2022

Bioinformatic Pipeline for Determining Terminal Repeats in the Human Cytomegalovirus Genome Assembled with PacBio Long Read Sequences

Ahmed Al Qaffas

Bioinformatic Pipeline for Determining Terminal Repeats in the Human Cytomegalovirus Genome Assembled with PacBio Long Read Sequences

By

Ahmed Ali Al Qaffas, Masters in Bioinformatics

Advisor:Michael McVoy, Professor of Pediatrics, Department of Pediatrics

Virginia Commonwealth University, Richmond VA,

August 2022

Acknowledgment

I wish to express my sincere gratitude to my mom, Dr. Sameerah Zaid Al Modhi, for her everlasting love and support. She is the light in my life. I would also like to thank my grammy, Hakeemah Al Qaffas, for being the best grammy ever and teaching me how to be a better human. I would also like to thank Dr. Michael Mcvoy, Jeffery Elhai, Dr. Allison Johnson, and Dr. Luiz Ozaki for guiding me through the years. I pray to God to enlighten your days and let your shine forever stay. Thank you! Finally, I would like to say to Leo, Marraowi, Meme, Fto, Big Bro Mo, Darren, Danny, John and Roshni, Vasu, Z, Red, Jared Mann, Ruairidh, and LeClare that I love you guys and I really appreciate everyone of you. I wouldn’t be the man I am today without you.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

HCMV	Human Cytomegalovirus
TRL	Terminal Repeat Long
TRS	Terminal Repeat Short
IRL	Inverted Repeat Long
IRS	Inverted Repeat Short
UL	Unique Long
US	Unique Short
CCS	Circular Consensus Sequencing read
P	Prototype (isoform)
IL	Inverted Long (isoform)
IS	Inverted Short (isoform)
ILIS	Inverted Long-Inverted Short (isoform)
ORF	Open Reading Frame

ABSTRACT

BIOINFORMATIC PIPELINE FOR DETERMINING TERMINAL REPEATS IN THE HUMAN CYTOMEGALOVIRUS GENOME ASSEMBLED WITH PACBIO LONG READ SEQUENCES

By Ahemd Ali Al Qaffas, M.B.

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Bioinformatics at Virginia Commonwealth University.
Virginia Commonwealth University, 2022.
Major Advisor: Michael Mcvoy, Professor of Pediatrics, Department of Pediatrics

Human Cytomegalovirus (HCMV) is a member of the betaherpesvirinae subfamily of the Herpesvirus family. HCMV infection is common among adults worldwide, with an estimated seroprevalence of 66 to 95%, depending on the geographic region (Zuhair et al., 2019). Although most of the virus genomic content has been studied extensively, the terminal repeating region sequences remain understudied. Two main challenges hindered the study of the region: a) limitations of sequencing technologies; and b) misassembly of the repeats due to its complex nature. Here I show a novel bioinformatics pipeline that takes advantage of PacBio's long reads to resolve the challenges mentioned earlier. Implementation of the pipeline yielded results that supported previous assumptions of the terminal region, showed evidence of new findings, and provided in-depth analysis of the terminal repeat known as the a sequence.

Keywords: Human Cytomegalovirus, HCMV, PacBio, Third-Generation Sequencing, Bioinformatics.

INTRODUCTION

1.1) Background

Human Cytomegalovirus (HCMV) is a member of the betaherpesvirinae subfamily of the Herpesvirus family. HCMV infection is common among adults worldwide, with an estimated seroprevalence of 66 to 95%, depending on the geographic region (Zuhair et. al, 2019), and seroprevalence ranging from 45 to 100% among women of reproductive age (Cannon et. al, 2010). Overall, HCMV infection is typically asymptomatic but considered life-threatening in immunocompromised individuals such as AIDS patients, people who have gone through organ transplantation, and the elderly (Griffiths, 2015). HCMV infection can be classified in three types: primary, non-primary, and congenital. Most people are infected through exposure to body fluids, sexual contact, and organ transplantation. The primary infection has mild flu-like symptoms such as fever, cough, and sore throat. Non-primary infection is of two subtypes; the first, infection with a different strain of an already infected host. The second, a lifetime recurring infection from reactivated latent viruses. Recurring non-primary infection occurs due to the failure of the immune system to completely clear the primary viral infection resulting in viral latency (Wang and Zhao, 2020). Finally, congenital HCMV infection of a fetus follows either primary infection, reinfection, or recurrent infection of the mother. About 10% of all newborns with congenital HCMV infection show disabilities such as hearing loss and mental disabilities (Leung et.al, 2003). Congintal HCMV infection is the most frequent among all congenital infections with an estimated prevalence of 2.5% of all newborns am other congenital infections (Stagno et. al, 1986). HCMV can spread via the bloodstream. It has a wide range of cell tropism that includes epithelial cells, endothelial cells, fibroblasts, and leukocytes (Gerna, 2019; Sinzger et.al, 2008).

HCMV is associated with an increase in morbidity and mortality in people with cardiovascular diseases (Simanek et al., 2011), different cancerous malignancies (Fulkerson et al., 2021), or when reactivated during sepsis (Kalil and Florescu, 2011). The virus is prone to establishing life-long latency (Reeves and Sinclair, 2008). The impact HCMV has on people's health worldwide makes it one of the most highly studied viruses for the past sixty years. Still, a wide knowledge gap exists concerning its behavior in human hosts and the structure of its complex genome.

A total of 334 complete genomic DNA sequences have been submitted to GenBank. Most of the HCMV genomic sequences have been reported accurately except for the terminal and junctional repeats. An accurate reconstruction of these regions is hindered by the limitation of sequencing technologies as well as the complexity within the viral terminal sequence. For these reasons, the functional relevance of the repeats remains mostly unknown. Here I show a novel bioinformatical pipeline that utilizes third-generation sequencing and a multi-mapping process to generate an accurate structure prediction of the terminal repeat nucleotide sequences, their frequency and prevalence in the virus genome assembled with long reads sequencing technology (PacBio).

1.2) Viral structure and genome

HCMV is a double-stranded DNA virus with high guanine-to-cytosine content. The ~230-kb genome with about 170 open reading frames is encapsidated within a capsid which is surrounded by tegument proteins and covered by a lipid bilayer envelope (Figure 1). The envelope contains different glycoproteins that are essential for viral attachment and entry into cells (Crough and Khanna, 2009).

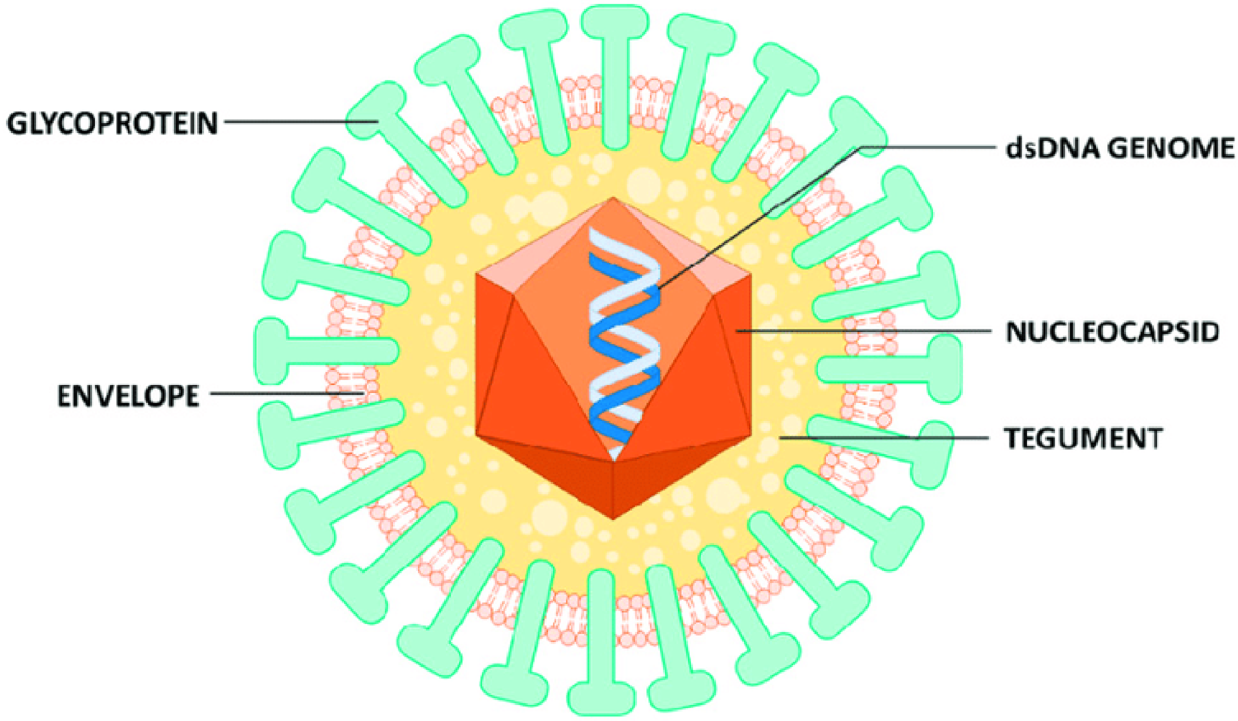


Figure 1) HCMV structure (from Gugliesi et. al., 2020). Mature virion with its genome enclosed by three different layers: nucleocapsid, tegument proteins, and envelope.

Herpesviruses have six different genomic structures that are classified from A to F (Figure 2). The HCMV genome belongs to class E. Its genome contains long (L) and short (S) segments, defined by terminal and internal inverted repeats (TRL/IRL, IRS/TRS) flanking unique long (UL) and unique short (US) regions with the following pattern: TRL-UL-IRL-IRS-US-TRS. Homologous recombination during genome replication gives rise to four different genome isoforms in which UL and US regions are inverted. TRL is composed of two repeating sequences, *a* and *b,* while IRL is composed of the inverted sequence of *a* and *b* known as *a'* and *b'.* Similarly, TRS is composed of *c* and *a sequences* while IRS is composed of the inverted sequences known as *c'* and *a'.* The region of the genome where all three repeats are present is known as the junction region (Figure 3). Some genomes have multiple *a sequences* on the left end and/or at the junction region at lower prevalence. The right end of the genome can either have one *a sequence* copy or entirely lack an *a sequence.*

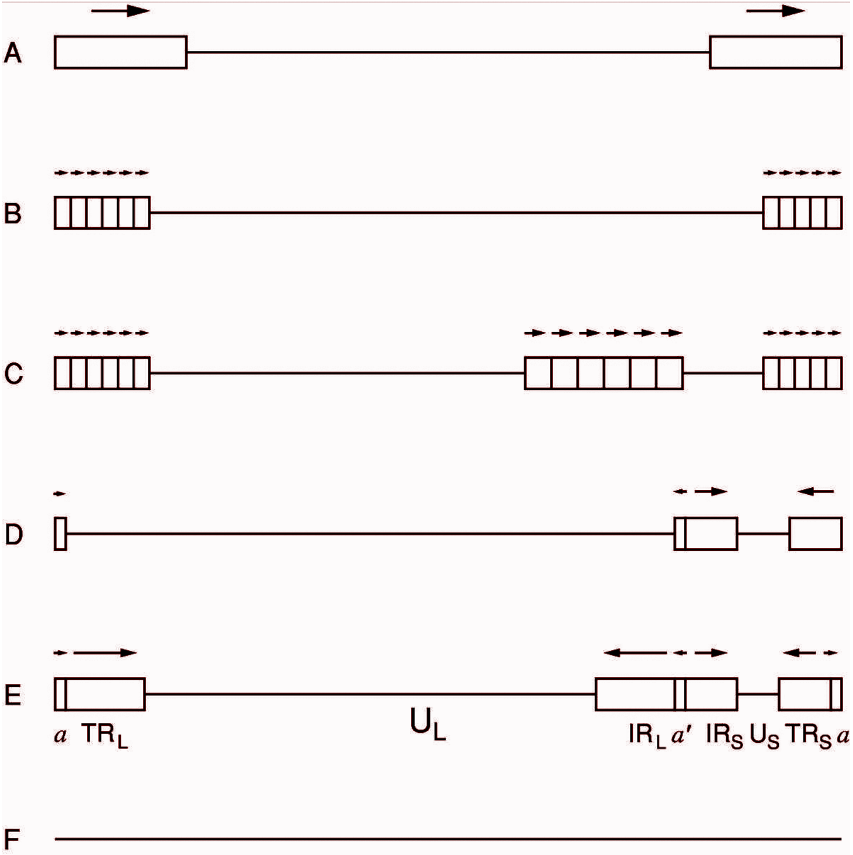


Figure 2) Classes of *Herpesvirus* genome structures (not to scale) as defined by Roizman and Pellett (2001). Unique and repeat regions are shown as horizontal lines and rectangles, respectively. The orientations of repeats are shown by arrows. The nomenclature of unique and repeat regions, including the terminal redundancy (*a*) and its internal, inverted copy (*a'*), is indicated for the class E genome.

