and AF1 were identified as the top two matches, with 61.35 % of reads mapping to U11 and 39.03 % of reads mapping to AF1 (**Table** 4-6).

In a second investigation, minion_Genotyper was used to assign strain codes for 13 hypervariable genes and correctly identified the strains in the single-strain datasets from Experiments 1-3 and the simulated multiple-strain dataset from Experiment 4 (**Table** 4-7). As described above, an input read length threshold of ≥25,000 nt was used to increase specificity but not at the detriment of unduly decreasing the absolute number of reads available. This approach correctly identified the strain in Experiments 1-3 and also AF1 and U11 in a 2:3 ratio (**Table** 4-7) in Experiment 4, in concordance with the ratio reported from the Minimap2 investigation described above.

Both investigations reliably differentiated the presence of multiple strains from the read data and thus demonstrated the ease by which multiple strains even in a roughly equimolar ratio can be identified and distinguished using nanopore data.

Table 4-4. Most closely matched strains on alignment of reads from Experiments 1-3 to the published collection of 244 HCMV genomes.
The top match in each case is identified in bold font, as are the corresponding percentages of mapped reads. Supplementary alignments account for the number of mapped reads exceeding the total number of reads when input read length was ≥100,000 nt, and this occurred more commonly in longer reads.

Experiment	Strain	Input read length (nt)	Total reads (no.)	Matched strains	Mapped reads (no.)	Mapped reads (%)
1	Merlin	Reads ≥ 10k	59,024	**Merlin**	17,592	**29.80**
				UK/Lon3/Plasma/2012	2,076	3.52
		Reads ≥ 25k	6,273	**Merlin**	5,631	**89.8**
				CINCY	173	2.76
		Reads ≥ 100k	240	**Merlin**	382	**159.17**
				CINCY	21	8.75
2	U11	Reads ≥ 10k	97,062	**U11**	23,923	**24.65**
				Pat_G	610	0.63
		Reads ≥ 25k	14,152	**U11**	11,474	**81.1**
				CINCY	335	2.37
		Reads ≥ 100k	643	**U11**	922	**143.39**
				CINCY	53	8.24
3	AF1	Reads ≥ 10k	104,235	**AF1**	56,932	**54.62**
				CINCY	915	0.88
		Reads ≥ 25k	28,351	**AF1**	29,405	**103.7**
				CINCY	793	2.80
		Reads ≥ 100k	883	**AF1**	1306	**147.90**
				CINCY	75	8.49

Table 4-5. Most closely matched strains on alignment of reads from Experiment 4 to the published collection of 244 HCMV genomes..
The top matches were identified by alignment of all reads ≥25,000 nt.

Experiment	Simulated multiple strain	Total no. reads ≥25,000 nt	Matched strains	Mapped reads (no.)	Mapped reads (%)
4	AF1 and U11	23,244	U11	14,260	61.35
			AF1	9,073	39.03

Table 4-6. Strains identified by minion_Genotyper from Experiments 1-4.
The 13-string codes are shown for the corresponding strain for single-strain sequencing runs (Experiments 1-3) and constituent strains from the multiple-strain simulation (Experiment 4).

Experiment	Strain(s) identified	13-string code
1	Merlin	AANNNHACIPYHD
2	AF1	AHHHHAQVHLDRN
3	U11	ACDAARAQADDHD
4	AF1:U11 = 2:3	AHHHHAQVHLDRN
		ACDAARAQADDHD

4.4.5 Detection of interstrain recombinants

In Experiment 5, two strains (U11 and AF1) were co-cultured to allow recombination. After sequencing, only a single strain was detected by inputting reads ≥10,000 nt into minion_Genotyper. A single 13-string code (AHHHHAQVHLDRN) identified the sole strain, corresponding to AF1. In support of this, alignment of all reads from this dataset by BLASTn against the AF1 and U11 reference genomes revealed only matches to AF1. This experiment was unsuccessful in stimulating recombination between AF1 and U11 in cell culture, as AF1 was the only strain propagated.

In the subsequent recombination experiment (Experiment 6A, RecR1 AF1/ΔRNA2.7), in which AF1 and ΔRNA2.7 were used to co-infect HFFF2 cells in culture, inputting reads of ≥10,000 nt to minion_Genotyper outputted single reads with genotypic code combinations corresponding to AF1 or ΔRNA2.7 or to AF1/ΔRNA2.7 recombinants. The 13-string strain code assigned to ΔRNA2.7 (AANNNHACIPYHD) is identical to that of Merlin, and that of AF1 is AHHHHAQVHLDRN. Recombinant reads were identified when the 13-string code was a hybrid of codes from ΔRNA2.7 and AF1 (**Table** 4-8**)**. Discarding read combinations that only appeared only once, a total of 14,755 reads matched AF1, 1,326 matched ΔRNA2.7 and 114 had evidence of recombination. The proportion of potentially recombinant reads was therefore only 0.64 %. Additionally, as seen previously in Experiment 5 (co-culture of AF1 and U11), AF1 is apparently better adapted for growth in HFFF2 cell culture.

The dataset from Experiment 4, in which DNA extracts from AF1 and U11 were mixed prior to sequencing, served as a negative control as it could not contain recombinant reads (**Table** 4-9). The number of reads with hybrid codes from both AF1 and U11 were much lower, accounting for only 0.21 % of reads of ≥10,000 nt. In addition, there were fewer reads (≤5) supporting each recombinant code, and these reads exhibited substantial gaps between hypervariable genes from the two strains. This is exemplified by the most frequent apparent recombinant read code (---H------DHD), which was detected in five reads (**Table** 4-9). The fourth gene in the code (RL13) was identified as having originated from AF1 and was followed by a gap of 157,901 nt before the last three genes (UL120, UL139 and UL146) that were identified as having

originated from U11. However, the five reads with this combination ranged in length from 36,492 to 108,178 nt. This suggests that the read error rate was sufficient for minion_Genotyper occasionally to misidentify reads as having RL13 motifs. As a result, the threshold for identifying potentially recombinant reads was set at ≥5 reads sharing a specific recombinant code and an absence of gaps between consecutive hypervariable genes.

The top three profiles identified by minion_Genotyper indicating recombination events between UL120 and UL146, RL6 and RL12, and UL1 and UL9, corresponding to the 13-string codes of "**----------**YRN", "-HNNNHAC-----" and "AHHHHHAC-----", respectively (**Table** 4-8; for reference to hypervariable gene positions, see **Table** 4-10 and **Figure** 3-3). The reads were aligned individually to an alignment of the reference sequences for AF1 and ΔRNA2.7. The proportional similarity of the read to each reference was then calculated and plotted within 1000 nt sliding windows. Examples of similarity plots are shown in **Figure** 4-6. The similarity plot in **Figure** 4-6A shows a recombinant read with the 13-string code "**----------**YRN", which includes UL120, UL139 and UL146, aligned against the relevant part of the reference. Between approximately 167,000 and 172,000, the read is more similar to ΔRNA2.7, but from 172,000 onwards it is more similar to AF1, thus suggesting that recombination occurred at a point between UL120 and UL146. The similarity plot in **Figure** 4-6B shows evidence of recombination between RL6 and RL12 ("-HNNNHAC-----"). It is notable that minion_Genotyper failed to identify a RL5A genotype in this read, probably as a result of the high error-rate in single reads. This is also the likely explanation of why similarity to the references is less than perfect along the alignments generally. The similarity plot in **Figure** 4-6C similarly suggests that recombination occurred between UL1 and UL9.

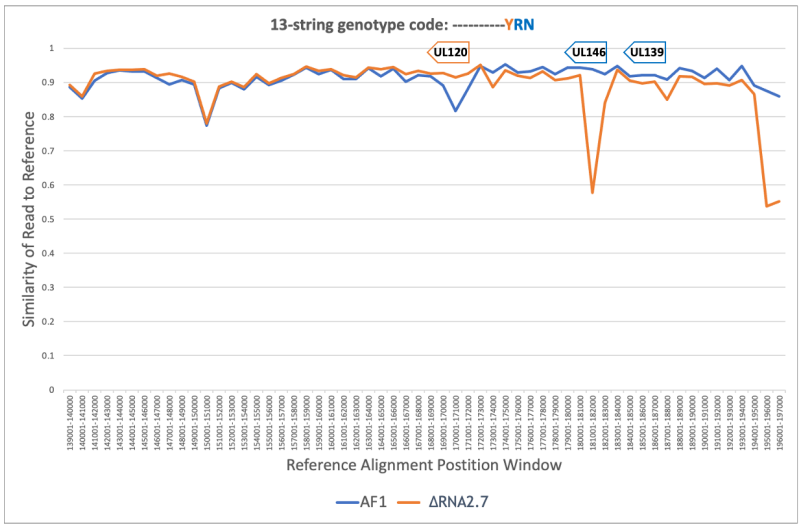
Table 4-7. Reads identified as potential recombinants in Experiment 6A (Rec AF1/ΔRNA2.7) by minion_Genotyper.
The 13-string code for AF1 is AHHHHAQVHLDRN (in blue font) and that for Merlin/ΔRNA2.7 is AANNNHACIPYHD (in orange font). A dash represents the position of a hypervariable gene where no motifs matched any genotype. Both AF1 and Merlin gene RL5A have the same genotypic code, A (in purple font). All potential recombinant reads are shown (n=14). (Proportion of all reads = 0.64 %.)

13-string genotype code	Reads (no.)
----------YRN	16
-HNNNHAC-----	13
AHNNNHAC-----	7
AHHHHHAC-----	5
----------HN	5
AHNNN--------	5
-AHHHAQV-----	5
-ANNNHAV-----	4
AHNNNHA------	4
AAHHHAQ------	4
-----AQC-----	4
-HNNN--------	3
AAHHHAQV-----	3
-HHHHHA------	3
AANNNHAV-----	3
-HNNNHA------	3
----------DRD	3
-----HQV-----	3
---N----HL---	3
-HHHHHAC-----	2
-HHHHHQ------	2
-HNN---------	2
----------DHN	2
AHN----------	2
AHNNNAQV-----	2
AHNNNH-------	2
--NNNHAV-----	2
--HHHAQC-----	2

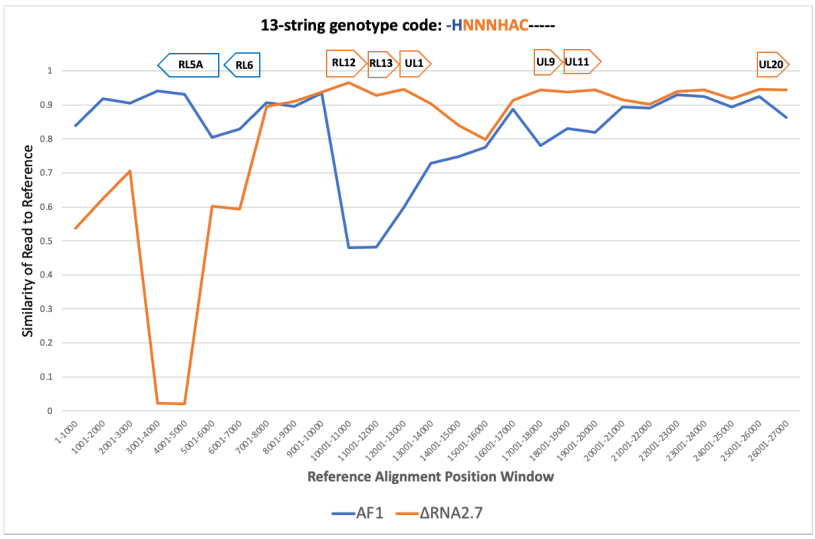
Table 4-8. Reads identified erroneously as AF1/U11 recombinant reads in Experiment 4 (simulated multiple-strain infection).
The 13-string code for U11 is ACDAARAQADDHD (in green font), and that for AF1 is AHHHHAQVHLDRN (in blue font). A dash represents the position of a hypervariable gene where no motifs matched any genotype. (Proportion of all reads = 0.21%).

13-string genotype code	Reads (no.)
---H------DHD	5
-CDAARAQ-L---	4
ACDAARAQ-L---	4
-CDAA----L---	3
-CHHHAQV-----	2
-CDAARAV-----	2
---H----ADDHD	2
ACDAA----L---	2

A



B



C

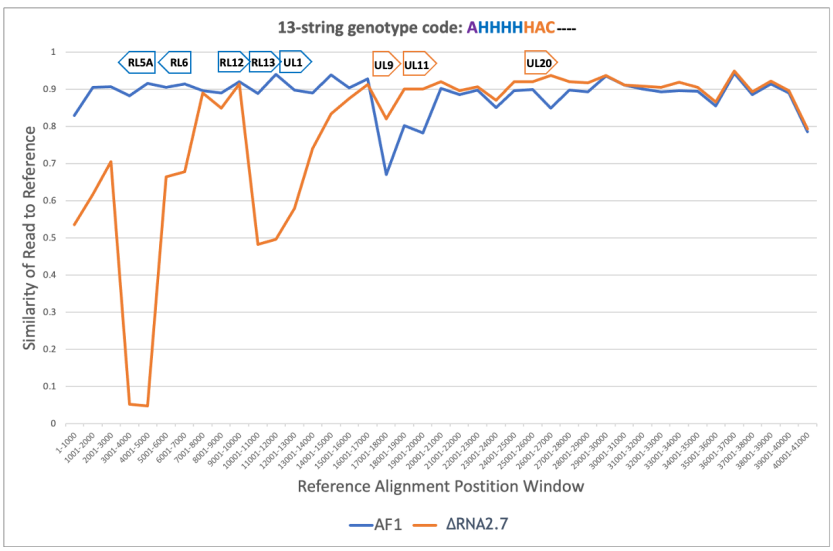


Figure 4-6. Similarity plots of potential recombinant reads aligned to the AF1 and ΔRNA2.7 reference sequences.
The examples shown represent the three profiles with the highest number of reads in **Table** 4-8. A) Similarity plot of a read containing UL120, UL139 and UL146 aligned against part (139,000-197,000 nt) of the references. B) and C) Similarity plots of reads containing RL5A to UL20 aligned against parts (1-27,000 nt and 1-41,000 nt) of the references, respectively. The positions and orientations of the hypervariable genes are shown above the similarity plots.

Table 4-9. Table of 13 hypervariable genes and their positions in the HCMV genome.
The corresponding genotypic codes assigned to each gene for AF1 and ΔRNA2.7 are shown. The start and end positions of each gene shown are relative to the Merlin genome.

	RL5A	RL6	RL12	RL13	UL1	UL9	UL11	UL20	UL73	UL74	UL120	UL146	UL139
AF1	A	H	H	H	H	A	Q	V	H	L	D	R	N
ΔRNA2.7	A	A	N	N	N	H	A	C	I	P	Y	H	D
Start	5,819	6,504	9,850	11,189	12,174	16,947	18,642	25,640	106,937	108,848	169,971	181,292	186,878
End	2,489	5,959	11,094	12,070	12,830	17,648	19,460	26,656	109,088	107,430	169,366	180,930	186,462

4.5 Discussion

The impact of the genotypes of HCMV strains and of multiple-strain infections on the clinical severity of HCMV disease has been difficult to study, given the limitations of short-read sequencing technology and the need to analyse entire HCMV genomes simultaneously. Long-read sequencing technology has opened a window on enabling such studies, but the limitations of this technology, which is still in its infancy, need to be ironed out. I started by exploring the ability of the ONT sequencing platform to determine the accurate, complete genomes of three single, well-characterised, high-titre cultured HCMV strains (Experiments 1-3 involving Merlin, AF1 and U11). I did this using a reference-based assembly pipeline (RDA), producing polished HCMV genome sequences for comparison with the published references. I also demonstrated a reference-independent assembly pipeline (RIA) using the same nanopore datasets. Next, I simulated a multiple-strain infection by mixing DNA extracts from two strains in a 50:50 ratio, which is precisely the situation that proves technically challenging for short-read sequencing (Experiment 4, involving AF1 and U11). Finally, I tested the ability of long-read sequencing to detect recombination between two strains grown together in cell culture (Experiment 6A, involving ΔRNA2.7 and AF1).

Initially, I planned to sequence high-titre cultured HCMV strains and obtain complete genomes using a hybrid approach involving *de novo* read assembly (to obtain the closest match strain to use as reference) and reference-based read assembly. However, as the nanopore datasets were being generated, I found that an end-to-end workflow, involving cell-free virus propagation, high-molecular weight genomic DNA extraction and the use of improved nanopore sequencing hardware and basecallers, was capable of sequencing almost complete HCMV genomes in single reads, thus removing the need for this hybrid approach. In addition, the capture of near complete HCMV genomes in single long reads confirmed the existence of four genome isomers differing in the relative orientations of U_L and U_S (Section 1.3.1). Previous use of nanopore sequencing on a cultured HCMV strain also confirmed the presence of genome isomers, but the evidence was partial because the reads only included incomplete genomes (Karamitros et al., 2018). As the published record length for a single nanopore read was 2.3 Mb at the time of my study (Payne et al.,

2018), which is well in excess of the 236 kb length of the HCMV genome, the ability of nanopore sequencing to capture complete HCMV genomes in single reads was not in doubt. The challenge was in obtaining intact, high-molecular weight genomic DNA from HCMV propagated in cell culture. This was achieved by careful extraction of HCMV genomes from cultured samples, thus fulfilling the potential of long-read sequencing.

Another hurdle to nanopore sequencing lies with obtaining DNA in sufficient quantities, since PCR amplification was not involved in library preparation in my study; micrograms are required, versus nanograms for Illumina sequencing. Previous studies using nanopore technology to sequence HCMV genomes have used transposase fragmentation with PCR amplification (Karamitros et al., 2018) or bait-based enrichment as part of library preparation (Eckert et al., 2016), both of which limit DNA fragment length. The maximum length of a single HCMV read prior to my study was 56,310 nt (Karamitros et al., 2018). In addition to isolating high-molecular weight DNA, I utilised the genomic DNA by ligation protocol to minimise the number of downstream steps potentially promoting DNA fragmentation while still successfully generating datasets.

Much focus has been placed on the relatively high error-rate of nanopore sequencing in comparison with traditional Sanger or short-read platforms. I demonstrated that with the contemporary iteration of flow-cells and basecallers, nanopore read data are of sufficient quality to identify HCMV genomes to the level of the relevant strain in single-strain infections (Experiments 1-3) and a simulated multiple-strain infection (Experiment 4). Furthermore, recombination is recognised as having occurred during HCMV evolution, as evidenced by the isolated regions of hypervariability in a largely well-conserved genome (Lurain et al., 2006, Bradley et al., 2008, Sijmons et al., 2015, Lassalle et al., 2016, Suárez et al., 2019b). In Experiment 6A, read-level data were sufficient to indicate the occurrence of recombination. However, potentially recombinant reads were rare, being detected in only 0.64 % of all reads HCMV reads sequenced. Evidence for recombination of HCMV during co-infection *in vivo* has been documented previously. One study obtained multiple-strain isolates from transplant recipients and demonstrated recombinant strains using restriction enzyme digest profiles after propagation by passage in culture (Chou, 1989). Another identified the presence of recombinants using RFLP and

DNA sequencing of gB genes directly from samples in three of 14 immunosuppressed patients (Haberland et al., 1999). I have shown the relative ease and sensitivity with which recombination events may be detected directly using nanopore technology, even though these events are rare.

My demonstration that a single long read can encompass the entire HCMV genome may obviate the requirement for genome reconstruction in situations where the primary aim is to identify constituent pathogens. This could be particularly helpful in the field of agnostic diagnostics in clinical practice. I found that the long-read assembler, Canu, required considerable restriction and subsampling (to a read coverage depth of 200 reads/nt) in order to reconstruct the HCMV genome from successive ONT sequencing runs, as the number of reads of >100 kb increased and therefore already encompassed most of the HCMV genome (**Table** 4-2). As Canu relies on an OLC method, HCMV genome assembly can be complicated by the presence of many whole-genome reads spanning the four genome isomers and cause this program to assemble contigs larger than a full-length genome.

In addition to exploring the quality of ONT sequencing in relation to individual reads, I investigated the accuracy of ONT sequencing at the consensus genome level. Although there has been a previous study using nanopore sequencing to reconstruct the genome of an HCMV strain grown in cell culture (Karamitros et al., 2018), this made use of a formulated transposase-based kit (the ONT rapid sequencing kit), which required fragmentation of genomic DNA during simultaneous attachment of adaptors, thus restricting read length as outlined above. These authors developed a self-correcting loop using only nanopore reads to reconstruct a full HCMV genome without supplementary reads from Illumina sequencing. Moreover, the strain used in this study (TB/40E) had undergone substantial passage in various cell types and exhibited a substantial degree of heterogeneity (Sinzger et al., 1999, Sinzger et al., 2000, Dolan et al., 2004). Nevertheless, the authors claimed to have managed to annotate 98.8 % of the recognised genome features correctly, demonstrating the utility of longer reads despite their higher error rate compared to short-read platforms. With the availability of improved basecallers, nanopores and sequencing chemistry, I ascertained the error rate of a polished, reference-dependent assembly (mRDA) sequence by sequencing single well-characterised strains (Experiments 1-3). I

chose three low-passage strains: Merlin (Dolan et al., 2004), U11 and AF1 (Dargan et al., 2010) to assess the consensus error rate from nanopore datasets. The RDA sequence was constructed using reference-based assembly from each of Experiments 1-3, and two consensus-polishing programs were assessed independently for their ability to improve the RDA sequences. Medaka, which is based on neural network improvement by pileup of individual reads against the genome assembly, was found to outperform Nanopolish, which carries out signal-level based correction and did not alter the final genome sequences. The Medaka-polished (mRDA) sequence was compared for each dataset to the reference sequences of Merlin, U11 and AF1 published in GenBank (**Table** 4-3**)**. Although Medaka correction of the genomes reduced the number of differences from those of the reference strains, the overall percentage accuracy increased by only 0.1 % as the final consensus genomes were already >99.5 % accurate. Thus, despite the raw error-rate associated with individual reads remaining high (modal accuracy of 97.6 %, equivalent to a Phred score of Q16, in the iteration of the ONT R9.4.1 flow cell and basecaller model used), the error-rate at consensus level approached that of Illumina sequencing. A caveat remains for insertions and deletions (indels) in homopolymer tracts, which can cause frameshift errors in gene annotation (Delahaye and Nicolas, 2021). The latest ONT updates to the flow cell (R10.4) and basecalling have tackled homopolymer calling accuracy to the extent that the majority of homopolymers at the consensus level are reportedly correctly resolved at lengths of <11 nt in R10.4 data, which is on a par with Illumina data (Sereika et al., 2022). It would be worth revisiting my experiments with the latest versions of the improved nanopore hardware and software.

In summary, I successfully harnessed the long-read sequencing capability of the nanopore platform to determine complete HCMV genomes and showcased the utility of this technology in separating the genomes of multiple strains, demonstrating overall genome structure (in terms of the four genome isomers), and detecting recombination events. In Chapter 5, I progressed to test my protocol on clinical samples with high HCMV loads. For this purpose, modifications to my pipeline were necessary, not least because I did not have *a priori* knowledge of the number of strains or their sequences in these samples, However, having demonstrated that *de novo* assembly can be used to generate

HCMV genomes from nanopore data, I could test the hybrid approach of using the RIA sequence as reference for RDA sequence assembly. I could further use the Illumina sequencing pipeline used to acquire HCMV genome sequences from FFPE (Chapter 3) to obtain the corresponding genomes for comparison. The utility of the published database of HCMV strains was also demonstrated in my work focused on obtaining HCMV genomes from high-titre cultured strains, and in identifying constituent strains (Section 4.3.7.2). Last, but most importantly, the challenge of implementing my workflow to clinical samples was the low starting concentration of DNA from these samples, without propagation in cell culture. An enrichment of HCMV DNA over human or bacterial contaminant DNA in order to obtain sufficient material for nanopore sequencing was therefore required.

5 Nanopore sequencing of HCMV strains from clinical material

5.1 Background

Building on the workflow I established in Chapter 4 to construct complete HCMV genomes from high-titre cultured HCMV samples using nanopore sequencing, demonstrating both a low consensus error rate and the ability to reconstruct individual genomes from multiple-strain infections, I progressed to test the capability of the nanopore platform to sequence HCMV genomes directly from clinical samples. As I demonstrated in Chapter 3, there is an established Illumina workflow for the sequencing of HCMV directly from clinical samples. However, the short reads from Illumina sequencing still struggle with the accurate characterisation of multiple-strain infections, especially when strains are present in approximately equal ratios (**Figure** 1-9). The longer reads afforded by nanopore sequencing enable multiple-strain infections to be separated readily, as I showed in Chapter 4 by using nanopore data to sequence a mixture of high-titre cultured AF1 and U11 DNAs (Section 4.4.4, Experiment 4). Even so, an enrichment process for HCMV DNA is still required from non-culture-propagated clinical samples, and to sequence HCMV from FFPE samples on the Illumina platform I used custom-designed oligonucleotide baits (Hage et al., 2017, Suárez et al., 2019b). Alternative methods, such as PCR amplification of the whole HCMV genome by tiled, overlapping amplicons (Cunningham et al., 2010) or by multiple-displacement amplification (MDA) (Sijmons et al., 2014) have also been used successfully for HTS. Nanopore sequencing offers an ideal means of investigating multiple-strain infections and their possible association with more severe disease. Additionally, its small footprint and low capital cost makes it an attractive option for sequencing HCMV in clinical practice. However, the major obstacles remain of a relatively low concentration of viral DNA in clinical samples and a relatively high amount of input DNA required in comparison with the Illumina platform. In order to capitalise on the long reads afforded by nanopore sequencing, I investigated a means of enriching viral DNA from clinical material while maintaining intact viral genomes.

Enriching HCMV DNA for sequencing on the nanopore platform has previously been achieved with a degree of success from cultured samples by using

transposase fragmentation and PCR (Karamitros et al., 2018) or oligonucleotide bait-based enrichment (Eckert et al., 2016). However, these studies still necessitated HCMV passage in cell culture, which is associated with mutations, gene loss and rearrangement of the wild-type HCMV genome (Akter et al., 2003, Cha et al., 1996, Dolan et al., 2004) (Section 1.8). Furthermore, cell culture has been superseded by real-time PCR for diagnostics in most clinical laboratories, prompting adaptation of a workflow that avoids this step. Also, these enrichment processes still fragment HCMV DNA and therefore diminish the potential of long-read sequencing.

I attempted to enrich viral DNA directly from clinical samples without using cell culture. I tested whole-genome amplification (WGA) of HCMV DNA using molecular methods while aiming to preserve intact genomic DNA. In the first method, I designed long-range PCR primers to amplify a large section of the HCMV genome. Theoretically, the entire HCMV genome could be amplified by using approximately 24 sets of tiled amplicons of approximately 10 kb generated using long-range PCR with primers in conserved regions. The goal was therefore to amplify the whole HCMV genome by designing large tiled amplicons. However, this would still inescapably shorten HCMV sequences to approximately 10 kb. In the second method, I used MDA with phi-29 DNA polymerase, which claims to achieve DNA lengths of up to 100 kb. MDA use was also encouraged by the fact that it had been incorporated into the Premium WGA protocol (ONT).

Finally, to test the limits of the ONT platform, I sequenced clinical samples directly without any enrichment step, using samples with high DNA concentration and HCMV load. To assess the accuracy of HCMV genomes determined using nanopore data, I also sequenced the same genomes on the Illumina platform using target enrichment. No HCMV genomes had been sequenced from clinical material using nanopore technology at the initiation of this project.

5.2 Objectives

1. To obtain whole HCMV genomes from clinical material by nanopore sequencing, with and without an enrichment step.

2. To compare the accuracy of HCMV consensus genomes derived from clinical samples by nanopore sequencing to those obtained by Illumina sequencing.

5.3 Materials and methods

Four fully anonymised residual DNA extracts and four residual primary samples with high HCMV loads were obtained for sequencing. Extracts from the West of Scotland Specialist Virology Centre (WoSSVC) were acquired from the NUCLISENS EASYMAG (bioMérieux, Marcy-l'Étoile, France) or Abbott m2000 (Abbott, Chicago, Illinois, USA) automated extraction platforms. The residual primary samples (two urines, one mouth swab and one vitreous humour) were extracted using QIAmp Mini DNA extraction kits and Genomic-tip extraction kits (QIAGEN) (Section 2.5). The DNA yields from Genomic-tip extractions were too low to proceed for downstream processing due to the small sample volume and low DNA concentration of the starting material. The residual extracts had higher DNA concentrations compared to the re-extracted samples, possibly due to the degradation of nucleic acid in the residual primary samples, which were stored at 4 ºC for at least a week prior to being re-extracted. qPCR (Section 2.7) was used to select samples for HCMV enrichment by WGA or direct sequencing. Two samples with sufficient DNA concentration (8.28 and 16.6 ng) and highest HCMV load to human DNA ratio were chosen to proceed to direct nanopore sequencing (sample 1: urine and 2: lung; **Table** 5-1). Two samples with low DNA concentration but high HCMV load to human DNA ratio were selected for WGA prior to nanopore sequencing (samples 5: VH and 8: U533; Table 5-1). Additionally, a collection of residual extracts positive for HCMV were obtained from WoSSVC to validate long-range primers designed for WGA of HCMV DNA. Ethical approval was not required for the use of fully anonymised residual diagnostic samples in the validation of the sequencing protocol.

Table 5-1. HCMV-positive residual samples and extracts that were assessed for nanopore sequencing.
Samples that proceeded to nanopore sequencing are highlighted in bold font.

Sample ID	Sample origin[a]	Qubit (ng/µL)	A260/A280	qPCR UL97 (IU/µL)	qPCR human FOXP2 (cp/µL)	HCMV:human DNA	HCMV viral load (log10 IU/µL)
1	**Urine**	**8.28**	1.79	258,199	323	**799.38**	**5.41**
2	**Lung**	**16.6**	1.83	172,259	9,636	**17.88**	**5.24**
3	Biopsy	16.3	1.62	0	6,151	0.00	undetectable
4	Biopsy	31.6	1.72	1,379	6,766	0.20	3.14
5	**Vitreous humour (VH)**	**0.14**	0.93	7,725	162	**47.69**	**3.89**
6	Mouth swab	0.392	3.91	383	199	1.92	2.58
7	Urine (U781)	too low	1.14	1,598	21	76.10	3.20
8	Urine (U533)	too low	0.72	19,500	11	1772.73	4.29

[a] Samples 1-4 were residual DNA extracts, and samples 5-8 were residual primary samples that were re-extracted manually.

5.3.1 HCMV enrichment by WGA

5.3.1.1 Trial of long-range PCR amplification

Primers were designed for a conserved region of the HCMV genome spanning from near the end of UL54 to near the beginning of UL57 (73,358 – 80,000 nt in the Merlin reference genome). The Primer-BLAST module at the National Center for Biotechnology Information (NCBI) website was used to design primers with amplicon sizes ranging from 1,000 to 10,000 nt (**Table** 5-2) (Ye et al., 2012). Primer parameters were input for minimum, optimum and maximum primer sizes of 15, 20 and 25 nt: primer melting settings of temperature (T_m) minimum 53 °C, optimum 53 °C, maximum 62 °C and a maximum Tm difference of 3 °C, avoiding regions of low complexity. Sigma OligoEvaluator (http://www.oligoevaluator.com/LoginServlet) was used subsequently to tailor the T_m of the primers for the GeneAmp PCR system 9700 (Applied Biosystems). Custom-designed primers were received lyophilised and reconstituted in NFW to 100 µM (PCR Oligos, Sigma-Aldrich, St. Louis, Missouri, USA). Primers were tested on a range of DNA extracts from cultured HCMV strains prior to amplification of HCMV DNA from clinical material.

Table 5-2. Summary of primers for long-range amplification of HCMV genomes.
Start and stop positions are relative to the Merlin reference genome.

Primer	Sequence	Start	Stop	Tm °C
F1	GCTTATAGTTGGGCGAGTTA	80421	80441	59.5
F2	GGCAGGTTAGATTGACGGTA	78698	78718	61.5
R1	TGATCAAGCATAAAACGGGA	81753	81733	62.9
R2	GAAAATACCGACTTCAGGGT	82821	82801	59.5
R3	AATGACCGCCACTTTCTTAT	83680	83660	60.0
R4	GGTACGGATCTTATTCGCTT	84495	84475	59.6
R5	CAATTTTTCGGACTGTCAGG	85265	85245	62.1
R6	ATGTGAATATCCAGACGGTG	86616	86596	59.9
R7	CTAAAGTACTGCGATCCGAA	87229	87249	59.5
R10	CGCTTTTCAGACCGCAACAA	87790	87770	68.4
R11	TACTACGGCTTCAAGGACTA	88854	88834	56.7

The Expand long template PCR system (Roche; cat. no. 11 681 834 001) was used to amplify HCMV DNA using the designed primers. The kit contains the Taq and Tgo DNA polymerases and offers improved accuracy due to the inherent 3'-5' exonuclease proofreading activity of the latter enzyme. Amplification of DNA fragments of up to 20 kb from human genomic DNA has been demonstrated using this kit (https://www.sigmaaldrich.com/deepweb/assets/sigmaaldrich/product/documents/362/912/elongrobul.pdf). Buffer 2 from the kit, which is optimised for amplification of fragments of 9-12 kb, was used. Using the recommended input of template DNA of 500 ng, stock virus concentrates from cultured extracts were diluted to 500 ng/µL, and 1 µL was added per reaction. Residual clinical extracts invariably had very low DNA concentrations (measured by Qubit fluorometry; Section 2.6) and thus required a higher input volume (5-10 µL). The composition of the mastermix is shown in **Table** 5-3.

Amplified DNA products were detected by agarose gel electrophoresis. DNA fragments in the size range 1-10 kb were detected using a 0.7 % (w/v) agarose gel containing 0.005 % (w/v) ethidium bromide, with the Quick-Load 1 kb Plus DNA ladder (New England Biolabs; cat. no. N3239S) as a marker. The gel was visualised under UV light in the Syngene U:GENIUS3 EZ Image Capture system (Syngene, Cambridge, UK).

Working stocks of Merlin, AF1 and U11 DNA (Section 2.3) were used to assess the long-range primers prior to their use on clinical material.

Table 5-3. Concentration of components required per reaction for long range PCR using the Expand long template PCR system for amplification of genomic DNA.
Template DNA volume varied according to the concentration of DNA in the extract (cultured or clinical sample).

	9-12 kb System 2
Components	**Volume (Final Concentration)**
Distilled water	up to 50 µL
dNTP	1 µL (500 µM)
Forward primer	2.5 µL (300 nM)
Reverse primer	2.5 µL (300 nM)
10x PCR Buffer	5 µL (2.75 mM $MgCl_2$)
Template DNA	x µL (up to 500 ng genomic DNA)
Expand long-template enzyme mix	0.75 µL

5.3.1.2 WGA using MDA

Samples 8 (urine, U533) and 5 (vitreous humour, VH) were re-extracted and amplified using MDA (**Table** 5-1). The DNA concentrations of these samples as measured by Qubit fluorometry was 0.14 ng/µL for VH and too low to quantify (<0.1 ng/µL) for U533. The REPLI-g kit (QIAGEN; cat. no. 150043) was used for WGA from these samples (Dean et al., 2002). The aim was to amplify whole genomic DNA uniformly from these samples and subject it to downstream nanopore sequencing, which requires a minimum input of 1 µg of DNA. The REPLI-g kit was chosen because it provides uniform amplification across the entire genome with negligible sequence bias (Hosono et al., 2003). The kit relies on isothermal genome amplification by Phi29 DNA polymerase and can replicate up to 100 kb without dissociating from the template. Typical DNA yields are approximately 40 µg per 50 µl reaction, and the average product length is >10 kb with a range of 2-100 kb. Phi29 polymerase also has a 3'–5' exonuclease proofreading activity that should maintain high fidelity during replication. After amplification, the samples were treated with T7 endonuclease I to resolve the hyperbranched structure of the WGA products, which results in shorter fragment lengths (<5 kb).