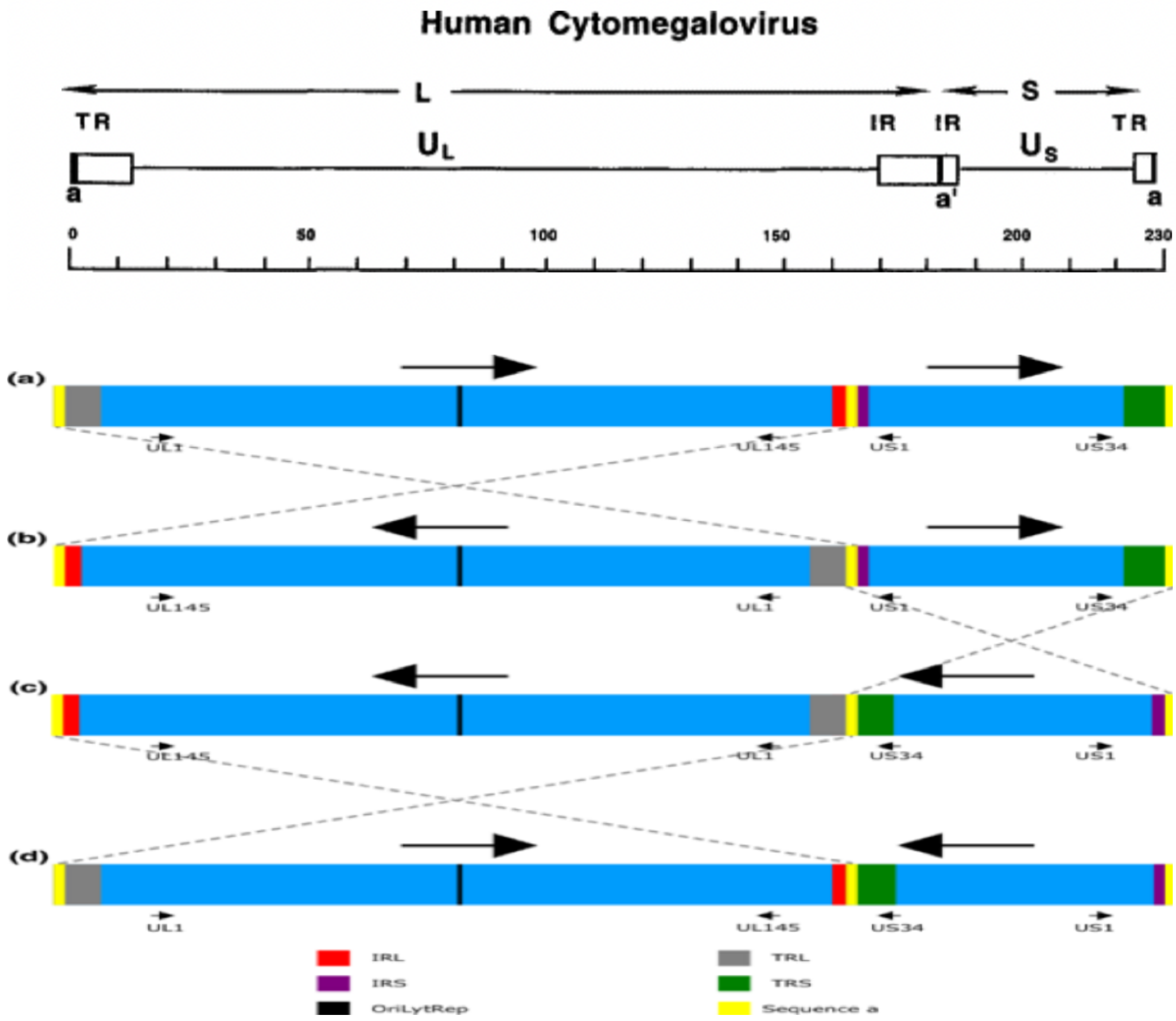


Figure 3) HCMV genomic structure and isoforms. The four isoforms: (a) Prototype; P, (b) inverted long; IL, (c) inverted long inverted short; ILIS, and (d) inverted short; IS. The genome has three repeat sequences: the *a sequence* (yellow) present as *a (direct)* or *a'* (inverted) copies; the *b sequence* (red or gray), present as *b* (direct) or *b'* (inverted) copies, and the *c sequence* (purple or green) present as *c* (direct) or *c'* (inverted) copies. Recombination between inverted repeats generates the four isoforms with a predicted equimolar ratio where a complete inversion of L (*a*-*b*-UL-*b'*-*a'*) or S segments (*a'*-c'-US-*c*-*[a]*) occurs (Shanley, 2000; Martí-Carreras and

Maes, 2019). The area that contains all three repeats in between the L and S segments is known as the junction region.

HCMV strains aren't defined by a phenotypic profile, unique surface proteins, or the total variation of the genomic composition. HCMV strains are defined by either analyzing the lengths of their endonuclease fragments at polymorphic regions (Adler, 1988), by polymorphisms within the *a sequence* (Bale et. al., 1993), or by analysis of short tandem repeating sequences across the viral genome (Walker et. al., 2001).

Genomes of different HCMV strains and their genetic features have been compared before (Sijmons et al., 2014) to aid in the understanding of the functions of genes located within the UL and US regions. It has been found that only 45 out of about 170 genes are essential for viral replication in cell culture, all of which are located within the UL region, leaving the whole US genomic region dispensable (Dunn et.al, 2003). Most genes that are located within the UL region are abbreviated with Unique Long followed by a numeric reference (e.g., *UL1*). Similarly, genes that are located within the US region are abbreviated with Unique Short followed by a numeric reference.

Certain genes and their subsequent protein products are essential in targeting special cell types. For instance, a trimeric complex of glycoproteins H, L, and O are believed to be required for entry into all cell types, while and additional pentameric complex of gH and gL with UL128, UL120, and UL131A is required for entry into endothelial and epithelial cells. Other genes are essential for DNA replication such as those encoding the Immediate Early protein 2 and UL84 (Vanarsdall and Johnson, 2012). The essentiality of a gene, however, can be dependent on the target cell. For example, genes encoding UL128, UL130, and UL131A are critical for viral replication in endothelial and epithelial cells but are not essential in fibroblast replication. Wild-type HCMV has a wider range of cell tropism while laboratory strains often have lost genes needed to infect certain cell types (Wilkinson et. al., 2015).

One of the most understudied regions of the HCMV genome are the terminal repeat regions. The terminal repeats TRL and TRS are composed of the short repetitive sequences called *a* - *b* and *c* - *a* sequences respectively (Mcvoy and Nixon, 2005). The functional significance of the terminal repeats of HCMV remains largely unknown. One reason that accounts for such ambiguity is the nature of the genomic regions. The regions are composed of multiple short repeats and unique segments in between these repeats of different lengths. In addition, the regions as a whole are highly polymorphic between strains. The terminal repeat sequences vary in length; the *a sequence* is about 700 bp long, the *b sequence* \varies from a single base to 11 kb, while the *c sequence* consists of ~ 2000 bp. The repeats were originally defined by restriction cleavage which showed three distinct DNA fragments of different sizes. The naming was made based on the size of the fragment.

The *a sequence* is poorly annotated in many genomic sequences due to its complex nature which consists of short repeating sequences ranging from 8 to 23 bp (Figure 4). The *a sequence*, in particular, holds a greater significance for it is the only repeating sequence that is present at both termini and the junctional region, has a multiple copy number with different frequencies, and inner variation within each sample. The copy number of *a sequences at* left termini is most often one but left ends with two or three *a sequences* occur with decreasing prevalence. On the right end the copy number of *a sequences* is either one or zero. Both *b* and *c sequences* have a similar repetitive short nucleic region but each has a single copy in their respective genomic locations. Recalling that *a* and *b sequences* are highly polymorphic between strains adds to the hurdle in analyzing these regions.

Like bacteriophages, herpesviruses reproduce their genomes as concatemers that are packaged into premade procapsids. Other genes are essential for DNA packaging such as UL51, UL56, and UL89 which encode subunits of the HCMV terminase complex. Additional components are required for DNA packaging, including the *cis*-acting DNA packaging signals *pac1* and *pac2,* located near the terminal regions of the genome (Theiß et.al, 2019). Although based on herpesvirus-conserved sequence motifs a *pac1* sequence is present near the left end of the HCMV *a sequence*, the HCMV *pac2*, presumed to be located near the end of the *a sequence*, lacks canonical features of herpesvirus *pac2* elements and remains cryptic. Moreover, that the right ends of some HCMV genomes lack an *a sequence* and instead end with a *c sequence* suggests that formation of these ends may rely on a second *pac2* element near the end of the *c sequence*. However, the functional relevance of these putative cryptic *pac2* sequences has not be confirmed by mutagenesis. The current understanding of functional regions within the HCMV genome across all its strains remains underdeveloped. Generating a good annotation of the terminal repeats will help in a better understanding of the virus.

Although the *a sequence* appears important for genome isomerization, the functional relevance of the process remains unknown (Sauer et. al, 2010). Most of the repeats and their sequences remain functionally ambiguous. The entire *b sequence* is functionally ambiguous due to strain-to-strain polymorphism and the number of internal repeats. This is also the case for most of the *a* and *c sequences.*

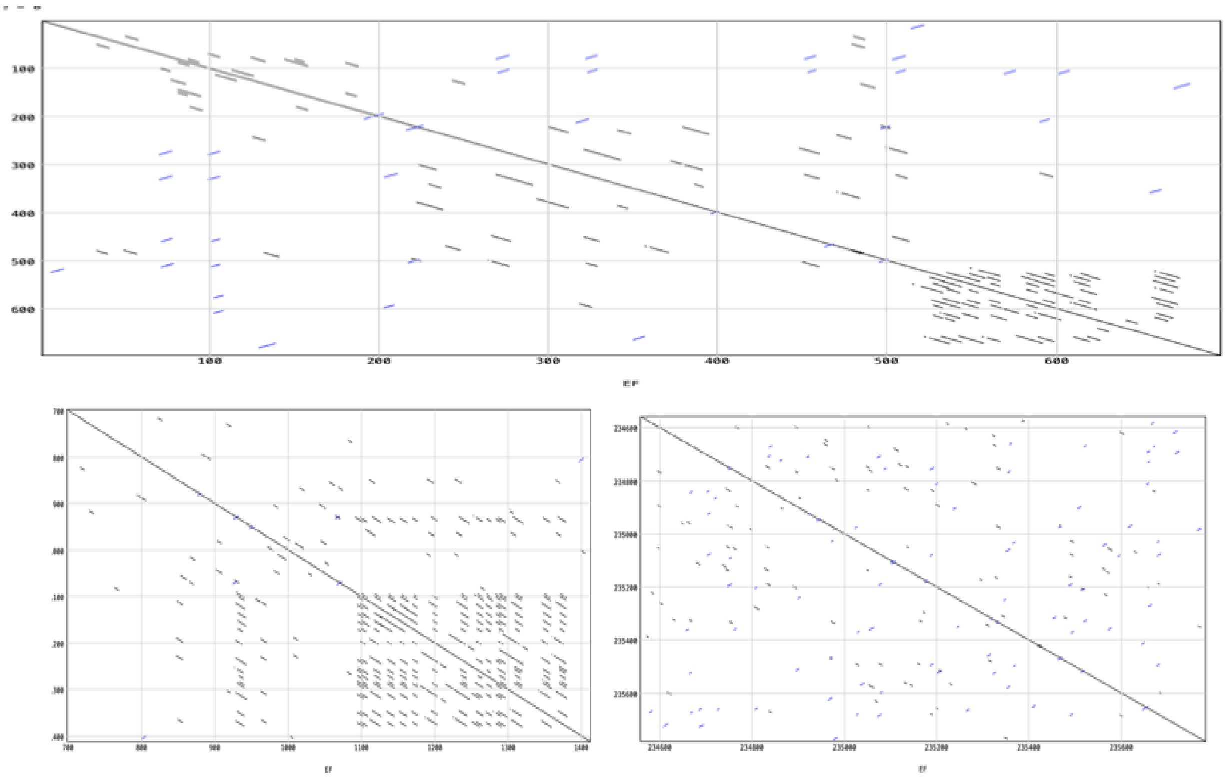


Figure 4) Dot-Plot of terminal repeats from HCMV strain TB40-EF (ASN: MW439038.1) against themselves. The *a sequence* (top) is composed of small repeating sequences ranging from 8 bp to 23 bp. The *b sequence* (bottom left) contains short repeats ranging from 6 to about 30 bp. The *c sequence* (bottom right) is less complex but contains multiple short sequences of ~ 10 bp long.

The functional importance of the terminal repeats remains, for the most part, unknown due to the complexity of the region and the limitations of sequencing technologies that hinders the proper analysis of the repeats. A known function of the repeat region is differentiation between types of HCMV strains. For example, Guinea pig cytomegalovirus (GPCMV) has a more simplistic arrangement of a 1 kb terminal repeat sequence compared to HCMV. Their known function is to divide GPCMV genomes into two types: type 1 contains a single copy on the left end of the genome while type 2 genomes have a single copy on each end (Gao and Isom, 1984). It was found (Nixon and McVoy, 2002) that circularization of type 2 GCMV genome leads to deletion of one copy of the terminal repeat but those copies are duplicated and restored during cleavage events. Such occurrence indicates that replication of type 1 genomes has a piece of different machinery that leads to the formation of either type simultaneously. This is the case across different cytomegaloviruses. It’s worth noting that *a sequences* counterparts within other herpesvirus genomes contain a hypothetical signal that is essential in the inversion of the L and S genome components during recombination events. This DNA signal is located within a region of multiplicity of heterogeneous short-tandem-repeats. In HCMV, the terminal repeats are more complex, have a greater length, and greater polymorphism between different strains.

1.3) Sequencing Technologies and Their Challenges

Analyzing the terminal repeats starts with choosing the appropriate sequencing technology. Illumina (Cronn et.al, 2008) is the most frequently used next-generation sequencing technology (NGS; Segerman, 2020). Like many other NGS technologies, Illumina calling of bases is done with each nucleotide incorporation in the DNA by DNA-polymerases in a parallel high throughput fashion (Hu et. al., 2021). Illumina reads are typically 50 bp to 300 bp long with read quality of 1 error every 1000 bases sequenced (Ewing et.al, 1998). Although Illumina sequencing is ideal for some organisms, it is not a good candidate for sequencing any viral complete genomes of higher complexity. Illumina's short reads are too short to be used with assembly nor dealing with complex long repeating sequences. In particular, the nature of TRL, junctional region, and TRS short repeats and their variable length is both a computational and logical challenge. Assembling HCMV regions of repeats typically results in a consensus that contains wide gaps in each region or a low mapping quality score [MQ = 0] (figure 5).

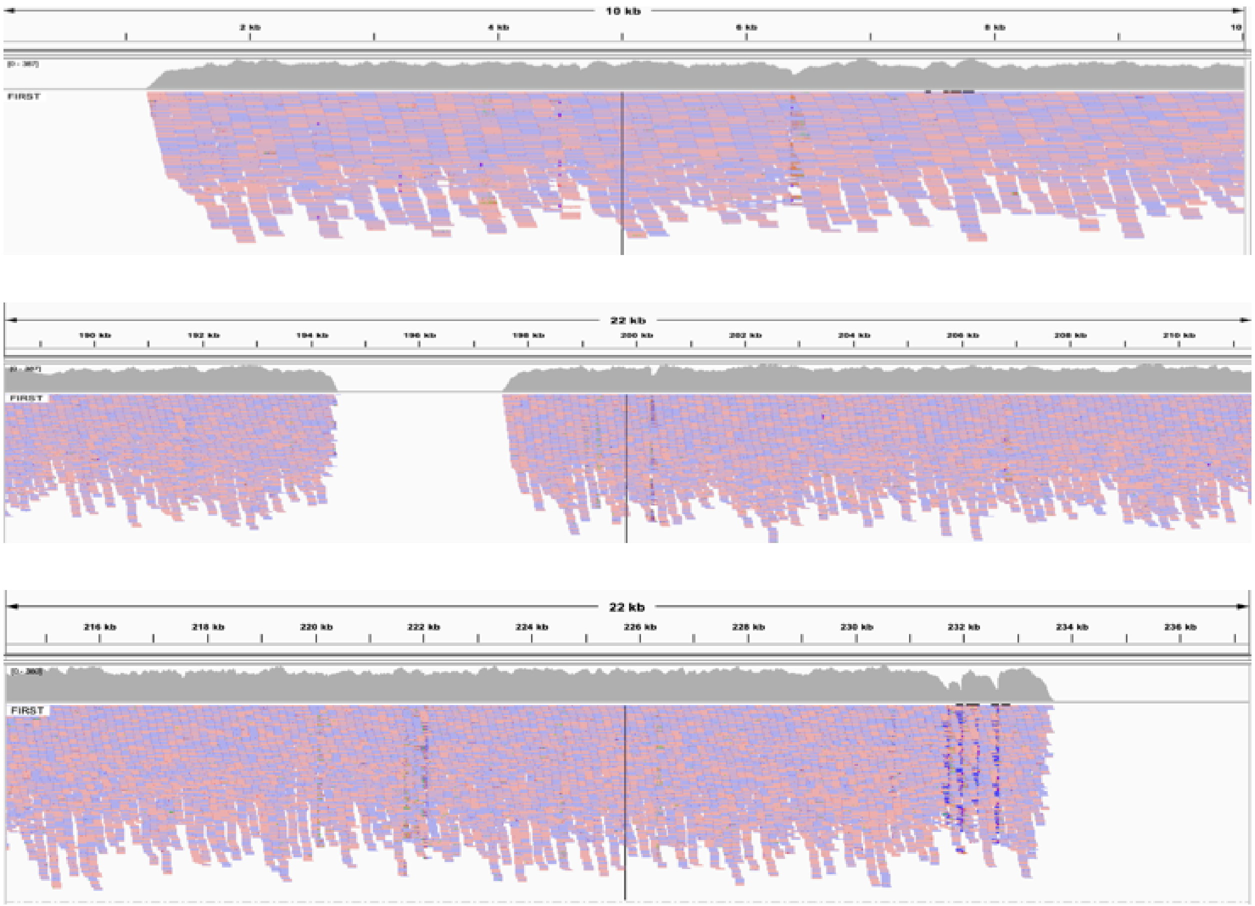


Figure 5) HCMV strain Ig-KG-H2p28S (H2p28; unpublished) Illumina read mapping. Illumina reads were used to assemble strain H2p28 using its parent strain (MT894141.2). Reads with low mapping quality were removed. Mapping results show a gap between TRL (top panel), junctional region (middle panel), and TRS (bottom panel). Both parent and offspring ought to have very low genomic variance. However, the nature of HCMV repeats and the limitation of NGS have posed a hurdle in mapping and assembling the reads to the terminal repeating region of the genome.

Except for a few labs that have specific techniques to deal with the repeating regions, most of the submitted whole genomic sequences of HCMV have a poor annotation of these regions. To deal with the issue of assembling the repeats using Illumina reads, we collaborated with Dr. Andrew Davison's lab in Glasgow, UK. The technique used in their lab is a stepwise elongation of the junctional repeats in multiple sequencing runs. After generating a consensus of the junctional repeats, *the b'-a' sequence* and the *c'-a' sequence* are reverse- complemented, and 'pasted' in their respective genomic regions. This approach has two main issues: it assumes that the repeats are identical across the genome and it fails to predict the exact copy number of the *a sequence* across the genome. Although Illumina sequencing technology generates great quality reads with high coverage, the complexity of repeat sequences and the short length of the reads prevents a correct assembly of the terminal repeat sequences.

Another approach to reconstructing the terminal repeats would be using third-generation sequencing (3GS). 3GS can produce long reads with lengths in kilobases without the need for the amplification step required by second-generation sequencing. Single-Molecule, Real-Time (SMRT) sequencing technology, developed by Pacific Bioscience (PacBio; http://www.pacificbiosciences.com/), is the most frequently used 3GS (Mehdi et.al, 2017). SMRT sequencing starts by segmenting a whole genomic DNA . Each segment gets isolated in a single cell where two DNA adapter sequences are ligated to it. Later, annealing of the DNA molecule occurs then two primers are added to the adapters. The newly generated complex is then bound to a DNA polymerase. At this stage, the circular template will get sequenced over multiple times, generating subreads. A consensus of the sub-reads is then generated, representing the initial DNA segment. SMRT sequencing can produce reads with an average of 10kb in length. A hurdle of using PacBio reads would be the higher error rate of 13-15% generated as either deletions or insertions (Kulski, 2016). However, using the circular consensus sequencing (CCS) technique reduces the error rate to less than 1%. CCS reads (also known as HiFi reads) are generated by creating a consensus of subreads sequenced in a single circularized DNA molecule by passing DNA polymerase over the segment multiple times. Each read is the consensus of both forward and reverse strand (figure 6).

Having longer and more accurate reads that span the terminal region is useful in reconstructing a more accurate representation of the HCMV genome. Using PacBio CCS reads has one major issue: the average read length is dependent on the fragmented DNA molecules. Meaning, that read length is affected by either random fragmentation or by reaching the end of a genome. With HCMV, reads of opposite termini may map wrongfully to the other end of the junction region due to mapping error during the assembly of reads. Such errors are generated when the sequence within reads is similar enough to multiple parts of the used reference. For example, reads that contain *a* and *b sequences* of the left termini can map to the junction region due to the high similarity of their sequences and that of the reference sequence or due to clipping preferences used by mapping tools. It is worth mentioning that most of the commonly used assembly tools produce lower quality viral genome assemblies when compared with larger organisms' genome assemblies (Sutton et.al, 2019). With the use of CCS reads, assembling and analyzing the repeats can be made with great precision. Table 1 summarizes common advantages and disadvantages between NGS and 3GS. Choosing CCS reads is a more reasonable approach to correctly identifying the terminal repeats.

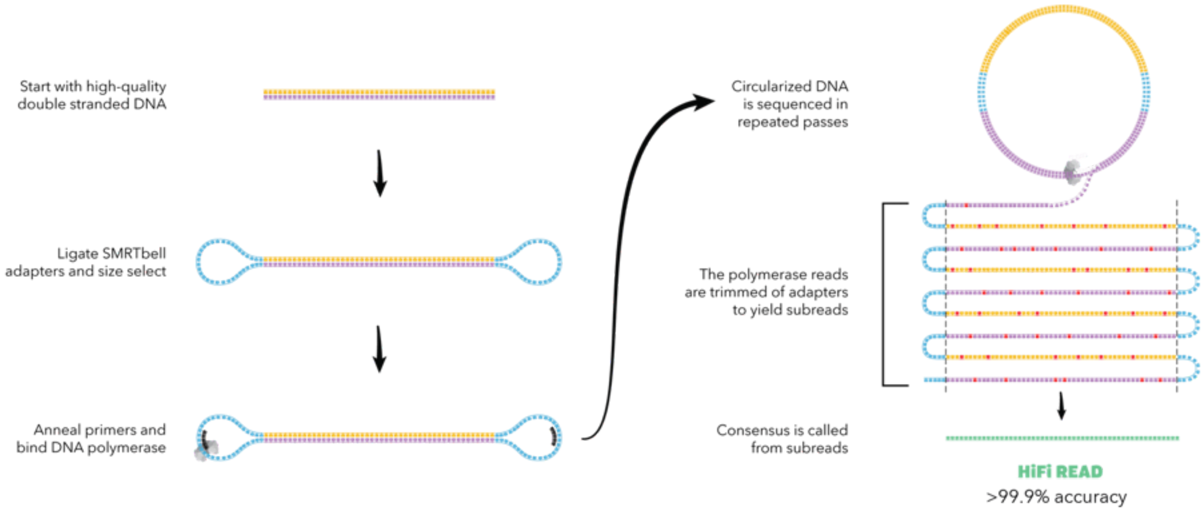

Start with high-quality double stranded DNA
Ligate SMRTbell adapters and size select
Anneal primers and bind DNA polymerase
Circularized DNA is sequenced in repeated passes
The polymerase reads are trimmed of adapters to yield subreads
Consensus is called from subreads
HiFi READ
>99.9% accuracy

Figure 6) Illustration of SMRT sequencing. Each DNA segment is represented by a single PacBio read (HiFi read). Adapters are added to the segment alongside primers. DNA annealing results in the circulation of the newly formed complex.

Table 1) Common advantages and disadvantages between NGS and 3GS.

	Read depth	Error Rate	Cost	Read length
NGS	> 1000	Error < 1/1000	++	200bp
3GS	< 500	11 - 15%	++++	15kb on average

1.4) Misalignment of terminal repeats

Correct mapping of the reads is required to analyze the terminal repeats effectively. The length of PacBio reads poses an issue during the alignment of reads to their reference. Most mapping tools use a matching percentage score from a position-specific matrix or a graph algorithm to align reads to the most likely genomic region. Since PacBio reads have an average of about 15kb in length, mapping of reads results in misalignment if a repeat sequence is long enough to be flagged as a valid mapping segment. For example, reads that include parts or all of the *a sequence* would possibly be mapped in three different genomic regions; once in each terminus and once in the junction. The latter is considered a misalignment because the correct reads belong to the *b'-a'* sequence which is the reverse complement of the *a-b* sequence. Another reason for such misalignment is that some reads belong to another isoform. The genomic orientation of that isoform terminates with the reverse complement of the *a-b* sequence due to the inversion of the UL genomic segment. Similarly, reads from the right end that contain the *c-a* sequence will map to the junctional region due to the inversion of the S segment. Noting that since right ends contain either on or zero *a sequences*, it is expected to see less of wrongly mapped reads containing *c-a sequences* (figure 7). To perform any sort of quality analysis, this issue has to be resolved.