After extraction of DNA using a QIAamp Mini kit (QIAGEN), the eluate was first concentrated using AMPure XP beads (Section 2.6) prior to WGA using the REPLI-g kit. Preparation for WGA included reconstitution of lysis buffer (DLB) in 500 µL of NFW. Buffers D1 and N1 were prepared with enough for seven reactions.

Buffer D1 was made by mixing 9 µL of reconstituted DLB in 32 µL of NFW, and buffer N1 comprised 12 µL of stop-solution and 68 µL of NFW. Both buffers were stored at -20 ºC until required. Concentrated eluates from samples VH and U533 were amplified by combining 5 µL of DNA extract with 5 µL of buffer D1, incubating at room temperature for 3 min, and adding 10 µL of buffer N1. REPLI-g Midi polymerase was thawed on ice. The reaction buffer (27 µL) and 1 µL of REPLI-g Midi polymerase were mixed (total volume 48 µL), and the reaction was incubated in a thermocycler at 30 ºC (lid set to 70 ºC) for 16 h.

Post-MDA clean up and T7 endonuclease digestion were performed prior to nanopore sequencing. The 48 µL of amplified product was transferred into a fresh 1.5 mL Eppendorf tube, and 90 µL of AMPure XP beads was added, mixed and incubated for 5 min. The sample was centrifuged and placed on a magnet until clear. The supernatant was removed, 70 % (v/v) ethanol was used to wash the sample, and then, without disturbing the beads, the ethanol was discarded and the pellet was air-dried for 30 s. The sample was removed from the magnet and resuspended in 100 µL of NFW, incubated for 3 min (at room temperature), and then allowed to pellet on a magnet. The eluate was removed and retained in a new 1.5 mL Eppendorf tube. DNA quantitation by Qubit fluorometry was performed (Section 2.7), with final DNA concentrations of 166 ng/µL and 216 ng/µL for VH and U533, respectively.

T7 endonuclease digestion was performed by incubating the recommended input DNA (1.5 µg) with 3 µL of NEB buffer 2, 1.5 µL of T7 Endonuclease I, and NFW added to a final volume of 30 µL at 37 ºC for 10 min. Custom buffer was prepared according to the Premium WGA protocol (ONT; SQK-LSK109, Version 14 Aug 2019), mixing 20 µL of 1 M Tris-HCl (pH 8.5), 4 µL of 0.5 M EDTA, 640 µL of 5 M NaCl and 440 µL of PEG 8000 into 888 µL of NFW. AMPure XP beads (2 mL) were cleaned in NFW and resuspended in 200 µL of this custom buffer, before completing transfer into the remaining custom buffer to make the custom bead suspension.

1 M Tris-HCl (pH 8.5) was added to the T7-digested amplified DNA sample to a final volume of 50 µL. Custom bead suspension (35 µL) was added to this and mixed by flicking the tube, which was then incubated for 20 min at room temperature. The beads were washed twice with freshly prepared 70 % (v/v)

ethanol as described above. The sample was centrifuged, air-dried for 30s, resuspended in 49 µL of NFW, and incubated for 1 min at 50 ºC and then for 5 min at room temperature. The eluate was retained in a fresh 1.5 mL Eppendorf tube after a final bead wash. The final DNA concentrations were measured by Qubit fluorometry to be 28 ng/µL and 12.1 ng/µL from VH and U533, respectively. Therefore, the former sample was chosen for nanopore sequencing.

5.3.2 Sequencing

A summary of the clinical samples which were sequenced using both the Illumina and ONT pipelines are illustrated in **Figure** 5-1. The successfully enriched sample (sample 5: VH) and the two samples with sufficient DNA concentration and the highest HCMV load to human DNA ratios from the residual diagnostic extracts (samples 1: urine and 2: lung) were chosen for direct nanopore sequencing and for reference, Illumina sequencing (**Table** 5-1).

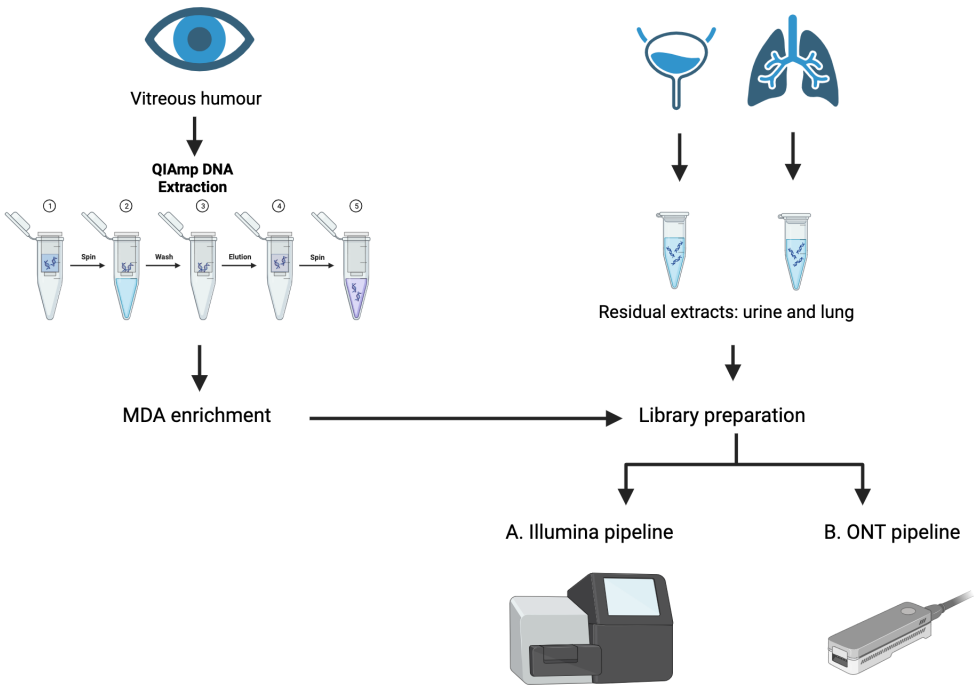


Figure 5-1. Workflow of the three clinical samples from which HCMV was sequenced. The vitreous humour sample was extracted manually (QIAmp DNA extraction kit), prior to MDA enrichment and library preparation for sequencing. Library preparation for sequencing was performed directly from the residual extracts of the other two clinical samples. The two sequencing pipelines, A. Illumina and B. ONT, are further detailed in Figure 5-2.

5.3.2.1 Illumina sequencing

Input DNA amounts of 0.29, 0.58 and 1.3 µg from the non-enriched urine and lung samples and the MDA-enriched VH sample, respectively, were sheared to an average size of 373 bp using a Covaris LE220 sonicator. The pipelines used to generate the sequence data from the samples and assemble the HCMV genomes are summarised in (**Figure** 5-2)**.** HCMV DNA enrichment using SureSelectXT custom biotinylated RNA baits (Agilent) and sequencing library preparation (KAPA Biosystems) were carried out as described (**Figure** 5-2, step A1; Section 2.12). Libraries were indexed using ultrapure (TruGrade) oligonucleotides and sequenced on an Illumina MiSeq sequencer, generating 4,652,420, 5,699,896 and 18,158,230 paired-ended reads of 150 nt for the urine, lung and VH samples respectively (**Figure** 5-2, steps A2 and A3; summary statistics are shown in **Table** 5-4).

5.3.2.2 Nanopore sequencing

The urine extract had viral and human DNA loads of 258,199 IU/µL (1 IU is approximately equivalent to 1 genome copy) and 323 copies/µL, respectively. The lung extract had viral and human DNA loads of 172,259 IU/µl and 9,636 copies/µl, respectively. The urine sample was therefore approximately 40 times more enriched in viral DNA than the lung sample (**Table** 5-1). These two samples and the MDA-enriched VH sample were prepared for nanopore sequencing. (**Figure** 5-2; Section 2.11).

The input DNA amounts for processing into the sequencing libraries were 1.5, 0.4 and 0.8 µg from the VH, urine and lung samples, respectively (**Figure** 5-2, step B1). Briefly, and as detailed in Section 2.11, this involved treatment with NEBNext FFPE DNA repair mix and an NEBNext Ultra II end repair/dA-tailing module purification using AMPure XP beads, ligation of adaptors using an NEBNext Quick ligation module, enrichment of large (>3 kb) fragments using long fragment buffer, and final purification in elution buffer. The libraries were loaded onto primed R9.4.1 flow cells at concentrations of 8.45, 37.4 and 17.7 ng/µL for the VH, urine and lung samples, respectively. Sequencing was carried out for 48 h on an GridION instrument (**Figure** 5-2, step B2). The reads were acquired using MinKNOW v.4.3.11 (https://nanoporetech.com), and FAST5-

formatted files were base-called using Guppy v.4.0 (https://nanoporetech.com) in high accuracy mode with a minimum quality score of 7. The reads were assessed using MinIONQC v.1.3.5 (Lanfear et al., 2019), and adaptors were trimmed and chimeric reads with internal adaptors removed using PoreChop v.0.2.3 (https://github.com/rrwick/Porechop) (step B3).

The non-enriched urine and lung samples generated 10,616,519 and 8,731,192 reads, respectively, whereas the MDA-enriched VH sample generated 1,900,980 reads. These were mapped to a collated set of 265 HCMV genomes representing all HCMV strains that had been sequenced at the time of analysis (April 2021; Appendix A2. List of published 265 HCMV genomes, this list originated from A. Davison, (Camiolo et al., 2022)) using Minimap2 v.2.17 (Li, 2018), and mapped reads were extracted using Samtools v.1.9 (Li et al., 2009). The 36,616 and 11,434 HCMV-enriched reads obtained from the urine and lung samples and the 47,364 HCMV-enriched reads from the VH sample formed the input nanopore data for analysis (**Figure** 5-2, step B4; summary statistics are shown in **Table** 5-5). The longest reads were obtained from the urine sample. For the urine and lung samples, maximum read lengths were 169,079 and 92,239 nt, respectively, and mean read lengths were 1,431 and 735 nt. The longest read from the VH sample was only 87,726 nt, and the mean read length was 3,795 nt.

Table 5-4. Summary statistics from Illumina sequencing runs involving clinical samples.

Statistics	Urine	Lung	VH
Original reads	4,652,420	5,699,896	18,158,230
Reads after host reads removal; no. (%)	4,629,630 (99.5)	3,808,640 (66.8)	16,194,628 (89.0)
Reads after adaptor trimming; no. (%)	4,623,352 (99.3)	3,802,260 (66.7)	16,176,568 (89.0)
Reads after deduplication; no. (%)	2,853,792 (61.3)	3,585,394 (62.9)	7,666,798 (42.0)
Merlin coverage for original reads	2,406.51	2265.39	9,401.10
Original reads mapping to Merlin; no. (%)	3,729,531 (80.1)	3,467,060 (60.8)	14,661,087 (80.7)
Bases covered by original reads; no. (%)	231,919 (98.0)	228,337 (96.0)	233,082 (98)
Merlin coverage for deduplicated reads	1364.36	2132.95	4,458.21
Deduplicated reads mapping to Merlin; no. (%)	2,118,644 (45.5)	3,265,267 (57.2)	6,975,103 (38.4)
Bases covered by deduplicated reads; no. (%)	231,917 (98.0)	228,337 (96)	233,079 (98)

Table 5-5. Summary statistics from nanopore sequencing runs involving clinical samples.

	Non-enriched		Enriched
Run no.	**1**	**2**	**3**
Clinical sample	Urine	Lung	VH
Total run time (h)	48	48	48
Total gigabases	6.52	15.24	7.21
Total reads (no.)	8,875,801	10,645,687	1,900,980
N50 length (bp)	915	3,550	5,658
Mean length (bp)	735	1,431	3,796
Median length (bp)	534	596	2,582
Max length (bp)	92,239	169,079	87,726
Mean q-score	11.40	12.40	11.3
Median q-score	11.50	12.70	11.5
Read length (no.)			
>10 kb	11,014	144,877	120,107
>20 kb	1,981	21,587	17,113
>50 kb	82	243	326
>100 kb	0	2	0
>200 kb	0	0	0

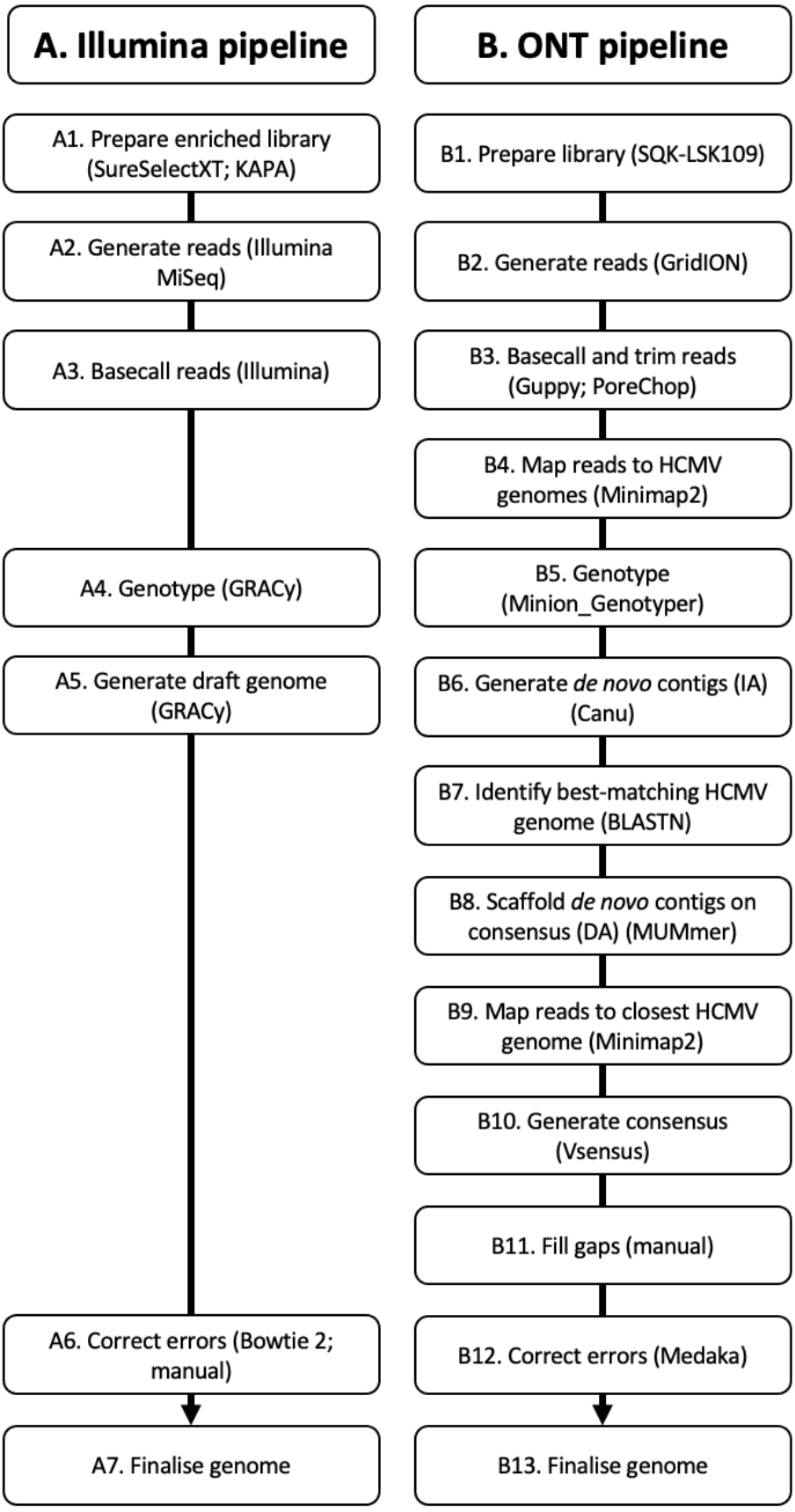


Figure 5-2. Sequencing and bioinformatics pipelines.
The steps in (A) the Illumina pipeline (A1–A7) and (B) the ONT pipeline (B1–B13) are described in the text.

5.3.3 Bioinformatic analysis

5.3.3.1 Genotyping using Illumina and nanopore data

The genotypes of 13 hypervariable HCMV genes were determined from the input Illumina data using the genotyping module in the HCMV genome assembly program, GRACy v0.4.4 (Camiolo et al., 2021) (**Figure** 5-2, step A4; Section 3.4) and from the input nanopore data using Minion_Genotyper v1.0 (https://github.com/salvocamiolo/minion_Genotyper/) (**Figure** 5-2, step B5; Section 4.3.7.1).

5.3.3.2 Genome determination using Illumina data

The Illumina input data were assembled using GRACy (**Figure** 5-2, step A5). The GRACy draft genomes were assessed by trimming and quality-filtering the reads using Trim Galore v.0.4.0 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) under parameters *--illumina --paired*, aligning the filtered reads to the sequences using Bowtie 2 v.2.3.1 (Langmead and Salzberg, 2012) and Samtools v.1.3 (Li et al., 2009), inspecting the alignments visually using Tablet v.1.21.02.08 (Milne et al., 2016), and making any necessary amendments (**Figure** 5-2, step A6; Section 3.4.2). Coverage statistics were generated using Tablet. These steps resulted in properly configured, complete HCMV genomes (referred to from here on as the final genomes) (**Figure** 5-2, step A7).

5.3.3.3 Genome determination using nanopore data

For the non-enriched urine and lung samples, draft genome sequences were determined using parallel approaches. In the first approach (**Figure** 5-2, steps B6-B8), the nanopore input data were assembled *de novo* using Canu v.1.9 (Koren et al., 2017) under the parameters *genomeSize=240000 minReadLength=500 minOverlapLength=50 -Nanopore-raw* (**Figure** 5-2, step B6; Section 4.3.6.2). For each sample, the best-matching genome in the collated set of 265 genomes that matched the longest contig was identified using BLASTN v.2.4.0 (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download) (**Figure** 5-2, step B7). The GenBank accession numbers of these genomes

were KY90081.1 and KP75641.1 for the urine and lung genomes, respectively. The contigs were compared with the best-matched genome using MUMmer v.3.23 and Mummerplot v.3.5 (Marçais et al., 2018) (http://mummer.sourceforge.net/manual/#mummerplot) and rearranged manually via inspection of an alignment in AliView v.1.27 (Larsson, 2014) (**Figure** 5-1, step B8). To accomplish this, a text editor was used to order and divide the contigs to ensure that the genome termini were at the ends of the sequences, locate any gaps, and represent the unique regions U_L and U_S and the inverted repeats *ab*/*b'a'* and *a'c'*/*ac* appropriately (unlike the situation with the Illumina genomes, the inverted repeats were not assumed to be identical).

In the second approach (**Figure** 5-2, steps B9-B10), the best-matching genome found as above (from step B7) was used as the reference for mapping the input data using Minimap2 (step B9). BAM and mpileup files were generated from the alignment using Samtools v.1.9 (Li et al., 2009), and the mpileup files were parsed to output a consensus genome using VSensus (https://github.com/rjorton/VSensus) with the minimum coverage and minimum base quality parameters set at 20 and 0, respectively (**Figure** 5-2, step B10). Coverage statistics were generated using weeSAM v.1.5 (https://github.com/centre-for-virus-research/weeSAM/blob/master/weeSAM). The genomes derived by the two approaches were then aligned, using MAFFT v.7.310 (Katoh et al., 2002), and any gaps in the former were filled with data from the latter (**Figure** 5-2, step B11). Finally, this final genome was corrected using Medaka v.1.3.3 (**Figure** 5-2, step B12 and Section 4.4.2). These steps resulted in properly configured, complete HCMV genomes that did not depend at any stage on the corresponding Illumina data (the final genome; **Figure** 5-2, step B13). With the MDA-enriched VH sample, an extra step to remove palindromic reads was performed using Pacasus (Warris et al., 2018), a tool for detecting and cleaning nanopore long reads after WGA, prior to inputting reads for assembly as described above for the urine and lung nanopore datasets.

5.3.3.4 Analysis of resistance mutations

To test the possible clinical utility of nanopore data from these samples, the final polished genome sequence was checked for common resistance-associated mutations. The coding sequences of genes UL97 and UL54 were analysed using a

web-based mutation resistance analyzer (MRA, accessed June 2022; https://www.informatik.uni-ulm.de/ni/mitarbeiter/HKestler/mra/app/index.php?plugin=updatepol) (Chevillotte et al., 2010).

5.4 Results

5.4.1 Trial of long-range tiled PCR amplicons

The results obtained using the test primer pairs are summarised in **Table** 5-6. The primers were used initially to amplify high-titre cultured HCMV extracts (Merlin, AF1 and U11) and if successful were used on residual clinical extracts (cultured strains in **Figures** 5-2 and 5-3 and clinical extracts in **Figure** 5-4). However, the required input volume of clinical extract for PCR exceeded its availability, requiring at least 10 µL per reaction per amplicon, leaving insufficient material eventually to amplify the entire genome using tiled amplicons (the minimum volume of extract required would be 240 µL, and likely more given that not all amplicons would span 10 kb). Given the precious and limited nature of the residual clinical extracts (the maximum volume obtained per sample was 50 µL), this outcome meant that this enrichment method had to be abandoned. Additionally, although products of the expected sizes were amplified from some clinical samples, multiple bands were detected in some cases. This could have been secondary to non-specific primer amplification of human genomic or bacterial DNA (**Table** 5-6 and **Figure** 5-4). Blood-derived samples were also found to be suboptimal for long fragment amplification; although PCR successfully amplified smaller HCMV targets (approximately 100 bp), no fragments longer than this were amplified (**Figures** 5-5 and 5-6). This may be in keeping with the previous finding that HCMV genomes are highly fragmented in blood (Boom et al., 2002).

Table 5-6. Summary of long-range PCR amplification using forward primers with paired reverse primers.

Forward primer	Reverse Primer	Expected product length (nt)	Cultured HCMV samples[a]	High HCMV viral load clinical samples[b]
F1	R1	1,332	++	NA
	R2	2,400	++	+
	R3	3,240	++	NA
	R4	4,075	++	+
	R5	4,844	++	NA
	R6	6,195	++	NA
	R7	6,808	++	+/–
	R10	7,369	++	NA
	R11	8,433	++	NA
	R8	9,084	++	NA
	R9	10,218	++	–
F2	R12	1,375	++	NA
	R1	3,104	++	NA
	R2	4,123	++	NA
	R3	5,012	++	Multiple bands
	R4	5,797	++	NA
	R5	6,566	++	Multiple bands
	R6	7,917	Multiple bands	NA
	R7	8,530	++	NA
	R10	9,091	++	Multiple bands
	R11	10,155	+	NA
	R8	10,806	+	–
	R9	11,940	++	–

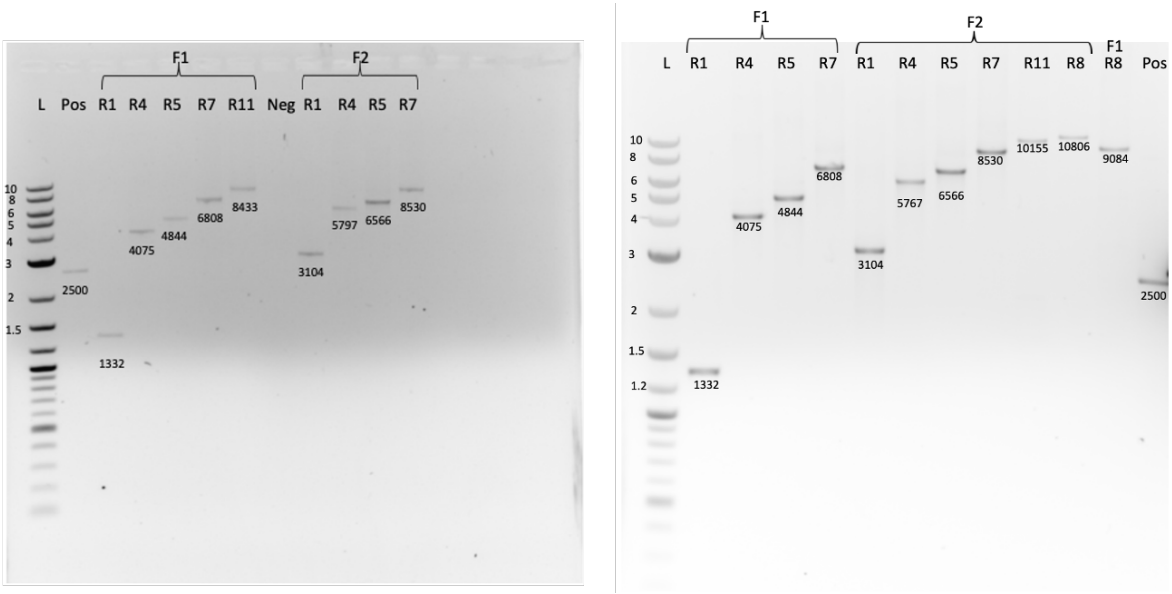


Figure 5-3. Agarose gel images showing amplified PCR products using long-range primers on high-titre cultured HCMV Merlin DNA.
Forward primers F1 and F2 were paired with reverse primers (prefixed R) as grouped by brackets above the corresponding lanes. The PCR product was diluted 1:10 before running on the agarose gel. The expected product size (nt) is shown below each band. L, ladder (sizes in kb); Neg, NFW negative control.

[a] The presence and strength of the band for the PCR product is denoted by "+".

[b] The absence of a band for the expected PCR product is denoted by "–"; NA, not available.

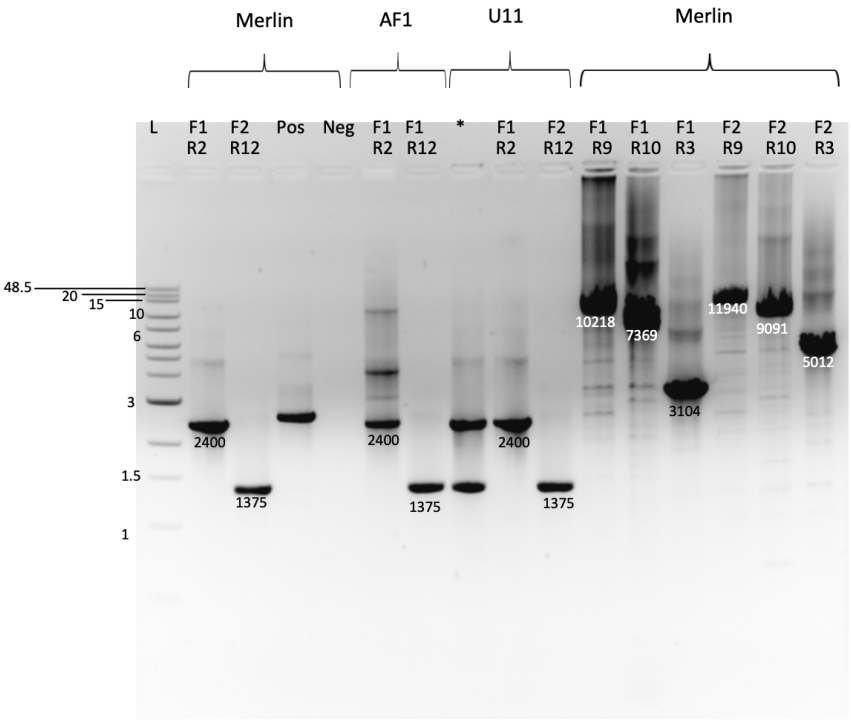


Figure 5-4. Agarose gel image showing amplified PCR products using long-range primers on high-titre cultured HCMV Merlin, AF1 and U11 DNA.
Forward primers F1 and F2 were paired with reverse primers (prefixed R). The expected product size (nt) is shown below each band. *Both amplified products from primer set F1/R2 and F2/R12 were added to this well in error, hence the presence of two bands. The PCR product was diluted 1:10 before running on the agarose gel. L, ladder (sizes in kb); Neg, NFW negative control.

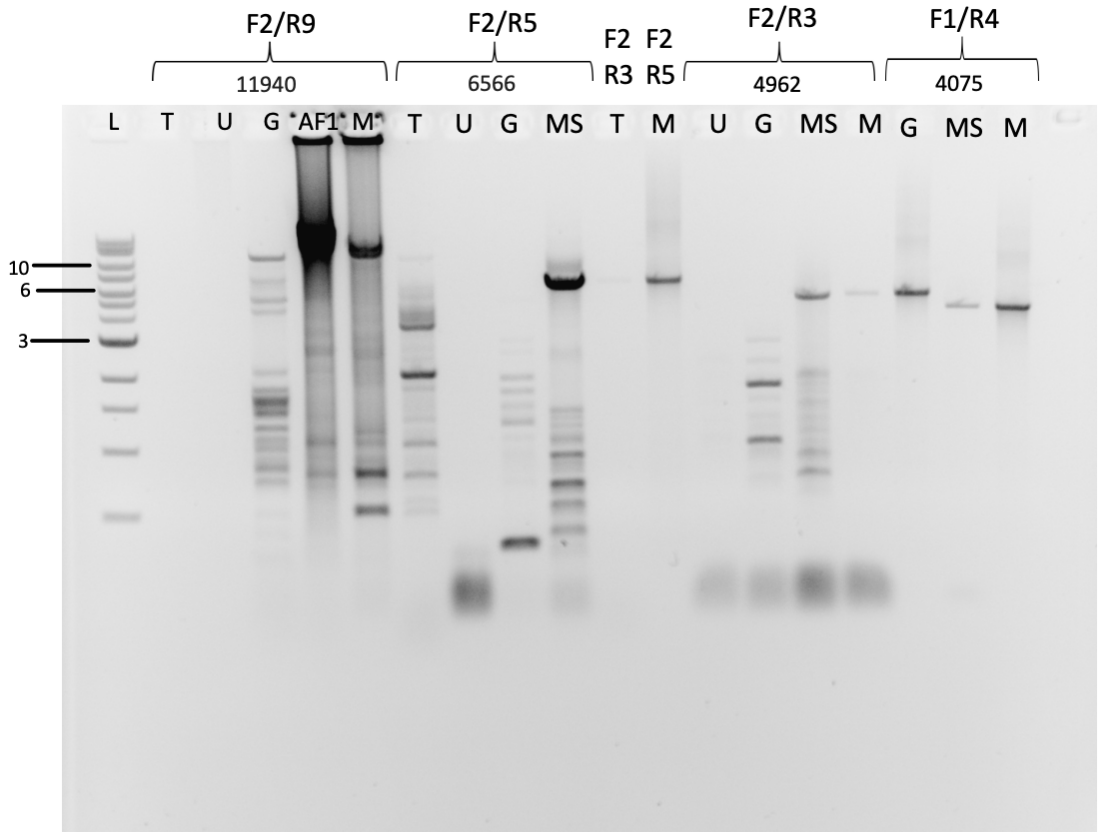


Figure 5-5. Agarose gel image showing amplification of PCR products using long-range primers on clinical sample DNA.
Forward primers F1 and F2 were paired with reverse primers (prefixed R). The expected product size (nt) is shown in brackets. Multiple bands can be seen in several samples. Sample types: T, throat swab; U, urine; G, gastrointestinal biopsy; MS, mouth swab. DNA from high-titre cultured AF1 and Merlin (M) were diluted 1:10 prior to loading as positive controls. The primer pairs used are stated above the brackets. L, ladder (sizes in kb).

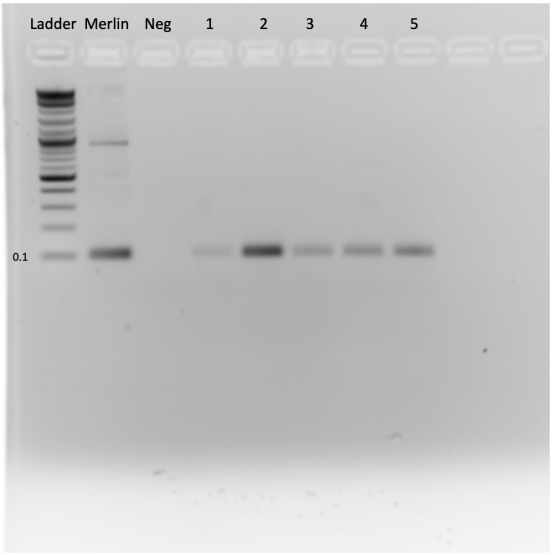


Figure 5-6. Agarose gel image showing amplification of small amplicons from clinical sample DNA.
Sample types: 1, 2 and 5, plasma; 3, mouth swab; 4, urine. The primers amplified part of the RNA1.2 gene and were as follows: RNA1.2_FWD (CCCATATAGAGAAATAATGATAGTTTGACAAC) and RNA1.2_REV (AATTCAATTGTTGAAAGTCTCTCCCT). The 0.1 kb marker in the ladder is indicated.

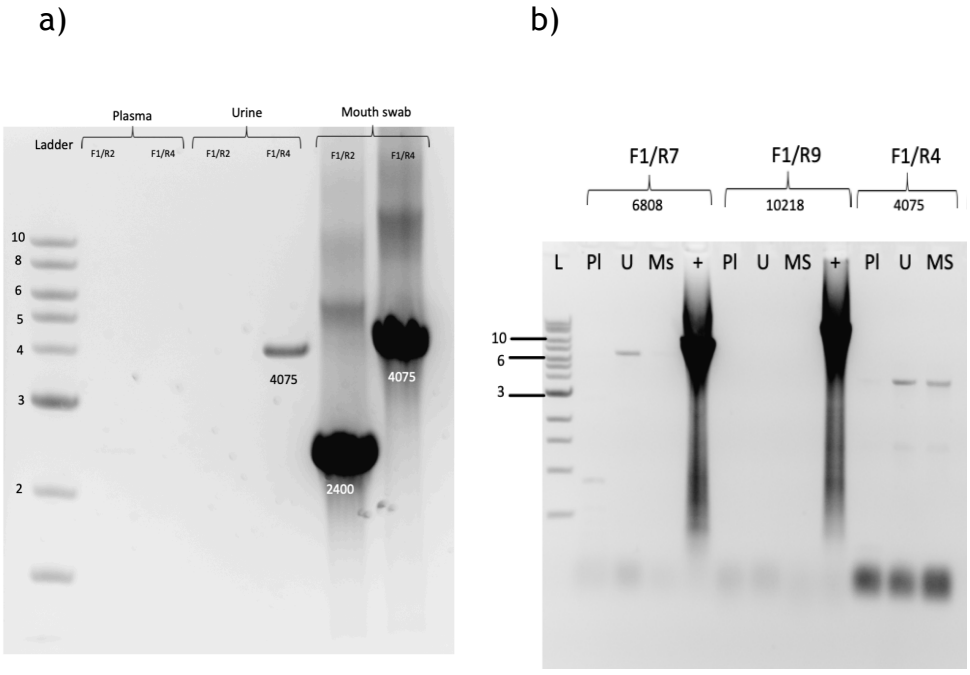


Figure 5-7. Agarose gel images showing the inability of long-range primers to amplify HCMV fragments from blood.
a) Gel image of PCR products obtained using primer pairs F1/R2 and F1/R4 on clinical samples. Sizes in the ladder are kb.b) Gel image of PCR products obtained using primer pairs F1/R7, F1/R9 and F1/R4 on plasma (Pl), urine (U) and mouth swab (MS) samples, with Merlin DNA as the positive control (+).The expected product sizes (nt) are shown below the brackets. L, ladder (sizes in kb).