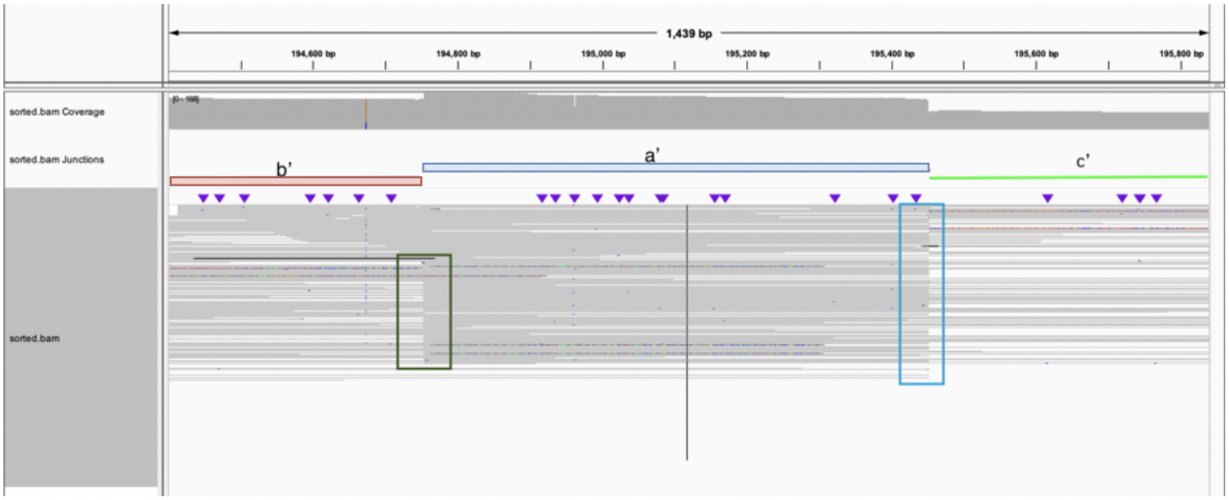


Figure 7) Misassembly of reads containing terminal repeats to the junctional region. Part of the *b' sequence* (red line) and *c' sequence* (green line) are shown. Reads that belong to the left terminal region of an isoform which have a hard stop at the right end of the *a sequence* are shown in the sky-blue box and reads that belong to the right end of the genome have a hard stop at the left end of the a sequence are shown in dark green box. Both sets of reads are wrongly mapped to the junctional region. Note that the frequency of wrongly mapped reads in the sky blue box is higher than the ones in the dark green box which is a direct indication of the prevalence of right end reads that have a *c sequence* without the *a sequence.*

The functional and structural ambiguity of the terminal repeats, the limitation sequencing technologies have, and the complexity of the regions raises a problem that can be solved bioinformatically. Here, I will demonstrate a computational approach to annotating the terminal repeats using PacBio CCS reads and a novel pipeline that will provide a better reconstruction of the repeats and an in-depth analysis of the region. Upon successful construction of the termini, the analysis of the regions will help in defining the ratio of genomic isoforms and validate the current assumption of equimolar presence, estimate the total copy number and possible frequencies of the a sequences within a sample, quantify the frequency of the a sequence at the right end, investigate evidence for heterogeneity within *a sequences* at different locations in the genome, locate cleavage sites at the right end, and infer *pac2* sequence and location. Gathering such information will help in developing a better understanding of HCMV and provide a possible antiviral approach.

Methods

2.1) Used tools

Reads were obtained from PacBio sequencing using Sequel 2 System 2 and HiFi SMRTbell library with P2/C2 chemistry according to the manufacturer's instruction at the Genomic Core at Virginia Commonwealth University. Evaluation of DNA quality was done using Agilent's Femto Pulse System. The Qubit fluorometer was used to quantify DNA concentration. Concentration and purification of the DNA samples were done by adding 0.45 volumes of AMPure PB magnetic beds. SMRTbell™ Express Template Prep Kit 2.0 was used for sequencing libraries preparation. BluePippin Systems was used to combine and size-selected electrophoretically with the 0.75% DF Marker S1 high-pass 6-10kb vs3 cassette. Qubit 1X dsDNA HS assay kit was used to measure the concentration of the library pool, and the Femto Pulse System was used to confirm the final size distribution. A total of 3 different raw reads of HCMV from the obtained reads will be used as follows: TB40EF (S1; ASN#MW439038), TB40EE (S2; ASN#MW439039), and Ig-KG-H2P14S (S3; ASN# MT894141.2). CCS of the reads set was obtained using PacBio SMRT Link software package ccs package version 4.2.0 (commit v4.2.0-1-g450908e4). The bam files containing the raw reads of S1, S2, and S3 were processed using the *CCS tool* with default arguments and the numbers of reads obtained are shown in table 2. Mapping was performed with minimap2 Version 2.18-r1015 (Li, 2018). Clipping minimization was performed by Samclip version 0.4. SAM/BAM file modifications were made using SAMTools and BCFtools version 1.13 (Danecek et.al, 2021). Minimap2 with arguments (-H -N -ax map-hifi), Samclip argument (--max 0), and samtools argument (view -Sbh -F 4) was used to generate a filtered alignment in a BAM format. The generated BAM that contains only the reads mapped to the reference was converted to fastQ format using samtools fastq package. Similarly, filtered reads were mapped to the complete genomic sequence of their respective whole-genome references for the next step alignment. Visualization of the alignment used beIGV2 version 10.3. BBMap tools, package 'filterbyname' version Sept,1, 2016. R studio (RStudio Team, 2020) was used to generate different plots with the aid of the ggplot2 package (Wickham 2016). Python (Van Rossum and Drake, 2009) was used to script a novel pipeline that performs mapping, format conversion, and calling of variants via command line.

Table 2) Number of CCS reads obtained from TB40EF (S1), TB40EE (S2), and Ig-KG-H2P14S (S3). CCS reads are statistically generated from multiple subreads generated from the same DNA fragment used in the sequencing machine.

Sample	Number of Raw Reads	Number of CCS Reads
S1	530635	16920
S2	270024	6985
S3	1118396	29340

2.2) Pipeline Logic and Description

2.2.A) Resolving read misalignment

To tackle the issue of read misalignment, two computational approaches were used: the first was to filter out reads known to map to the junctional region based on the reference coordinates. To filter out reads that belong to the junction region, reads that include parts of *b',* *a', and c'* sequences were extracted from the CCS file. The second step of this approach was to delete the entire junctional sequence from the reference FASTA file. Finally, filtered reads in the newly obtained CCS files were mapped to the modified reference file. The first approach runs in three mapping steps as follows:

- On the first mapping step, a small segment (SS) of length 80 bp was selected as a reference to align against all reads. The small segments contain either a right-end sequence of UL and left-end sequence of *b'* (UL-b') or right-end of *c'* and left-end of US sequence (c'-US). Reads that don't align to either one of the small segments will be filtered out leaving only mapped reads (Figure 8).

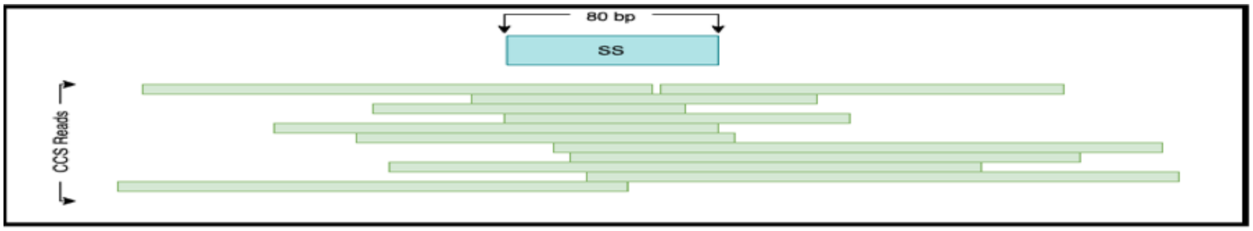


Figure 8) Mapped reads with either UL-b' or c'-US. Blue; small segment used as a reference. Green; CCS reads mapped to the sequence.

- In the next step, reads that are mapped to small segments are aligned against WGS as a reference. Such mapping results in two files, one for *UL-b'* and one for *c'-US*, each of which contains the alignment of reads in the terminal and the junction region. In theory, removing reads that are known to map to the junctional region would spare reads that map to the terminal sequences (Figure 9). The remaining reads are either terminated by random cuts during library preparation or by reaching the end of a genome. This approach filters reads according to the *a' sequence* coordinates. In the case of the *UL-b'* alignment file, reads that are aligned to the junctional region contain some that cross *b'-a'-c'* region and some that terminate before the a' sequence. Separating the two reads that are at least N base-pair (15> N >= 1) into the opposite end of a sequence ensures that selected reads contain at least N bp of *c' sequence*. Similarly, *c'-US* alignment file is filtered based on the opposite end of the a' sequence ensuring that at least N bp includes b' sequence. Validation of the output of this step can be made by visualizing the alignment. The filtered reads (*i.e.*, containing *b'-a'-c'*) are removed from the initial CCS reads.

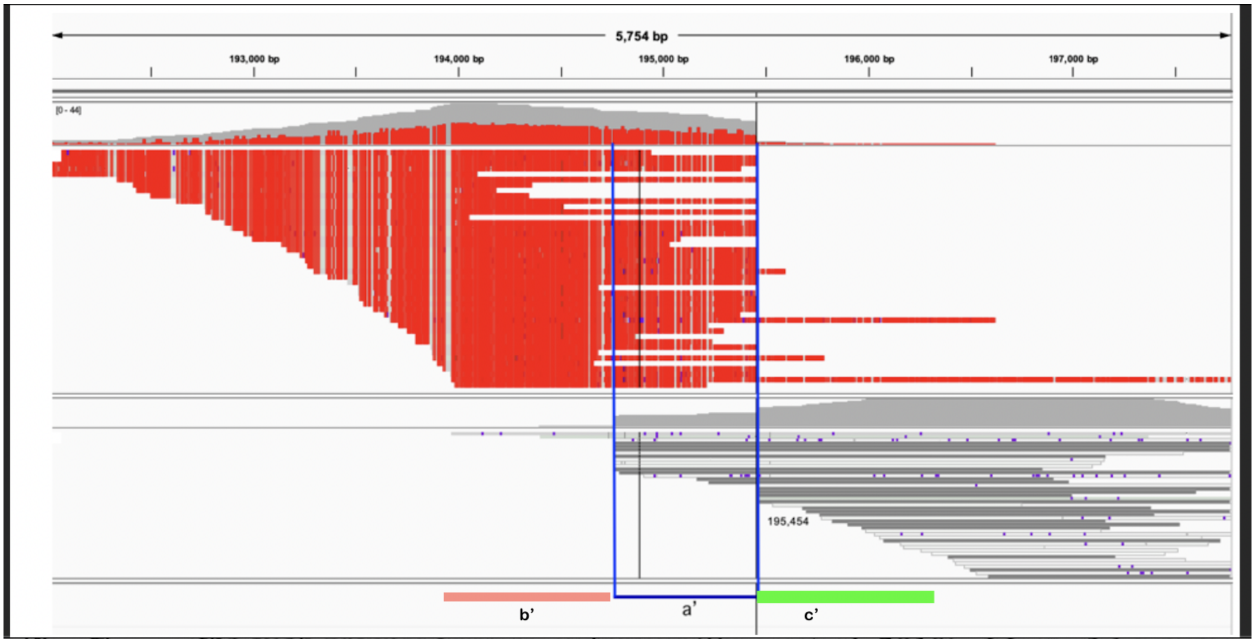


Figure 9) Second step mapping of raw reads to the complete genomic sequence. Target reads to be extracted are defined by any read that spans across the entire junctional region. Those are determined by a' sequence (blue line). Reads that are aligned to the *UL-b'* region (red) and those that are aligned to *c-US* (gray) are shown to have reads that contain parts of the *b'-a'-c' sequence.* Reads that fall outside of the opposite end of the blue box and most of the aligned reads in their respective file are the target reads for extraction.