5.4.2 WGA using MDA

The VH sample, with its low overall DNA concentration but high HCMV to human DNA ratio, was chosen for WGA using MDA (Section 5.3.1.2), prior to nanopore sequencing. The longest HCMV read in the dataset (78,404 nt) yielded a top match using BLASTn to strain SYD-SCT1 (GenBank accession MT044485.1). When the longest read was aligned to this reference sequence using Minimap2, it resulted in the read being split up into over 300 supplementary or additional alignments due to highly repeated sequences within the read. The resulting coverage plot showed that this single read consisted of >300 repeats of the same region (approximately 20,255 – 20,375 nt in the reference genome) (**Figure** 5-7). There also appeared to be a lower number (approximately 30 copies) of a slightly longer repeat from 20,185-20,255 nt. Palindromic reads were therefore removed using Pacasus (Warris et al., 2018), and the remaining reads were mapped to the 265 published HCMV genomes and extracted. Minimap2 was used to align these HCMV-mapped reads to the VH genome assembled by GRACy using reads from the Illumina dataset. The maximum depth of coverage was >277,000 reads/nt (at approximately 75,000 nt in the reference genome), and simultaneously only 87 % of the genome was covered by reads (**Figure** 5-8). This suggested that MDA had produced biased amplification of certain regions of the genome and that Pacasus was unable to resolve some of the chimaeric reads created by MDA. Regardless, the extracted reads were subjected to *de novo* assembly using Canu, limiting the *readSamplingCoverage* parameter to 200x. A total of 41 contigs between 1,086 to 27,814 nt in length were assembled. However, the assembled contigs mapped repetitively to the same short region of the HCMV genome when aligned using BLASTn (**Figure** 5-8). Again, this was an indication that MDA had amplified the HCMV genome through erroneous elongation of newly amplified DNA fragments to produce long DNA fragments of short repeated sequences (Section 5.5). MDA was unsuccessful in equally amplifying HCMV from this clinical sample to reconstruct a complete genome sequence and was not used further in this study.

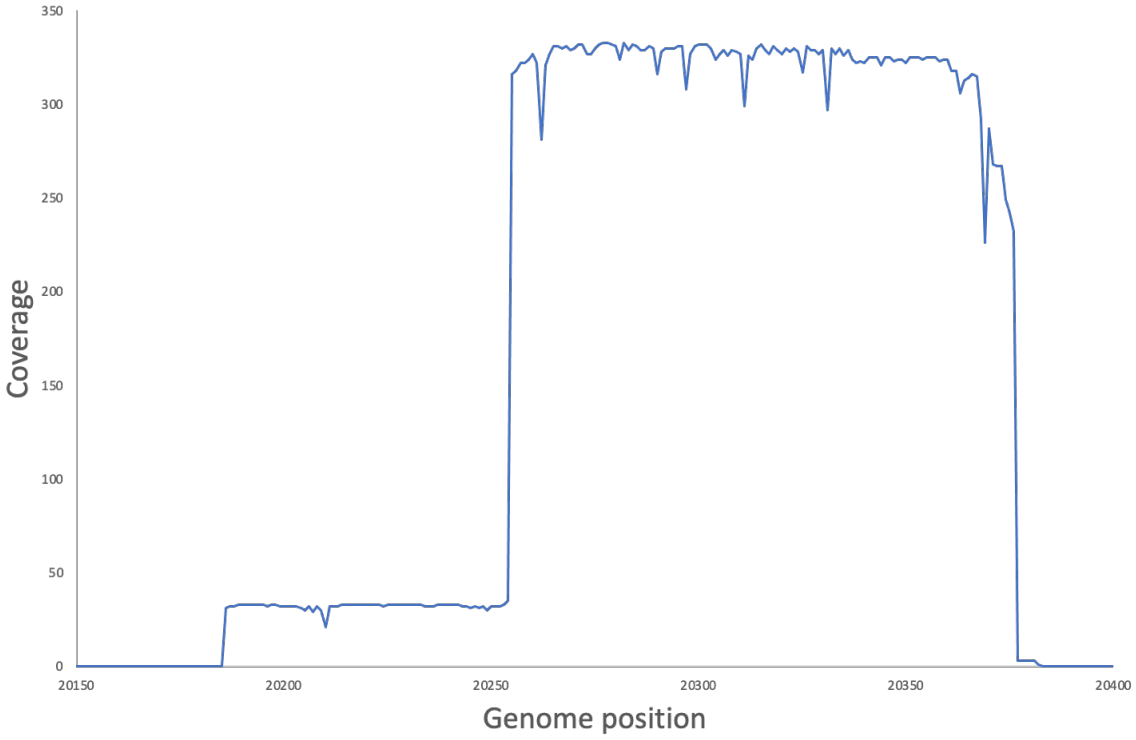


Figure 5-8. Coverage plot of the longest nanopore read obtained from the VH sample amplified by MDA.
The alignment of the longest read (78,404 nt) against the strain SYD-SCT1 genome covered only a single region of approximately 190 nt (20,185-20,375 nt) repetitively.

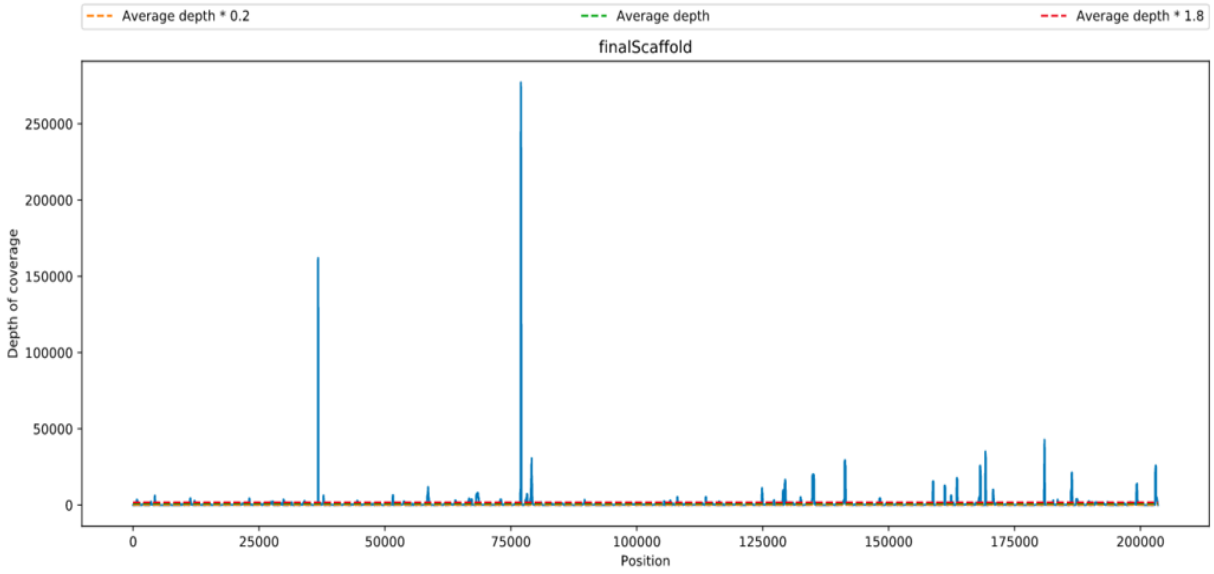


Figure 5-9. Coverage plot of all Pacasus-processed, HCMV-mapped reads aligned to the Illumina-assembled reference.
Depth of coverage across the whole genome shows that some regions had biased coverage compared to other regions that had little or no coverage (https://github.com/centre-for-virus-research/weeSAM).

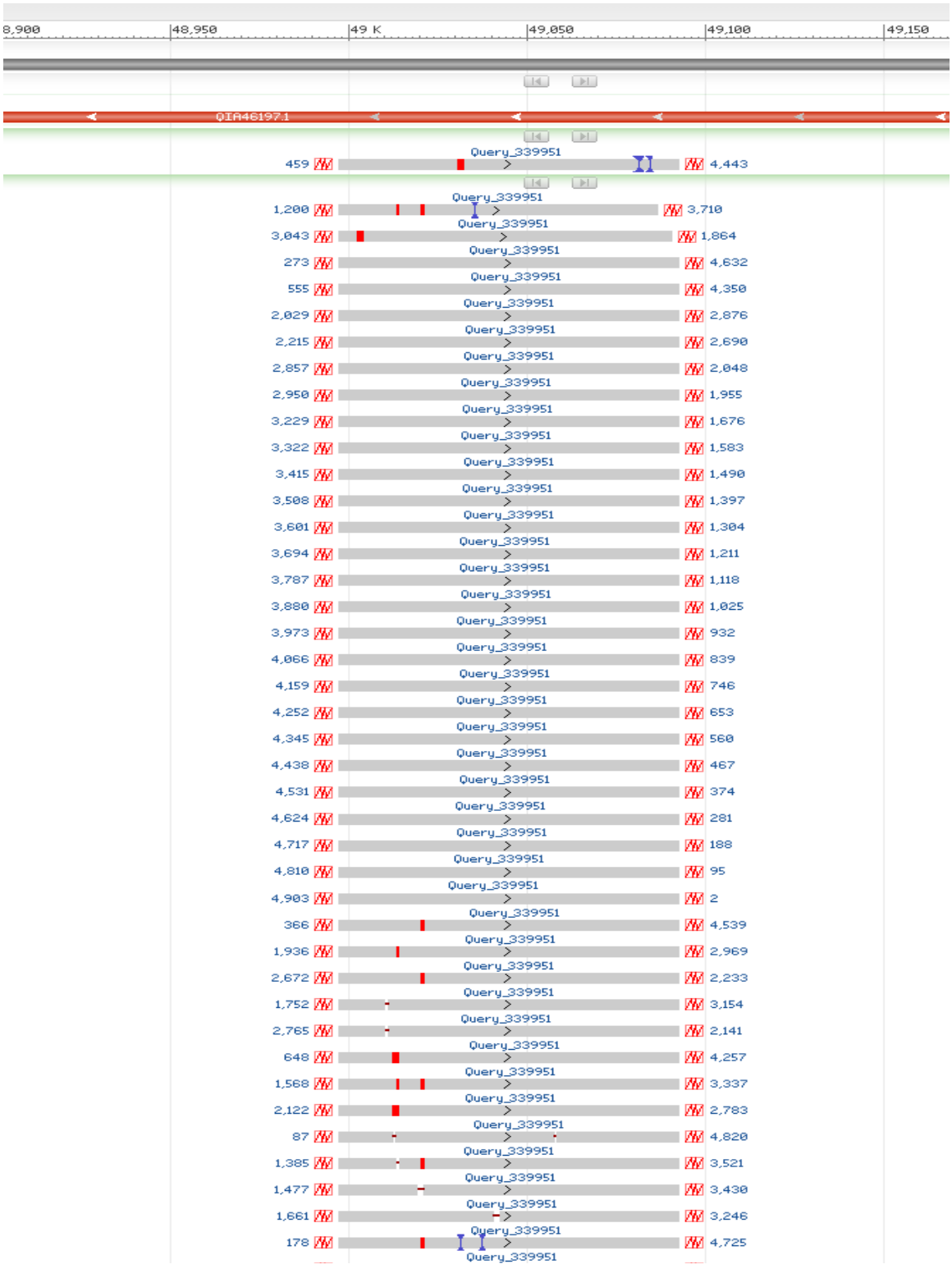


Figure 5-10. Graphical view of the alignment of a *de novo*-assembled contig against the strain SYD-SCT1 genome.
The entire 27,814 nt contig aligned repetitively to the same region (49,000-49,990 nt, scale displayed in the top panel).

5.4.3 Genotyping results

Initial screening of the Illumina data to determine the number of HCMV strains present in the original samples was carried out using GRACy (Section 3.4), and that of the nanopore data was carried out using Minion_Genotyper (Section 4.3.7.1). Neither tool requires a genome but rather analyses the reads using a kmer-based approach, and each identified the same genotypes for the 13 hypervariable genes analysed (**Table** 5-7). This again confirmed the ability of minion_Genotyper to genotype HCMV strains accurately from nanopore data and indicated that each sample contained a single HCMV strain. The threshold length established for reads from the high-titre cultured HCMV samples (Section 4.4.5) did not work with for these clinical samples because there were insufficient reads of >10,000 nt, let alone >25,000 nt. Therefore, all reads, regardless of length, were required in order for minion_Genotyper to extract genotype data from these clinical datasets. Closer inspection of the genotyping results from the VH dataset showed that most reads contained only one of the hypervariable genes, another indication that MDA may not have amplified complete HCMV genomes. As found for the genotypic combinations of HCMV strains detected in clinical FFPE samples (Chapter 3), the genotypic combinations of the urine and lung samples genomes were unique in comparison with those reported previously.

Table 5-7. Genotypes of 13 hypervariable HCMV genes in the urine and lung samples.

Gene	Urine	Lung
RL5A	G1	G2
RL6	G6	G4
RL12	G7	G1B
RL13	G7	G1
UL1	G7	G1
UL9	G1	G4
UL11	G1	G1
UL20	G6	G6
UL73	G4D	G1
UL74	G5	G1A
UL120	G1A	G4B
UL146	G2	G10
UL139	G1A	G2

5.4.4 Genome determination using Illumina data

Determination of the final Illumina genomes was accomplished using the GRACy draft genomes, which were curated manually (Section 3.3.4) and ended with 86 % of reads aligning to the urine genome (234,678 nt) at an average coverage depth of 2,554 reads/nt and 66 % of reads aligning to the lung genome (235,215 nt) at an average coverage depth of 2,368 reads/nt. The genomes for the lung and urine samples were deposited in GenBank (accessions OQ466311 and OQ466312, respectively).

5.4.5 Genome determination using nanopore data

The final nanopore genomes were determined by a more complex series of steps involving parallel reference-independent, *de novo* assembly (RIA) of the input data and alignment of the input data to the best-matching published genome for reference-dependent assembly (RDA) (Section 4.3.6). RIA using Canu resulted in a single contig (255,416 nt) corresponding to the entire HCMV genome and a partial repeat for the urine sample but in 20 contigs for the lung sample, with the four largest originating from the HCMV genome. The three largest of these contigs (37,585, 40,716 and 156,323 nt) represented the whole genome except for one gap of 1,849 nt. The subsequent steps led to a complete nanopore genome for both samples, each with complete read coverage and determined independently of the corresponding Illumina genome. The longest RIA contig was compared to the database of published HCMV genomes using BLASTn, and the top matched HCMV strain was then used for RDA. For the urine sample, this strain was HANSCTR2 (GenBank accession KY490081.1), and for the lung sample it was BE/31/2011 (GenBank accession KP75641.1). The orientation and length of the RIA contigs was amended using the RDA genome as a scaffold; the multiple sequence alignment program, MAFFT, was used to align the RIA consensus to the RDA consensus and the alignment was visualised in AliView. Where a gap was present in the RIA contig, the corresponding sequence from the RDA consensus was used to fill the gap and complete the genome.

The consensus polishing tool, Medaka, was used on the final assembled genomes. A custom script, AlignMuts.sh (https://github.com/rjorton) using samtools to generate an mpileup file and Vsensus (https://github.com/rjorton/VSensus)

were used to output a text file with all differences (insertions, deletions and substitutions) compared with the Illumina-generated genome as reference (**Figure** 5-2, A7). This analysis ended with 39,122 reads (99.8 %) aligning to the urine genome at an average coverage depth of 203 reads/nt, and 12,488 reads (99.4 %) aligning to the lung genome at an average coverage depth of 26 reads/nt.

Comparisons between the Illumina and Nanopore genomes revealed differences at 62 nt (0.03 %) and 164 nt (0.07 %) for the urine and lung samples, respectively. The great majority of differences (97 and 90 %, respectively) were due to insertions or deletions (indels) associated with homopolymeric tracts. The Illumina version at the position of each difference was validated by visual inspection of the read alignment using Tablet. Moreover, comparisons with the corresponding regions in the collated set of 265 genomes supported the Illumina version much more frequently than the nanopore version (95 and 96 %, respectively).

5.4.6 Identification of clinically significant resistance-associated mutations

In clinical practice, the sequences of genes UL54 and UL97 are typically examined for resistance mutations when patients fail to respond to antiviral therapy. As a test of the utility of nanopore genomes for such investigations, the coding sequences of these genes were submitted to the web-based tool MRA, which hosts a curated database of published resistance mutations from a range of viruses including HCMV (**Figure** 5-10). The UL97 sequences of the Illumina and nanopore genomes from the urine sample were identical, whereas the UL54 sequence contained an erroneous deletion of 1 nt in the nanopore genome from the lung sample. MRA successfully registered the frameshift caused by this deletion (**Figure** 5-10B). Differences associated with strain polymorphisms or mutations not associated with resistance were registered in both genes in both samples, but no mutations associated with drug resistance were found in either gene in either sample (**Figure** 5-10). As the clinical samples were fully anonymised, no information was available on whether the patients had received antiviral therapy.

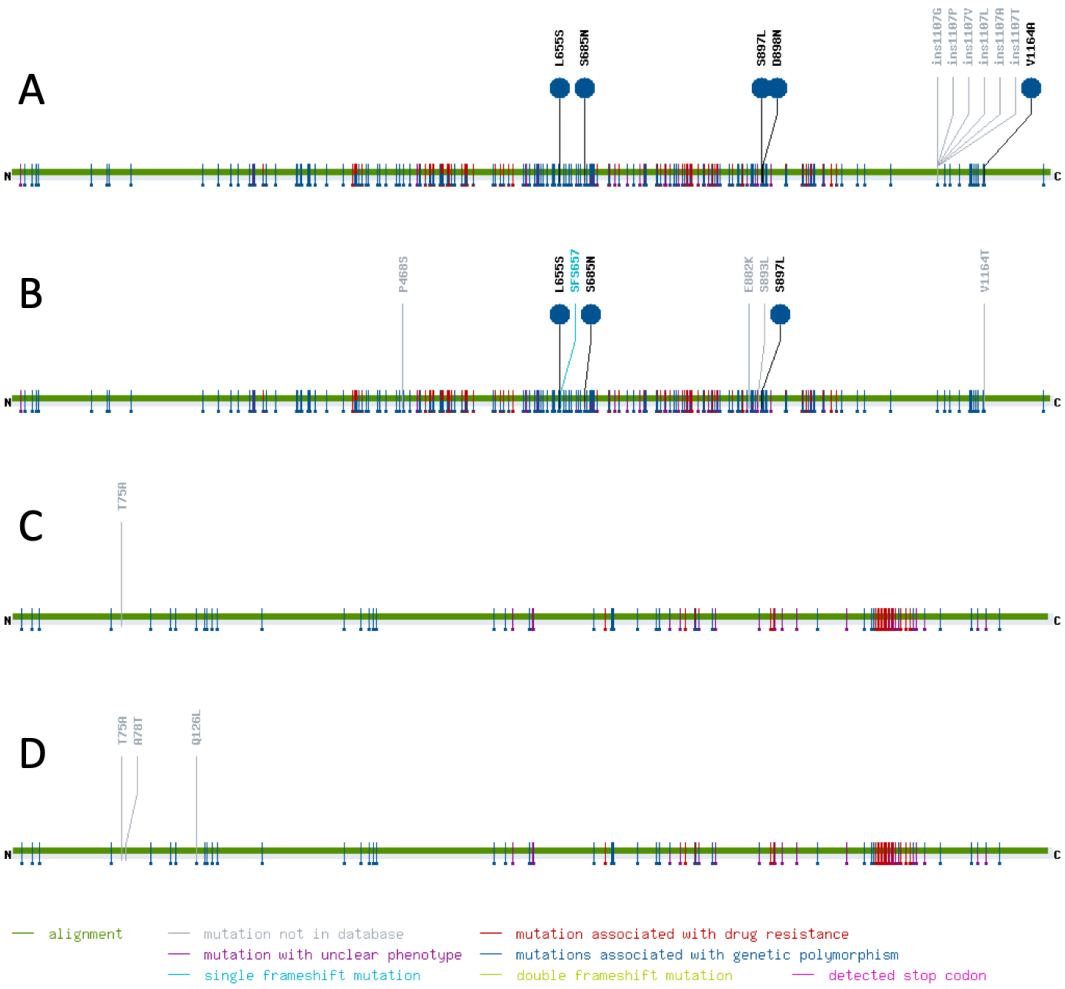


Figure 5-11. Graphical output from MRA for nanopore sequences.
(A) urine genome UL54, (B) lung genome UL54, (C) urine genome UL97 and (D) lung genome UL97. A sequence alignment is represented for each gene from the 5'-end encoding the N terminus of the protein (N) to the 3'-end encoding the C terminus (C). The features of the alignment and the sequence are shown by the flags on and above the alignment and are coloured according to the key at the foot. The coordinates relate to encoded amino acid residues.

5.5 Discussion

Having demonstrated the capacity of nanopore sequencing to produce complete HCMV genomes from high-titre cultured samples, I accepted the challenge of obtaining complete genomes directly from clinical samples. Initially, I performed an enrichment step using two molecular approaches based on WGA. A test of WGA using tiled PCR amplicons proved impractical due to the volume of sample required to analyse the whole genome, and WGA using MDA led to highly biased replication of the HCMV genome and obfuscated genome reconstruction.

Although the test of WGA using tiled PCR amplicons indicated that application of this approach across the whole genome was impractical, preliminary results using primers designed to amplify the conserved UL54-UL57 region had some promising results, amplifying the expected PCR product in mouth swab and gastrointestinal biopsy samples up to a size of 9,091 bp. However, multiple-band detection on agarose gel electrophoresis of PCR products from these clinical samples was seen commonly. The general absence of multiple bands in PCR products from high-titre cultured HCMV extracts suggested that this phenomenon was due to the lower levels of HCMV DNA and to non-specific amplification of human or bacterial DNA contaminants in the clinical samples.

Although the WGA using MDA experiments yielded evidence of highly biased amplification, this approach had been successful previously in enriching HCMV DNA from high-titre cultured samples that were sequenced by short-read HTS (Sijmons et al., 2014). Amplification bias has been documented to be more problematic when the starting material has miniscule amounts of DNA, as is the case in single-cell sequencing (Dean et al., 2002, Nurk et al., 2013). A high sequencing depth can overcome this bias in some cases, although chimaera formation and non-uniform amplification of linear genomes is well documented (Dean et al., 2002, Yan et al., 2004, Lasken and Stockwell, 2007, Nurk et al., 2013). Indeed, analysis of the reads from the nanopore sequencing dataset from the MDA-enriched VH sample showed evidence of abundant chimaeric rearrangements. MDA relies on multiple priming by random hexamers on each template strand and phi29 polymerase extending the 3'-termini of the primers and displacing downstream extending primers (Lasken and Stockwell, 2007) (**Figure** 5-11). This results in a branched DNA molecule with numerous single-

stranded 5'-ends, which should be converted to dsDNA. However, in some cases, branch migration displaces the 3'-ends and chimaeras are generated (**Figure** 5-12), as was the case with the VH sample (**Figure** 5-7). It would be of interest to investigate further the performance of MDA on samples with higher starting DNA concentrations, as this technology has been used successfully in single-cell genomics and in environmental sequencing in the field of microbial ecology (Marine et al., 2014).

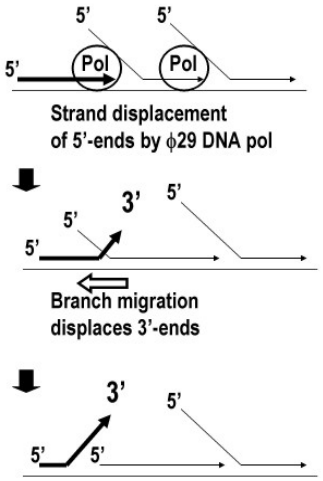


Figure 5-12. Mechanism of 5'-end displacement by phi29 DNA polymerase and 3'-end displacement by branch migration during MDA.
(Reproduced with modification under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0) (Lasken and Stockwell, 2007).)

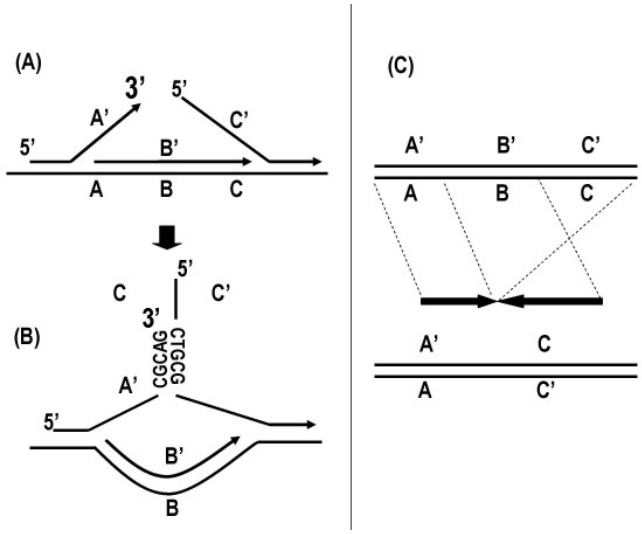


Figure 5-13. Mechanism of chimaera formation with inverted sequences during MDA.
A) Simple branch migration reaction leading to 3'-termini displacement and mispriming events. B) The displaced 3'-terminus is free to reanneal, preferentially at randomly occurring complementary segments on nearby 5'-strands. C) The two sequences in inverted orientation join with the deletion of B, forming the chimaeric sequence A'C. (Reproduced without modification under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0) (Lasken and Stockwell, 2007)

Fortuitously, two of the clinical samples had sufficiently high HCMV loads to enable direct nanopore sequencing. My study is the first to use nanopore technology to determine complete HCMV genomes from clinical samples without involving cell culture, bait-based target enrichment, PCR amplification, amendments based on prior knowledge of the target sequence, or supplementation by data for the same genome acquired on other platforms. Earlier published studies using the nascent ONT MinION protocols showed promising results, obtaining full HCMV genomes from high-titre cultured HCMV samples. However, these protocols required fragmentation of the DNA (by mechanical shearing or the action of transposase), and one included an enrichment step by bait hybridisation, thus also restricting the length of input DNA molecules (Eckert et al., 2016, Karamitros et al., 2018). Moreover, as transposase fragmentation is non-random, it may have led to sequencing biases (Marine et al., 2011). From my work, it is evident that the current ONT hardware (i.e. flow-cells and pores) and software (i.e. basecallers and consensus-correcting software) are able to support the direct sequencing of HCMV from high-viral load clinical samples without the need for enrichment. The urine sample sequenced had only 40 % of the protocol-recommended input DNA amount, demonstrating the feasibility of a lower input DNA amount for successful nanopore sequencing. Additionally, I was able to assess in detail the nanopore errors compared to a “gold-standard” genome obtained on the Illumina platform with bait-based target enrichment. Independent genotyping of the Illumina and nanopore data from two high-titre samples from urine and lung demonstrated the presence of a single HCMV strain in each sample. The nanopore genomes were determined by parallel approaches (RIA and RDA), with the outputs being incorporated together and polished to produce the final genomes. As anticipated from the approximately 40-fold greater ratio of viral to non-viral DNA in the urine sample, the urine genome was assembled more easily than the lung genome and at greater final read depth.

Sequence comparisons suggested that the differences between the Illumina and nanopore genomes were largely or entirely due to errors in the latter. Nonetheless, the nanopore genomes were highly accurate (99.97 and 99.93 % for the urine and lung genomes, respectively), with the errors distributed thinly (on

average, one approximately every 3.8 and 1.5 kb, respectively) and located mostly in polynucleotide tracts in non-coding regions. Thus, although improvements in the ONT hardware and the software for analysing the data have steadily boosted read accuracy for single read base-calling to >99 % (https://nanoporetech.com/accuracy), in this study they did not achieve the level of accuracy of the Illumina platform. For completely accurate characterisation of HCMV strains on the ONT platform, supplementary data from another platform is still required, with the longer nanopore reads providing support for scaffolding and evidence for recombination. The implication is that we are currently unable to rely solely on nanopore sequencing alone to determine the sequence of any novel large DNA virus for which no prior sequence information is available. However, for the purposes of resolution of HCMV genomes on the basis of the 13 hypervariable genes to enable an overview of whether a single-strain or multiple-strain infection is involved, the pipeline developed here may be useful. This situation may change in future as ONT technology continues to improve. Indeed, recent advances made after my study was carried out are claimed to have achieved quality scores in excess of Q30 (https://nanoporetech.com/accuracy, accessed 30 August 2023).

The main limitation of my study was that the two clinical samples, urine and lung, were chosen because they contained a high concentration of DNA and a high ratio of HCMV to human DNA. In typical clinical settings, the mainstay sample type for post-transplant monitoring of HCMV reactivation or re-infection is blood. I encountered the previous observation that HCMV genomes in blood are highly fragmented (Boom et al., 2002) and thus an inappropriate sample type for nanopore sequencing. Blood also contains high levels of host DNA, which can lead to low-efficiency enrichment for HCMV genomes, although use of plasma rather than whole blood can reduce this (Boom et al., 2002, Suárez et al., 2020). In contrast, urine is normally depleted of host cells and was shown in my study to be suitable for sequencing the HCMV genome with high accuracy on the ONT platform. In instances of post-transplant HCMV disease refractory to treatment, urine is a readily obtained, non-invasive sample type. Further studies are needed to assess whether HCMV is shed in urine during viraemia in these patients. Outwith these specific circumstances, the protocol I developed for direct nanopore sequencing would require samples with high HCMV to human DNA

ratios. Therefore, for broader application, the incorporation of a target enrichment step that is capable of preserving large DNA fragments will still be required. Oligonucleotide bait-based enrichment has been shown to be successful at capturing HCMV DNA from cultured samples, and it is worth further investigating its direct application to sequencing clinical samples; however, the cost may be prohibitive. MDA may also be worth revisiting for this purpose, as, although the HCMV genome was not uniformly sequenced from the VH sample and chimaeric reads were common, the reads covered >85 % of the genome. Future studies could investigate the efficacy of using MDA on samples with higher DNA concentrations. Various mechanical and bioinformatic ways of minimising chimaera formation during MDA have been developed (Nurk et al., 2013) and could be tested. Another means of enrichment widely used with WGA is multiple annealing and looping-based amplification cycles (MALBAC), and a recent study has suggested that it offers a greater uniformity and reproducibility of product compared to MDA (Zhou et al., 2020).

Despite the current limitations in its application, nanopore technology potentially offers much to HCMV testing in clinical practice. This is particularly true where antiviral resistance is the focus, at the level of a few genes where nanopore sequencing is almost as accurate as Illumina sequencing. In the UK, HCMV antiviral resistance testing is currently referred to specialist reference laboratories in which analysis is based on Sanger sequencing of the PCR amplified UL97 and UL54 genes applied on a timescale that is not conducive to prompt clinical action. Recently, Suárez et al. established an Illumina sequencing pipeline enabling whole HCMV genome-based analysis for resistance mutations, which could pre-empt resistance testing against novel drug targets (Suárez et al., 2020). A nanopore protocol similar to my approach could provide a means of resistance testing on a shorter timescale. Indeed, this has been demonstrated for samples with lower HCMV loads, in which PCR products covering UL97 and UL54 were sequenced using nanopore technology, thus enabling more rapid analysis (Chorlton et al., 2021).

Direct sequencing of viruses from clinical samples without enrichment has been applied to the recent outbreak of MPox (the monkeypox virus genome is approximately 197 kb), where there is abundant virus in vesicle fluid (Isidro et al., 2022, Selhorst et al., 2022). Therefore, the application of nanopore

technology could be envisaged for investigating congenital infection, where viral loads are typically high, of the order of that of the urine sample used in my study. Indeed, urine is the diagnostic sample of choice for cCMV diagnosis, and infants with cCMV generally shed high viral loads in their urine. The protocol developed in my study might therefore be useful in the study of HCMV strains in congenitally infected babies and infants. As the management of congenitally infected infants is nuanced, with treatment reserved for only a proportion with severe disease (Rawlinson et al., 2017), a more widely accessible sequencing system may prove advantageous, especially if it transpires that there are links between certain HCMV strains, or mixtures of strains, and disease outcome. However, further work to establish thresholds for the detection of low-level variants, as has been conducted for Illumina sequence data, will be required (Suárez et al., 2020).

In conclusion, I was able to determine two complete HCMV genomes by direct nanopore sequencing from high-titre clinical samples at an accuracy that was close enough to that of Illumina sequencing to greatly encourage further investigation. In future, additional improvements in the hardware and software in nanopore technology, as well as appropriate target enrichment techniques, may contribute to the clinical management of HCMV infections.

6 Direct RNA sequencing of the HCMV lytic transcriptome

6.1 Background

Having utilised nanopore sequencing technology to characterise whole HCMV genomes from high-titre cultured and clinical samples, I proceeded to use the same technology to sequence the HCMV lytic transcriptome. Characterisation of the HCMV transcriptome by annotation of protein-coding transcripts and non-coding RNAs is an essential part of understanding the full genetic potential of HCMV and how mutations or variation may impact gene function. This information may eventually inform novel strategies and targets for treating HCMV disease.

The first comprehensive map of the wild-type HCMV genome annotated 165 canonical functional protein-coding genes (Dolan et al., 2004). Later work increased this number to 170 (Davison et al., 2013). Analyses of the lytic HCMV transcriptome using RNA-Seq on Illumina platforms and then ribosome profiling (sequencing ribosome-protected transcripts) identified many more transcripts than are needed to encode these proteins (Gatherer et al., 2011, Stern-Ginossar et al., 2012). RNA-Seq requires no *a priori* knowledge of gene sequences, in comparison to microarrays or more traditional techniques such as Northern blotting or quantitative RT-PCR, which are highly labour-intensive and depend on targeting preselected regions of the genome (Kondo et al., 1996, Ma et al., 2011). Moreover, proteomic analysis has also demonstrated novel HCMV proteins that are not associated with canonical genes (Varnum et al., 2004, Weekes et al., 2014).

The use of short-read RNA-Seq has identified a full complement of architecturally complex HCMV transcripts. These include transcripts with alternative TSSs but co-terminal TESs, co-terminal TSSs but alternative TESs, and nested transcripts differing at both the TSSs and the TESs with partial sequence identity among them. Alternative splicing may also form polycistronic transcripts. However, the pre-requisite generation of a cDNA library for RNA-Seq entails reverse transcription and PCR (RT-PCR) steps, both of which can lead to template-switching artefacts (Kanagawa, 2003, Geiszt et al., 2004). This occurs

when elongation along the initial template stops and reinitiation occurs at a different but homologous site, creating artefactual chimaeric cDNAs that may be misinterpreted as due to novel splicing or antisense transcription (Kanagawa, 2003, Cocquet et al., 2006). Another limitation of short-read sequencing is its inability to detect multiple introns within a single transcript (Steijger et al., 2013) (**Figure** 1-10).