- The third and final mapping step is to modify the reference sequence file by deleting the junction region altogether and mapping it against the newly filtered reads file. (Figure 10). The modified genomic sequence is used for the final alignment against the reads, resulting in mapping to the ends of reads that contain some variation in their predicted alignment. For the left termini, the mapping will show four categorizations of reads:
    1. Reads that contain *a-b sequence* and continue with UL of the P isoform and r eads that contain *a-b sequence* and continue with UL of the IL isoform.These

reads can be used to quantify the frequency of the isoforms and to predict variant differences within *a sequences*.

2. Reads that are short enough to be mapped to *a-b sequence* only and lack UL.
3. Reads that contain *a-b sequence* and have a hard stop at the end of the *b sequence*. These reads would require further analysis and validation for these are the first encounters of evidence of inverted b-a segment as a terminal end.

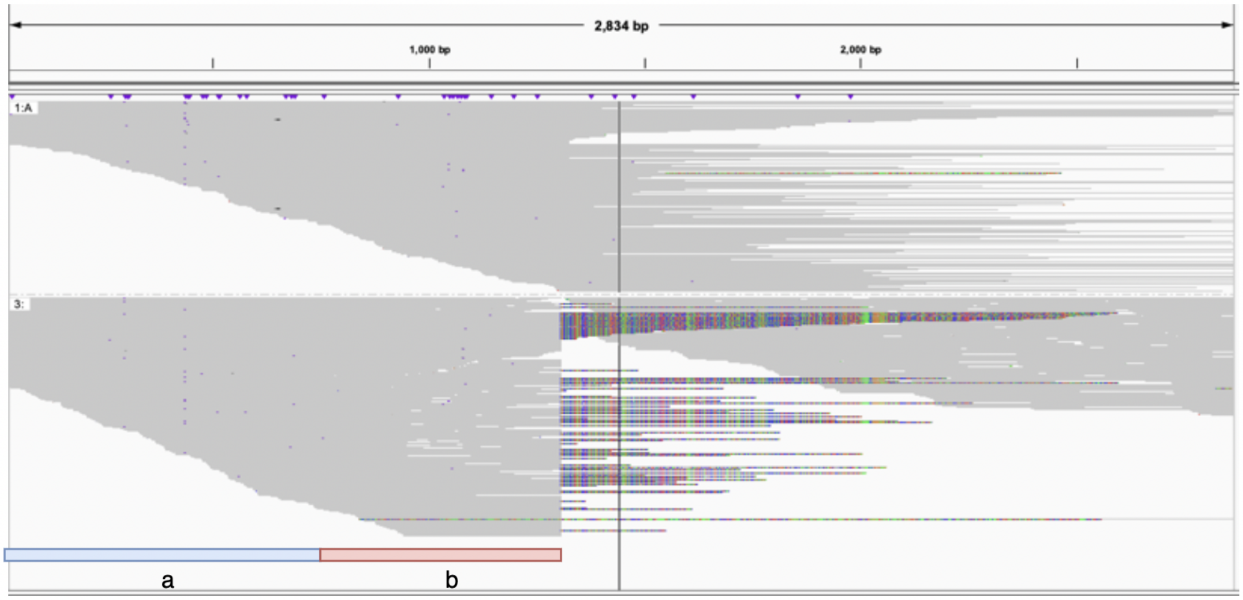


Figure 10) Visualization of reads mapped to the left terminal post alignments steps. Alignment between modified genomic sequence and filtered reads is shown. Reads that contain *a-b sequence* and carry on with the P isomer UL region are grouped in the upper panel. The bottom panel contains reads from the IL isoform (colored).

Similarly, and with respect to the right termini, the mapping shows three different sets (Figure 11):

A) The set of reads that contain US that matches the reference orientation and *c-a sequence. This set contains two subsets:*

a. Reads that end with a single copy of *a sequence.*

b. Reads that contain extra segments (figure 10; panel 1:T).

B) The set of reads that doesn't contain any single copy of the *a sequence.*

C) The set of reads that contain US from a different isoform and *c-a sequence.*

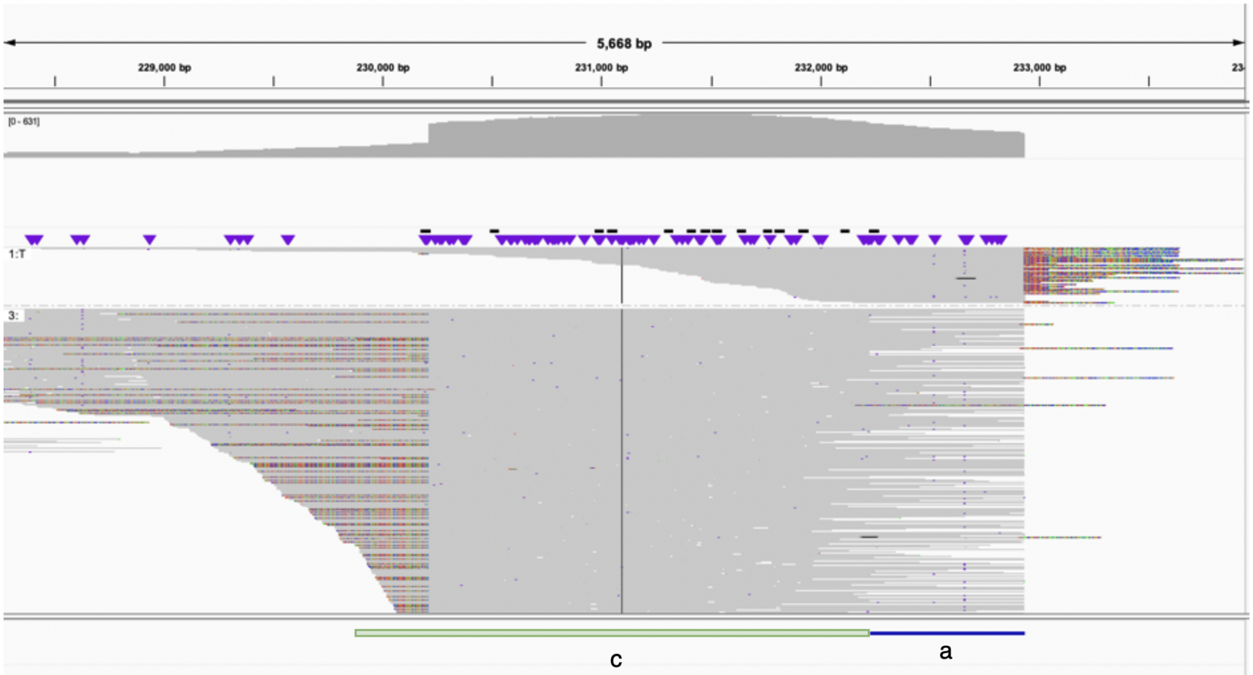


Figure 11) Visualization of reads mapped to the right terminal post alignments steps. The *a sequence* (blue line) and alignment between modified genomic sequence and filtered reads are shown. The top panel (1:T) shows reads that contain *US-c-a* and extra copies or partial copies of the a sequence. Bottom panel (3:) shows reads that contain *US-c-a* segments from both P and IS orientations of the S segment. Reads containing parts that didn't map to the reference sequence (colored) had either an extra copy of the a sequence (right bottom panel)

The second approach used to resolve the misalignment of reads belonging to the repeats that were mapped to the wrong genomic regions is to collectively map the reads to a distinct region of the WGS in a multi-alignment fashion. That is, to use different regions of the WGS, one at a time, to extract reads that are mapped to that region in a separate file. The filtered reads files (FRF; figure 12) contain filtered reads that are aligned to different references; Initially, CCS reads are aligned twice; once to the *a string* and once to the *c string* fasta files. The term string is equivalent to ‘sequence’ computationally. The a, b, and c strings are obtained from whole genome assembly using LoReTTA (Al Qaffas et.al, 2021; version 1.0) which provides a good starting point for a better prediction of the string inquiry. The result of the mapping is two BAM files containing tags of reads that are mapped to their respective strings. Tags are used to give information about the alignment type, direction, and quality. Most notably the primary alignment tag is defined in minimap by the alignment with the highest score. Disabling any other alignment (with option -N) provides a more accurate mapping of the repeats. Reads tagged in the two bam files are then extracted as a.fastq and c.fastq files respectively. The two files will branch out as follows:

- The a.fastq file is mapped to the *b string* generating ab.fastq file which contains both the a and b reads. The ab.fastq file is then mapped to the *c string* generating two files; abc.fastq file which contains reads that are certain to belong to the junction region since the file contains parts of all the repeating strings that occur only in that region; ab_no_c.fastq, which contains reads that don't include any of the *c string* in them. The latter file is used to separate reads of the right termini of each isoform from reads of the junction region. That is, to map the reads in ab_no_c.fastq three times; once with part (~

2kb) of UL left-end region which belongs to the P isoform string (UL_P); and once with the reverse and complement of the same segment of UL to represent the same part in the IL isoform (UL_IL). Merging the two files results in the fastq file UL_All which includes all the reads from both UL_P and UL__IL.

- Likewise, the c.fsatq file is mapped twice; once against the *a string* excluding any reads that include the *a sequence* (c_no_a.fastq); and once against the *b string* excluding reads that include the *b sequence.* Reads within the latter fastq file, c_no_abfastq, are mapped twice to the US either to P (US_P) or the IS (US_IS) isoform right end with each string being 2 kb in length. Merging of the two files US_P and US_IS results in US_ALL fastq file.

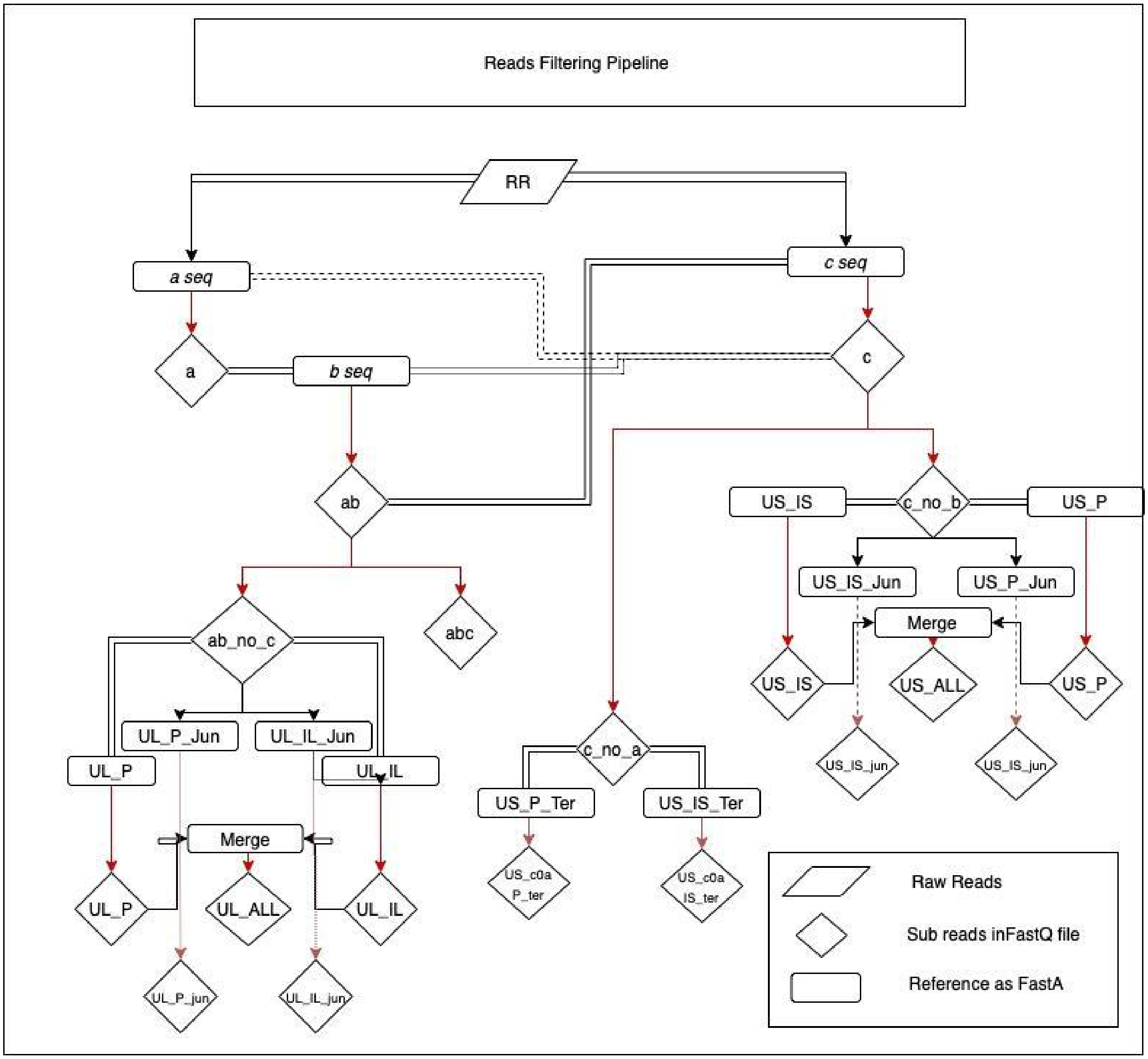


Figure 12) FRF pipeline; Each sequence in fasta format (square) is mapped to sequencing reads file fastq format (diamond).