In contrast, long-read sequencing of intact transcripts offers an opportunity to capture much fuller information on how splice sites are connected in transcripts and on the structure of isoforms. The ONT dRNA-Seq workflow allows for direct sequencing of polyadenylated (polyA) RNAs without requiring RT-PCR steps prior to sequencing (**Figure** 6-1). RNA is translocated through the nanopore in the 3'-5' direction, with the polyA tail and TES entering first. However, the basecalling algorithms automatically output reads in the 5'-3' direction. Reverse transcription is not required, but is recommended in order to generate a scaffold for the RNA and ensure higher throughput by helping to resolve secondary structures (**Figure** 6-1, step ii). However, the motor protein is attached only to the RNA strand, and it is only this strand that enters the nanopore and is sequenced (Workman et al., 2019). One of the major disadvantages of dRNA-Seq is its inability to accurately sequence the TSS at the 5’-end of a transcript, which enters the nanopore last. This is because the ratcheting effect of the motor protein as the RNA strand is pushed through the nanopore breaks down during the final 10-15 nt and gives uninterpretable basecalls (Workman et al., 2019) (**Figure** 6-1, step iv). Although dRNA-Seq of native polyA RNA can be used to map TESs and transcript isoforms, my focus in the time available was to use long-read data to capture splice junctions, including multiple splice junctions that occur in single transcripts. I then compared the splice junctions identified in my study to those published in two previous studies that used Illumina RNA-Seq to characterise the lytic transcriptome of Merlin (Gatherer et al., 2011, Stern-Ginossar et al., 2012).

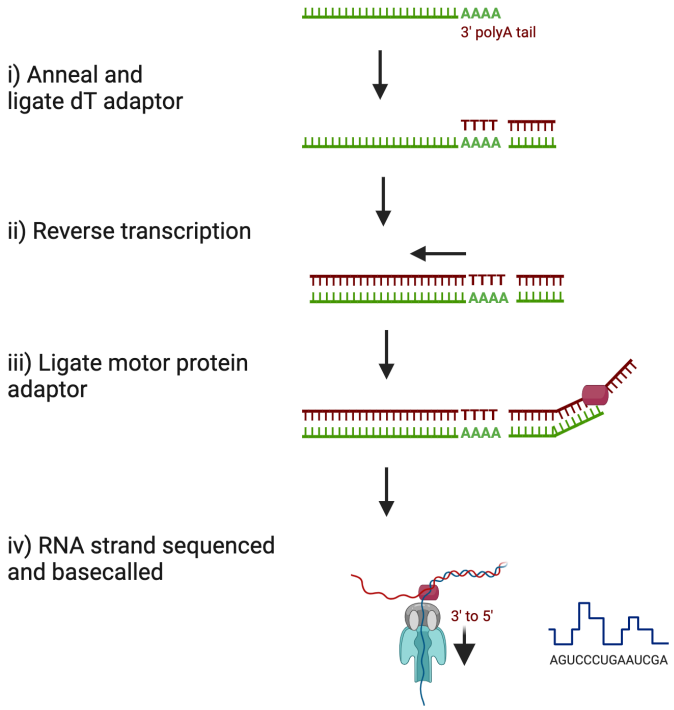


Figure 6-1. ds-RNA sequencing steps.
i) Polyadenylated (polyA) RNA is prepared for nanopore sequencing by annealing an adaptor with a poly(dT) overhang to the RNA poly(A) tail, followed by ligation. ii) The poly(dT) complement is extended by reverse transcription. iii) An adaptor with a motor enzyme is ligated to the first adaptor. (iv) The library is loaded onto the ONT flow cell for sequencing. (Created in BioRender.com, based on the ONT dRNA-Seq protocol.)

6.2 Objective

To use the ONT dRNA-Seq protocol to sequence the lytic transcriptome of HCMV strain Merlin, in order to compare the splice junctions to those published previously and to identify novel transcripts.

6.3 Materials and Methods

The virus reconstituted from the Merlin BAC was used in the previous RNA-Seq studies and therefore was used in my study (Gatherer et al., 2011, Stern-Ginossar et al., 2012). Like the parental virus, this virus is stable in cell culture due to point mutations in two genes (RL13 and UL128) (pAL1111; GenBank accession GU179001.1) (Stanton et al., 2010). The transcriptome generated by this virus in infected HFFF2 cells at 72 h post-infection was investigated using

long-read dRNA-Seq. Three biological replicates were performed by repeating the infection, extraction and sequencing processes on separate days (Runs 1-3).

6.3.1 RNA preparation

HFFF2 cells were cultured in GM in six-well plates ($2x10^5$ cells/well) (Section 2.3) and infected with the virus reconstituted from the Merlin BAC at a MOI of 5 PFU/cell. The inoculum was removed after 1 h, and the monolayers were washed and replenished with GM and incubated for a total of 72 h.

Infected cell RNA was extracted from the drained monolayers using a Direct-zol RNA MiniPrep kit (Zymo Research, Irvine, California, USA). To avoid contamination, this process was carried out in a microbiological safety cabinet, and RNAse-ZAP was used to clean the hood and associated equipment prior to extraction. Trizol (300 µL) was added to each well, a pipette was used vigorously to detach and lyse the cells, and an equal volume of 100 % (v/v) ethanol was added and mixed. The suspension was transferred to a Zymo-spin IICR column and centrifuged into a collection tube at 8000x g for 1 min. RNA wash buffer (400 µL) was added and the tube was centrifuged again at 8000x g for 1 min. A 5 µL aliquot of DNAse I (6 U/µL) in 75 µL of DNA digestion buffer was added to the column and incubated at room temperature for 15 min. The column was then washed twice with 500 µL of Direct-Zol RNA prewash, and the flow-through was discarded. RNA wash buffer (700 µL) was added and centrifuged at 8000x g for 2 min to ensure complete removal of the wash buffer, and the column was transferred to a fresh RNAse-free tube. NFW was added to the column matrix, incubated for 1 min at room temperature, and centrifuged at 8000x g for 2 min. The eluted RNA was kept on ice and handled with blunt-tip pipettes from this stage. RNA integrity was assessed using High Sensitivity RNA ScreenTape (Agilent Technologies; catalogue no. 5067-5579) on the 4200 TapeStation system (Agilent Technologies). RNA concentration was assessed by Qubit fluorometry using a Invitrogen Qubit RNA HS Assay kit (ThermoFisher Scientific; catalogue no. Q32852).

6.3.2 RNA library preparation

A dRNA-Seq kit (ONT; protocol SQK-RNA002 version 22 November 2018) was used to prepare the extracted RNA for nanopore sequencing. RNA library preparation was conducted according to the protocol, except that the RNA control (RNA CS) was diluted 1:10 for Runs 2-3.

6.3.2.1 cDNA synthesis

Extracted RNA (500 ng in 9 µL) was mixed by gentle pipetting in a 0.2 mL thin-walled PCR tube with 3 µL of NEBNext Quick Ligation reaction buffer (New England Biolabs; catalogue no. B6058), 0.5 µL of RNA CS (diluted 1:10 for Runs 2-3), 1 µL of RT Adaptor and 1.5 µL of T4 DNA ligase (New England Biolabs; catalogue no. M0202). The reaction was incubated at room temperature for 10 min. Meanwhile, the reverse transcription mix was prepared by mixing 2 µL of 10 mM dNTPs, 8 µL of 5x first-strand buffer, 4 µL of 0.1 M DTT and 9 µL of NFW. This master mix was added to the 0.2 mL PCR tube containing the RT adaptor-ligated RNA, mixed by pipetting, and supplemented with 2 µl of SuperScript III reverse transcriptase (ThermoFisher Scientific; catalogue no. 18080044). The reaction was again mixed and incubated at 50 ºC for 50 min and 70 ºC for 10 min, and then held at 4 ºC.

The sample was transferred to a fresh 1.5 mL DNA LoBind Eppendorf tube, and 72 µL of resuspended RNAClean XP beads was added and mixed by pipetting. The reaction was incubated for 5 min at room temperature, centrifuged, and pelleted on a magnet, and the supernatant was discarded. Keeping the tube on the magnet, the beads were washed with 150 µL of freshly prepared 70 % (v/v) ethanol without disturbing the pellet. The sample was again centrifuged, and residual ethanol was removed. The tube was removed from the magnet, and the sample was resuspended in 20 µL of NFW, incubated for 5 min at room temperature, and pelleted on a magnet until the eluate was clear. The eluate was transferred into a fresh 1.5 mL Eppendorf DNA LoBind tube.

6.3.2.2 Attachment of sequencing adaptors to the ends of RNA-cDNA hybrids

The reverse-transcribed RNA was mixed with 8 µL of NEBNext Quick Ligation reaction buffer, 6 µL of RNA adaptor (RMX), 3 µL of NFW and 3 µL of T4 DNA ligase, and incubated at room temperature for 10 min. Clean-up was performed by adding 40 µL of resuspended RNA Clean XP beads to the adaptor ligation reaction, mixing, and incubating for 5 min at room temperature. The sample was centrifuged, pelleted on a magnet, washed twice by resuspending in 150 µL of Wash Buffer (WSB) by flicking the tube, and returned to the magnet, and the supernatant was removed after the beads had pelleted. Finally, the beads were resuspended in 21 µL of elution buffer and incubated at room temperature for 10 min prior to pelleting on a magnet, and the clear eluate was transferred into a fresh 1.5mL Eppendorf DNA LoBind tube.

6.3.2.3 RNA library sequencing on the MinION flow cell

After priming the flow cell as described in Section 2.11.3, 20 µl of the prepared RNA library was mixed with 17.5 µl of NFW (total 37.5 µL), and an equal volume of RNA running buffer (RRB) was added and mixed by gentle pipetting. An aliquot (75 µL) of the prepared RNA library was loaded onto a primed flow cell in dropwise fashion via the SpotOn sample port (**Figure** 2-2). A final amount of 100 ng of reverse-transcribed and adapted RNA was loaded onto the flow cell and run for 48 h or until all pores were used. Reads were acquired for processing as described in Sections 2.11.3 and 4.3.5.

6.3.3 Bioinformatic analysis

Three datasets were obtained from the biological replicates (Runs 1, 2 and 3). After the first dRNA-Seq experiment (Run 1), the kit RNA control was diluted 1:10 as it was found that 60 % of reads had mapped to the kit control from the yeast *Saccharomyces cerevisiae*, and only 13 % mapped to HCMV using DIAMOND, which allows rapid alignment against a protein database (Buchfink et al., 2015). This indicated that Run 1 was overwhelmed by the RNA CS control. Trimming of end adaptors was performed using Porechop v.0.2.3 (Wick et al., 2017). All reads from each of the datasets were aligned to the reference Merlin genome (GenBank accession AY446894.2) using Minimap2 v.2.17 (Li, 2018) with the *splice*

option, which allows for the mapping of spliced reads indicated by apparent deletions in comparison with the Merlin reference. The line command used was as follows.

```
$ minimap2 -ax splice -uf -k14 merlin.fasta all_rnaseq.fastq > aln_mini.sam
```

Samtools v.1.9 (Li et al., 2009) was then used to create an indexed BAM file for further processing. Subsequently, a custom program, ShaSam (https://github.com/rjorton/ShaSam) was used to identify splice sites. ShaSAM is Java-executable and works by examining the alignment of each read in the BAM file and detecting the co-ordinates of deletions of >50 nt with respect to the reference. These deletions may correspond to introns. ShaSAM then reports the number of reads in which each deletion is observed, the length of the deletion, the dinucleotides at each end of the deletion (most introns have GT at the start and AG at the end), and the number of times multiple deletions occur on the same reads (indicative of a multiply-spliced transcript). ShaSAM also reports the frequency of potential TSSs and TESs at each genome position, and outputs all results as tab-delimited text files. The line command used was as follows, with ShaSAM taking an aligned BAM file and the corresponding reference FASTA file as input.

```
$ java -jar ~orto01r/dist/ShaSAM/ShaSAM.jar aln_mini.bam merlin.fasta
```

The ShaSAM output files listing the splice junctions from the three replicate datasets were merged into a single text file that was then processed by a custom python script, splice_compare.py (https://github.com/rjorton/ShaSam) that links the splice junctions in multiple files together based on their names (given as the location of the intron ends). This compilation was then compared to the published lists of spliced Merlin transcripts (Gatherer et al., 2011, Stern-Ginossar et al., 2012), and the intersections of the five datasets (from my nanopore experiments (Runs 1-3), the Gatherer et al. dataset (DSG) and the Stern-Ginossar et al. dataset (DSS)), were visualised using UpSetR (https://gehlenborglab.shinyapps.io/upsetr/) (Lex et al., 2014). This application shows the number of splice junctions identified uniquely in each of the datasets and the number of splice junctions in common in the various intersections of the datasets. The annotation file (in gff format) of the Merlin genome was obtained

from GenBank to locate the splice junction sites in relation to the canonical genes. An alternative notation had been used to describe splice junctions in DSG, involving stating the locations of the ends of the flanking exons rather than the location of the intron. For example, D+27528^A+27612 corresponds to a splice junction and represents a splice donor (D) at 27528 on the forward strand (+) joined (^) to a splice acceptor (A) at 27612 also on the forward strand. This difference in notation numbering was factored into the comparisons of the nanopore datasets.

6.4 Results

6.4.1 Summary of sequencing statistics

A summary of the data obtained from the three nanopore runs (Runs 1-3) using MinIONQC (Lanfear et al., 2019) is shown in **Table** 6-1. As a result of the higher frequency of RNA CS in Run 1, the total number of reads mapping to Merlin was greater in Runs 2-3 (**Table** 6-1). The proportion of mapped reads containing at least one splice junction was high and ranged between 23 and 30 % in the three runs. The number of different splice junctions observed (at any frequency) was 1,845 for Run 1 (which had far fewer mapped reads) and approximately 5,000 for Runs 2 and 3. However, discounting splice junctions that were observed in <3 reads led to an approximately 10-fold reduction in the number of unique splice junctions observed. Splice junctions observed at low frequency may be indicative of erroneous sequences resulting from the high nanopore read error rate.

The splice junctions observed in the three nanopore datasets were then combined with the DSG and DSS splice junctions. **Table** 6-2 shows the top 50 observed splice junctions in the nanopore data. By far the most frequent splice junction observed was due to deletion of 27,529-27,611 (corresponding to an intron in UL22A) on the forward strand. This intron was also the most frequently detected in DSG and DSS. Overall, **Table** 6-2 shows a strong agreement between the nanopore and previous datasets, with the most frequently observed splice junctions in the nanopore datasets being typically observed at high frequency in one or both of DSG and DSS.

Table 6-1. Summary of sequencing statistics for nanopore Runs 1-3.

Run	Run 1	Run 2	Run 3
Total gigabases	1.30	1.32	1.77
Total reads (no.)	1,133,088	1,233,791	1,587,417
N50 length (bp)	1,360	1,370	1,386
Mean length (bp)	1,148	1,072	1,118
Median length (bp)	1,317	882	954
Max length (bp)	12,962	107,285	14,626
Mean q-score	9.3	8.7	8.5
Mapped reads (no.)	284,999	555,398	746,336
Mapped reads with a splice junction (no.)[a]	64,269	161,855	225,540
Proportion of mapped reads with a splice junction	0.23	0.29	0.30
Unique splice junctions (no.)[b]	1,845	5,290	4,819
Unique splice junctions observed ≥3 times (no.)[c]	231	519	506

[a] Mapped reads with a splice junction correspond to the total number of reads with a deletion of >50 nt.

[b] Unique splice junctions count the number of unique splice junctions observed at any frequency.

[c] Unique splice junctions count the number of unique splice junctions observed at any frequency.

Table 6-2. Top 50 observed splice junctions for nanopore runs 1-3.

No.	Splice junction[a]	Run 1 (no.)	Run 2 (no.)	Run 3 (no.)	Total (no.)	Donor site[b]	Acceptor site[b]	Length	DSG (no.)	DSS (no.)
1	27529-27611	50881	127565	183786	362232	GT	AG	83	131826	35347
2	107497-108853	2888	4474	7415	14777	GT	AG	1357	2574	587
3	176522-176403	1320	3209	4287	8816	GT	AG	120	2612	535
4	160956-156414	1121	2637	3303	7061	GT	AG	4543	2432	369
5	193543-193653	233	1234	1279	2746	GT	AG	111	3118	242
6	107497-108858	395	834	1460	2689	GT	AG	1362	0	0
7	224025-224099	516	556	1149	2221	GT	AG	75	1557	465
8	108156-108853	473	697	1046	2216	GT	AG	698	788	214
9	161186-161261	275	704	1014	1993	GT	AG	76	94	219
10	105813-108853	327	702	805	1834	GT	AG	3041	592	98
11	179162-181562	226	1024	494	1744	GT	AG	2401	0	82
12	162674-162830	122	418	760	1300	GT	AG	157	316	334
13	177910-177803	178	314	590	1082	GT	AG	108	200	115
14	180804-181562	127	629	304	1060	GT	AG	759	15	119
15	174019-173904	112	363	530	1005	GT	AG	116	299	751
16	163110-163390	98	346	552	996	GT	AG	281	135	155
17	49926-49824	114	473	406	993	GT	AG	103	240	770
18	192429-193171	72	328	429	829	GT	AG	743	286	97
19	192429-193168	65	337	388	790	GT	AG	740	224	108
20	176780-176658	95	303	392	790	GT	AG	123	295	61
21	149568-149832	57	397	318	772	GT	AG	265	0	10
22	174928-174108	89	240	441	770	GT	AG	821	414	1044
23	173718-172176	71	284	403	758	GT	AG	1543	302	80
24	171875-174080	114	207	336	657	GT	AG	2206	66	49
25	103437-108853	108	306	230	644	GT	AG	5417	295	20
26	106021-108853	126	189	253	568	GT	AG	2833	248	42
27	43442-43561	48	151	357	556	GT	AG	120	252	120
28	163110-163512	44	168	298	510	GT	AG	403	63	72
29	101642-108853	81	206	211	498	GT	AG	7212	203	33
30	204853-204447	29	210	166	405	GT	AG	407	29	41
31	12991-8197	64	111	200	375	GT	AG	4795	51	28
32	161463-161545	55	146	167	368	GT	AG	83	24	41
33	103286-108853	34	198	124	356	GT	AG	5568	641	10
34	9029-8197	76	108	171	355	GT	AG	833	10	20
35	173718-173549	61	104	174	339	GT	AG	170	222	1375
36	160952-156401	74	140	124	338	GT	AG	4552	0	0
37	107741-108853	57	105	158	320	GT	AG	1113	142	66
38	107484-108853	46	114	121	281	GT	AG	1370	0	0
39	228333-228419	39	94	139	272	GT	AG	87	69	19
40	188074-187869	30	103	131	264	GT	AG	206	100	40
41	107497-108861	36	61	132	229	GT	AG	1365	0	0
42	169675-170081	44	99	76	219	GT	AG	407	77	8
43	118698-141711	28	85	104	217	GT	AG	23014	0	0
44	170158-174080	36	112	66	214	GT	AG	3923	151	0
45	3204-3308	21	110	54	185	GT	AG	105	0	15
46	102292-108853	25	82	71	178	GT	AG	6562	189	15
47	6748-3241	28	57	90	175	GT	CT	3508	0	0
48	7609-7397	34	70	68	172	GT	AG	213	24	0
49	12653-8197	35	54	72	161	GT	AG	4457	14	8
50	106578-108853	13	76	68	157	GT	AG	2276	45	7

[a] The splice junction is expressed as the first and last nucleotide in the intron relative to the Merlin genome. If the second number is greater than the first, the deletion maps on the reverse strand.

[b] The dinucleotides observed at the 5' (donor) and 3' (acceptor) ends of the intron.

A further comparison of the splice junctions observed between the nanopore datasets and DSG and DSS was explored using an UpSetR plot (**Figure** 6-2). This type of plot enables efficient visualisation of the intersections of multiple datasets (i.e. the degree to which datasets in various combinations agree with each other). The UpSetR plot was used to demonstrate the extent if intersection of the five datasets and highlighted some important aspects.

First, the number of splice junctions identified in the nanopore datasets was larger than in the previous datasets. However, a large proportion was observed only within one dataset (Run 1: 1202/1845, Run 2: 3675/5290, Run 3: 4169/4819), suggesting low abundance of many spliced transcripts and stochastic appearance in the datasets probably due to the high error rate and noise of nanopore reads. Consistent with this, only a minority of observed nanopore splice variants were detected ≥3 times (**Table** 6-1). Second, 295 splice junctions detected in all three nanopore runs were not reported in DSG or DSS. These are potentially novel splice junctions as they were observed in all three biological replicates, and they were investigated further (Section 6.4.2). Third, there were 118 unique splice junctions in DSG and one unique splice junction in DSS. This suggests a degree of biological variability or perhaps different thresholds across the studies for managing potential artefacts.

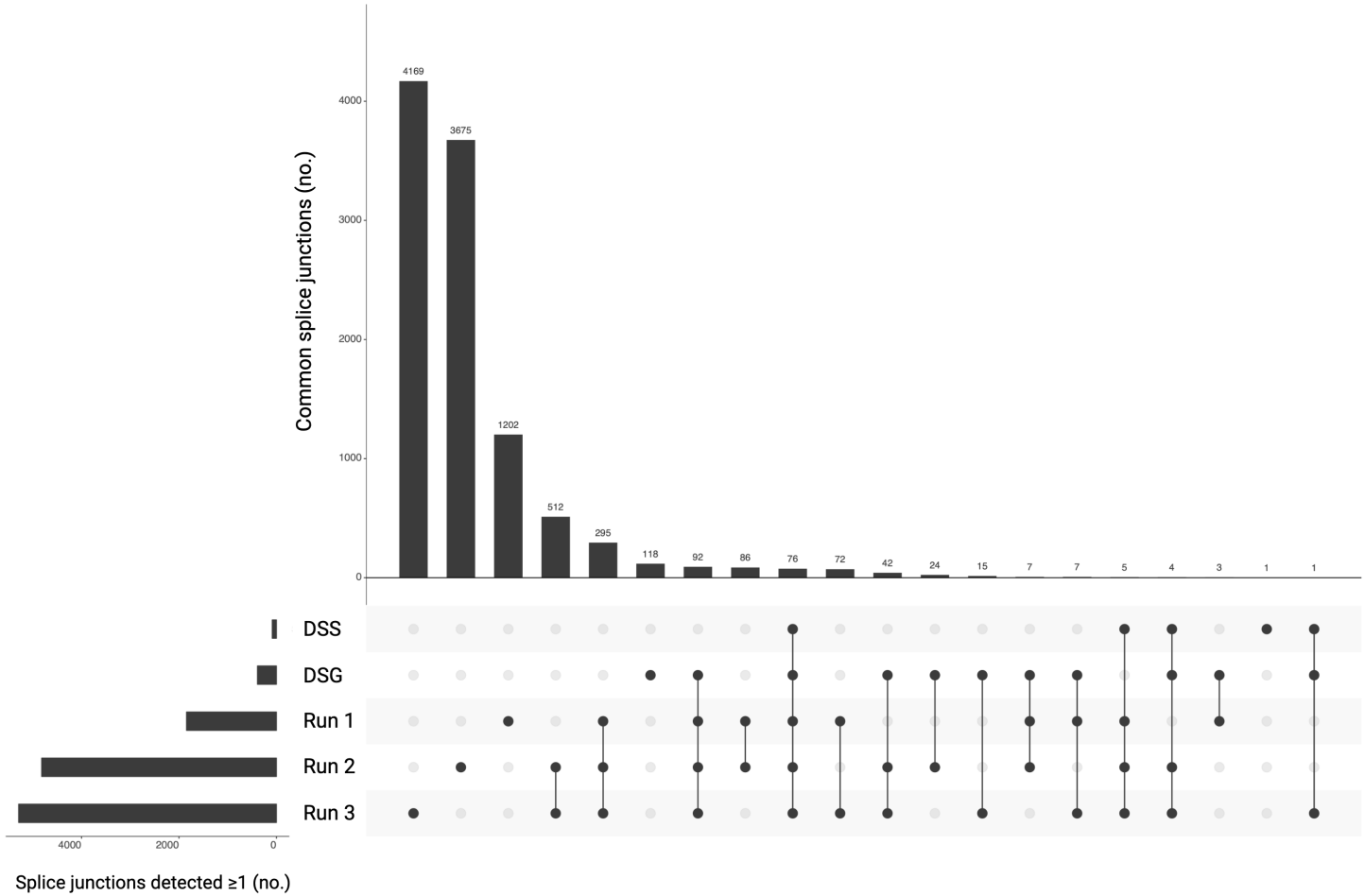


Figure 6-2. Plot showing the number of splice junctions present in each of five datasets and the number of splice junctions that were found in common between different datasets.
The splice junctions detected in the three nanopore replicates (Runs 1-3) and in DSG and DSS were compared. The numbers of splice junctions that appeared at least once in each dataset are tallied in the histogram on the bottom left. Each dataset is represented by a dot in the rows below the histogram showing the number of common splice junctions between datasets. A line connecting the dots indicating that these datasets share common splice junctions.

6.4.2 Novel splice junctions

As mentioned above, a total of 295 splice junctions unique to the three nanopore datasets were not observed in DSG or DSS (**Figure** 6-2). To increase confidence, introns that did not start and end with the canonical GT and AG dinucleotides, respectively, were discounted, and a threshold of ≥10 reads was set. The 103 novel splice junctions in this category were further filtered to identify potential duplicates that may have been caused by nanopore sequencing errors. As an example, four splice junctions within the coding regions of UL73/UL74A were observed in the top 50 splice junctions (**Table** 6-2): 107497-108853 (no. 2) was detected 14,777 times in the nanopore datasets and also in DSG and DSS; 107497-108858 (no. 6) was detected 2,689 times only in the nanopore datasets; 107484-108853 (no. 38) was detected 281 times only in the nanopore datasets; and 107497-108861 (no. 41) was detected 229 times only in the nanopore datasets. Although the co-ordinates of each intron are different and the introns contain the canonical GT/AG dinucleotides (**Table** 6-2), they clearly have very similar co-ordinates. This suggests that they are less likely to represent genuine alternative splice junctions and rather may be due to nanopore sequencing errors. As a result, the 103 potentially novel splice junctions were filtered further by discounting those located within 10 nt of the start or end coordinates of another splice junction, retaining the splice junction with the highest count. This resulted in a final filtered set of 57 potentially novel splice junctions (**Table** 6-3). Due to this stringent level of filtering, these splice junctions are likely to be genuine. Nevertheless, these splice junctions, in particular those represented by larger numbers of reads, would benefit in future from validation by additional data from the Illumina platform and by RT-PCR.

To investigate the any potential impact on coding and downstream protein products, the coordinates of the novel splice junctions were compared to the published gene annotation of the HCMV genome strain Merlin.

While the majority of the novel splice junctions were in non-coding regions of the genome, some of these were associated with the following two genes: RL8A (8,184-7,912 on the negative strand) and UL111A (three coding exons 161,003 – 161,185, 161, 262 - 161,462 and 161,546 - 161,692 on the positive strand). However, these terminate upstream of the start codon (at position 8,197 for

RL8A) or disrupt the reading frame of the first exon (ending at position 161,054 for UL111A). Gatherer and co-workers noted similarly that alternate splicing was frequent and particularly evident for RL8A (in addition to UL74A, UL124A and UL150A) and that upstream exons may form untranslated leaders (Gatherer et al., 2011).

Some novel splice junctions were identified which encode potentially novel protein products (**Figure** 6.3). Chimeric transcripts of US14 and UL145 were identified in 20 reads, with an intron size of 26,935 nt (**Figure** 6.3a). Another chimera of US32/US33A was identified in 12 reads, a transcript which appended a truncated US33A to US32 (**Figure** 6.3b). Furthermore, a truncated form of UL22A (**Figure** 6.3c middle panel) and a shortened ORF with premature stop in UL22A (**Figure** 6.3c bottom panel) were seen in 102 and 17 reads respectively. These novel splice junctions and potential protein products would need further investigation by RT-PCR or protein-purification, especially for those observed at low frequency.

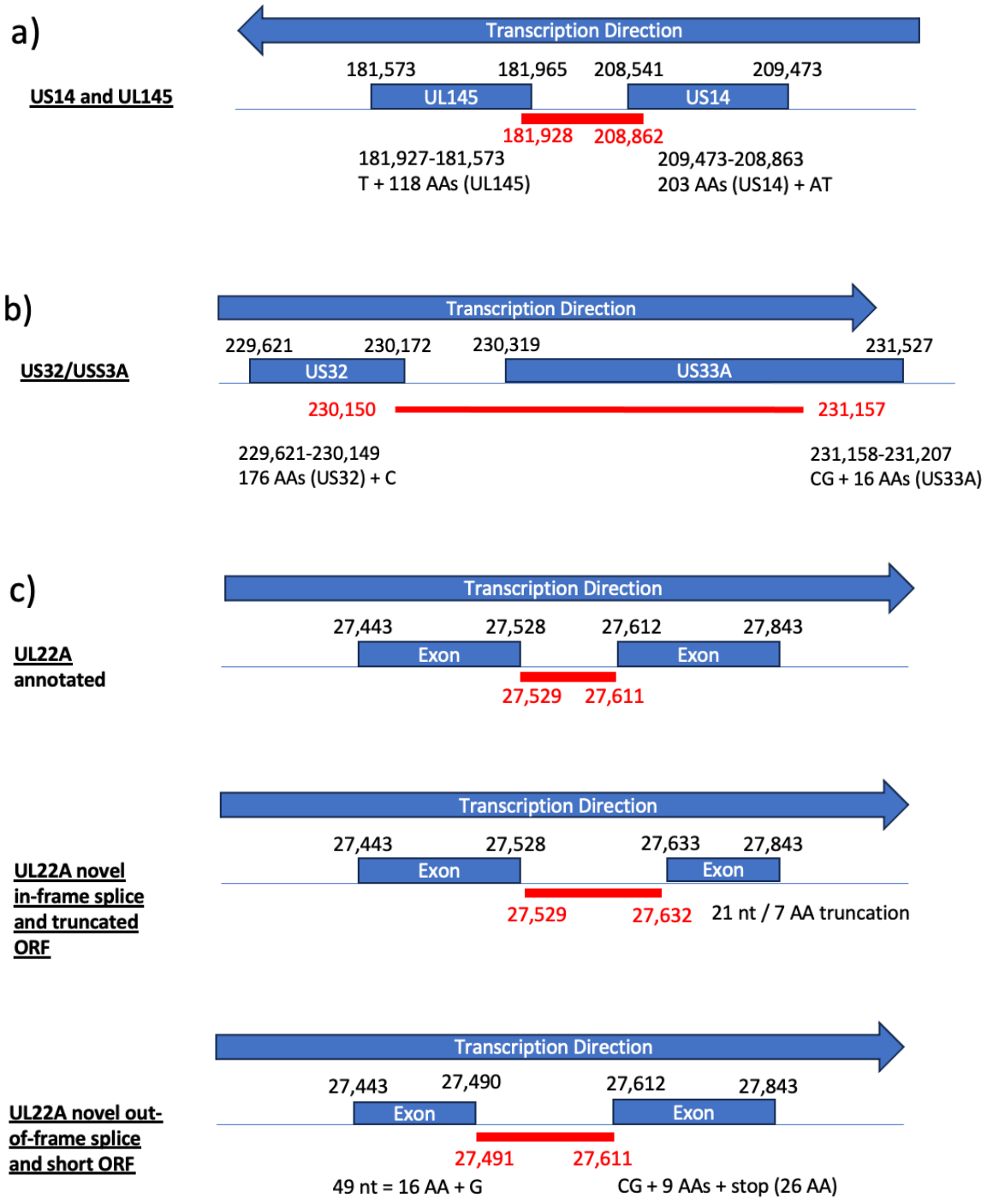


Figure 6-3. Novel splice junction and impact on gene annotation.
The blue block arrow indicates transcription direction of the gene. The position of the exons and introns are shown in black and the splice junctions in red, and introns are represented by the red line. Disruption to coding annotated below the gene. AA, amino acid.

Table 6-3. The 57 potentially novel splice junctions identified from the nanopore sequencing datasets.

Splice junction	Run 1 (no.)	Run 2 (no.)	Run 3 (no.)	Total (no.)	Intron length	Start genes	End genes
160952-156401	74	140	124	338	4552	RNA5.0	RNA5.0
107484-108853	46	114	121	281	1370	UL73	UL73/UL74A
118698-141711	28	85	104	217	23014		
171153-170872	19	56	79	154	282	UL122	UL122
25752-25523	25	48	62	135	230		
180373-181562	16	85	31	132	1190		
27529-27632	10	35	57	102	104	UL22A	UL22A
3482-3612	9	50	18	77	131		
192597-193171	3	38	21	62	575	UL150A/UL148D	UL150A/UL148D
167658-174080	3	27	27	57	6423		UL124
180779-181562	8	32	14	54	784		
118672-141711	4	25	17	46	23040		
19819-8197	3	12	26	41	11623		RL8A/RL9A
171336-171220	4	27	9	40	117	UL122	UL122
7573-7522	1	14	23	38	52	RL8A/RL9A/RNA1.2	RL8A/RL9A/RNA1.2
27558-27668	6	17	9	32	111	UL22A	UL22A
99747-108853	4	14	14	32	9107		UL73/UL74A
178138-157540	4	8	19	31	20599	UL131A	RNA5.0
177815-181562	5	16	9	30	3748		
21052-8197	4	11	12	27	12856		RL8A/RL9A
176396-176494	5	14	8	27	99		
6748-3893	1	10	14	25	2856		RL5A/RNA2.7
181076-181562	3	17	5	25	487		
157431-161054	4	9	10	23	3624		UL111A
6748-4333	4	4	14	22	2416		RL5A/RNA2.7
10633-8197	6	7	8	21	2437	RL8A	RL8A/RL9A
208862-181928	2	8	10	20	26935	US14	UL145
160358-161054	6	7	6	19	697		UL111A
29995-30362	2	6	10	18	368		
3256-3199	3	9	5	17	58	RL5A/RNA2.7	RL5A/RNA2.7
27491-27611	2	9	6	17	121	UL22A	UL22A
26192-25523	1	8	6	15	670		
157545-161054	5	2	8	15	3510		UL111A
12991-9814	2	7	5	14	3178	RL8A	RL8A
25036-8197	3	2	9	14	16840		RL8A/RL9A
123179-123099	1	8	5	14	81	UL84	UL84
149568-161054	1	7	6	14	11487	UL102	UL111A
176522-176389	2	2	10	14	134	UL128	UL128
985-2844	1	11	1	13	1860		
6748-4321	3	2	8	13	2428		RL5A/RNA2.7
27909-28123	1	4	8	13	215		
167658-170081	1	5	7	13	2424		
17464-8197	1	6	5	12	9268		RL8A/RL9A
162674-163390	1	6	5	12	717	UL112	UL112
171845-174080	3	6	3	12	2236	UL124	UL124
230150-231157	4	5	3	12	1008	US32	US33A/US34
4630-4564	2	3	6	11	67	RL5A/RNA2.7	RL5A/RNA2.7
7609-7522	1	3	7	11	88	RL8A/RL9A/RNA1.2	RL8A/RL9A/RNA1.2
32419-3210	1	4	6	11	29210		RL5A/RNA2.7
76456-27540	1	4	6	11	48917		
157128-161054	4	1	6	11	3927		UL111A
180414-181562	4	4	3	11	1149		
184399-177803	1	8	2	11	6597	UL141	UL131A
54167-54054	2	4	4	10	114	UL41A/UL40	UL41A/UL40
139180-147375	2	3	5	10	8196		
147073-146862	3	1	6	10	212	UL100	UL100
204553-204447	1	2	7	10	107	US9	US9

6.4.3 Previously identified splice junctions

A total of 76 splice junctions were detected in the nanopore datasets that were also listed in both DSG and DSS (**Figure** 6-2). The most highly expressed transcripts represented in the nanopore datasets, in descending order of magnitude, were associated with UL22A, UL73, UL128, RNA5.0, UL150A/UL148D, US27, UL73/UL74A, UL111A, UL112, UL131A and UL122/UL123. A single splice junction was detected in 51 reads in DSS (27525-27611) but not in the nanopore datasets or DSG. This is located in the UL22A coding region and represents the use of an alternative splice donor site. Conversely, DSG reported an additional 118 splice junctions that were not found in either the nanopore datasets or DSS. However, only three of these were detected in >10 reads: D+76458^A+108853 (37 reads; supported by RT-PCR), D+134992^A+165962 (19 reads) and D-183243^A-108855 (15 reads).