The second approach takes into account the misalignment of reads by defining the terminal regions of P, IL, and US isoforms (Figure 13). It is reasonable to say that if many reads start or end at a given X position and those reads terminate at the *a sequence* ends, they reach one of the isoform genomic ends. This is supported by the IL string being the reverse complement of the P isoform. That is, IL's terminal region is a reverse complement of L and starts from a, b, and UL150A. Similarly and with the same relation to the P isoform, IS is formed. The strategy of this definition, overall, is to filter reads obtained from end-stage FRF based on the definition of each terminal region of HCMV.

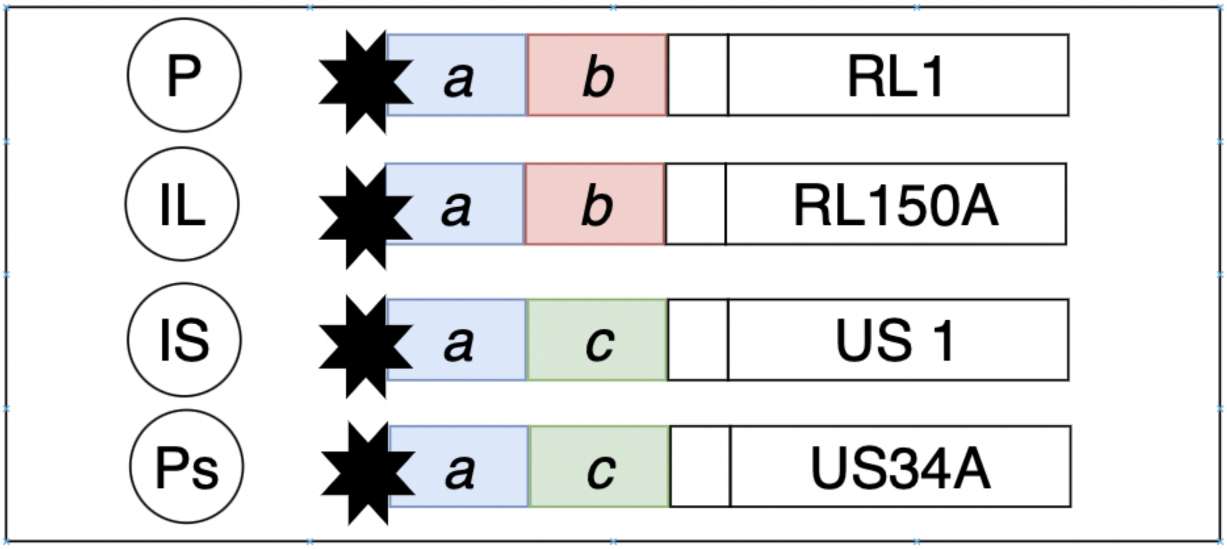


Figure 13) Definition of terminal regions based on the content of reads; each definition consists of a hard stop (star) two repeating sequences followed by a gene that is located around the end of the UL or US segments. (P) the expected sequence of the prototype's terminal region which includes ORF of RL1 gene. (IL) the expected sequence of the IL's terminal region which includes the *RL150A* ORF. (IS) the expected sequence of the IS's terminal region which includes US1 ORF. (Ps) the expected sequence of the P isoform right terminus including *US34A*. Note that IS and P termini containing one *a sequence* are shown.

2.2.B) Bioinformatics Scripts

Resolving the misalignment of reads is done by the Python script '***Terminal_Analyzer.py***' which generates the FRF reads described earlier. The Python script (Supplementary Data S1) takes 8 fasta files (*a, b, c, UL_p/IS, US_P/Is,* and reference) and the CCS reads with the minimal required options being the a and the reference fasta files. For each fasta file input and with the exclusion of the reference fasta, the following will be generated:

1- BAM file of alignments for each reference used according to the pipeline.

2- FASTQ file for each filtered from (1).

3- Calculations of:

I- Read filtering percentile which is obtained using the following formula:

%FRF = (filtered reads/ total CCS reads) * 100

II- Within-repeat variants in case of the *a, b,* and *c* sequences analysis.

Shortening the length of reads was used to resolve the issue of validating stops within repeated regions. Extracting the first and last 20 bp from each read and its reverse-complement provides position validation by counts of repeats. The arbitrary number of 20 bp is long enough to be used by mapping tools when scoring. It provides, also, that the string belongs to the used repeat sequence not another (P = (¼)^20 = 9.094947e-13). The approach is carried out using "***stat_cuts.py***" Python script and visualization of the alignment is obtained from R script with name " ***HCMV_Analysis_plots.r*** " (supplementary Data S2 and S3 respectively) . The script outputs four text files corresponding to each case mentioned earlier. The text file contains the position of the read mapping placement and counts the number of total reads at the same position (figure 14).

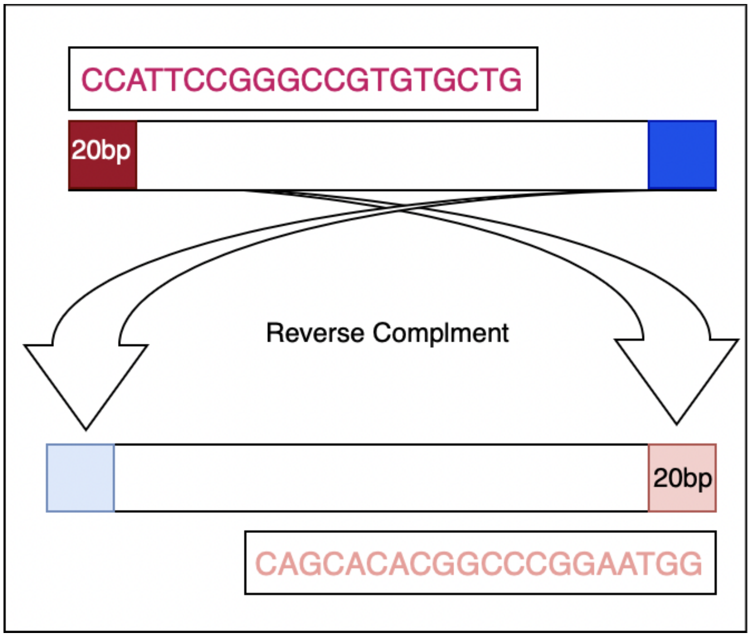


Figure 14) Subread extraction. 20 bp of each read's starting string (red) and ending string (blue) is extracted and saved as a fasta file for mapping with the reference. The reverse and complement of each read's starting string (light blue) and ending string (light red) are similarly extracted for mapping.

2.2.C) Notes

- The used alinger, minimap2, can suppress multi-alignment of reads using the argument '--secondary=no'. However, there is no way provided to prevent splitting of reads to suppress secondary alignment. In order to count the number of unique reads in any FRF, samtools flag -F0x900 was used.

- Generating the alignment plot based on the definition mentioned above (figure 12) was used individually using the samtools and coordinates of the *a* sequence across the genome. The command used was:

```
samtools view -Sbh {reads} "{ref_name}:1-10" > out.bam
```

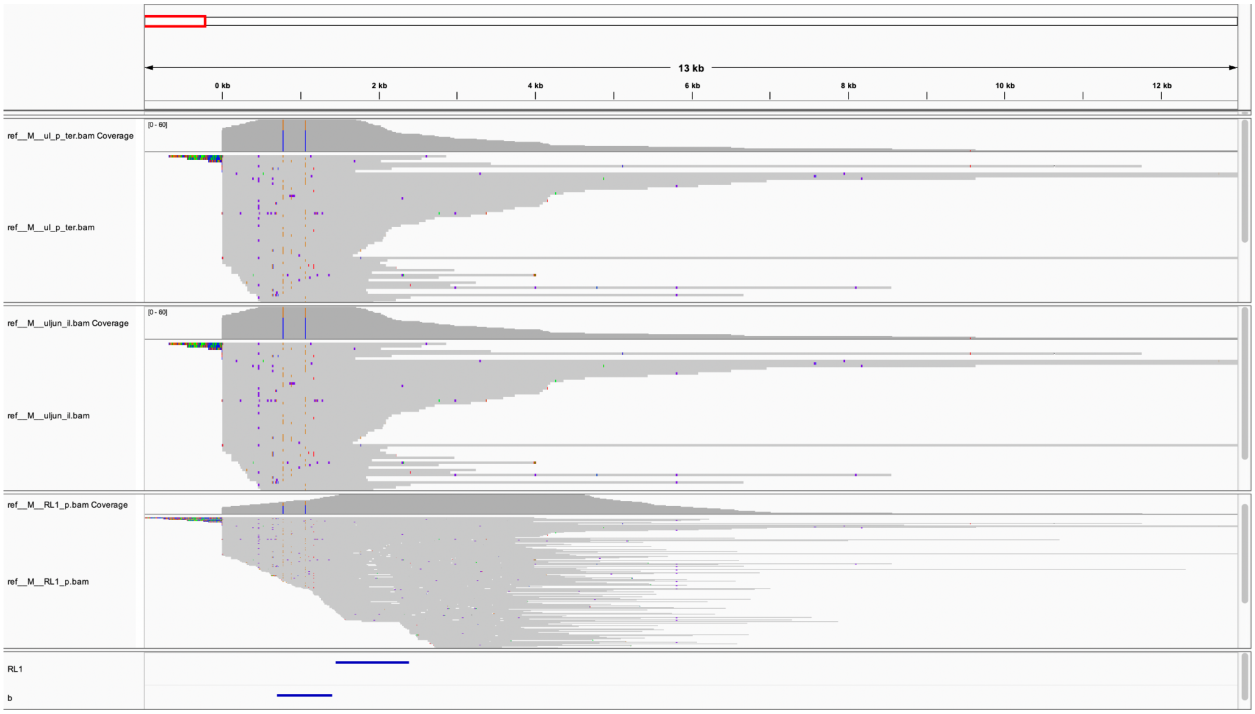
Results

Results from the first approach to deal with misalignment issues are not reported here due to failure in the filtering reads from the alignment files using common bioinformatics tools. Specifically, the approach relies on the Venn diagram exclusion principle which is not implemented in tested tools but with faulty results. The results include multiple reads that aligned in small part with the target region. Hence, the pipeline implementation included only the second approach that dealt with misalignment issues and its results. The pipeline was run using S1 and S3 successfully. Generating a total of 21 FRFs. S2, however, had a low number of reads (n=6985) which hindered a good internal representation of the sample genomic composition and structural orientation.