6.4.4 Multiply-spliced transcripts

dRNA-Seq combined with the long-read capability of the nanopore platform offers the potential not only to detect individual splice junctions but also multiple splice junctions in individual reads. In contrast, short-read Illumina reads are restricted to detecting single transcripts with multiple introns that occur within a few hundred nt of each other. The software tool ShaSAM was used to report the number of times splice junctions co-occurred on reads.

Multiply-spliced transcripts were first filtered for those that were observed in all three nanopore data sets, resulting in 116 different transcripts: 94 with two different splice junctions, 20 with three different splice junctions, and 2 with four different splice junctions (49926-49824,51301-51198,77368-51345,77622-77505 observed nine times, and 168876-169008,169136-169506,169675-170081,170176-170507 observed six times).

The splice junction co-ordinates were compared to the published gene annotation of the HCMV Merlin genome to predict the potential functional effects. Most of the transcripts were associated with UL122/UL123. In total, 11 transcripts contained three introns, five of which involve UL122 (170689-174090) and UL123 (172329-174090), both on the reverse strand **(Table** 6-4). In the HCMV

Merlin GenBank accession, the UL122 transcript contains two introns at 173718-172176 and 174019-173904 on the reverse strand, and UL123 contains the introns 173718-173549 and 174019-173904 (top panel of **Figure** 6-4a and b and **Figure** 6-5). The nanopore data suggest that these known UL122 splice variants can occur with a third intron at 174928-174018 (Upstream Splice1, n=448), 174933-174108 (Upstream Splice2, n=15) or 174241-174108 (Upstream Splice3, n=60) (**Figure** 6-4a, **Table** 6-5). Given that the coding sequence for UL122 begins at 174090, these introns are located in an upstream non-coding region of UL122 and UL123 (**Table** 6-5,). This region of the genome has also been shown previously to be a hotspot for highly spliced antisense transcripts (Gatherer et al., 2011). Indeed, the introns seen in these multiply-spliced transcripts were present in DSG and DSS apart from the intron between 174933-174108 (Upstream Splice2), but this was only present in 15 reads in my dataset (**Table** 6-5). However, there were no apparent effects on the known exons of the genes involved (**Figure** 6-4). These may alter the start of the mRNA but would have no impact on the exons of the UL122 or UL123 genes.

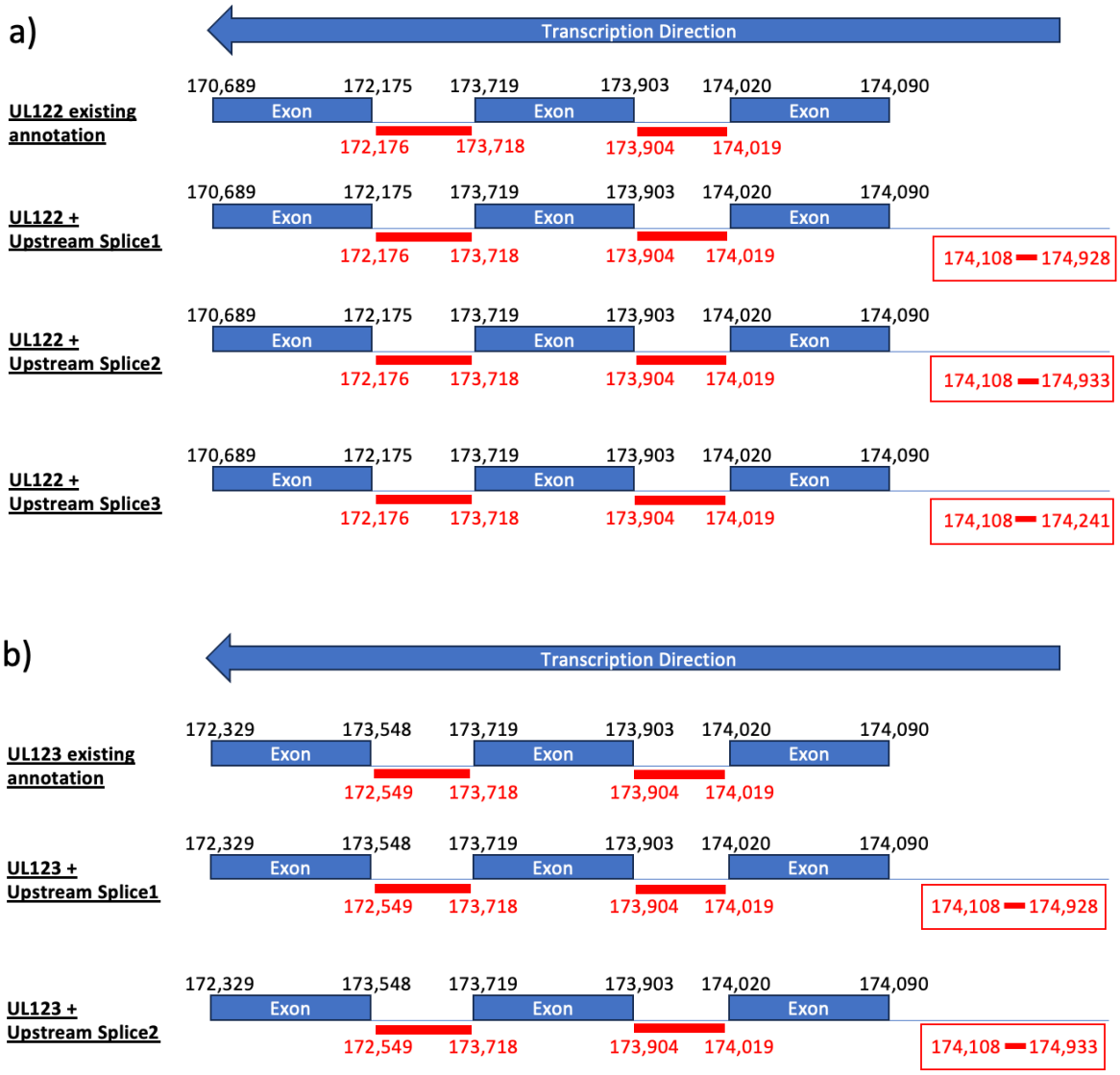


Figure 6-4. Multiply-spliced transcripts and novel upstream splice sites.
The blue block arrow indicates transcription direction of the gene. The position of the exons and introns are shown in black and the splice junctions in red, and introns are represented by the red line. The novel splice sites upstream of known transcript is highlight in boxes outlined in red.

Transcripts containing two introns occurring at high frequency were also found in UL112 (**Figure** 6-6). Transcripts with two splice junctions were commonly detected, consisting of intron 162674-162830 spliced to 163110-163390, 163110-163512, 163110-163655 or 163110-163509 and detected in 696, 344, 27 and 23 reads, respectively (**Figure** 6-6). These splice junctions, except for 163110-163509, have been described in both DSG and DSS. Again, these potentially novel transcripts would warrant further investigation by RT-PCR or protein analysis to identify their relevance in the HCMV life-cycle.

Table 6-4. Multiply-spliced transcripts with three introns.

Introns[a]	Run 1 (no.)	Run 2 (no.)	Run 3 (no.)	Total
173718-172176, 174019-173904, 174928-174108	35	137	276	448
173718-173549, 174019-173904, 174928-174108	45	67	132	244
173718-172176, 174019-173904, 174241-174108	4	29	27	60
176522-176403, 176780-176658, 177910-177803	5	9	15	29
173718-173549, 174019-173904, 174241-174108	4	5	10	19
173718-172176, 174019-173904, 174933-174108	2	5	10	17
165822-165959, 166063-166160, 166438-166522	1	3	13	17
191909-192071, 192429-193171, 193543-193653	1	9	3	13
191909-192071, 192429-193168, 193543-193653	2	4	6	12
165822-165962, 166063-166160, 166438-166522	2	2	8	12
49926-49824, 51301-51198, 52572-51345	2	6	3	11

Table 6-5. Most commonly detected transcripts with three introns.
Corresponding to genes encoding UL122 and UL123.

Reads (no.)	Intron 1	Intron 2	Intron 3
448	173718-172176	174019-173904	174928-174108
60	"	"	174241-174108
15	"	"	174933-174108
244	173718-173549	174019-173904	174928-174108
19	"	"	174241-174108

[a] The splice junctions indicate the bases deleted (inclusive), and each intron is separated by a comma.

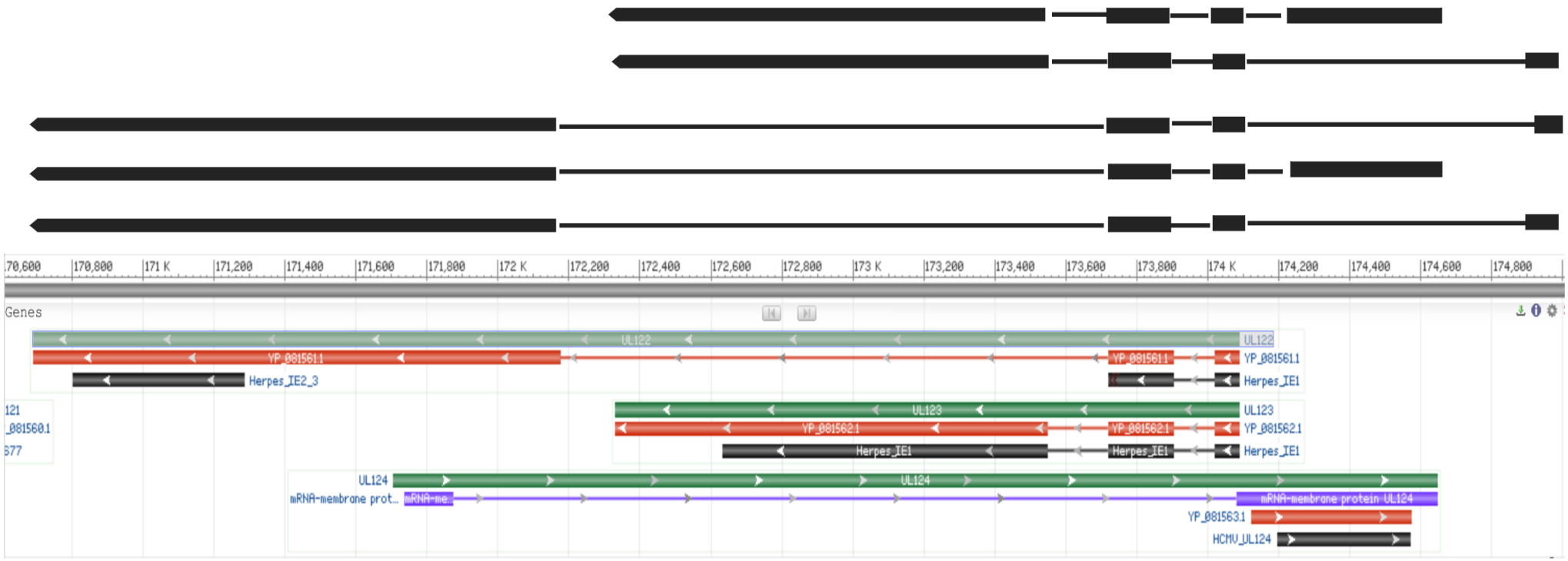


Figure 6-5. Multiply-spliced transcripts with three introns.
Transcripts spanning UL122/UL123 encompassing three introns are shown above the sequence view (https://www.ncbi.nlm.nih.gov/gene/3077563). Thick black lines, exons; thin lines, introns; block arrowheads, TSSs. Annotated transcripts are shown in each of the lower panels and the region of genome is shown in the scale. Green, gene; red, coding region; purple, RNA; black, protein feature.

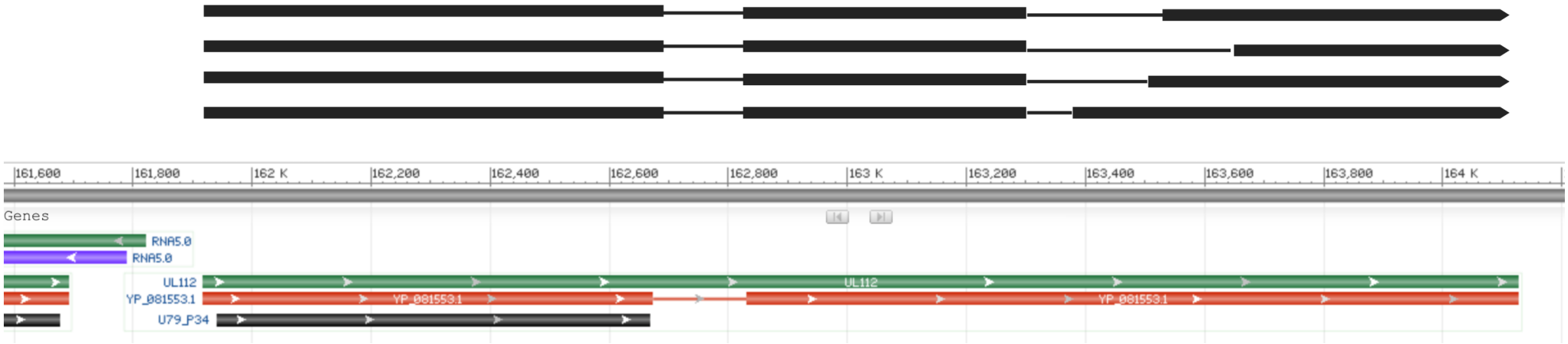


Figure 6-6. Multiply-spliced transcripts with two introns.
Transcripts spanning the UL112 gene encompassing two introns. The novel polycistronic transcripts identified in the nanopore datasets are shown above the sequence view (https://www.ncbi.nlm.nih.gov/gene/3077495). Thick black lines, exons; thin lines, introns; block arrowheads, TSSs. Known transcripts are shown in each of the bottom panels and the region of genome is shown by the scale. Green, gene; red, coding region; purple, RNA; black, protein feature.

6.5 Discussion

I characterised the lytic HCMV strain Merlin transcriptome in infected HFFF2 cells by sequencing RNA libraries on the ONT platform. Previous work on this subject using RNASeq on short-read sequencing platforms was limited due to the inability to detect multiple splicing within single transcripts (**Figure** 1-10). Moreover, RT-PCR predisposes to false detection of chimaeric reads or artefactual splice sites which can be due to template-switching artefacts, even in regions with as few as three consecutive adenine bases (Cocquet et al., 2006, Balázs et al., 2019). In contrast, dRNA-Seq of native RNA omits RT-PCR steps and should not generate such artefacts. However, dRNA-Seq cannot reliably basecall the TSSs to the release from the motor protein of the RNA upon reaching the 5'-end. The aim of my work in the time available was therefore only to confirm the existence of previously reported splice junctions and to identify novel transcripts and transcript isoforms. Future work could be done to investigate the TESs.

In accord with the previously published RNA-Seq experiments, the most highly expressed transcript during lytic HCMV infection detected by dRNA-Seq was UL22A. This small, spliced gene encodes an 83 amino acid residue chemokine-binding protein that is critical for binding the chemokine RANTES (regulated on activation, normal T cell expressed and secreted) (Wang et al., 2017). The most commonly identified novel alternative splice junctions in the nanopore datasets were found in UL74A, which encodes envelope glycoprotein 24 and has previously been shown to be spliced at into 108853 from one or other of 46 upstream exons (Gatherer et al., 2011). 107497-108861 was another novel splice junction within this coding region.

One of the limitations of dRNA-Seq is that the reads cannot reliably identify TSSs. An alternative ONT protocol is available that adds adaptors onto cDNA ends and allows for the detection of 5'-ends, but this is technically not sequencing native RNA. A separate means of mapping TSSs, such as cap analysis of gene expression (CAGE), could be utilised in conjunction with the dRNA-Seq protocol and would have been performed if time had permitted (Kodzius et al., 2006). Another limitation of dRNA-Seq is the high read error-rate of nanopore sequencing and the relatively low-throughput compared to RNASeq on the Illumina platform. The higher throughput and larger number of reads obtained

from Illumina sequencing allows for more stringent filtering of artefacts, as many of these will occur in <1 % of all reads. The higher error-rate of nanopore sequencing may limit interpretation of apparently novel splice junctions. However, the criteria set to identify these were high, including detection in more than one read per dataset and in more than 100 reads in total, with the presence of canonical splice sites.

In summary, my experiments have added to the databases describing the complex lytic transcription of HCMV. The ability to sequence unmodified RNA using the dRNA-Seq protocol on ONT sequencers avoids the artefacts associated with cDNA sequencing. However, disadvantages remain in the relatively low throughput and the inability to characterise transcript 5'-ends. One of the difficulties in interpreting the numerous novel transcripts described in my study was the lack of a comparable long-read dataset. Recent work to sequence the HCMV lytic transcriptome using a combination of short- and long-read datasets led to the development of the software LoRTIA (Long-read RNA-Seq Transcript Isoform Annotator toolkit, https://github.com/zsolt-balazs/LoRTIA) (Kakuk et al., 2021). This was used to characterise the lytic transcriptome of HCMV strain Towne, which is a highly passaged strain containing numerous, and in some instances drastic, mutations (Dolan et al., 2004, Kakuk et al., 2021). In future, it would be appropriate to analyse my datasets using LoRTIA and compare the results with those described above.

It is possible that many HCMV transcripts that appear not to be related directly to the expression of canonical genes may have regulatory roles (as proposed for antisense transcripts across the orthoherpesviruses (Bego et al., 2005, Zhang et al., 2007, Xu and Ganem, 2010)). It may also transpire that some novel transcripts encode hitherto unsuspected functional proteins (Xu and Ganem, 2010). Overall, the functional relevance of the complex transcriptome of HCMV revealed by advanced sequencing technologies remains to be determined.

7 Final summary

The study of HCMV genomes provides basic insights into the biology of viral replication, virulence factors and virus-host interactions, and enables the identification of potential antiviral targets. Historically, HCMV from clinical tissues required propagation in cell culture prior to the isolation and sequencing of viral DNA conducted directly or on PCR amplicons. However, HCMV genomes obtained by these means are not necessarily reflective of wild-type clinical strains because of mutations occurring in cell culture and biases imposed by the molecular techniques involved. The advent of advanced sequencing technologies, including short-read and long-read sequencing, has facilitated the direct sequencing of HCMV genomes from clinical samples and with a much higher throughput. I utilised the Illumina and ONT platforms to characterise HCMV genomes directly from clinical samples.

First, I demonstrated the capability of an Illumina protocol to sequence HCMV genomes directly from archived FFPE samples of tissues from cCMV infants by using extraction kits designed for this material, an established bait-based target enrichment protocol, and downstream data processing using the bioinformatics tool GRACy. I obtained complete HCMV genomes from five of ten cCMV cases, and genotyping of the sequence datasets identified a unique strain from each case. This confirmed the diversity of HCMV strains seen in previous studies. My findings reiterated the need to consider the geographical prevalence of HCMV genotypes when analysing inference studies of the effects of viral genetic variation on clinical outcomes. Importantly, evaluating and setting thresholds to avoid false variant discoveries will be required for investigating HCMV from FFPE samples to bring future studies to the standard set recently for HCMV (Camiolo et al., 2022). As cCMV is a relatively rare disease, FFPE repositories of pathology specimens have the potential to be a source of clinical material able to provide statistical power to investigations into whether there are associations between multiple-strain infection or individual genotypes with severe disease.

Next, I used the ONT platform to sequence the genomes of well-characterised high-titre cultured HCMV strains and established that the consensus error rate was on a par with that obtained using Illumina datasets. The much greater read lengths achievable using nanopore sequencing easily provided single-read

identification of the characteristic isomeric genome structure of HCMV, resolution of two strains in an artificial equimolar mix, and identification of potential recombination events occurring during infection by two strains in cell culture. The relative rarity of such events echoes previous findings suggesting that recombination during an individual co-infection is uncommon, although recombination has clearly played a significant role in shaping the HCMV genome on an evolutionary timescale.

Subsequently, I modified the nanopore protocol for use directly with clinical samples having high HCMV loads, with no *a priori* knowledge of the number of strains present or their sequences and no input from data for these strains obtained on other sequencing platforms. The analysis was aided by use of a large number of published HCMV genomes, which allowed the constituent strains in the samples to be identified readily. In each case, I used a hybrid bioinformatic approach involving RIA to identify the best matching published HCMV sequence and then RDA to assemble the clinical HCMV genomes. Moreover, I confirmed the accuracy of the nanopore sequences by independent Illumina sequencing of the same samples. In parallel work, I highlighted the pitfalls of using MDA for WGA of HCMV DNA present in low amounts in clinical samples, the most prominent of which was biased chimaeric DNA amplification. Excitingly, the implementation of a nanopore sequencing workflow to capture complete HCMV genomes from clinical samples has not been demonstrated previously and heralds the potential for using this instrument in a clinical setting, for example in the rapid detection of resistance-associated mutations.

Finally, I utilise the ONT dRNA-Seq protocol to sequence the native RNA populations in the HCMV lytic transcriptome. These datasets detected most of the splice junctions previously published from Illumina RNA-Seq experiments, and also identified many novel splice sites and multiply-spliced transcripts. The multiply-spliced transcripts included previously described introns together with novel ones. My analysis occupied the time available for investigating the HCMV transcriptome, and the datasets are likely to be useful in future investigations.

In summary, I utilised the latest available advanced sequencing platforms in my study of the clinically significant pathogen, HCMV. The pace at which technological advances have enabled sequencing at a higher throughput, greater

depth, faster turnaround and cheaper cost portends the transition of pathogen genomics and transcriptomics from a area focusing on research to one that can be used clinically on a relevant timescale to impact patient care. This is particularly the case for the portable sequencer, MinION. Additional improvements in the hardware and software in nanopore technology may, in time, thus facilitate the clinical management of HCMV infections. There is still much to be discovered in the large dsDNA genome of HCMV, and the combined efforts of long-read and more robust short-read sequencing is continuing to add to our understanding. The workflows and pipelines that I have established are intended to be a contribution to this endeavour.

Appendices

Table A1. SNPs identified using the GRACy SNP calling module on FFPE sequences.
All results are summarised for cases 35, 150, 184, 239 and 413 as output by GRACy, using Lofreq. (Section 3.5.2)

Dataset	Gene	Position in genome	Position in protein	Frequency (%)	Coverage	Nucleotide change	Codon Change	Amino acid change	C:G to T:A
35C_bc	RL1	1393	185	2.36	254	G → A	CGC → CAC	R→ H	y
35C_bc	RL13	11164	277	2.17	323	C → T	CAT → TAT	H→ Y	y
35C_bc	UL9	17261	560	2.51	358	C → T	GCC → GTC	A→ V	y
35C_bc	UL11	19065	685	1.71	350	G → A	GCC → ACC	A→ T	y
35C_gr	UL25	31445	1294	2.2	227	C → T	CGC → TGC	R→ C	y
35C_bc	UL35	46582	396	2.08	240	G → A	GCG → GCA	synonymous	y
35C_gr	UL45	59133	1272	2.62	229	C → T	AAC → AAT	synonymous	y
35C_gr	UL48	64970	602	2.38	210	T → C	CTT → CCT	L→ P	n
35C_gr	UL48	68640	4272	2.07	241	C → T	GGC → GGT	synonymous	y
35C_gr	UL48	70151	5783	2.43	247	C → T	ACG → ATG	T→ M	y
35C_bc	UL57	89843	1474	3.5	143	C → T	CGC → TGC	R→ C	y
35C_gr	UL69	99797	1684	1.87	268	C → A	CAA → AAA	Q→ K	n
35C_gr	UL74	108063	498	2.33	257	C → T	TGC → TGT	synonymous	y
35C_gr	UL86	126687	2630	2.86	245	G → A	CGT → CAT	R→ H	y
35C_bc	UL94	137973	600	2.62	267	C → T	TGC → TGT	synonymous	y
35C_bc	UL116	165962	591	4.03	149	C → T	GAC → GAT	synonymous	y
35C_bc	UL147A	179974	180	3.4	147	C → T	GCC → GCT	synonymous	y
35C_gr	UL139	186309	328	1.85	325	G → A	GCA → ACA	A→ T	y
35C_bc	IRS1	197912	2476	2.17	276	C → A	CAA → AAA	Q→ K	n
35C_gr	US15	209178	635	2.45	327	C → T	GCG → GTG	A→ V	y
35C_gr	US15	209798	15	3.11	289	A → G	AAA → AAG	synonymous	n
35C_gr	US22	216606	730	2.62	267	C → T	CGC → TGC	R→ C	y
35C_gr	US27	223863	223	1.7	352	G → A	GTA → ATA	V→ I	y
150C_gr	UL82	118375	1513	1.68	417	C→ T	CAG→ TAG	stop	y
150C_gr	UL98	143515	77	1.78	337	C→ T	CCC→ CTC	P→ L	y
150C_gr	UL102	147539	280	3.90	282	G→ A	GAG→ AAG	E→ K	y
150C_gr	UL102	149146	1887	4.56	263	C→ T	GGC→ GGT	synonymous	y
150C_gr	UL105	153548	879	1.51	463	C→ T	CGC→ CGT	synonymous	y
150C_gr	UL140	185406	317	1.75	456	C→ A	TCC→ TAC	S→ Y	n
150C_gr	UL148A	190362	144	1.44	487	C→ T	TGC→ TGT	C→ C	y
150C_bc	UL5	13936	28	1.25	638	G→ A	GAT→ AAT	D→ N	y
150C_bc	UL44	56170	1173	3.08	325	C→ A	GGC→ GGA	G→ G	n
150C_bc	UL52	74897	268	6.06	66	G→ A	GCC→ ACC	A→ T	y
150C_bc	UL57	89641	1481	7.14	84	G→ A	CGC→ CAC	R→ H	y
150C_bc	UL97	142390	1169	1.41	426	C→ T	ACG→ ATG	T→ M	y
150C_bc	UL102	147539	280	2.18	275	G→ A	GAG→ AAG	E→ K	y
150C_bc	UL105	154824	2155	1.56	384	T→ C	TTT→ CTT	F→ L	n
150C_bc	US2	198956	319	1.21	580	C→ T	CGG→ TGG	R→ W	y
150C_bc	US7	202607	4	1.50	668	C→ T	CGG→ TGG	R→ W	y
150C_bc	US8	202928	566	1.35	594	C→ T	GCG→ GTG	A→ V	y
150C_bc	US15	209634	15	1.61	373	A→ G	AAA→ AAG	K→ K	n
150C_bc	US16	209946	692	1.88	425	C→ T	TCG→ TTG	S→ L	y
150C_bc	US26	221381	1729	2.61	307	G→ A	GCC→ ACC	A→ T	y
150C_bc	US32	229122	160	1.39	433	G→ A	GCC→ ACC	A→ T	y
184C_bc	RL1	1561	263	2.65	189	G→ T	TGC→ TTC	C→ F	N
184C_bc	RL13	11585	549	1.57	383	C→ T	TAC→ TAT	Y→ Y	Y
184C_bc	UL4	13877	341	1.55	387	G→ A	CGC→ CAC	R→ H	Y
184C_gr	UL6	15245	565	0.94	1906	C→ G	CTG→ GTG	L→ V	N
184C_bc	UL9	17359	697	1.8	444	G→ A	GCT→ ACT	A→ T	Y
184C_gr	UL24	29233	836	1.65	846	G→ C	GGC→ GCC	G→ A	N
184C_gr	UL32	40654	2394	1.12	1072	G→ T	CCG→ CCT	P→ P	N
184C_bc	UL33	43887	649	2.51	199	G→ A	GTG→ ATG	V→ M	Y
184C_gr	UL44	56329	1173	1.39	721	C→ A	GGC→ GGA	G→ G	Y
184C_bc	UL44	56611	891	4.26	141	C→ T	GGC→ GGT	G→ G	Y
184C_gr	UL48	66695	2369	0.97	824	A→ T	CAG→ CTG	Q→ L	N
184C_bc	UL54	78200	3423	3.27	153	C→ T	GGC→ GGT	G→ G	Y
184C_bc	UL54	80957	666	2.7	185	C→ T	ATC→ ATT	I→ I	Y
184C_bc	UL54	81371	252	3.87	155	G→ A	CCG→ CCA	P→ P	Y
184C_gr	UL57	90171	1105	1.25	638	T→ C	TCG→ CCG	S→ P	N
184C_gr	UL69	99275	2127	2.32	734	G→ A	CCG→ CCA	P→ P	Y
184C_bc	UL69	99605	1797	4.14	145	C→ T	CCC→ CCT	P→ P	Y
184C_gr	UL69	100413	989	0.88	904	A→ G	TAC→ TGC	Y→ C	N
184C_bc	UL74	107820	668	2.02	346	G→ A	CGC→ CAC	R→ H	Y
184C_bc	UL74	107978	510	1.93	311	C→ T	TGC→ TGT	C→ C	Y
184C_bc	UL77	112330	455	10	40	C→ T	GCG→ GTG	A→ V	Y
184C_bc	UL85	124864	105	5.13	117	G→ A	CCG→ CCA	P→ P	Y
184C_gr	UL86	125845	3395	1.26	714	G→ T	CGG→ CTG	R→ L	N
184C_gr	UL86	127765	1475	1.62	557	C→ T	GCG→ GTG	A→ V	Y

184C_bc	UL88	132139	19	2.59	193	G→ A	GCT→ ACT	A→ T	Y
184C_bc	UL92	135590	507	3.19	188	G→ A	GCG→ GCA	A→ A	Y
184C_gr	UL99	145707	370	0.67	1193	C→ A	CAA→ AAA	Q→ K	N
184C_bc	UL123	172451	723	3.25	154	C→ T	TGC→ TGT	C→ C	Y
184C_gr	UL128	175952	74	0.56	1621	G→ A	GTA→ GAA	V→ E	Y
184C_gr	UL144	181830	506	0.65	1389	G→ A	GGC→ GAC	G→ D	Y
184C_gr	UL144	182305	31	0.8	1381	G→ T	GCT→ TCT	A→ S	N
184C_gr	UL135	189178	223	0.99	704	G→ C	GGG→ CGG	G→ R	N
184C_bc	IRS1	197611	1959	2.91	206	G→ A	CCG→ CCA	P→ P	Y
184C_bc	US2	199282	338	2.36	212	C→ T	CCG→ CTG	P→ L	Y
184C_bc	US11	205649	609	2.37	253	G→ A	CTG→ CTA	L→ L	Y
184C_gr	US15	209904	84	1.03	1166	G→ A	GCG→ GCA	A→ A	Y
184C_gr	US15	209906	82	0.76	1184	G→ C	GCG→ CCG	A→ P	N
184C_bc	US21	215244	360	2.2	227	G→ A	ACG→ ACA	T→ T	Y
184C_bc	US23	218934	469	2.78	180	G→ A	GTA→ ATA	V→ I	Y
184C_gr	US29	226394	60	0.87	1264	T→ C	GCT→ GCC	A→ A	N
184C_gr	US29	226433	99	0.73	1364	G→ A	CGG→ CGA	R→ R	Y
184C_gr	US29	226444	110	0.79	1394	C→ A	ACC→ AAC	T→ N	N
239C_bc	RL11	9232	444	4.88	82	T→ C	CGT→ CGC	synonomous	n
239C_bc	UL45	58950	1403	5.26	76	C→ T	GCT→ GTT	A→ V	y
239C_gr	UL49	72579	517	4.12	97	C→ T	CGG→ TGG	R→ W	y
239C_gr	UL53	77165	386	4.03	149	G→ A	CGC→ CAC	R→ H	y
239C_gr	UL104	152097	984	3.82	131	C→ A	TGC→ TGA	stop	n
239C_gr	UL105	154188	1281	3.64	110	C→ T	AGC→ AGT	synonomous	y
239C_gr	UL128	176501	457	3.4	147	C→ A	GTT→ ATT	V→ I	n
239C_bc	UL147	180263	272	36.36	99	A→ G	TAT→ TGT	Y→ C	n
239C_gr	UL147	180263	272	38.46	143	A→ G	TAT→ TGT	Y→ C	n
239C_gr	UL147	180357	178	4	175	C→ T	CCA→ TCA	P→ S	y
239C_gr	IRS1	198599	2503	4.44	90	C→ T	CGG→ TGG	R→ W	y
413A_gr	UL8	15903	470	1.87	268	C→ T	ACG→ ATG	T→ M	y
413A_gr	UL7	15903	470	1.87	268	C→ T	ACG→ ATG	T→ M	y
413A_gr	UL102	149238	1884	4.59	109	C→ T	GGC→ GGT	synonymous	y
413A_gr	US1	198113	449	4.44	90	C→ T	GCG→ GTG	A→ V	y
413C_gr	UL35	46461	442	13.79	29	G→ A	GAC→ AAC	D→ N	y
413C_gr	UL43	55406	630	12.5	32	C→ T	GGC→ GGT	synonymous	y
413C_gr	UL44	56639	745	9.3	43	G→ A	GAC→ AAC	D→ N	y
413C_gr	UL48	68477	4269	11.43	35	C→ T	GGC→ GGT	synonymous	y
413C_gr	UL50	73261	875	15.79	19	C→ T	GCG→ GTG	A→ V	y
413C_gr	UL55	82104	2257	16.67	24	T→ C	TTC→ CTC	F→ L	n
413C_gr	UL55	82162	2199	11.54	26	C→ T	GGC→ GGT	synonymous	y
413C_gr	UL56	85777	1096	12.2	41	G→ A	GCC→ ACC	A→ T	y
413C_gr	UL70	102020	2212	11.54	26	C→ T	CGC→ TGC	R→ C	y
413C_gr	UL72	105711	869	10.71	28	C→ T	TCG→ TTG	S→ L	y
413C_gr	UL74	107530	845	7.41	54	C→ T	ACA→ ATA	T→ I	y
413C_gr	UL76	112002	827	11.76	34	G→ A	CGT→ CAT	R→ H	y
413C_gr	UL77	112002	247	11.76	34	G→ A	GTG→ ATG	V→ M	y
413C_gr	UL82	119561	435	12	25	C→ T	AGC→ AGT	synonymous	y
413C_gr	UL86	125845	3285	12.9	31	C→ T	TGC→ TGT	synonymous	y
413C_gr	UL87	129385	197	10.34	29	C→ T	GCG→ GTG	A→ V	y
413C_gr	UL89	134084	350	8.16	49	T→ A	GAG→ GAG	synonymous	n
413C_gr	UL94	138126	940	13.79	29	C→ T	CCC→ TCC	P→ S	y
413C_gr	UL97	142527	1210	11.54	26	C→ T	CTG→ TTG	synonymous	y
413C_gr	UL102	148766	1412	12	25	C→ T	GCG→ GTG	A→ V	y
413C_gr	UL102	148971	1617	9.68	31	C→ T	TGC→ TGT	C→ C	y
413C_gr	UL105	155403	2644	15.63	32	A→ G	ATC→ GTC	I→ V	n
413C_gr	UL112	162462	884	18.75	16	G→ A	AGC→ AAC	S→ N	y
413C_gr	UL119	168789	645	8.7	46	G→ A	TTG→ TTA	synonymous	y
413C_gr	UL123	172012	1046	7.02	57	A→ G	AAG→ AGG	K→ R	n
413C_gr	UL128	175882	34	8.2	61	C→ T	GCG→ TCG	A→ S	y
413C_gr	US16	210774	63	10.81	37	C→ T	AGC→ AGT	S→ S	y
413C_gr	US20	214118	508	8.7	46	C→ T	CAG→ TAG	stop	y
413C_gr	US23	219014	237	8.82	34	G→ A	ATG→ ATA	M→ I	y
413C_gr	US24	220840	12	8	50	G→ A	CCG→ CCA	P→ P	y
413C_gr	US28	225284	327	8	50	C→ T	GCC→ GCT	synonymous	y

A2. List of 265 published HCMV genomes

Accession[a]	Strain
AY446894	Merlin
FJ527563	AD169
GQ221973	HAN13
GQ221974	3157
GQ221975	JP
GQ396662	HAN38
GQ396663	HAN20
GQ466044	3301
GU179288	U8
GU179289	VR1814
GU179290	U11
GU179291	AF1
HQ380895	JHC
JX512197	6397
JX512198	Davis
JX512199	HAN1
JX512200	HAN2
JX512201	HAN3
JX512202	HAN8
JX512203	HAN12
JX512204	HAN16
JX512205	HAN19
JX512206	HAN22
JX512207	HAN28
JX512208	HAN31
KC519319	BE/9/2010
KC519320	BE/10/2010
KC519321	BE/11/2010
KC519322	BE/21/2010
KC519323	BE/27/2010
KF021605	TR
KF297339	TB40/E
KJ361946	2CEN2
KJ361947	2CEN5
KJ361948	2CEN15
KJ361949	2CEN30
KJ361950	HAN11
KJ361951	HAN21
KJ361952	HAN27
KJ361953	HAN30
KJ361954	HAN32
KJ361955	HAN33
KJ361956	HAN36
KJ361957	HAN39
KJ361958	HAN40
KJ361959	PAV1
KJ361960	PAV4
KJ361961	PAV5
KJ361962	PAV6
KJ361963	PAV7
KJ361964	PAV8
KJ361965	PAV11
KJ361966	PAV12
KJ361967	PAV23
KJ361968	PAV24
KJ361969	PAV25
KJ361970	PAV26
KJ361971	UKNEQAS1
KJ426589	HAN
KJ872539	PAV16
KJ872540	PAV18

KJ872541	PAV20
KJ872542	PAV21
KP745633	BE/45/2011
KP745634	BE/32/2010
KP745635	BE/5/2012
KP745636	BE/7/2011
KP745637	BE/9/2011
KP745638	BE/15/2010
KP745639	BE/10/2011
KP745640	BE/22/2010
KP745641	BE/31/2011
KP745642	CZ/1/2012
KP745643	CZ/2/2012
KP745644	BE/31/2010
KP745645	BE/13/2010
KP745646	BE/8/2012
KP745647	BE/18/2010
KP745648	BE/8/2011
KP745649	BE/10/2012
KP745650	BE/1/2011
KP745651	BE/9/2012
KP745652	BE/2/2011
KP745653	BE/22/2011
KP745654	BE/19/2011
KP745655	BE/3/2010
KP745656	BE/2/2013
KP745657	BE/13/2011
KP745658	BE/14/2012
KP745659	BE/3/2011
KP745660	BE/6/2011
KP745661	BE/33/2010
KP745662	BE/20/2010
KP745663	BE/5/2010
KP745664	CZ/2/2013
KP745665	BE/16/2012
KP745666	BE/7/2012
KP745667	BE/5/2011
KP745668	BE/18/2011
KP745669	BE/28/2011
KP745670	BE/30/2011
KP745671	BE/14/2011
KP745672	BE/29/2011
KP745673	BE/42/2011
KP745674	BE/33/2011
KP745675	BE/23/2011
KP745676	BE/28/2010
KP745677	BE/1/2010
KP745678	BE/25/2010
KP745679	BE/24/2010
KP745680	BE/11/2012
KP745681	BE/43/2011
KP745682	BE/46/2011
KP745683	BE/12/2011
KP745684	BE/11/2011
KP745686	BE/39/2011
KP745687	BE/36/2011
KP745688	BE/12/2012
KP745689	BE/17/2011
KP745690	BE/34/2011
KP745691	CZ/1/2013
KP745692	BE/3/2012
KP745693	BE/15/2012
KP745694	BE/12/2010

KP745695	BE/6/2012
KP745696	BE/27/2011
KP745697	BE/23/2010
KP745698	BE/20/2011
KP745699	BE/1/2012
KP745700	BE/4/2011
KP745701	BE/6/2010
KP745702	BE/21/2011
KP745703	BE/26/2011
KP745704	BE/32/2011
KP745705	BE/38/2011
KP745706	BE/41/2011
KP745707	BE/13/2012
KP745708	BE/8/2010
KP745709	BE/48/2011
KP745710	BE/2/2012
KP745711	BE/24/2011
KP745712	BE/19/2010
KP745713	BE/35/2011
KP745714	BE/29/2010
KP745715	BE/44/2011
KP745716	BE/16/2010
KP745717	BE/2/2010
KP745718	CZ/1/2011
KP745719	BE/26/2010
KP745720	BE/15/2011
KP745721	BE/14/2010
KP745722	BE/40/2011
KP745723	BE/37/2011
KP745724	BE/4/2012
KP745725	BE/49/2011
KP745726	BE/30/2010
KP745727	BE/17/2010
KP745728	BE/4/2010
KR534196	JER847
KR534197	JER851
KR534198	JER893
KR534199	JER1070
KR534200	JER1289
KR534201	JER2002
KR534202	JER2282
KR534203	JER3230
KR534204	JER3855
KR534205	JER4035
KR534206	JER4041
KR534207	JER4053
KR534208	JER4559
KR534209	JER4755
KR534210	JER5268
KR534211	JER5409
KR534212	JER5550
KR534213	JER5695
KT634296	UKNEQAS2
KT726940	NL/Rot1/Urine/2012
KT726941	NL/Rot2/Urine/2012
KT726942	NL/Rot3/Nasal/2012
KT726943	NL/Rot4/Nasal/2012
KT726944	NL/Rot5/Urine/2012
KT726945	NL/Rot6/Nasal/2012
KT726946	NL/Rot7/Urine/2012
KT726947	UK/Lon1/Blood/2013
KT726948	UK/Lon2/Blood/2013
KT726949	UK/Lon6/Urine/2011

KT726950	UK/Lon7/Urine/2011
KT726951	UK/Lon8/Urine/2012
KT726952	UK/Lon3/Plasma/2012
KT726953	UK/Lon9/Urine/2012
KT726954	UK/Lon4/Bile/2011
KT726955	UK/Lon5/Blood/2010
KT959235	DB
KU550087	NAN1LA
KU550088	NAN2LA
KU550089	NAN4LA
KU550090	NANU
KY123649	HANChild4
KY123650	HANRTR2
KY123651	HANRTR4
KY123652	HANRTR5
KY123653	HANSCTR4
KY490061	PAV31
KY490062	PAV32
KY490063	PRA1
KY490064	PRA2
KY490065	PRA3
KY490066	PRA4
KY490067	PRA5
KY490068	PRA6
KY490069	PRA7
KY490070	PRA8
KY490071	HANChild1
KY490072	HANChild2&3
KY490073	HANRTR1A
KY490074	HANRTR1B
KY490075	HANRTR6
KY490076	HANRTR8
KY490077	HANRTR9
KY490078	HANRTR10
KY490079	HANSCTR1A
KY490080	HANSCTR1B
KY490081	HANSCTR2
KY490082	HANSCTR8
KY490083	HANSCTR9
KY490084	HANSCTR10
KY490085	HANSCTR11A
KY490086	HANSCTR11B
KY490087	HANSCTR12
KY490088	HANSCTR13
MF084223	LON1
MF084224	HER1
MK290742	LUS193
MK290743	LUS248
MK290744	LUS283
MK422176	LUS243
MN274568	Ig-KG-H2
MT044476	GLA-SOT1
MT044477	GLA-SOT4
MT044478	HAN-SOT1
MT044479	HAN-SOT5
MT044480	SYD-SCT2
MT044481	GLA-SOT2
MT044482	GLA-SOT3
MT044483	HAN-SOT3
MT044484	HAN-SOT4
MT044485	SYD-SCT1
MT070138	P4
MT070139	P6

MT070140	P10
MT070141	P14
MT070142	P15
MW197154	P-083
MW197155	P-225
MW197156	P-226
MW197157	P-232
LR131270	Pat_A
LR131940	Pat_C
LR131941	Pat_D
LR131942	Pat_E
LR131943	Pat_F
LR131944	Pat_G
LR131945	Pat_H
LR131946	Pat_K

[a] From GenBank except for the last eight, which are from the European Nucleotide Archive (ENA).

List of References

ABATE, D. A., WATANABE, S. & MOCARSKI, E. S. 2004. Major human cytomegalovirus structural protein pp65 (ppUL83) prevents interferon response factor 3 activation in the interferon response. *J Virol,* 78**,** 10995-1006.

ADLER, B., SCRIVANO, L., RUZCICS, Z., RUPP, B., SINZGER, C. & KOSZINOWSKI, U. 2006. Role of human cytomegalovirus UL131A in cell type-specific virus entry and release. *J Gen Virol,* 87**,** 2451-2460.

ADLER, S. P., MANGANELLO, A. M., LEE, R., MCVOY, M. A., NIXON, D. E., PLOTKIN, S., MOCARSKI, E., COX, J. H., FAST, P. E., NESTERENKO, P. A., MURRAY, S. E., HILL, A. B. & KEMBLE, G. 2016. A Phase 1 Study of 4 Live, Recombinant Human Cytomegalovirus Towne/Toledo Chimera Vaccines in Cytomegalovirus-Seronegative Men. *J Infect Dis,* 214**,** 1341-1348.

AKTER, P., CUNNINGHAM, C., MCSHARRY, B. P., DOLAN, A., ADDISON, C., DARGAN, D. J., HASSAN-WALKER, A. F., EMERY, V. C., GRIFFITHS, P. D., WILKINSON, G. W. G. & DAVISON, A. J. 2003. Two novel spliced genes in human cytomegalovirus. *J Gen Virol,* 84**,** 1117-1122.

ALCALAY, M., TOMASSONI, L., COLOMBO, E., STOLDT, S., GRIGNANI, F., FAGIOLI, M., SZEKELY, L., HELIN, K. & PELICCI, P. G. 1998. The promyelocytic leukemia gene product (PML) forms stable complexes with the retinoblastoma protein. *Mol Cell Biol,* 18**,** 1084-93.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *J Mol Biol,* 215**,** 403-10.

ANDERS, D. G., KACICA, M. A., PARI, G. & PUNTURIERI, S. M. 1992. Boundaries and structure of human cytomegalovirus oriLyt, a complex origin for lytic-phase DNA replication. *J Virol,* 66**,** 3373-84.

ARAV-BOGER, R. 2015. Strain Variation and disease severity in congenital CMV infection – in search of a viral marker. *Infectious disease clinics of North America,* 29**,** 401-414.

ARAV-BOGER, R., BOGER, Y. S., FOSTER, C. B. & BOGER, Z. 2008. The use of artificial neural networks in prediction of congenital CMV outcome from sequence data. *Bioinform Biol Insights,* 2**,** 281-9.

ARAV-BOGER, R., WILLOUGHBY, R. E., PASS, R. F., ZONG, J. C., JANG, W. J., ALCENDOR, D. & HAYWARD, G. S. 2002. Polymorphisms of the cytomegalovirus (CMV)-encoded tumor necrosis factor-alpha and beta-chemokine receptors in congenital CMV disease. *J Infect Dis,* 186**,** 1057-64.

ARAV-BOGER, R., ZONG, J. C. & FOSTER, C. B. 2005. Loss of linkage disequilibrium and accelerated protein divergence in duplicated cytomegalovirus chemokine genes. *Virus Genes,* 31**,** 65-72.

ARELLANO-GALINDO, J., VILLANUEVA-GARCÍA, D., CRUZ-RAMIREZ, J. L., YALAUPARI-MEJÌA, J. P., URIBE-GUTIÉRREZ, G., VELAZQUEZ-GUADARRAMA, N., NAVA-FRIAS, M., MUNOZ-HERNÁNDEZ, O. & MEJÍA-ARANGURE, J. M. 2014. Detection and gB genotyping of CMV in Mexican preterm infants in the context of maternal seropositivity. *J Infect Dev Ctries,* 8**,** 758-67.

ARNON, T. I., ACHDOUT, H., LEVI, O., MARKEL, G., SALEH, N., KATZ, G., GAZIT, R., GONEN-GROSS, T., HANNA, J., NAHARI, E., PORGADOR, A., HONIGMAN, A., PLACHTER, B., MEVORACH, D., WOLF, D. G. & MANDELBOIM, O. 2005. Inhibition of the NKp30 activating receptor by pp65 of human cytomegalovirus. *Nat Immunol,* 6**,** 515-23.

AUCOIN, D. P., SMITH, G. B., MEIERING, C. D. & MOCARSKI, E. S. 2006. Betaherpesvirus-conserved cytomegalovirus tegument protein ppUL32 (pp150) controls cytoplasmic events during virion maturation. *J Virol,* 80**,** 8199-210.

BALÁZS, Z., TOMBÁCZ, D., CSABAI, Z., MOLDOVÁN, N., SNYDER, M. & BOLDOGKŐI, Z. 2019. Template-switching artifacts resemble alternative polyadenylation. *BMC Genomics,* 20**,** 824.

BALÁZS, Z., TOMBÁCZ, D., SZŰCS, A., CSABAI, Z., MEGYERI, K., PETROV, A. N., SNYDER, M. & BOLDOGKŐI, Z. 2017. Long-Read Sequencing of Human Cytomegalovirus Transcriptome Reveals RNA Isoforms Carrying Distinct Coding Potentials. *Sci Rep,* 7**,** 15989.

BALÁZS, Z., TOMBÁCZ, D., SZŰCS, A., SNYDER, M. & BOLDOGKŐI, Z. 2018. Dual Platform Long-Read RNA-Sequencing Dataset of the Human Cytomegalovirus Lytic Transcriptome. *Front Genet,* 9**,** 432.

BALDANTI, F., SARASINI, A., FURIONE, M., GATTI, M., COMOLLI, G., REVELLO, M. G. & GERNA, G. 1998. Coinfection of the immunocompromised but not the immunocompetent host by multiple human cytomegalovirus strains. *Arch Virol,* 143**,** 1701-9.

BALDICK, C. J., MARCHINI, A., PATTERSON, C. E. & SHENK, T. 1997. Human cytomegalovirus tegument protein pp71 (ppUL82) enhances the infectivity of viral DNA and accelerates the infectious cycle. *J Virol,* 71**,** 4400-8.

BALDICK, C. J. & SHENK, T. 1996. Proteins associated with purified human cytomegalovirus particles. *J Virol,* 70**,** 6097-105.

BALE, J. F., MURPH, J. R., DEMMLER, G. J., DAWSON, J., MILLER, J. E. & PETHERAM, S. J. 2000. Intrauterine cytomegalovirus infection and glycoprotein B genotypes. *Journal of Infectious Diseases,* 182**,** 933-936.

BANKEVICH, A., NURK, S., ANTIPOV, D., GUREVICH, A. A., DVORKIN, M., KULIKOV, A. S., LESIN, V. M., NIKOLENKO, S. I., PHAM, S., PRJIBELSKI, A. D., PYSHKIN, A. V., SIROTKIN, A. V., VYAHHI, N., TESLER, G., ALEKSEYEV, M. A. & PEVZNER, P. A. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol,* 19**,** 455-77.

BARBI, M., BINDA, S., CAROPPO, S., PRIMACHE, V., DIDÒ, P., GUIDOTTI, P., CORBETTA, C. & MELOTTI, D. 2001. CMV gB genotypes and outcome of vertical transmission: study on dried blood spots of congenitally infected babies. *J Clin Virol,* 21**,** 75-9.

BECHTEL, J. T. & SHENK, T. 2002. Human cytomegalovirus UL47 tegument protein functions after entry and before immediate-early gene expression. *J Virol,* 76**,** 1043-50.

BEGO, M., MACIEJEWSKI, J., KHAIBOULLINA, S., PARI, G. & ST JEOR, S. 2005. Characterization of an antisense transcript spanning the UL81-82 locus of human cytomegalovirus. *J Virol,* 79**,** 11022-34.

BERG, C., ROSENKILDE, M. M., BENFIELD, T., NIELSEN, L., SUNDELIN, T. & LÜTTICHAU, H. R. 2021. The frequency of cytomegalovirus non-ELR UL146 genotypes in neonates with congenital CMV disease is comparable to strains in the background population. *BMC Infect Dis,* 21**,** 386.

BHAGWATE, A. V., LIU, Y., WINHAM, S. J., MCDONOUGH, S. J., STALLINGS-MANN, M. L., HEINZEN, E. P., DAVILA, J. I., VIERKANT, R. A., HOSKIN, T. L., FROST, M., CARTER, J. M., RADISKY, D. C., CUNNINGHAM, J. M., DEGNIM, A. C. & WANG, C. 2019. Bioinformatics and DNA-extraction strategies to reliably detect genetic variants from FFPE breast tissue samples. *BMC Genomics,* 20**,** 689.

BHELLA, D., RIXON, F. J. & DARGAN, D. J. 2000. Cryomicroscopy of human cytomegalovirus virions reveals more densely packed genomic DNA than in herpes simplex virus type 1. *J Mol Biol,* 295**,** 155-61.

BIRON, K. K., HARVEY, R. J., CHAMBERLAIN, S. C., GOOD, S. S., SMITH, A. A., DAVIS, M. G., TALARICO, C. L., MILLER, W. H., FERRIS, R., DORNSIFE, R. E., STANAT, S. C., DRACH, J. C., TOWNSEND, L. B. & KOSZALKA, G. W. 2002. Potent and selective inhibition of human cytomegalovirus replication by 1263W94, a benzimidazole L-riboside with a unique mode of action. *Antimicrob Agents Chemother,* 46**,** 2365-72.

BOECKH, M., NICHOLS, W. G., PAPANICOLAOU, G., RUBIN, R., WINGARD, J. R. & ZAIA, J. 2003. Cytomegalovirus in hematopoietic stem cell transplant recipients: Current status, known challenges, and future strategies. *Biol Blood Marrow Transplant,* 9**,** 543-58.

BOETZER, M. & PIROVANO, W. 2012. Toward almost closed genomes with GapFiller. *Genome Biol,* 13**,** R56.

BOLOGNESI, C., FORCATO, C., BUSON, G., FONTANA, F., MANGANO, C., DOFFINI, A., SERO, V., LANZELLOTTO, R., SIGNORINI, G., CALANCA, A., SERGIO, M., ROMANO, R., GIANNI, S., MEDORO, G., GIORGINI, G., MORREAU, H., BARBERIS, M., CORVER, W. E. & MANARESI, N. 2016. Digital Sorting of Pure Cell Populations Enables Unambiguous Genetic Analysis of Heterogeneous Formalin-Fixed Paraffin-Embedded Tumors by Next Generation Sequencing. *Sci Rep,* 6**,** 20944.

BOOM, R., SOL, C. J., SCHUURMAN, T., VAN BREDA, A., WEEL, J. F., BELD, M., TEN BERGE, I. J., WERTHEIM-VAN DILLEN, P. M. & DE JONG, M. D. 2002. Human cytomegalovirus DNA in plasma and serum specimens of renal transplant recipients is highly fragmented. *J Clin Microbiol,* 40**,** 4105-13.

BOPPANA, S. B., RIVERA, L. B., FOWLER, K. B., MACH, M. & BRITT, W. J. 2001. Intrauterine transmission of cytomegalovirus to infants of women with preconceptional immunity. *N Engl J Med,* 344**,** 1366-71.

BORST, E. M., BAUERFEIND, R., BINZ, A., STEPHAN, T. M., NEUBER, S., WAGNER, K., STEINBRÜCK, L., SODEIK, B., LENAC ROVIŠ, T., JONJIĆ, S. & MESSERLE, M. 2016. The Essential Human Cytomegalovirus Proteins pUL77 and pUL93 Are Structural Components Necessary for Viral Genome Encapsidation. *J Virol,* 90**,** 5860-5875.

BORST, E. M., WAGNER, K., BINZ, A., SODEIK, B. & MESSERLE, M. 2008. The essential human cytomegalovirus gene UL52 is required for cleavage-packaging of the viral genome. *J Virol,* 82**,** 2065-78.

BOYLE, K. A. & COMPTON, T. 1998. Receptor-binding properties of a soluble form of human cytomegalovirus glycoprotein B. *J Virol,* 72**,** 1826-33.

BRADLEY, A. J., KOVÁCS, I. J., GATHERER, D., DARGAN, D. J., ALKHARSAH, K. R., CHAN, P. K., CARMAN, W. F., DEDICOAT, M., EMERY, V. C., GEDDES, C. C., GERNA, G., BEN-ISMAEIL, B., KAYE, S., MCGREGOR, A., MOSS, P. A., PUSZTAI, R., RAWLINSON, W. D., SCOTT, G. M., WILKINSON, G. W., SCHULZ, T. F. & DAVISON, A. J. 2008. Genotypic analysis of two hypervariable human cytomegalovirus genes. *J Med Virol,* 80**,** 1615-23.

BRADLEY, A. J., LURAIN, N. S., GHAZAL, P., TRIVEDI, U., CUNNINGHAM, C., BALUCHOVA, K., GATHERER, D., WILKINSON, G. W., DARGAN, D. J. & DAVISON, A. J. 2009. High-throughput sequence analysis of variants of human cytomegalovirus strains Towne and AD169. *J Gen Virol,* 90**,** 2375-80.

BRITT, W. J., VUGLER, L., BUTFILOSKI, E. J. & STEPHENS, E. B. 1990. Cell surface expression of human cytomegalovirus (HCMV) gp55-116 (gB): use

of HCMV-recombinant vaccinia virus-infected cells in analysis of the human neutralizing antibody response. *J Virol,* 64**,** 1079-85.

BROWN, A. C. & CHRISTIANSEN, M. T. 2017. Whole-Genome Enrichment Using RNA Probes and Sequencing of Chlamydia trachomatis Directly from Clinical Samples. *Methods Mol Biol,* 1616**,** 1-22.

BUCHFINK, B., XIE, C. & HUSON, D. H. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods,* 12**,** 59-60.

BUTCHER, S. J., AITKEN, J., MITCHELL, J., GOWEN, B. & DARGAN, D. J. 1998. Structure of the human cytomegalovirus B capsid by electron cryomicroscopy and image reconstruction. *J Struct Biol,* 124**,** 70-6.

CAMIOLO, S., HUGHES, J., BALDANTI, F., FURIONE, M., LILLERI, D., LOMBARDI, G., ANGELINI, M., GERNA, G., ZAVATTONI, M., DAVISON, A. J. & SUÁREZ, N. M. 2022. Identifying high-confidence variants in human cytomegalovirus genomes sequenced from clinical samples. *Virus Evolution,* 8.

CAMIOLO, S., SUÁREZ, N. M., CHALKA, A., VENTURINI, C., BREUER, J. & DAVISON, A. J. 2021. GRACy: A tool for analysing human cytomegalovirus sequence data. *Virus Evol,* 7**,** veaa099.

CANTRELL, S. R. & BRESNAHAN, W. A. 2005. Interaction between the human cytomegalovirus UL82 gene product (pp71) and hDaxx regulates immediate-early gene expression and viral replication. *J Virol,* 79**,** 7792-802.

CARLSSON, J., DAVIDSSON, S., FRIDFELDT, J., GIUNCHI, F., FIANO, V., GRASSO, C., ZELIC, R., RICHIARDI, L., ANDRÉN, O., PETTERSSON, A., FIORENTINO, M. & AKRE, O. 2018. Quantity and quality of nucleic acids extracted from archival formalin fixed paraffin embedded prostate biopsies. *BMC Med Res Methodol,* 18**,** 161.

CHA, T. A., TOM, E., KEMBLE, G. W., DUKE, G. M., MOCARSKI, E. S. & SPAETE, R. R. 1996. Human cytomegalovirus clinical isolates carry at least 19 genes not found in laboratory strains. *J Virol,* 70**,** 78-83.

CHAMBERS, J., ANGULO, A., AMARATUNGA, D., GUO, H., JIANG, Y., WAN, J. S., BITTNER, A., FRUEH, K., JACKSON, M. R., PETERSON, P. A., ERLANDER, M. G. & GHAZAL, P. 1999. DNA microarrays of the complex human cytomegalovirus genome: profiling kinetic class with drug sensitivity of viral gene expression. *J Virol,* 73**,** 5757-66.

CHEE, M. S., BANKIER, A. T., BECK, S., BOHNI, R., BROWN, C. M., CERNY, R., HORSNELL, T., HUTCHISON, C. A., KOUZARIDES, T. & MARTIGNETTI, J. A. 1990. Analysis of the protein-coding content of the sequence of human cytomegalovirus strain AD169. *Curr Top Microbiol Immunol,* 154**,** 125-69.

CHEN, D. H., JIANG, H., LEE, M., LIU, F. & ZHOU, Z. H. 1999. Three-dimensional visualization of tegument/capsid interactions in the intact human cytomegalovirus. *Virology,* 260**,** 10-6.

CHEN, G., MOSIER, S., GOCKE, C. D., LIN, M. T. & ESHLEMAN, J. R. 2014. Cytosine deamination is a major cause of baseline noise in next-generation sequencing. *Mol Diagn Ther,* 18**,** 587-93.

CHENG, S., CAVINESS, K., BUEHLER, J., SMITHEY, M., NIKOLICH-ŽUGICH, J. & GOODRUM, F. 2017. Transcriptome-wide characterization of human cytomegalovirus in natural infection and experimental latency. *Proc Natl Acad Sci U S A,* 114**,** E10586-E10595.

CHERN, K. C., CHANDLER, D. B., MARTIN, D. F., KUPPERMANN, B. D., WOLITZ, R. A. & MARGOLIS, T. P. 1998. Glycoprotein B subtyping of cytomegalovirus (CMV) in the vitreous of patients with AIDS and CMV retinitis. *J Infect Dis,* 178**,** 1149-53.

CHEUNG, A. K., GOTTLIEB, D. J., PLACHTER, B., PEPPERL-KLINDWORTH, S., AVDIC, S., CUNNINGHAM, A. L., ABENDROTH, A. & SLOBEDMAN, B. 2009. The role of the human cytomegalovirus UL111A gene in down-regulating CD4+ T-cell recognition of latently infected cells: implications for virus elimination during latency. *Blood,* 114**,** 4128-37.

CHEVILLOTTE, M., VON EINEM, J., MEIER, B. M., LIN, F. M., KESTLER, H. A. & MERTENS, T. 2010. A new tool linking human cytomegalovirus drug resistance mutations to resistance phenotypes. *Antiviral Res,* 85**,** 318-27.

CHILD, S. J., HAKKI, M., DE NIRO, K. L. & GEBALLE, A. P. 2004. Evasion of cellular antiviral responses by human cytomegalovirus TRS1 and IRS1. *J Virol,* 78**,** 197-205.

CHORLTON, S. D., RITCHIE, G., LAWSON, T., MCLACHLAN, E., ROMNEY, M. G., MATIC, N. & LOWE, C. F. 2021. Next-generation sequencing for cytomegalovirus antiviral resistance genotyping in a clinical virology laboratory. *Antiviral Res,* 192**,** 105123.

CHOU, S. W. 1989. Reactivation and recombination of multiple cytomegalovirus strains from individual organ donors. *J Infect Dis,* 160**,** 11-5.

CHOU, S. W. & DENNISON, K. M. 1991. Analysis of interstrain variation in cytomegalovirus glycoprotein B sequences encoding neutralization-related epitopes. *J Infect Dis,* 163**,** 1229-34.

CICIN-SAIN, L., PODLECH, J., MESSERLE, M., REDDEHASE, M. J. & KOSZINOWSKI, U. H. 2005. Frequent coinfection of cells explains functional in vivo complementation between cytomegalovirus variants in the multiply infected host. *J Virol,* 79**,** 9492-502.

COAQUETTE, A., BOURGEOIS, A., DIRAND, C., VARIN, A., CHEN, W. & HERBEIN, G. 2004. Mixed cytomegalovirus glycoprotein B genotypes in immunocompromised patients. *Clin Infect Dis,* 39**,** 155-61.

COCQUET, J., CHONG, A., ZHANG, G. & VEITIA, R. A. 2006. Reverse transcriptase template switching and false alternative transcripts. *Genomics,* 88**,** 127-31.

CUNNINGHAM, C., GATHERER, D., HILFRICH, B., BALUCHOVA, K., DARGAN, D. J., THOMSON, M., GRIFFITHS, P. D., WILKINSON, G. W., SCHULZ, T. F. & DAVISON, A. J. 2010. Sequences of complete human cytomegalovirus genomes from infected cell cultures and clinical specimens. *J Gen Virol,* 91**,** 605-15.

DAI, X., YU, X., GONG, H., JIANG, X., ABENES, G., LIU, H., SHIVAKOTI, S., BRITT, W. J., ZHU, H., LIU, F. & ZHOU, Z. H. 2013. The smallest capsid protein mediates binding of the essential tegument protein pp150 to stabilize DNA-containing capsids in human cytomegalovirus. *PLoS Pathog,* 9**,** e1003525.

DARGAN, D. J., DOUGLAS, E., CUNNINGHAM, C., JAMIESON, F., STANTON, R. J., BALUCHOVA, K., MCSHARRY, B. P., TOMASEC, P., EMERY, V. C., PERCIVALLE, E., SARASINI, A., GERNA, G., WILKINSON, G. W. & DAVISON, A. J. 2010. Sequential mutations associated with adaptation of human cytomegalovirus to growth in cell culture. *J Gen Virol,* 91**,** 1535-46.

DAVISON, A. J., AKTER, P., CUNNINGHAM, C., DOLAN, A., ADDISON, C., DARGAN, D. J., HASSAN-WALKER, A. F., EMERY, V. C., GRIFFITHS, P. D. & WILKINSON, G. W. 2003a. Homology between the human cytomegalovirus RL11 gene family and human adenovirus E3 genes. *J Gen Virol,* 84**,** 657-63.

DAVISON, A. J., DOLAN, A., AKTER, P., ADDISON, C., DARGAN, D. J., ALCENDOR, D. J., MCGEOCH, D. J. & HAYWARD, G. S. 2003b. The human

cytomegalovirus genome revisited: comparison with the chimpanzee cytomegalovirus genome. *J Gen Virol,* 84**,** 17-28.

DAVISON, A. J., HOLTON, M., DOLAN, A., DARGAN, D. J., GATHERER, D. & HAYWARD, G. S. 2013. Comparative Genomics of Primate Cytomegaloviruses. *In:* REDDEHASE, M. J. (ed.) *Cytomegaloviruses: from Molecular Pathogenesis to Intervention.* 2nd ed. Malta: Caister Academic Press.

DE LA CRUZ-DE LA CRUZ, A., MORENO-VERDUZCO, E. R., MARTÍNEZ-ALARCÓN, O., GONZÁLEZ-ALVAREZ, D. L., VALDESPINO-VÁZQUEZ, M. Y., HELGUERA-REPETTO, A. C., FONSECA-CORONADO, S., LOZANO-CUENCA, J., RAMÍREZ-RAMÍREZ, A., SORIANO-BECERRIL, D., MANCILLA-HERRERA, I., FIGUEROA-DAMIÁN, R. & HERRERA-SALAZAR, A. 2020. Utility of two DNA extraction methods using formalin-fixed paraffin-embedded tissues in identifying congenital cytomegalovirus infection by polymerase chain reaction. *Diagn Microbiol Infect Dis,* 97**,** 115075.

DE VRIES, J. J., WESSELS, E., KORVER, A. M., VAN DER EIJK, A. A., RUSMAN, L. G., KROES, A. C. & VOSSEN, A. C. 2012. Rapid genotyping of cytomegalovirus in dried blood spots by multiplex real-time PCR assays targeting the envelope glycoprotein gB and gH genes. *J Clin Microbiol,* 50**,** 232-7.

DEAN, F. B., HOSONO, S., FANG, L., WU, X., FARUQI, A. F., BRAY-WARD, P., SUN, Z., ZONG, Q., DU, Y., DU, J., DRISCOLL, M., SONG, W., KINGSMORE, S. F., EGHOLM, M. & LASKEN, R. S. 2002. Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A,* 99**,** 5261-6.

DELAHAYE, C. & NICOLAS, J. 2021. Sequencing DNA with nanopores: Troubles and biases. *PLoS One,* 16**,** e0257521.

DEPLEDGE, D. P., MOHR, I. & WILSON, A. C. 2019a. Going the Distance: Optimizing RNA-Seq Strategies for Transcriptomic Analysis of Complex Viral Genomes. *J Virol,* 93.

DEPLEDGE, D. P., SRINIVAS, K. P., SADAOKA, T., BREADY, D., MORI, Y., PLACANTONAKIS, D. G., MOHR, I. & WILSON, A. C. 2019b. Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nat Commun,* 10**,** 754.

DO, H. & DOBROVIC, A. 2012. Dramatic reduction of sequence artefacts from DNA isolated from formalin-fixed cancer biopsies by treatment with uracil- DNA glycosylase. *Oncotarget,* 3**,** 546-58.

DO, H. & DOBROVIC, A. 2015. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. *Clin Chem,* 61**,** 64-71.

DOLAN, A., CUNNINGHAM, C., HECTOR, R. D., HASSAN-WALKER, A. F., LEE, L., ADDISON, C., DARGAN, D. J., MCGEOCH, D. J., GATHERER, D., EMERY, V. C., GRIFFITHS, P. D., SINZGER, C., MCSHARRY, B. P., WILKINSON, G. W. & DAVISON, A. J. 2004. Genetic content of wild-type human cytomegalovirus. *J Gen Virol,* 85**,** 1301-12.

DOLLARD, S. C., GROSSE, S. D. & ROSS, D. S. 2007. New estimates of the prevalence of neurological and sensory sequelae and mortality associated with congenital cytomegalovirus infection. *Rev Med Virol,* 17**,** 355-63.

DREW, W. L., CHOU, S., MINER, R. C., MOHR, B. A., BUSCH, M. P., VAN DER HORST, C. M., ASMUTH, D. M. & KALISH, L. A. 2002. Cytomegalovirus glycoprotein B groups in human immunodeficiency virus-infected patients with incident retinitis. *J Infect Dis,* 186**,** 114-7.

DUNN, W., CHOU, C., LI, H., HAI, R., PATTERSON, D., STOLC, V., ZHU, H. & LIU, F. 2003. Functional profiling of a human cytomegalovirus genome. *Proc Natl Acad Sci U S A,* 100**,** 14223-8.

ECKERT, S. E., CHAN, J. Z., HOUNIET, D., BREUER, J., SPEIGHT, G. & PATHSEEK CONSORTIUM 2016. Enrichment by hybridisation of long DNA fragments for Nanopore sequencing. *Microb Genom,* 2**,** e000087.