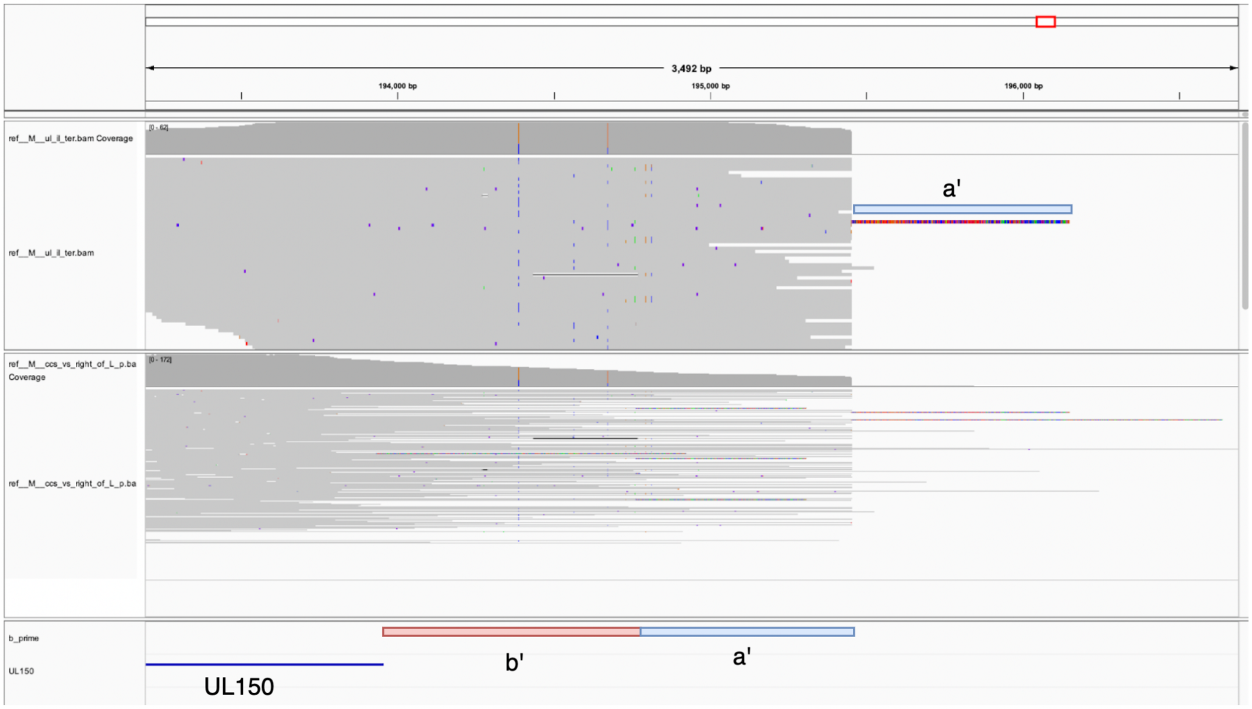
3.1) FRFs Alignment and Analysis

Analysis of TB40EF

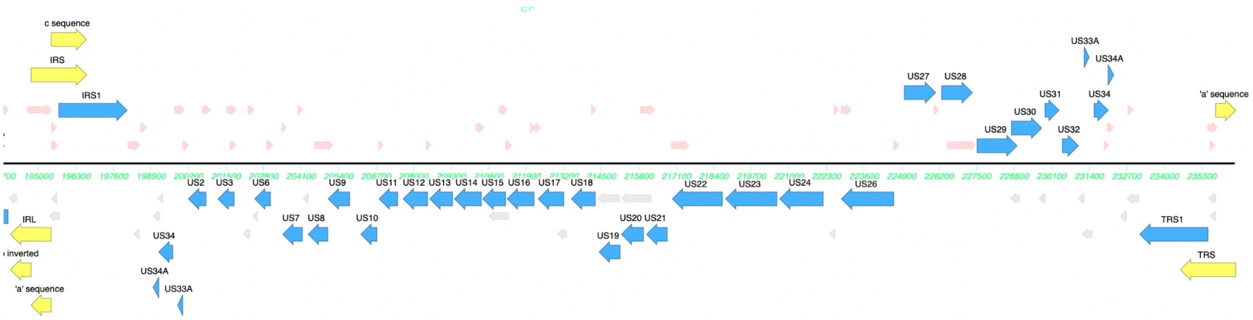
A total of 16920 reads were used to generate end-stage FRFs that were mapped against S1 WGS as a reference. The following plots show the mapping of end-stage FRFs against the prototype as a reference. For each plot, the top red box represents the WGS region and the ruler represents the window size. S1.plot.1.a shows reads that are mapped to the reference from UL_P_Ter (top panel), UL_IL_Jun (middle panel), and all reads that contain RL1 ORF (bottom panel). Two blue lines, RL1 and the b sequence are shown for reference. Both UL_P and UL_IL_Jun had the same reads mapping to both (n=60). A total of 3 reads had an extra segment before the reference's first nucleotide (position = 0); the top two reads contained an extra copy of the *a sequence.* The bottom read contained a small segment (~174bp) of the c sequence. The total number of reads mapped to the left of the L segment -excluding a and b- is 299 reads. S1.plot.1.b shows the mapping overview of the right end of the L segment at the junction region from the files UL_IL_Ter (top panel) and all CSS reads that contained UL150/UL150A (bottom panel). Two Blue lines showing UL150 and the b sequence for reference. Both UL_P_jun and UL_IL_Ter had the same reads mapping to both (n=62). The total number of reads mapped to the right of the L segment -excluding a and b- is 334. A single read that extends with an extra copy of the *a sequence*. One read contained part of the c sequence. For the S segment, S1.plot.2.a shows duplicate ORF of genes US33, US34, and US34A. On the other hand, S1.plot2.b shows two reads containing an extra copy of the *a sequence.*



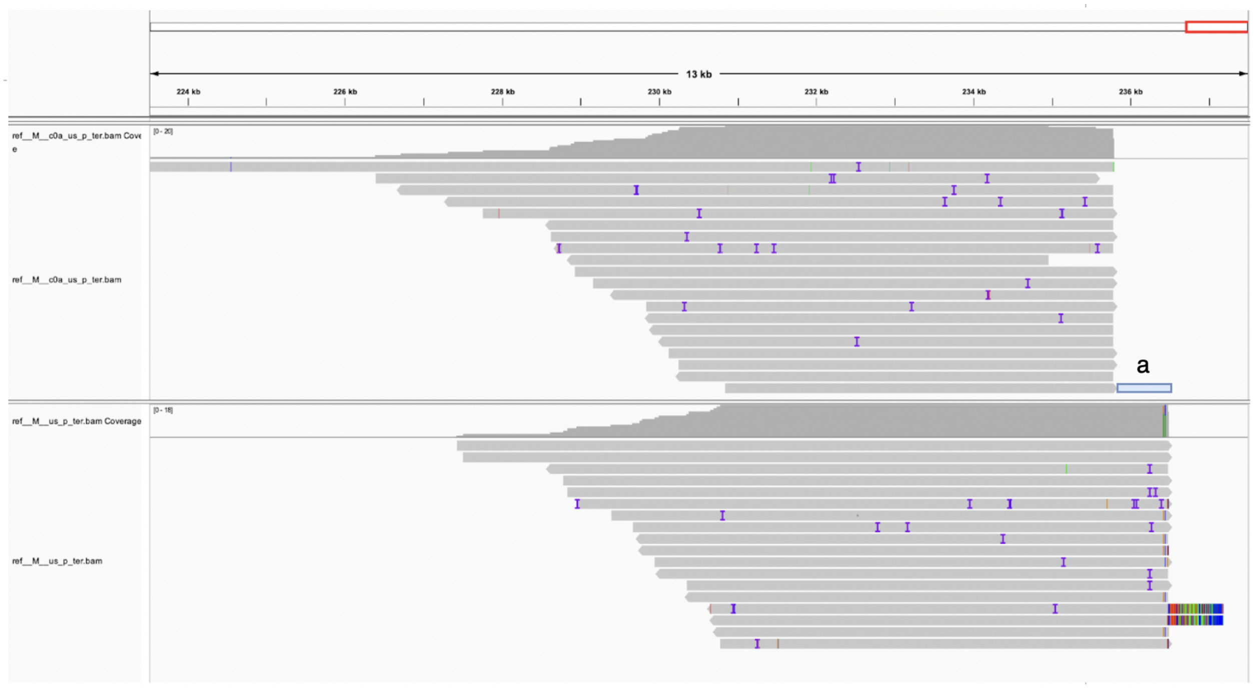
S1.plot.1.a) Mapping overview of left of L segment. FRFs UL_P_Ter (top panel), UL_IL_Jun (middle panel), and all reads that contain RL1 ORF (bottom panel) alignment are shown. All the three files alignment showing two reads containing parts of an additional *a sequence* (top colored) and one read containing part of the *c sequence* (bottom colored).


3,492 bp
194,000 bp
195,000 bp
196,000 bp
ref__M__ul_il_ter.bam Coverage
[0 - 62]
ref__M__ul_il_ter.bam
a'
ref__M__ccs_vs_right_of_L_p.ba Coverage
[0 - 172]
ref__M__ccs_vs_right_of_L_p.ba
b_prime
UL150
UL150
b'
a'

S1.plot.1.b) Mapping overview of right end of L segment. FRF UL_IL_Ter (top panel) shows an additional *a sequence* to the right. Mapping of all CSS reads that contained UL150/UL150A (bottom panel) shows a read containing an additional *a sequence* (colored top) and one read containing an additional *c sequence* (colored bottom) .



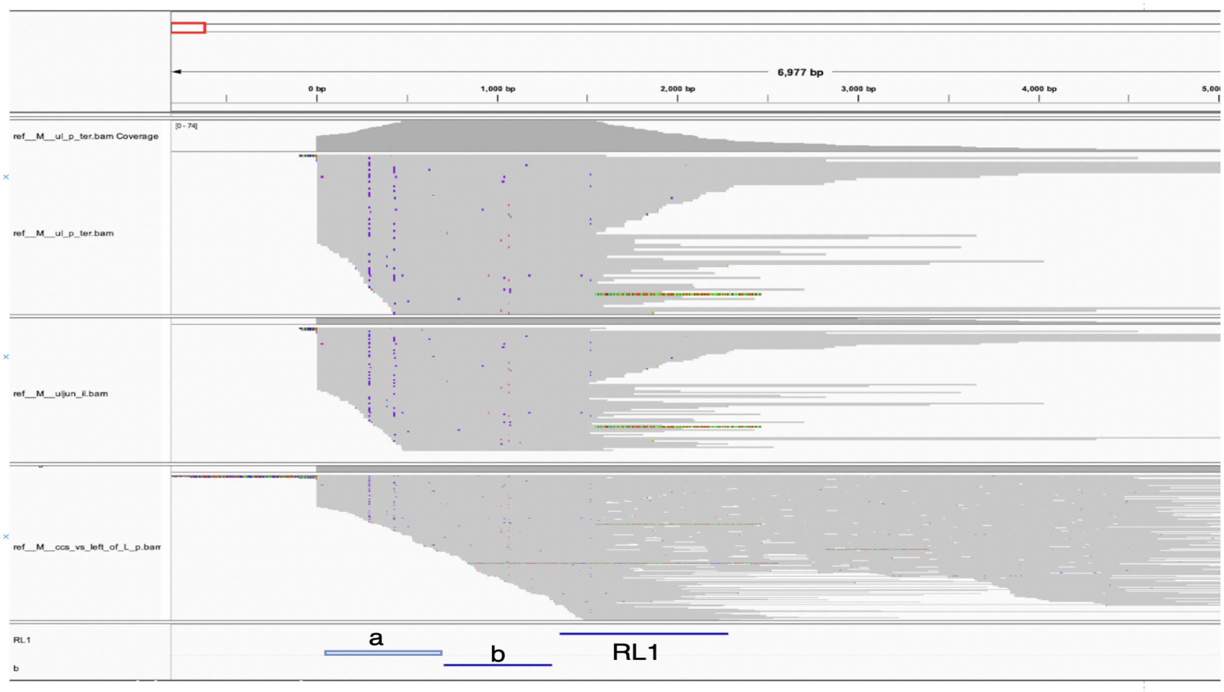
S1.plot.2.a) Mapping overview of S segment alignment with FRFst; the S1 genome includes a duplication of US33A, US34, and US34A on both ends of the US region, which in turn hinders an accurate filtration of reads based on their position.



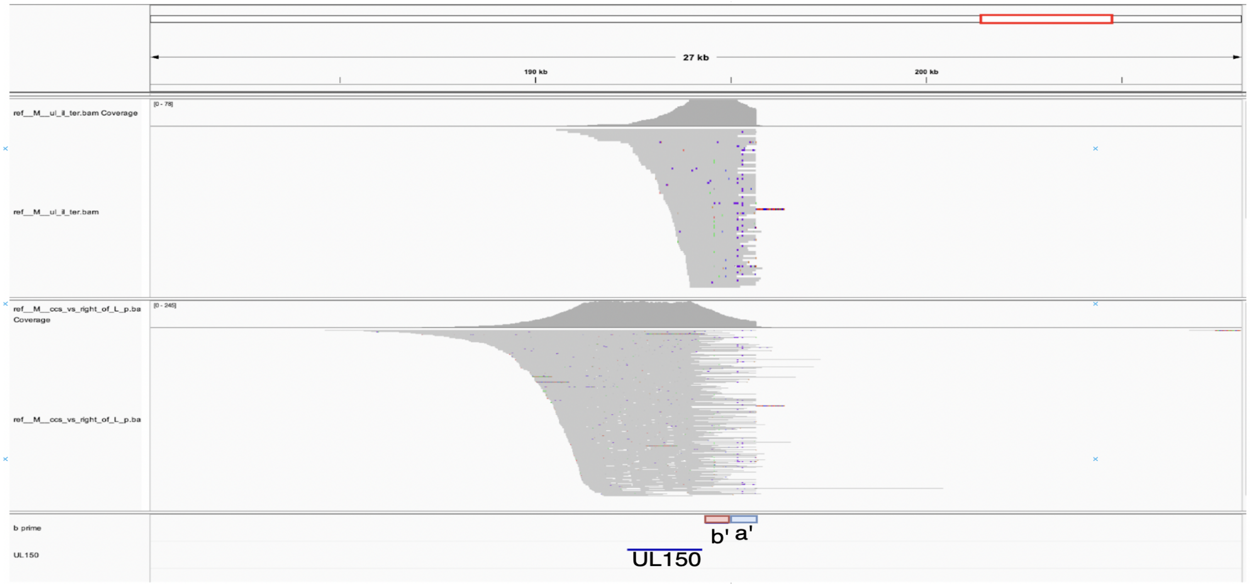
S1.plot.2.b) Mapping overview of right of the S segment alignment with FRFs US_P_c_no_a (top panel) and US_P_Ter (bottom panel). Alignment with US_P_Ter shows two reads with an extra copy of the *a sequence* (colored).