ELEK, S. D. & STERN, H. 1974. Development of a vaccine against mental retardation caused by cytomegalovirus infection in utero. *Lancet,* 1**,** 1-5.

ENDERS, J. F., WELLER, T. H. & ROBBINS, F. C. 1949. Cultivation of the Lansing Strain of Poliomyelitis Virus in Cultures of Various Human Embryonic Tissues. *Science,* 109**,** 85-7.

ERGUNER, B., USTEK, D. & SAGIROGLU, M. S. 2015. Performance comparison of Next Generation sequencing platforms. *Conf Proc IEEE Eng Med Biol Soc,* 2015**,** 6453-6.

FANG, W., MORI, T. & COBRINIK, D. 2002. Regulation of PML-dependent transcriptional repression by pRB and low penetrance pRB mutants. *Oncogene,* 21**,** 5557-65.

FEIRE, A. L., ROY, R. M., MANLEY, K. & COMPTON, T. 2010. The glycoprotein B disintegrin-like domain binds beta 1 integrin to mediate cytomegalovirus entry. *J Virol,* 84**,** 10026-37.

FIDOUH-HOUHOU, N., DUVAL, X., BISSUEL, F., BOURBONNEUX, V., FLANDRE, P., ECOBICHON, J. L., JORDAN, M. C., VILDÉ, J. L., BRUN-VÉZINET, F. & LEPORT, C. 2001. Salivary cytomegalovirus (CMV) shedding, glycoprotein B genotype distribution, and CMV disease in human immunodeficiency virus-seropositive patients. *Clin Infect Dis,* 33**,** 1406-11.

FOLKINS, A. K., CHISHOLM, K. M., GUO, F. P., MCDOWELL, M., AZIZ, N. & PINSKY, B. A. 2013. Diagnosis of congenital CMV using PCR performed on formalin-fixed, paraffin-embedded placental tissue. *Am J Surg Pathol,* 37**,** 1413-20.

FRIES, B. C., CHOU, S., BOECKH, M. & TOROK-STORB, B. 1994. Frequency distribution of cytomegalovirus envelope glycoprotein genotypes in bone marrow transplant recipients. *J Infect Dis,* 169**,** 769-74.

GATHERER, D., DEPLEDGE, D. P., HARTLEY, C. A., SZPARA, M. L., VAZ, P. K., BENKŐ, M., BRANDT, C. R., BRYANT, N. A., DASTJERDI, A., DOSZPOLY, A., GOMPELS, U. A., INOUE, N., JAROSINSKI, K. W., KAUL, R., LACOSTE, V., NORBERG, P., ORIGGI, F. C., ORTON, R. J., PELLETT, P. E., SCHMID, D. S., SPATZ, S. J., STEWART, J. P., TRIMPERT, J., WALTZEK, T. B. & DAVISON, A. J. 2021. ICTV Virus Taxonomy Profile: *Herpesviridae* 2021. *J Gen Virol,* 102.

GATHERER, D., SEIRAFIAN, S., CUNNINGHAM, C., HOLTON, M., DARGAN, D. J., BALUCHOVA, K., HECTOR, R. D., GALBRAITH, J., HERZYK, P., WILKINSON, G. W. & DAVISON, A. J. 2011. High-resolution human cytomegalovirus transcriptome. *Proc Natl Acad Sci U S A,* 108**,** 19755-60.

GEISZT, M., LEKSTROM, K. & LETO, T. L. 2004. Analysis of mRNA transcripts from the NAD(P)H oxidase 1 (Nox1) gene. Evidence against production of the NADPH oxidase homolog-1 short (NOH-1S) transcript variant. *J Biol Chem,* 279**,** 51661-8.

GIBSON, W., BAXTER, M. K. & CLOPPER, K. S. 1996a. Cytomegalovirus "missing" capsid protein identified as heat-aggregable product of human cytomegalovirus UL46. *J Virol,* 70**,** 7454-61.

GIBSON, W., CLOPPER, K. S., BRITT, W. J. & BAXTER, M. K. 1996b. Human cytomegalovirus (HCMV) smallest capsid protein identified as product of

short open reading frame located between HCMV UL48 and UL49. *J Virol,* 70**,** 5680-3.

GILBERT, M. J., RIDDELL, S. R., PLACHTER, B. & GREENBERG, P. D. 1996. Cytomegalovirus selectively blocks antigen processing and presentation of its immediate-early gene product. *Nature,* 383**,** 720-2.

GILBERT, M. T., HASELKORN, T., BUNCE, M., SANCHEZ, J. J., LUCAS, S. B., JEWELL, L. D., VAN MARCK, E. & WOROBEY, M. 2007. The isolation of nucleic acids from fixed, paraffin-embedded tissues-which methods are useful when? *PLoS One,* 2**,** e537.

GOLDNER, T., HEMPEL, C., RUEBSAMEN-SCHAEFF, H., ZIMMERMANN, H. & LISCHKA, P. 2014. Geno- and phenotypic characterization of human cytomegalovirus mutants selected in vitro after letermovir (AIC246) exposure. *Antimicrob Agents Chemother,* 58**,** 610-3.

GOODWIN, S., MCPHERSON, J. D. & MCCOMBIE, W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet,* 17**,** 333-51.

GORZER, I., GUELLY, C., TRAJANOSKI, S. & PUCHHAMMER-STOCKL, E. 2010a. Deep sequencing reveals highly complex dynamics of human cytomegalovirus genotypes in transplant patients over time. *J Virol,* 84**,** 7195-203.

GÖRZER, I., GUELLY, C., TRAJANOSKI, S. & PUCHHAMMER-STÖCKL, E. 2010. The impact of PCR-generated recombination on diversity estimation of mixed viral populations by deep sequencing. *J Virol Methods,* 169**,** 248-52.

GORZER, I., KERSCHNER, H., REDLBERGER-FRITZ, M. & PUCHHAMMER-STOCKL, E. 2010b. Human cytomegalovirus (HCMV) genotype populations in immunocompetent individuals during primary HCMV infection. *J Clin Virol,* 48**,** 100-3.

GÖRZER, I., TRAJANOSKI, S., POPOW-KRAUPP, T. & PUCHHAMMER-STÖCKL, E. 2015. Analysis of human cytomegalovirus strain populations in urine samples of newborns by ultra deep sequencing. *J Clin Virol,* 73**,** 101-104.

GOVENDER, K., PARBOOSING, R., CAMIOLO, S., HUBÁČEK, P., GÖRZER, I., PUCHHAMMER-STÖCKL, E. & SUÁREZ, N. M. 2022. Complexity of Human Cytomegalovirus Infection in South African HIV-Exposed Infants with Pneumonia. *Viruses,* 14.

GRIFFITHS, P. D., STANTON, A., MCCARRELL, E., SMITH, C., OSMAN, M., HARBER, M., DAVENPORT, A., JONES, G., WHEELER, D. C., O'BEIRNE, J., THORBURN, D., PATCH, D., ATKINSON, C. E., PICHON, S., SWENY, P., LANZMAN, M., WOODFORD, E., ROTHWELL, E., OLD, N., KINYANJUI, R., HAQUE, T., ATABANI, S., LUCK, S., PRIDEAUX, S., MILNE, R. S., EMERY, V. C. & BURROUGHS, A. K. 2011. Cytomegalovirus glycoprotein-B vaccine with MF59 adjuvant in transplant recipients: a phase 2 randomised placebo-controlled trial. *Lancet,* 377**,** 1256-63.

HABERLAND, M., MEYER-KÖNIG, U. & HUFERT, F. T. 1999. Variation within the glycoprotein B gene of human cytomegalovirus is due to homologous recombination. *J Gen Virol,* 80 (Pt 6)**,** 1495-500.

HAGE, E., WILKIE, G. S., LINNENWEBER-HELD, S., DHINGRA, A., SUÁREZ, N. M., SCHMIDT, J. J., KAY-FEDOROV, P. C., MISCHAK-WEISSINGER, E., HEIM, A., SCHWARZ, A., SCHULZ, T. F., DAVISON, A. J. & GANZENMUELLER, T. 2017. Characterization of Human Cytomegalovirus Genome Diversity in Immunocompromised Hosts by Whole-Genome Sequencing Directly From Clinical Specimens. *J Infect Dis,* 215**,** 1673-1683.

HAHN, G., JORES, R. & MOCARSKI, E. S. 1998. Cytomegalovirus remains latent in a common precursor of dendritic and myeloid cells. *Proc Natl Acad Sci U S A,* 95**,** 3937-42.

HAHN, G., REVELLO, M. G., PATRONE, M., PERCIVALLE, E., CAMPANINI, G., SARASINI, A., WAGNER, M., GALLINA, A., MILANESI, G., KOSZINOWSKI, U., BALDANTI, F. & GERNA, G. 2004. Human cytomegalovirus UL131-128 genes are indispensable for virus growth in endothelial cells and virus transfer to leukocytes. *J Virol,* 78**,** 10023-33.

HAILE, S., CORBETT, R. D., BILOBRAM, S., BYE, M. H., KIRK, H., PANDOH, P., TRINH, E., MACLEOD, T., MCDONALD, H., BALA, M., MILLER, D., NOVIK, K., COOPE, R. J., MOORE, R. A., ZHAO, Y., MUNGALL, A. J., MA, Y., HOLT, R. A., JONES, S. J. & MARRA, M. A. 2019. Sources of erroneous sequences and artifact chimeric reads in next generation sequencing of genomic DNA from formalin-fixed paraffin-embedded samples. *Nucleic Acids Res,* 47**,** e12.

HASHIMOTO, Y., SHENG, X., MURRAY-NERGER, L. A. & CRISTEA, I. M. 2020. Temporal dynamics of protein complex formation and dissociation during human cytomegalovirus infection. *Nat Commun,* 11**,** 806.

HOFMANN, H., SINDRE, H. & STAMMINGER, T. 2002. Functional interaction between the pp71 protein of human cytomegalovirus and the PML-interacting protein human Daxx. *J Virol,* 76**,** 5769-83.

HOLLENBACH, A. D., MCPHERSON, C. J., MIENTJES, E. J., IYENGAR, R. & GROSVELD, G. 2002. Daxx and histone deacetylase II associate with chromatin through an interaction with core histones and the chromatin-associated protein Dek. *J Cell Sci,* 115**,** 3319-30.

HOSONO, S., FARUQI, A. F., DEAN, F. B., DU, Y., SUN, Z., WU, X., DU, J., KINGSMORE, S. F., EGHOLM, M. & LASKEN, R. S. 2003. Unbiased whole-genome amplification directly from clinical samples. *Genome Res,* 13**,** 954-64.

HUBER, M. T. & COMPTON, T. 1998. The human cytomegalovirus UL74 gene encodes the third component of the glycoprotein H-glycoprotein L-containing envelope complex. *J Virol,* 72**,** 8191-7.

IBANEZ, C. E., SCHRIER, R., GHAZAL, P., WILEY, C. & NELSON, J. A. 1991. Human cytomegalovirus productively infects primary differentiated macrophages. *J Virol,* 65**,** 6581-8.

IKEGAMI, M., KOHSAKA, S., HIROSE, T., UENO, T., INOUE, S., KANOMATA, N., YAMAUCHI, H., MORI, T., SEKINE, S., INAMOTO, Y., YATABE, Y., KOBAYASHI, H., TANAKA, S. & MANO, H. 2021. MicroSEC filters sequence errors for formalin-fixed and paraffin-embedded samples. *Commun Biol,* 4**,** 1396.

ISAACSON, M. K. & COMPTON, T. 2009a. Human cytomegalovirus glycoprotein B is required for virus entry and cell-to-cell spread but not for virion attachment, assembly, or egress. *J Virol,* 83**,** 3891-903.

ISAACSON, M. K. & COMPTON, T. 2009b. Human Cytomegalovirus Glycoprotein B Is Required for Virus Entry and Cell-to-Cell Spread but Not for Virion Attachment, Assembly, or Egress. *Journal of Virology,* 83**,** 3891-3903.

ISAACSON, M. K., FEIRE, A. L. & COMPTON, T. 2007. Epidermal growth factor receptor is not required for human cytomegalovirus entry or signaling. *J Virol,* 81**,** 6241-7.

ISHOV, A. M., VLADIMIROVA, O. V. & MAUL, G. G. 2002. Daxx-mediated accumulation of human cytomegalovirus tegument protein pp71 at ND10

facilitates initiation of viral infection at these nuclear domains. *J Virol,* 76**,** 7705-12.

ISIDRO, J., BORGES, V., PINTO, M., FERREIRA, R., SOBRAL, D., NUNES, A., SANTOS, J. D., JOSÉ BORREGO, M., NÚNCIO, S., PELERITO, A., CORDEIRO, R. & GOMES, J. P. 2022. First draft genome sequence of Monkeypox virus associated with the suspected multi-country outbreak, May 2022 (confirmed case in Portugal). *Virological* [Online]. Available from: https://virological.org/t/first-draft-genome-sequence-of-monkeypox-virus-associated-with-the-suspected-multi-country-outbreak-may-2022-confirmed-case-in-portugal/799 2022].

JAIN, M., FIDDES, I. T., MIGA, K. H., OLSEN, H. E., PATEN, B. & AKESON, M. 2015. Improved data analysis for the MinION nanopore sequencer. *Nat Methods,* 12**,** 351-6.

JAIN, M., TYSON, J. R., LOOSE, M., IP, C. L. C., ECCLES, D. A., O'GRADY, J., MALLA, S., LEGGETT, R. M., WALLERMAN, O., JANSEN, H. J., ZALUNIN, V., BIRNEY, E., BROWN, B. L., SNUTCH, T. P., OLSEN, H. E. & CONSORTIUM, M. A. A. R. 2017. MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry. *F1000Res,* 6**,** 760.

JANECKA, A., ADAMCZYK, A. & GASIŃSKA, A. 2015. Comparison of eight commercially available kits for DNA extraction from formalin-fixed paraffin-embedded tissues. *Anal Biochem,* 476**,** 8-10.

JEAN BELTRAN, P. M. & CRISTEA, I. M. 2014. The life cycle and pathogenesis of human cytomegalovirus infection: lessons from proteomics. *Expert Rev Proteomics,* 11**,** 697-711.

JIANG, X. J., SAMPAIO, K. L., ETTISCHER, N., STIERHOF, Y. D., JAHN, G., KROPFF, B., MACH, M. & SINZGER, C. 2011. UL74 of human cytomegalovirus reduces the inhibitory effect of gH-specific and gB-specific antibodies. *Arch Virol,* 156**,** 2145-55.

JOHN, S., YUZHAKOV, O., WOODS, A., DETERLING, J., HASSETT, K., SHAW, C. A. & CIARAMELLA, G. 2018. Multi-antigenic human cytomegalovirus mRNA vaccines that elicit potent humoral and cell-mediated immunity. *Vaccine,* 36**,** 1689-1699.

KAKUK, B., TOMBÁCZ, D., BALÁZS, Z., MOLDOVÁN, N., CSABAI, Z., TORMA, G., MEGYERI, K., SNYDER, M. & BOLDOGKŐI, Z. 2021. Combined nanopore and single-molecule real-time sequencing survey of human betaherpesvirus 5 transcriptome. *Sci Rep,* 11**,** 14487.

KALEJTA, R. F. 2008. Tegument proteins of human cytomegalovirus. *Microbiol Mol Biol Rev,* 72**,** 249-65, table of contents.

KALEJTA, R. F. 2013. Pre-Immediate Early Tegument Protein Functions. *In:* REDDEHASE, M. J. (ed.) *Cytomegaloviruses: from Molecular Pathogenesis to Intervention.* 2nd ed. Malta: Caister Academic Press.

KANAGAWA, T. 2003. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J Biosci Bioeng,* 96**,** 317-23.

KARAMITROS, T., HARRISON, I., PIORKOWSKA, R., KATZOURAKIS, A., MAGIORKINIS, G. & MBISA, J. L. 2016. De Novo Assembly of Human Herpes Virus Type 1 (HHV-1) Genome, Mining of Non-Canonical Structures and Detection of Novel Drug-Resistance Mutations Using Short- and Long-Read Next Generation Sequencing Technologies. *PLoS One,* 11**,** e0157600.

KARAMITROS, T., VAN WILGENBURG, B., WILLS, M., KLENERMAN, P. & MAGIORKINIS, G. 2018. Nanopore sequencing and full genome de novo assembly of human cytomegalovirus TB40/E reveals clonal diversity and structural variations. *BMC Genomics,* 19**,** 577.

KARI, B. & GEHRZ, R. 1992. A human cytomegalovirus glycoprotein complex designated gC-II is a major heparin-binding component of the envelope. *J Virol,* 66**,** 1761-4.

KATOH, K., MISAWA, K., KUMA, K. & MIYATA, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res,* 30**,** 3059-66.

KATOH, K., ROZEWICKI, J. & YAMADA, K. D. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform,* 20**,** 1160-1166.

KIEHL, A., HUANG, L., FRANCHI, D. & ANDERS, D. G. 2003. Multiple 5' ends of human cytomegalovirus UL57 transcripts identify a complex, cycloheximide-resistant promoter region that activates oriLyt. *Virology,* 314**,** 410-22.

KIM, J. H., COLLINS-MCMILLEN, D., BUEHLER, J. C., GOODRUM, F. D. & YUROCHKO, A. D. 2017. Human Cytomegalovirus Requires Epidermal Growth Factor Receptor Signaling To Enter and Initiate the Early Steps in the Establishment of Latency in CD34. *J Virol,* 91.

KOBOLDT, D. C., ZHANG, Q., LARSON, D. E., SHEN, D., MCLELLAN, M. D., LIN, L., MILLER, C. A., MARDIS, E. R., DING, L. & WILSON, R. K. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res,* 22**,** 568-76.

KODZIUS, R., KOJIMA, M., NISHIYORI, H., NAKAMURA, M., FUKUDA, S., TAGAMI, M., SASAKI, D., IMAMURA, K., KAI, C., HARBERS, M., HAYASHIZAKI, Y. & CARNINCI, P. 2006. CAGE: cap analysis of gene expression. *Nat Methods,* 3**,** 211-22.

KOETSIER, G. C., E. 2019. A Practical Guide to Analyzing Nucleic Acid Concentration and Purity with Microvolume Spectrophotometers. *New England Biolabs Technical Note*.

KOLMOGOROV, M., RANEY, B., PATEN, B. & PHAM, S. 2014. Ragout-a reference-assisted assembly tool for bacterial genomes. *Bioinformatics,* 30**,** i302-9.

KONDO, K., KANESHIMA, H. & MOCARSKI, E. S. 1994. Human cytomegalovirus latent infection of granulocyte-macrophage progenitors. *Proc Natl Acad Sci U S A,* 91**,** 11879-83.

KONDO, K., XU, J. & MOCARSKI, E. S. 1996. Human cytomegalovirus latent gene expression in granulocyte-macrophage progenitors in culture and in seropositive individuals. *Proc Natl Acad Sci U S A,* 93**,** 11137-42.

KOREN, S., WALENZ, B. P., BERLIN, K., MILLER, J. R., BERGMAN, N. H. & PHILLIPPY, A. M. 2017. Canu: scalable and accurate long-read assembly via adaptive. *Genome Res,* 27**,** 722-736.

KOTTON, C. N., KUMAR, D., CALIENDO, A. M., ASBERG, A., CHOU, S., DANZIGER-ISAKOV, L., HUMAR, A. & GROUP, T. S. I. C. C. 2013. Updated international consensus guidelines on the management of cytomegalovirus in solid-organ transplantation. *Transplantation,* 96**,** 333-60.

KROSKY, P. M., BAEK, M. C. & COEN, D. M. 2003. The human cytomegalovirus UL97 protein kinase, an antiviral drug target, is required at the stage of nuclear egress. *J Virol,* 77**,** 905-14.

LANARI, M., CAPRETTI, M. G., LAZZAROTTO, T., GABRIELLI, L., PIGNATELLI, S., DAL MONTE, P., LANDINI, M. P. & FALDELLA, G. 2008. Cytomegalovirus infection via mother's milk: could distinct virus strains determine different disease patterns in preterm twins? *New Microbiol,* 31**,** 131-5.

LANFEAR, R., SCHALAMUN, M., KAINER, D., WANG, W. & SCHWESSINGER, B. 2019. MinIONQC: fast and simple quality control for MinION sequencing data. *Bioinformatics,* 35**,** 523-525.

LANGMEAD, B. & SALZBERG, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods,* 9**,** 357-9.

LARSSON, A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics,* 30**,** 3276-8.

LASKEN, R. S. & STOCKWELL, T. B. 2007. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol,* 7**,** 19.

LASSALLE, F., DEPLEDGE, D. P., REEVES, M. B., BROWN, A. C., CHRISTIANSEN, M. T., TUTILL, H. J., WILLIAMS, R. J., EINER-JENSEN, K., HOLDSTOCK, J., ATKINSON, C., BROWN, J. R., VAN LOENEN, F. B., CLARK, D. A., GRIFFITHS, P. D., VERJANS, G. M. G. M., SCHUTTEN, M., MILNE, R. S. B., BALLOUX, F. & BREUER, J. 2016. Islands of linkage in an ocean of pervasive recombination reveals two-speed evolution of human cytomegalovirus genomes. *Virus Evol,* 2**,** vew017.

LAU, B., KERR, K., CAMIOLO, S., NIGHTINGALE, K., GU, Q., ANTROBUS, R., SUÁREZ, N. M., LONEY, C., STANTON, R. J., WEEKES, M. P. & DAVISON, A. J. 2021. Human Cytomegalovirus RNA2.7 Is Required for Upregulating Multiple Cellular Genes To Promote Cell Motility and Viral Spread Late in Lytic Infection. *J Virol,* 95**,** e0069821.

LEE, J. H. & KALEJTA, R. F. 2019. Human Cytomegalovirus Enters the Primary CD34. *J Virol,* 93.

LEX, A., GEHLENBORG, N., STROBELT, H., VUILLEMOT, R. & PFISTER, H. 2014. UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph,* 20**,** 1983-92.

LI, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics,* 34**,** 3094-3100.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & SUBGROUP, G. P. D. P. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

LIGAT, G., CAZAL, R., HANTZ, S. & ALAIN, S. 2018. The human cytomegalovirus terminase complex as an antiviral target: a close-up view. *FEMS Microbiol Rev,* 42**,** 137-145.

LISBOA, L. F., TONG, Y., KUMAR, D., PANG, X. L., ASBERG, A., HARTMANN, A., ROLLAG, H., JARDINE, A. G., PESCOVITZ, M. D. & HUMAR, A. 2012. Analysis and clinical correlation of genetic variation in cytomegalovirus. *Transplant Infectious Disease,* 14**,** 132-140.

LITTLER, E., STUART, A. D. & CHEE, M. S. 1992. Human cytomegalovirus UL97 open reading frame encodes a protein that phosphorylates the antiviral nucleoside analogue ganciclovir. *Nature,* 358**,** 160-2.

LJUNGMAN, P., HAKKI, M. & BOECKH, M. 2011. Cytomegalovirus in hematopoietic stem cell transplant recipients. *Hematol Oncol Clin North Am,* 25**,** 151-69.

LOMAN, N. J., QUICK, J. & SIMPSON, J. T. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods,* 12**,** 733-5.

LOOSE, M., MALLA, S. & STOUT, M. 2016. Real-time selective sequencing using nanopore technology. *Nat Methods,* 13**,** 751-4.

LURAIN, N. S., FOX, A. M., LICHY, H. M., BHORADE, S. M., WARE, C. F., HUANG, D. D., KWAN, S. P., GARRITY, E. R. & CHOU, S. 2006. Analysis of the human cytomegalovirus genomic region from UL146 through UL147A

reveals sequence hypervariability, genotypic stability, and overlapping transcripts. *Virol J,* 3**,** 4.
LÜTTICHAU, H. R. 2010. The cytomegalovirus UL146 gene product vCXCL1 targets both CXCR1 and CXCR2 as an agonist. *J Biol Chem,* 285**,** 9137-46.
MA, Y. P., RUAN, Q., JI, Y. H., WANG, N., LI, M. L., QI, Y., HE, R., SUN, Z. R. & REN, G. W. 2011. Novel transcripts of human cytomegalovirus clinical strain found by cDNA library screening. *Genet Mol Res,* 10**,** 566-75.
MACH, M., KROPFF, B., DAL MONTE, P. & BRITT, W. 2000. Complex formation by human cytomegalovirus glycoproteins M (gpUL100) and N (gpUL73). *J Virol,* 74**,** 11881-92.
MADOUI, M. A., ENGELEN, S., CRUAUD, C., BELSER, C., BERTRAND, L., ALBERTI, A., LEMAINQUE, A., WINCKER, P. & AURY, J. M. 2015. Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics,* 16**,** 327.
MANICKLAL, S., EMERY, V. C., LAZZAROTTO, T., BOPPANA, S. B. & GUPTA, R. K. 2013. The "silent" global burden of congenital cytomegalovirus. *Clin Microbiol Rev,* 26**,** 86-102.
MANUEL, O., ASBERG, A., PANG, X., ROLLAG, H., EMERY, V. C., PREIKSAITIS, J. K., KUMAR, D., PESCOVITZ, M. D., BIGNAMINI, A. A., HARTMANN, A., JARDINE, A. G. & HUMAR, A. 2009. Impact of genetic polymorphisms in cytomegalovirus glycoprotein B on outcomes in solid-organ transplant recipients with cytomegalovirus disease. *Clin Infect Dis,* 49**,** 1160-6.
MARÇAIS, G., DELCHER, A. L., PHILLIPPY, A. M., COSTON, R., SALZBERG, S. L. & ZIMIN, A. 2018. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol,* 14**,** e1005944.
MARINE, R., MCCARREN, C., VORRASANE, V., NASKO, D., CROWGEY, E., POLSON, S. W. & WOMMACK, K. E. 2014. Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome. *Microbiome,* 2**,** 3.
MARINE, R., POLSON, S. W., RAVEL, J., HATFULL, G., RUSSELL, D., SULLIVAN, M., SYED, F., DUMAS, M. & WOMMACK, K. E. 2011. Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. *Appl Environ Microbiol,* 77**,** 8071-9.
MARTY, F. M., WINSTON, D. J., CHEMALY, R. F., MULLANE, K. M., SHORE, T. B., PAPANICOLAOU, G. A., CHITTICK, G., BRUNDAGE, T. M., WILSON, C., MORRISON, M. E., FOSTER, S. A., NICHOLS, W. G., BOECKH, M. J. & GROUP, S. T. C. S. 2019. A Randomized, Double-Blind, Placebo-Controlled Phase 3 Trial of Oral Brincidofovir for Cytomegalovirus Prophylaxis in Allogeneic Hematopoietic Cell Transplantation. *Biol Blood Marrow Transplant,* 25**,** 369-381.
MATHIESON, W. & THOMAS, G. A. 2020. Why Formalin-fixed, Paraffin-embedded Biospecimens Must Be Used in Genomic Medicine: An Evidence-based Review and Conclusion. *J Histochem Cytochem,* 68**,** 543-552.
MAUL, G. G. 2008. Initiation of cytomegalovirus infection at ND10. *Curr Top Microbiol Immunol,* 325**,** 117-32.
MCCORMICK, A. L., ROBACK, L., LIVINGSTON-ROSANOFF, D. & ST CLAIR, C. 2010. The human cytomegalovirus UL36 gene controls caspase-dependent and -independent cell death programs activated by infection of monocytes differentiating to macrophages. *J Virol,* 84**,** 5108-23.
MCDONOUGH, S. J., BHAGWATE, A., SUN, Z., WANG, C., ZSCHUNKE, M., GORMAN, J. A., KOPP, K. J. & CUNNINGHAM, J. M. 2019. Use of FFPE-

derived DNA in next generation sequencing: DNA extraction methods. *PLoS One,* 14**,** e0211400.
MCGEOCH, D. J., RIXON, F. J. & DAVISON, A. J. 2006. Topics in herpesvirus genomics and evolution. *Virus Res,* 117**,** 90-104.
MCSHARRY, B. P., AVDIC, S. & SLOBEDMAN, B. 2012. Human cytomegalovirus encoded homologs of cytokines, chemokines and their receptors: roles in immunomodulation. *Viruses,* 4**,** 2448-70.
MELNIKOV, A., GALINSKY, K., ROGOV, P., FENNELL, T., VAN TYNE, D., RUSS, C., DANIELS, R., BARNES, K. G., BOCHICCHIO, J., NDIAYE, D., SENE, P. D., WIRTH, D. F., NUSBAUM, C., VOLKMAN, S. K., BIRREN, B. W., GNIRKE, A. & NEAFSEY, D. E. 2011. Hybrid selection for sequencing pathogen genomes from clinical samples. *Genome Biol,* 12**,** R73.
MENDELSON, M., MONARD, S., SISSONS, P. & SINCLAIR, J. 1996. Detection of endogenous human cytomegalovirus in CD34+ bone marrow progenitors. *J Gen Virol,* 77 (Pt 12)**,** 3099-102.
MERCER, T. R., DINGER, M. E. & MATTICK, J. S. 2009. Long non-coding RNAs: insights into functions. *Nat Rev Genet,* 10**,** 155-9.
MESHESHA, M. K., VEKSLER-LUBLINSKY, I., ISAKOV, O., REICHENSTEIN, I., SHOMRON, N., KEDEM, K., ZIV-UKELSON, M., BENTWICH, Z. & AVNI, Y. S. 2012. The microRNA Transcriptome of Human Cytomegalovirus (HCMV). *Open Virol J,* 6**,** 38-48.
METTENLEITER, T. C., KLUPP, B. G. & GRANZOW, H. 2006. Herpesvirus assembly: a tale of two membranes. *Curr Opin Microbiol,* 9**,** 423-9.
MEYER-KÖNIG, U., VOGELBERG, C., BONGARTS, A., KAMPA, D., DELBRÜCK, R., WOLFF-VORBECK, G., KIRSTE, G., HABERLAND, M., HUFERT, F. T. & VON LAER, D. 1998. Glycoprotein B genotype correlates with cell tropism in vivo of human cytomegalovirus infection. *J Med Virol,* 55**,** 75-81.
MILNE, I., BAYER, M., STEPHEN, G., CARDLE, L. & MARSHALL, D. 2016. Tablet: Visualizing Next-Generation Sequence Assemblies and Mappings. *Methods Mol Biol,* 1374**,** 253-68.
MOCARSKI, E. S. 1996. *Cytomegaloviruses and their replication.,* Philadelphia, Lippincott-Raven.
MOREY, M., FERNÁNDEZ-MARMIESSE, A., CASTIÑEIRAS, D., FRAGA, J. M., COUCE, M. L. & COCHO, J. A. 2013. A glimpse into past, present, and future DNA sequencing. *Mol Genet Metab,* 110**,** 3-24.
MORGULIS, A., COULOURIS, G., RAYTSELIS, Y., MADDEN, T. L., AGARWALA, R. & SCHÄFFER, A. A. 2008. Database indexing for production MegaBLAST searches. *Bioinformatics,* 24**,** 1757-64.
MURPHY, E., RIGOUTSOS, I., SHIBUYA, T. & SHENK, T. E. 2003. Reevaluation of human cytomegalovirus coding potential. *Proc Natl Acad Sci U S A,* 100**,** 13585-90.
MURRELL, I., WILKIE, G. S., DAVISON, A. J., STATKUTE, E., FIELDING, C. A., TOMASEC, P., WILKINSON, G. W. & STANTON, R. J. 2016. Genetic Stability of Bacterial Artificial Chromosome-Derived Human Cytomegalovirus during Culture In Vitro. *J Virol,* 90**,** 3929-43.
MURTHY, S., HAYWARD, G. S., WHEELAN, S., FORMAN, M. S., AHN, J.-H., PASS, R. F. & ARAV-BOGER, R. 2011. Detection of a Single Identical Cytomegalovirus (CMV) Strain in Recently Seroconverted Young Women. *PLoS ONE,* 6**,** e15949.
MYERSON, D., HACKMAN, R. C., NELSON, J. A., WARD, D. C. & MCDOUGALL, J. K. 1984. Widespread presence of histologically occult cytomegalovirus. *Hum Pathol,* 15**,** 430-9.

NEFF, B. J., WEIBEL, R. E., BUYNAK, E. B., MCLEAN, A. A. & HILLEMAN, M. R. 1979. Clinical and laboratory studies of live cytomegalovirus vaccine Ad-169. *Proc Soc Exp Biol Med,* 160**,** 32-7.

NIJMAN, J., MANDEMAKER, F. S., VERBOON-MACIOLEK, M. A., AITKEN, S. C., VAN LOON, A. M., DE VRIES, L. S. & SCHUURMAN, R. 2014. Genotype distribution, viral load and clinical characteristics of infants with postnatal or congenital cytomegalovirus infection. *PLoS One,* 9**,** e108018.

NOBRE, L. V., NIGHTINGALE, K., RAVENHILL, B. J., ANTROBUS, R., SODAY, L., NICHOLS, J., DAVIES, J. A., SEIRAFIAN, S., WANG, E. C., DAVISON, A. J., WILKINSON, G. W., STANTON, R. J., HUTTLIN, E. L. & WEEKES, M. P. 2019. Human cytomegalovirus interactome analysis identifies degradation hubs, domain associations and viral protein functions. *Elife,* 8.

NOVAK, Z., ROSS, S. A., PATRO, R. K., PATI, S. K., KUMBLA, R. A., BRICE, S. & BOPPANA, S. B. 2008. Cytomegalovirus strain diversity in seropositive women. *J Clin Microbiol,* 46**,** 882-6.

NURK, S., BANKEVICH, A., ANTIPOV, D., GUREVICH, A. A., KOROBEYNIKOV, A., LAPIDUS, A., PRJIBELSKI, A. D., PYSHKIN, A., SIROTKIN, A., SIROTKIN, Y., STEPANAUSKAS, R., CLINGENPEEL, S. R., WOYKE, T., MCLEAN, J. S., LASKEN, R., TESLER, G., ALEKSEYEV, M. A. & PEVZNER, P. A. 2013. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J Comput Biol,* 20**,** 714-37.

ODEBERG, J., PLACHTER, B., BRANDÉN, L. & SÖDERBERG-NAUCLÉR, C. 2003. Human cytomegalovirus protein pp65 mediates accumulation of HLA-DR in lysosomes and destruction of the HLA-DR alpha-chain. *Blood,* 101**,** 4870-7.

OGAWA-GOTO, K., TANAKA, K., GIBSON, W., MORIISHI, E., MIURA, Y., KURATA, T., IRIE, S. & SATA, T. 2003. Microtubule network facilitates nuclear targeting of human cytomegalovirus capsid. *J Virol,* 77**,** 8541-7.

PANG, J., SLYKER, J. A., ROY, S., BRYANT, J., ATKINSON, C., CUDINI, J., FARQUHAR, C., GRIFFITHS, P., KIARIE, J., MORFOPOULOU, S., ROXBY, A. C., TUTIL, H., WILLIAMS, R., GANTT, S., GOLDSTEIN, R. A. & BREUER, J. 2020. Mixed cytomegalovirus genotypes in HIV-positive mothers show compartmentalization and distinct patterns of transmission to infants. *Elife,* 9.

PAPANICOLAOU, G. A., SILVEIRA, F. P., LANGSTON, A. A., PEREIRA, M. R., AVERY, R. K., UKNIS, M., WIJATYK, A., WU, J., BOECKH, M., MARTY, F. M. & VILLANO, S. 2019. Maribavir for Refractory or Resistant Cytomegalovirus Infections in Hematopoietic-cell or Solid-organ Transplant Recipients: A Randomized, Dose-ranging, Double-blind, Phase 2 Study. *Clin Infect Dis,* 68**,** 1255-1264.

PARADOWSKA, E., JABLONSKA, A., PLOCIENNIKOWSKA, A., STUDZINSKA, M., SUSKI, P., WISNIEWSKA-LIGIER, M., DZIERZANOWSKA-FANGRAT, K., KASZTELEWICZ, B., WOZNIAKOWSKA-GESICKA, T. & LESNIKOWSKI, Z. J. 2014a. Cytomegalovirus alpha-chemokine genotypes are associated with clinical manifestations in children with congenital or postnatal infections. *Virology,* 462**,** 207-217.

PARADOWSKA, E., JABLONSKA, A., STUDZINSKA, M., KASZTELEWICZ, B., ZAWILINSKA, B., WISNIEWSKA-LIGIER, M., DZIERZANOWSKA-FANGRAT, K., WOZNIAKOWSKA-GESICKA, T., KOSZ-VNENCHAK, M. & LESNIKOWSKI, Z. J. 2014b. Cytomegalovirus Glycoprotein H Genotype Distribution and the Relationship With Hearing Loss in Children. *Journal of Medical Virology,* 86**,** 1421-1427.

PARADOWSKA, E., STUDZIŃSKA, M., SUSKI, P., KASZTELEWICZ, B., WIŚNIEWSKA-LIGIER, M., ZAWILIŃSKA, B., GAJ, Z. & NOWAKOWSKA, D. 2015. Human cytomegalovirus UL55, UL144, and US28 genotype distribution in infants infected congenitally or postnatally. *J Med Virol,* 87**,** 1737-48.

PARI, G. S. & ANDERS, D. G. 1993. Eleven loci encoding trans-acting factors are required for transient complementation of human cytomegalovirus oriLyt-dependent DNA replication. *J Virol,* 67**,** 6979-88.

PASS, R. F., ZHANG, C., EVANS, A., SIMPSON, T., ANDREWS, W., HUANG, M. L., COREY, L., HILL, J., DAVIS, E., FLANIGAN, C. & CLOUD, G. 2009. Vaccine prevention of maternal cytomegalovirus infection. *N Engl J Med,* 360**,** 1191-9.

PATI, S. K., PINNINTI, S., NOVAK, Z., CHOWDHURY, N., PATRO, R. K., FOWLER, K., ROSS, S., BOPPANA, S. & INVESTIGATORS, N. C. S. 2013. Genotypic diversity and mixed infection in newborn disease and hearing loss in congenital cytomegalovirus infection. *Pediatr Infect Dis J,* 32**,** 1050-4.

PAYNE, A., HOLMES, N., RAKYAN, V. & LOOSE, M. 2018. Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. *bioRxiv***,** 312256.

PENFOLD, M. E., DAIRAGHI, D. J., DUKE, G. M., SAEDERUP, N., MOCARSKI, E. S., KEMBLE, G. W. & SCHALL, T. J. 1999. Cytomegalovirus encodes a potent alpha chemokine. *Proc Natl Acad Sci U S A,* 96**,** 9839-44.

PIETROPAOLO, R. & COMPTON, T. 1999. Interference with annexin II has no effect on entry of human cytomegalovirus into fibroblast cells. *J Gen Virol,* 80 (Pt 7)**,** 1807-1816.

PIGNATELLI, S., DAL MONTE, P. & LANDINI, M. P. 2001. gpUL73 (gN) genomic variants of human cytomegalovirus isolates are clustered into four distinct genotypes. *J Gen Virol,* 82**,** 2777-84.

PIGNATELLI, S., DAL MONTE, P., ROSSINI, G. & LANDINI, M. P. 2004. Genetic polymorphisms among human cytomegalovirus (HCMV) wild-type strains. *Rev Med Virol,* 14**,** 383-410.

PODLECH, J., REDDEHASE, M. J., ADLER, B. & LEMMERMANN, N. A. 2015. Principles for studying in vivo attenuation of virus mutants: defining the role of the cytomegalovirus gH/gL/gO complex as a paradigm. *Med Microbiol Immunol,* 204**,** 295-305.

POKALYUK, C., RENZETTE, N., IRWIN, K. K., PFEIFER, S. P., GIBSON, L., BRITT, W. J., YAMAMOTO, A. Y., MUSSI-PINHATA, M. M., KOWALIK, T. F. & JENSEN, J. D. 2017. Characterizing human cytomegalovirus reinfection in congenitally infected infants: an evolutionary perspective. *Mol Ecol,* 26**,** 1980-1990.

POOLE, E. L., KEW, V. G., LAU, J. C. H., MURRAY, M. J., STAMMINGER, T., SINCLAIR, J. H. & REEVES, M. B. 2018. A Virally Encoded DeSUMOylase Activity Is Required for Cytomegalovirus Reactivation from Latency. *Cell Rep,* 24**,** 594-606.

PÖTZSCH, S., SPINDLER, N., WIEGERS, A. K., FISCH, T., RÜCKER, P., STICHT, H., GRIEB, N., BAROTI, T., WEISEL, F., STAMMINGER, T., MARTIN-PARRAS, L., MACH, M. & WINKLER, T. H. 2011. B cell repertoire analysis identifies new antigenic domains on glycoprotein B of human cytomegalovirus which are target of neutralizing antibodies. *PLoS Pathog,* 7**,** e1002172.

PRENTICE, L. M., MILLER, R. R., KNAGGS, J., MAZLOOMIAN, A., AGUIRRE HERNANDEZ, R., FRANCHINI, P., PARSA, K., TESSIER-CLOUTIER, B., LAPUK, A., HUNTSMAN, D., SCHAEFFER, D. F. & SHEFFIELD, B. S. 2018. Formalin

fixation increases deamination mutation signature but should not lead to false positive mutations in clinical practice. *PLoS One,* 13**,** e0196434.

PUCHHAMMER-STÖCKL, E., GÖRZER, I., ZOUFALY, A., JAKSCH, P., BAUER, C. C., KLEPETKO, W. & POPOW-KRAUPP, T. 2006. Emergence of multiple cytomegalovirus strains in blood and lung of lung transplant recipients. *Transplantation,* 81**,** 187-94.

QUICK, J., LOMAN, N. J., DURAFFOUR, S., SIMPSON, J. T., SEVERI, E., COWLEY, L., BORE, J. A., KOUNDOUNO, R., DUDAS, G., MIKHAIL, A., OUÉDRAOGO, N., AFROUGH, B., BAH, A., BAUM, J. H., BECKER-ZIAJA, B., BOETTCHER, J. P., CABEZA-CABRERIZO, M., CAMINO-SANCHEZ, A., CARTER, L. L., DOERRBECKER, J., ENKIRCH, T., DORIVAL, I. G. G., HETZELT, N., HINZMANN, J., HOLM, T., KAFETZOPOULOU, L. E., KOROPOGUI, M., KOSGEY, A., KUISMA, E., LOGUE, C. H., MAZZARELLI, A., MEISEL, S., MERTENS, M., MICHEL, J., NGABO, D., NITZSCHE, K., PALLASH, E., PATRONO, L. V., PORTMANN, J., REPITS, J. G., RICKETT, N. Y., SACHSE, A., SINGETHAN, K., VITORIANO, I., YEMANABERHAN, R. L., ZEKENG, E. G., TRINA, R., BELLO, A., SALL, A. A., FAYE, O., MAGASSOUBA, N., WILLIAMS, C. V., AMBURGEY, V., WINONA, L., DAVIS, E., GERLACH, J., WASHINGTON, F., MONTEIL, V., JOURDAIN, M., BERERD, M., CAMARA, A., SOMLARE, H., GERARD, M., BADO, G., BAILLET, B., DELAUNE, D., NEBIE, K. Y., DIARRA, A., SAVANE, Y., PALLAWO, R. B., GUTIERREZ, G. J., MILHANO, N., ROGER, I., WILLIAMS, C. J., YATTARA, F., LEWANDOWSKI, K., TAYLOR, J., RACHWAL, P., TURNER, D., POLLAKIS, G., HISCOX, J. A., MATTHEWS, D. A., O'SHEA, M. K., JOHNSTON, A. M., WILSON, D., HUTLEY, E., SMIT, E., DI CARO, A., WOELFEL, R., STOECKER, K., FLEISCHMANN, E., GABRIEL, M., WELLER, S. A., KOIVOGUI, L., DIALLO, B., KEITA, S., RAMBAUT, A., FORMENTY, P., GUNTHER, S. & CARROLL, M. W. 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature,* 530**,** 228-232.

RAFAILIDIS, P. I., MOURTZOUKOU, E. G., VARBOBITIS, I. C. & FALAGAS, M. E. 2008. Severe cytomegalovirus infection in apparently immunocompetent patients: a systematic review. *Virology Journal,* 5**,** 47-47.

RASMUSSEN, L., GEISSLER, A. & WINTERS, M. 2003. Inter- and intragenic variations complicate the molecular epidemiology of human cytomegalovirus. *J Infect Dis,* 187**,** 809-19.

RAWLINSON, W. D., BOPPANA, S. B., FOWLER, K. B., KIMBERLIN, D. W., LAZZAROTTO, T., ALAIN, S., DALY, K., DOUTRÉ, S., GIBSON, L., GILES, M. L., GREENLEE, J., HAMILTON, S. T., HARRISON, G. J., HUI, L., JONES, C. A., PALASANTHIRAN, P., SCHLEISS, M. R., SHAND, A. W. & VAN ZUYLEN, W. J. 2017. Congenital cytomegalovirus infection in pregnancy and the neonate: consensus recommendations for prevention, diagnosis, and therapy. *Lancet Infect Dis,* 17**,** e177-e188.

REEVES, M. B., BREIDENSTEIN, A. & COMPTON, T. 2012. Human cytomegalovirus activation of ERK and myeloid cell leukemia-1 protein correlates with survival of latently infected cells. *Proc Natl Acad Sci U S A,* 109**,** 588-93.

REEVES, M. B. & SINCLAIR, J. H. 2013. Circulating dendritic cells isolated from healthy seropositive donors are sites of human cytomegalovirus reactivation in vivo. *J Virol,* 87**,** 10660-7.

RENZETTE, N., BHATTACHARJEE, B., JENSEN, J. D., GIBSON, L. & KOWALIK, T. F. 2011. Extensive genome-wide variability of human cytomegalovirus in congenitally infected infants. *PLoS Pathog,* 7**,** e1001344.

RENZETTE, N., GIBSON, L., BHATTACHARJEE, B., FISHER, D., SCHLEISS, M. R., JENSEN, J. D. & KOWALIK, T. F. 2013. Rapid intrahost evolution of human

cytomegalovirus is shaped by demography and positive selection. *PLoS Genet,* 9, e1003735.
RIBBERT, H. 1904. Ueber protozoenartige Zellen in der Niere eines syphilitischen Neugeborenen und in der Parotis von Kindern. *Zbl All Pathol* 15, 945-948.
ROMANOWSKI, M. J. & SHENK, T. 1997. Characterization of the human cytomegalovirus irs1 and trs1 genes: a second immediate-early transcription unit within irs1 whose product antagonizes transcriptional activation. *J Virol,* 71, 1485-96.
ROSEN, H. R., CORLESS, C. L., RABKIN, J. & CHOU, S. 1998. Association of cytomegalovirus genotype with graft rejection after liver transplantation. *Transplantation,* 66, 1627-31.
ROSSINI, G., PIGNATELLI, S., DAL MONTE, P., CAMOZZI, D., LAZZAROTTO, T., GABRIELLI, L., GATTO, M. R. & LANDINI, M. P. 2005. Monitoring for human cytomegalovirus infection in solid organ transplant recipients through antigenemia and glycoprotein N (gN) variants: evidence of correlation and potential prognostic value of gN genotypes. *Microbes Infect,* 7, 890-6.
ROUBALOVA, K., STRUNECKY, O., VITEK, A., ZUFANOVA, S. & PROCHAZKA, B. 2011. Genetic variability of cytomegalovirus glycoprotein O in hematopoietic stem cell transplant recipients. *Transpl Infect Dis,* 13, 237-43.
ROWE, W. P., HARTLEY, J. W., WATERMAN, S., TURNER, H. C. & HUEBNER, R. J. 1956. Cytopathogenic agent resembling human salivary gland virus recovered from tissue cultures of human adenoids. *Proc Soc Exp Biol Med,* 92, 418-24.
ROZMAN, B., NACHSHON, A., LEVI SAMIA, R., LAVI, M., SCHWARTZ, M. & STERN-GINOSSAR, N. 2022. Temporal dynamics of HCMV gene expression in lytic and latent infections. *Cell Rep,* 39, 110653.
RUZSICS, Z., M., B. E., B., B. J., BRUNE, W. & MESSERLE, M. 2013. Manipulating CMV Genomes by BAC Mutagenesis: Strategies and Applications. *In:* REDDEHASE, M. J. (ed.) *Cytomegaloviruses: from Molecular Pathogenesis to Intervention.* 2nd ed. Malta: Caister Academic Press.
RYCEL, M., WUJCICKA, W., ZAWILIŃSKA, B., PARADOWSKA, E., SUSKI, P., GAJ, Z., WILCZYŃSKI, J., LEŚNIKOWSKI, Z. & NOWAKOWSKA, D. 2015. Mixed infections with distinct cytomegalovirus glycoprotein B genotypes in Polish pregnant women, fetuses, and newborns. *Eur J Clin Microbiol Infect Dis,* 34, 585-91.
RYCKMAN, B. J., RAINISH, B. L., CHASE, M. C., BORTON, J. A., NELSON, J. A., JARVIS, M. A. & JOHNSON, D. C. 2008. Characterization of the human cytomegalovirus gH/gL/UL128-131 complex that mediates entry into epithelial and endothelial cells. *J Virol,* 82, 60-70.
SAFFERT, R. T. & KALEJTA, R. F. 2006. Inactivating a cellular intrinsic immune defense mediated by Daxx is the mechanism through which the human cytomegalovirus pp71 protein stimulates viral immediate-early gene expression. *J Virol,* 80, 3863-71.
SAHOO, M. K., LEFTEROVA, M. I., YAMAMOTO, F., WAGGONER, J. J., CHOU, S., HOLMES, S. P., ANDERSON, M. W. & PINSKY, B. A. 2013. Detection of cytomegalovirus drug resistance mutations by next-generation sequencing. *J Clin Microbiol,* 51, 3700-10.
SAITOU, N. & NEI, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol,* 4, 406-25.

SALSMAN, J., WANG, X. & FRAPPIER, L. 2011. Nuclear body formation and PML body remodeling by the human cytomegalovirus protein UL35. *Virology,* 414**,** 119-29.

SARCINELLA, L., MAZZULLI, T., WILLEY, B. & HUMAR, A. 2002. Cytomegalovirus glycoprotein B genotype does not correlate with outcomes in liver transplant patients. *J Clin Virol,* 24**,** 99-105.

SCHMIEDER, R. & EDWARDS, R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics,* 27**,** 863-4.

SEKULIN, K., GÖRZER, I., HEISS-CZEDIK, D. & PUCHHAMMER-STÖCKL, E. 2007. Analysis of the variability of CMV strains in the RL11D domain of the RL11 multigene family. *Virus Genes,* 35**,** 577-83.

SELHORST, P., REZENDE, A. M., DE BLOCK, T., COPPENS, S., HILDE SMET1, MARIËN, J., HAUNER, A., BROSIUS, I., LIESENBORGHS, L., BOTTIEAU, E., FLORENCE, E., BANGWEN, E., ARIËN, K. K., VAN ESBROECK, M., KENYON, C. & VERCAUTEREN, K. 2022. Belgian case of Monkeypox virus linked to outbreak in Portugal. *Virological* [Online].

SEREIKA, M., KIRKEGAARD, R. H., KARST, S. M., MICHAELSEN, T. Y., SØRENSEN, E. A., WOLLENBERG, R. D. & ALBERTSEN, M. 2022. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat Methods,* 19**,** 823-826.

SHEN, S., WANG, S., BRITT, W. J. & LU, S. 2007. DNA vaccines expressing glycoprotein complex II antigens gM and gN elicited neutralizing antibodies against multiple human cytomegalovirus (HCMV) isolates. *Vaccine,* 25**,** 3319-27.

SHEPP, D. H., MATCH, M. E., ASHRAF, A. B., LIPSON, S. M., MILLAN, C. & PERGOLIZZI, R. 1996. Cytomegalovirus glycoprotein B groups associated with retinitis in AIDS. *J Infect Dis,* 174**,** 184-7.

SHIMAMURA, M., MACH, M. & BRITT, W. J. 2006. Human cytomegalovirus infection elicits a glycoprotein M (gM)/gN-specific virus-neutralizing antibody response. *J Virol,* 80**,** 4591-600.

SHINDE, D., LAI, Y., SUN, F. & ARNHEIM, N. 2003. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. *Nucleic Acids Res,* 31**,** 974-80.

SHNAYDER, M., NACHSHON, A., KRISHNA, B., POOLE, E., BOSHKOV, A., BINYAMIN, A., MAZA, I., SINCLAIR, J., SCHWARTZ, M. & STERN-GINOSSAR, N. 2018. Defining the Transcriptional Landscape during Cytomegalovirus Latency with Single-Cell RNA Sequencing. *mBio,* 9.

SIJMONS, S., THYS, K., CORTHOUT, M., VAN DAMME, E., VAN LOOCK, M., BOLLEN, S., BAGUET, S., AERSSENS, J., VAN RANST, M. & MAES, P. 2014. A method enabling high-throughput sequencing of human cytomegalovirus complete genomes from clinical isolates. *PLoS One,* 9**,** e95501.

SIJMONS, S., THYS, K., MBONG NGWESE, M., VAN DAMME, E., DVORAK, J., VAN LOOCK, M., LI, G., TACHEZY, R., BUSSON, L., AERSSENS, J., VAN RANST, M. & MAES, P. 2015. High-throughput analysis of human cytomegalovirus genome diversity highlights the widespread occurrence of gene-disrupting mutations and pervasive recombination. *J Virol*.

SILVA, G. G., DUTILH, B. E., MATTHEWS, T. D., ELKINS, K., SCHMIEDER, R., DINSDALE, E. A. & EDWARDS, R. A. 2013. Combining de novo and reference-guided assembly with scaffold_builder. *Source Code Biol Med,* 8**,** 23.

SILVA, M. C., YU, Q. C., ENQUIST, L. & SHENK, T. 2003. Human cytomegalovirus UL99-encoded pp28 is required for the cytoplasmic envelopment of tegument-associated capsids. *J Virol,* 77**,** 10594-605.

SINCLAIR, J. & SISSONS, P. 2006. Latency and reactivation of human cytomegalovirus. *J Gen Virol,* 87**,** 1763-79.

SINZGER, C., KAHL, M., LAIB, K., KLINGEL, K., RIEGER, P., PLACHTER, B. & JAHN, G. 2000. Tropism of human cytomegalovirus for endothelial cells is determined by a post-entry step dependent on efficient translocation to the nucleus. *J Gen Virol,* 81**,** 3021-3035.

SINZGER, C., SCHMIDT, K., KNAPP, J., KAHL, M., BECK, R., WALDMAN, J., HEBART, H., EINSELE, H. & JAHN, G. 1999. Modification of human cytomegalovirus tropism through propagation in vitro is associated with changes in the viral genome. *J Gen Virol,* 80 (Pt 11)**,** 2867-2877.

SLAVOV, S. N., OTAGUIRI, K. K., DE FIGUEIREDO, G. G., YAMAMOTO, A. Y., MUSSI-PINHATA, M. M., KASHIMA, S. & COVAS, D. T. 2016. Development and optimization of a sensitive TaqMan® real-time PCR with synthetic homologous extrinsic control for quantitation of Human cytomegalovirus viral load. *J Med Virol,* 88**,** 1604-12.

SMITH, M. G. 1956. Propagation in tissue cultures of a cytopathogenic virus from human salivary gland virus (SGV) disease. *Proc Soc Exp Biol Med,* 92**,** 424-30.

SÖDERBERG, C., GIUGNI, T. D., ZAIA, J. A., LARSSON, S., WAHLBERG, J. M. & MÖLLER, E. 1993. CD13 (human aminopeptidase N) mediates human cytomegalovirus infection. *J Virol,* 67**,** 6576-85.

SOEJIMA, M., HIROSHIGE, K., YOSHIMOTO, J. & KODA, Y. 2012. Selective quantification of human DNA by real-time PCR of FOXP2. *Forensic Sci Int Genet,* 6**,** 447-51.

SOROCEANU, L., AKHAVAN, A. & COBBS, C. S. 2008. Platelet-derived growth factor-alpha receptor activation is required for human cytomegalovirus infection. *Nature,* 455**,** 391-5.

SPAETE, R. R., THAYER, R. M., PROBERT, W. S., MASIARZ, F. R., CHAMBERLAIN, S. H., RASMUSSEN, L., MERIGAN, T. C. & PACHL, C. 1988. Human cytomegalovirus strain Towne glycoprotein B is processed by proteolytic cleavage. *Virology,* 167**,** 207-25.

STANGHERLIN, L. M., DE PAULA, F. N., ICIMOTO, M. Y., RUIZ, L. G. P., NOGUEIRA, M. L., BRAZ, A. S. K., JULIANO, L. & DA SILVA, M. C. C. 2017. Positively Selected Sites at HCMV gB Furin Processing Region and Their Effects in Cleavage Efficiency. *Front Microbiol,* 8**,** 934.

STANTON, R., WESTMORELAND, D., FOX, J. D., DAVISON, A. J. & WILKINSON, G. W. 2005. Stability of human cytomegalovirus genotypes in persistently infected renal transplant recipients. *J Med Virol,* 75**,** 42-6.

STANTON, R. J., BALUCHOVA, K., DARGAN, D. J., CUNNINGHAM, C., SHEEHY, O., SEIRAFIAN, S., MCSHARRY, B. P., NEALE, M. L., DAVIES, J. A., TOMASEC, P., DAVISON, A. J. & WILKINSON, G. W. 2010. Reconstruction of the complete human cytomegalovirus genome in a BAC reveals RL13 to be a potent inhibitor of replication. *J Clin Invest,* 120**,** 3191-208.

STARK, T. J., ARNOLD, J. D., SPECTOR, D. H. & YEO, G. W. 2012. High-resolution profiling and analysis of viral and host small RNAs during human cytomegalovirus infection. *J Virol,* 86**,** 226-35.

STEIJGER, T., ABRIL, J. F., ENGSTRÖM, P. G., KOKOCINSKI, F., HUBBARD, T. J., GUIGÓ, R., HARROW, J., BERTONE, P. & CONSORTIUM, R. 2013.

Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods,* 10**,** 1177-84.

STERN-GINOSSAR, N., WEISBURD, B., MICHALSKI, A., LE, V. T., HEIN, M. Y., HUANG, S. X., MA, M., SHEN, B., QIAN, S. B., HENGEL, H., MANN, M., INGOLIA, N. T. & WEISSMAN, J. S. 2012. Decoding human cytomegalovirus. *Science,* 338**,** 1088-93.

SUÁREZ, N. M., BLYTH, E., LI, K., GANZENMUELLER, T., CAMIOLO, S., AVDIC, S., WITHERS, B., LINNENWEBER-HELD, S., GWINNER, W., DHINGRA, A., HEIM, A., SCHULZ, T. F., GUNSON, R., GOTTLIEB, D., SLOBEDMAN, B. & DAVISON, A. J. 2020. Whole-Genome Approach to Assessing Human Cytomegalovirus Dynamics in Transplant Patients Undergoing Antiviral Therapy. *Front Cell Infect Microbiol,* 10**,** 267.

SUÁREZ, N. M., LAU, B., KEMBLE, G. M., LEE, R., MOCARSKI, E. S., WILKINSON, G. W. G., ADLER, S. P., MCVOY, M. A. & DAVISON, A. J. 2017. Genomic analysis of chimeric human cytomegalovirus vaccine candidates derived from strains Towne and Toledo. *Virus Genes,* 53**,** 650-655.

SUÁREZ, N. M., MUSONDA, K. G., ESCRIVA, E., NJENGA, M., AGBUEZE, A., CAMIOLO, S., DAVISON, A. J. & GOMPELS, U. A. 2019a. Multiple-Strain Infections of Human Cytomegalovirus With High Genomic Diversity Are Common in Breast Milk From Human Immunodeficiency Virus-Infected Women in Zambia. *J Infect Dis,* 220**,** 792-801.

SUÁREZ, N. M., WILKIE, G. S., HAGE, E., CAMIOLO, S., HOLTON, M., HUGHES, J., MAABAR, M., VATTIPALLY, S. B., DHINGRA, A., GOMPELS, U. A., WILKINSON, G. W. G., BALDANTI, F., FURIONE, M., LILLERI, D., AROSSA, A., GANZENMUELLER, T., GERNA, G., HUBÁČEK, P., SCHULZ, T. F., WOLF, D., ZAVATTONI, M. & DAVISON, A. J. 2019b. Human Cytomegalovirus Genomes Sequenced Directly From Clinical Material: Variation, Multiple-Strain Infection, Recombination, and Gene Loss. *J Infect Dis,* 220**,** 781-791.

TAYLOR-WIEDEMAN, J., SISSONS, J. G., BORYSIEWICZ, L. K. & SINCLAIR, J. H. 1991. Monocytes are a major site of persistence of human cytomegalovirus in peripheral blood mononuclear cells. *J Gen Virol,* 72 (Pt 9)**,** 2059-64.

TOWLER, J. C., EBRAHIMI, B., LANE, B., DAVISON, A. J. & DARGAN, D. J. 2012. Human cytomegalovirus transcriptome activity differs during replication in human fibroblast, epithelial and astrocyte cell lines. *J Gen Virol,* 93**,** 1046-1058.

TRGOVCICH, J., CEBULLA, C., ZIMMERMAN, P. & SEDMAK, D. D. 2006. Human cytomegalovirus protein pp71 disrupts major histocompatibility complex class I cell surface expression. *J Virol,* 80**,** 951-63.

VARNUM, S. M., STREBLOW, D. N., MONROE, M. E., SMITH, P., AUBERRY, K. J., PASA-TOLIC, L., WANG, D., CAMP, D. G., RODLAND, K., WILEY, S., BRITT, W., SHENK, T., SMITH, R. D. & NELSON, J. A. 2004. Identification of proteins in human cytomegalovirus (HCMV) particles: the HCMV proteome. *J Virol,* 78**,** 10960-6.

VEY, M., SCHÄFER, W., REIS, B., OHUCHI, R., BRITT, W., GARTEN, W., KLENK, H. D. & RADSAK, K. 1995. Proteolytic processing of human cytomegalovirus glycoprotein B (gpUL55) is mediated by the human endoprotease furin. *Virology,* 206**,** 746-9.

VINUESA, V., BRACHO, M. A., ALBERT, E., SOLANO, C., TORRES-PUENTE, M., GIMÉNEZ, E., GONZÁLEZ-CANDELAS, F. & NAVARRO, D. 2017. The impact of virus population diversity on the dynamics of cytomegalovirus DNAemia in allogeneic stem cell transplant recipients. *J Gen Virol,* 98**,** 2530-2542.

VOGELBERG, C., MEYER-KÖNIG, U., HUFERT, F. T., KIRSTE, G. & VON LAER, D. 1996. Human cytomegalovirus glycoprotein B genotypes in renal transplant recipients. *J Med Virol,* 50**,** 31-4.

VONGLAHN, W. C. & PAPPENHEIMER, A. M. 1925. Intranuclear Inclusions in Visceral Disease. *Am J Pathol,* 1**,** 445-466.3.

WANG, D. & SHENK, T. 2005a. Human cytomegalovirus UL131 open reading frame is required for epithelial cell tropism. *J Virol,* 79**,** 10330-8.

WANG, D. & SHENK, T. 2005b. Human cytomegalovirus virion protein complex required for epithelial and endothelial cell tropism. *Proc Natl Acad Sci U S A,* 102**,** 18153-8.

WANG, X., SANCHEZ, J., STONE, M. J. & PAYNE, R. J. 2017. Sulfation of the Human Cytomegalovirus Protein UL22A Enhances Binding to the Chemokine RANTES. *Angew Chem Int Ed Engl,* 56**,** 8490-8494.

WARRIS, S., SCHIJLEN, E., VAN DE GEEST, H., VEGESNA, R., HESSELINK, T., TE LINTEL HEKKERT, B., SANCHEZ PEREZ, G., MEDVEDEV, P., MAKOVA, K. D. & DE RIDDER, D. 2018. Correcting palindromes in long reads after whole-genome amplification. *BMC Genomics,* 19**,** 798.

WEEKES, M. P., TOMASEC, P., HUTTLIN, E. L., FIELDING, C. A., NUSINOW, D., STANTON, R. J., WANG, E. C. Y., AICHELER, R., MURRELL, I., WILKINSON, G. W. G., LEHNER, P. J. & GYGI, S. P. 2014. Quantitative temporal viromics: an approach to investigate host-pathogen interaction. *Cell,* 157**,** 1460-1472.

WELLER, T. H., CRAIG, J. M., MACAULEY, J. C. & WIRTH, P. 1957. Isolation of intranuclear inclusion producing agents from infants with illnesses resembling cytomegalic inclusion disease. *Proc Soc Exp Biol Med,* 94**,** 4-12.

WELLER, T. H., HANSHAW, J. B. & SCOTT, D. E. 1960. Serologic differentiation of viruses responsible for cytomegalic inclusion disease. *Virology,* 12**,** 130-2.

WHITE, R., PELLEFIGUES, C., RONCHESE, F., LAMIABLE, O. & ECCLES, D. 2017. Investigation of chimeric reads using the MinION. *F1000Res,* 6**,** 631.

WICK, R. R., JUDD, L. M., GORRIE, C. L. & HOLT, K. E. 2017. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genom,* 3**,** e000132.

WILFINGER, W. W., MACKEY, K. & CHOMCZYNSKI, P. 1997. Effect of pH and ionic strength on the spectrophotometric assessment of nucleic acid purity. *Biotechniques,* 22**,** 474-6, 478-81.

WILKINSON, G. W. G., DAVISON, A. J., TOMASEC, P., FIELDING, C. A., AICHELER, R., MURRELL, I., SEIRAFIAN, S., WANG, E. C. Y., WEEKES, M., LEHNER, P. J., WILKIE, G. S. & STANTON, R. J. 2015. Human cytomegalovirus: taking the strain. *Medical Microbiology and Immunology,* 204**,** 273-284.

WILLIAMS, S. L., HARTLINE, C. B., KUSHNER, N. L., HARDEN, E. A., BIDANSET, D. J., DRACH, J. C., TOWNSEND, L. B., UNDERWOOD, M. R., BIRON, K. K. & KERN, E. R. 2003. In vitro activities of benzimidazole D- and L-ribonucleosides against herpesviruses. *Antimicrob Agents Chemother,* 47**,** 2186-92.

WILM, A., AW, P. P., BERTRAND, D., YEO, G. H., ONG, S. H., WONG, C. H., KHOR, C. C., PETRIC, R., HIBBERD, M. L. & NAGARAJAN, N. 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res,* 40**,** 11189-201.

WORKMAN, R. E., TANG, A. D., TANG, P. S., JAIN, M., TYSON, J. R., RAZAGHI, R., ZUZARTE, P. C., GILPATRICK, T., PAYNE, A., QUICK, J., SADOWSKI, N.,

HOLMES, N., DE JESUS, J. G., JONES, K. L., SOULETTE, C. M., SNUTCH, T. P., LOMAN, N., PATEN, B., LOOSE, M., SIMPSON, J. T., OLSEN, H. E., BROOKS, A. N., AKESON, M. & TIMP, W. 2019. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat Methods,* 16**,** 1297-1305.

WYATT, J. P., SAXTON, J., LEE, R. & PINKERTON, H. 1950. Generalized cytomegalic inclusion disease. *J Pediatr,* 36**,** 271-94, illust.

XU, H., LUO, X., QIAN, J., PANG, X., SONG, J., QIAN, G., CHEN, J. & CHEN, S. 2012. FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS One,* 7**,** e52249.

XU, Y. & GANEM, D. 2010. Making sense of antisense: seemingly noncoding RNAs antisense to the master regulator of Kaposi's sarcoma-associated herpesvirus lytic replication do not regulate that transcript but serve as mRNAs encoding small peptides. *J Virol,* 84**,** 5465-75.

XUAN, B., QIAN, Z., TORIGOI, E. & YU, D. 2009. Human cytomegalovirus protein pUL38 induces ATF4 expression, inhibits persistent JNK phosphorylation, and suppresses endoplasmic reticulum stress-induced cell death. *J Virol,* 83**,** 3463-74.

YAN, J., FENG, J., HOSONO, S. & SOMMER, S. S. 2004. Assessment of multiple displacement amplification in molecular epidemiology. *Biotechniques,* 37**,** 136-8, 140-3.

YANG, S., GHANNY, S., WANG, W., GALANTE, A., DUNN, W., LIU, F., SOTEROPOULOS, P. & ZHU, H. 2006. Using DNA microarray to study human cytomegalovirus gene expression. *J Virol Methods,* 131**,** 202-8.

YE, J., COULOURIS, G., ZARETSKAYA, I., CUTCUTACHE, I., ROZEN, S. & MADDEN, T. L. 2012. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics,* 13**,** 134.

YU, D., SILVA, M. C. & SHENK, T. 2003. Functional map of human cytomegalovirus AD169 defined by global mutational analysis. *Proc Natl Acad Sci U S A,* 100**,** 12396-401.

ZAWILINSKA, B., SZOSTEK, S., KOPEC, J., PIATKOWSKA-JAKUBAS, B. & KOSZ-VNENCHAK, M. 2016. Multiplex real-time PCR to identify a possible reinfection with different strains of human cytomegalovirus in allogeneic hematopoietic stem cell transplant recipients. *Acta Biochim Pol,* 63**,** 161-166.

ZHANG, G., RAGHAVAN, B., KOTUR, M., CHEATHAM, J., SEDMAK, D., COOK, C., WALDMAN, J. & TRGOVCICH, J. 2007. Antisense transcription in the human cytomegalovirus transcriptome. *J Virol,* 81**,** 11267-81.

ZHANG, L., YU, J. & LIU, Z. 2020. MicroRNAs expressed by human cytomegalovirus. *Virol J,* 17**,** 34.

ZHANG, Z., SCHWARTZ, S., WAGNER, L. & MILLER, W. 2000. A greedy algorithm for aligning DNA sequences. *J Comput Biol,* 7**,** 203-14.

ZHOU, X., XU, Y., ZHU, L., SU, Z., HAN, X., ZHANG, Z., HUANG, Y. & LIU, Q. 2020. Comparison of Multiple Displacement Amplification (MDA) and Multiple Annealing and Looping-Based Amplification Cycles (MALBAC) in Limited DNA Sequencing Based on Tube and Droplet. *Micromachines (Basel),* 11.

ZUHAIR, M., SMIT, G. S. A., WALLIS, G., JABBAR, F., SMITH, C., DEVLEESSCHAUWER, B. & GRIFFITHS, P. 2019. Estimation of the worldwide seroprevalence of cytomegalovirus: A systematic review and meta-analysis. *Rev Med Virol,* 29**,** e2034.