Analysis of Ig-KG-H2P14S

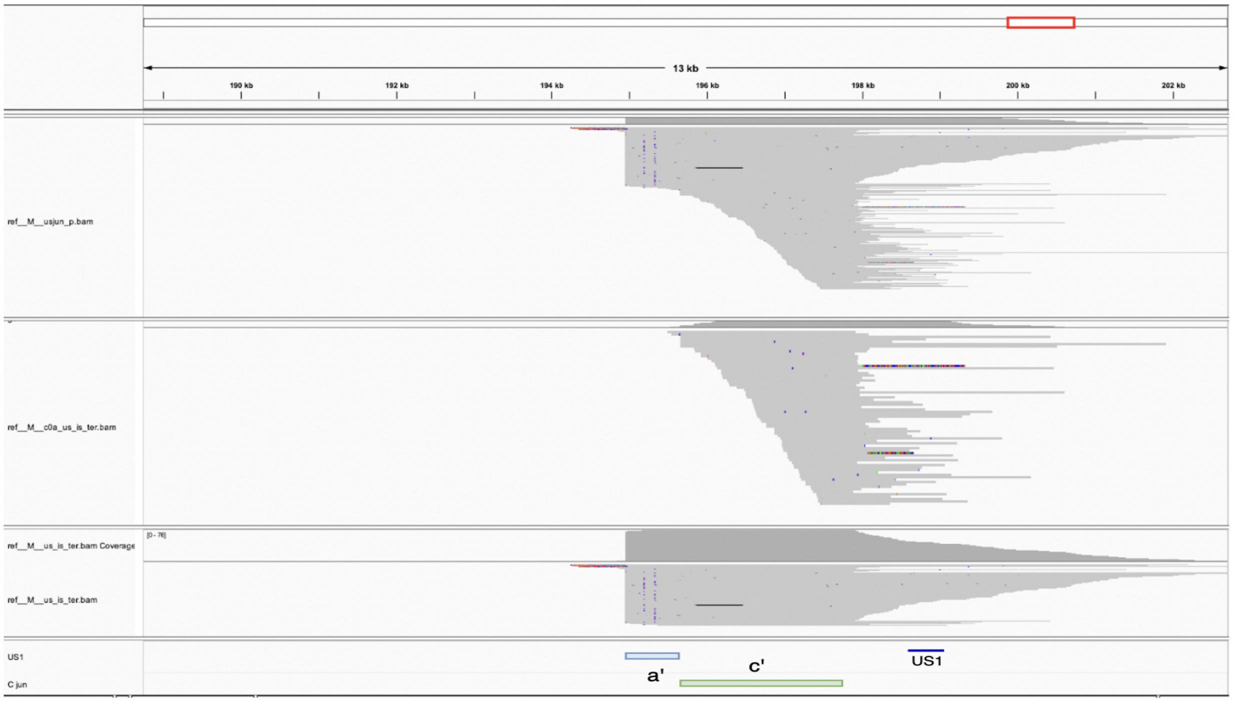
A total of 29340 reads were used to generate FRFs that were mapped against S3 WGS. The following plots show both ends of the L and S segments aligned with end-stage FRFs. For each plot, the top red box represents the WGS region and the ruler represents the window size. Concerning the L segment, S3.plot.1.a shows an overview of the left L segment alignment with FRFs UL_P and UL_IL_Jun. The plot shows UL_IL_Jun had 5 extra reads not present in UL_P_Ter (n=74). The non-viral sequence was found in a single read within RL1 ORF. S3.plot.1.b shows an overview of the right of L segment alignment with FRF UL_IL_Ter. The alignment shows five reads that contained part of the *c sequence* indicating that they belong to the junctional region of the prototype. In addition, it shows one read that contained an extra copy plus 22bp of the *a sequence.* With respect to the S segment, S3.plot.2.a shows an overview of the left of S segment alignment with FRFs US_P_Jun, US_IS and US_IS_ter_c_no_a. S3.plot.2.b shows an overview of the right end of S segment alignment with FRFs US_P_Ter, US_IS_Jun and US_P_Ter_c_no_a.



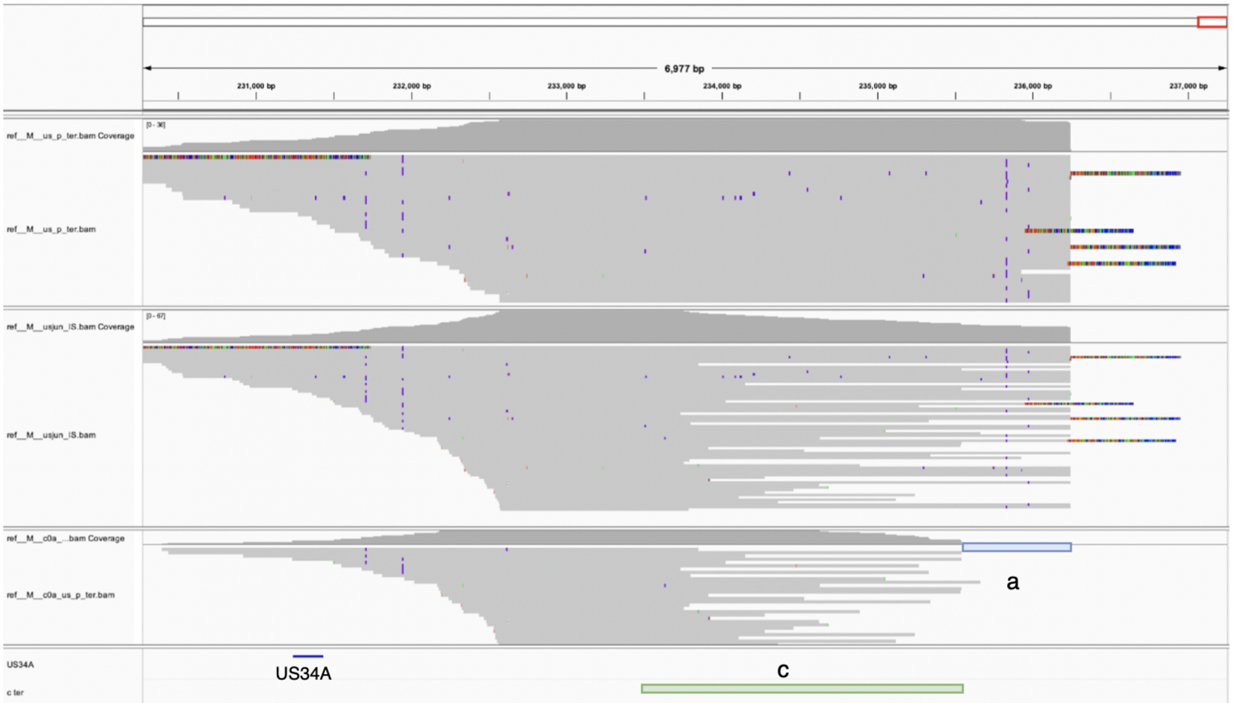
S3.plot.1.a) Mapping overview of left of L segment. UL_P_Ter (top panel), UL_IL_Jun (middle panel), and all reads that contain RL1 ORF (bottom panel). Two blue lines, RL1 and the *b sequence,* are shown for reference. Top panel shows one read containing part of the *c sequence* colored). Middle panel shows two reads with part of the *c sequence* (colored). Bottom panel shows 3 reads with part of the *c sequence (colored).* All of the FRF show a read that contains none-viral sequence within RL1 ORF.


27 kb
190 kb
200 kb
ref__M__ul_il_ter.bam Coverage
[0 - 78]
ref__M__ul_il_ter.bam
ref__M__ccs_vs_right_of_L_p.ba Coverage
[0 - 245]
ref__M__ccs_vs_right_of_L_p.ba
b prime
UL150
b'
a'
UL150

S3.plot.1.b) Mapping overview of right end of L segment at the junctional region. FRF UL_IL_Ter (top panel) and all CSS reads that contain UL150/UL150A ORF (bottom panel) alignment is shown. Both panels show an extra copy of the *a sequence* (colored).



S3.plot.2.a) Mapping overview of left end of S segment. Alignment of FRFs US_P_Jun (top panel), US_IS_Ter_c_no_a (middle panel), and US_IS_Ter (bottom panel) is shown. Both top and bottom panel show three reads containing extra copies of the *a sequence.* Middle panel does not contain a hard stop *c sequence* end on the left side.



S3.plot.2.b) Mapping overview of right end of S segment. Alignment of FRFs US_P_Ter (top panel), US_IS_Jun (middle panel), US_P_Ter_c_no_a (bottom panel) is shown. A total of four reads had an extra copy of the *a sequence* (colored).

3.2) FRFs Reads' count

Table 3) Initial FRFs counts of S3; Reads that contain the a (blue), b (red), and c (green) sequences are reported below. For the c file, additional filtration was made to remove reads that contained *a* sequence (c_no_a) or *b* sequence (c_no_b).

FRF	*Number of Reads*	*Percentage (total reads)*
a	*785*	*2.6755 %*
b	*498*	*1.6973 %*
c	*959*	*3.2686 %*
c_no_a	514	1.7519 %
c_no_b	911	3.105 %

Table 4) Conjoint FRFs of S3; the *a* file was mapped against the *b sequence.* The ab file was mapped to exclude reads continuing the *c sequence* (ab_no_c_) or include all three (abc). The *c_no_b* file was aligned to the *a sequence* to generate the *ca_no_b* file.

FRF	*Number of Reads*	*Percentage (total reads)*
a+b	337	1.1486 %
ab_no_c	289	0.985 %
abc	48	0.1636 %
ca_no_b	397	1.3531 %

Table 5) Regional FRFs of S3; alignments between FRFs and L/S/IL/IS unique regions were performed and end-stage FRFs were generated as follows: (UL_P_Ter) from ab_no_c and a sequence from the left end of P L segment, (UL_IL_Ter) from ab_no_c and the left end of IL L segment, (UL_P_Jun) ab_no_c and a sequence from the right end of P L segment, (UL_IL_Jun) from ab_no_c and the right end of IL L segment, (US_P_Jun) c_no_b and left end of P S segment post the junction region, (US_IS_Jun) c_no_b and left end of IS S segment post the junction region, (US_P_Ter) ca_no_b and the right end of P S segment, (US_P_Ter_no_a) c_no_a and right end of P S segment, (US_IS_Ter) ca_no_b and the right end of IS S segment, and (US_IS_Ter_no_a) c_no_a and right end of IS S segment.

FRF	*Number of Reads* *Out of parent FRF*	*Percentage (total reads)*
UL_P_Ter	*74*	*0.2522 %*
UL_IL_Ter	*78*	*0.2658%*
UL_P_Jun	*78*	*0.2658%*
UL_IL_Jun	*79*	*0.2693%*
US_P_Jun	*201*	*0.6851%*
US_IS_Jun	*67*	*0.2284%*
US_P_Ter	*36*	*0.1227%*
US_P_Ter_no_a	*31*	*0.1057%*
US_IS_Ter	*76*	*0.259%*
US_IS_Ter_no_a	*72*	*0.2454%*

* FRF frequency spectrum; counts of reads mapped to each compounding reference and their calculated percentage using the formula (% = (100*FRF)/total number of CCS reads).

** The reads reported include parts or the whole sequence. Including any read is subject to Minimap2 scoring.

3.3) Inner variants of repeats

Calling of variants within the *a, b, and c repeats* were performed using samtools, NGMLR (Sedlazeck, 2018), and bcftools. No variants were found within the c sequence in both S1 or S3. The following table shows the found variants within both *a* and *b sequences* after mapping reads that contain either direction of the repeats.

Table 6) Inner variants within S1(A) and S3 (B).

(A) Inner variants of repeats within S1:

Repeat	INDEL	SNP	Position	Note
a	A:AG	-	479	PloyG Tract
a	-	A:G	639	
a	-	A:C	646	
b	-	C:G	80	
b	-	C:G	189	
b	-	C:G	365	
b	C:CG	-	427	
b	-	G:T	477	

*position is reported relative to the sequence itself not the whole genomic coordinates.

** Length of the *a sequence* = 697 bp and *b sequence* = 715 bp

(B) Inner variants of repeats within S3.

Repeat	INDEL	SNP	Position	Note
a	A:AG,AGG	-	299	PloyG Tract
a	A:AGG	-	439	PloyG Tract
b	-	G:T	362	15/147 supporting reads

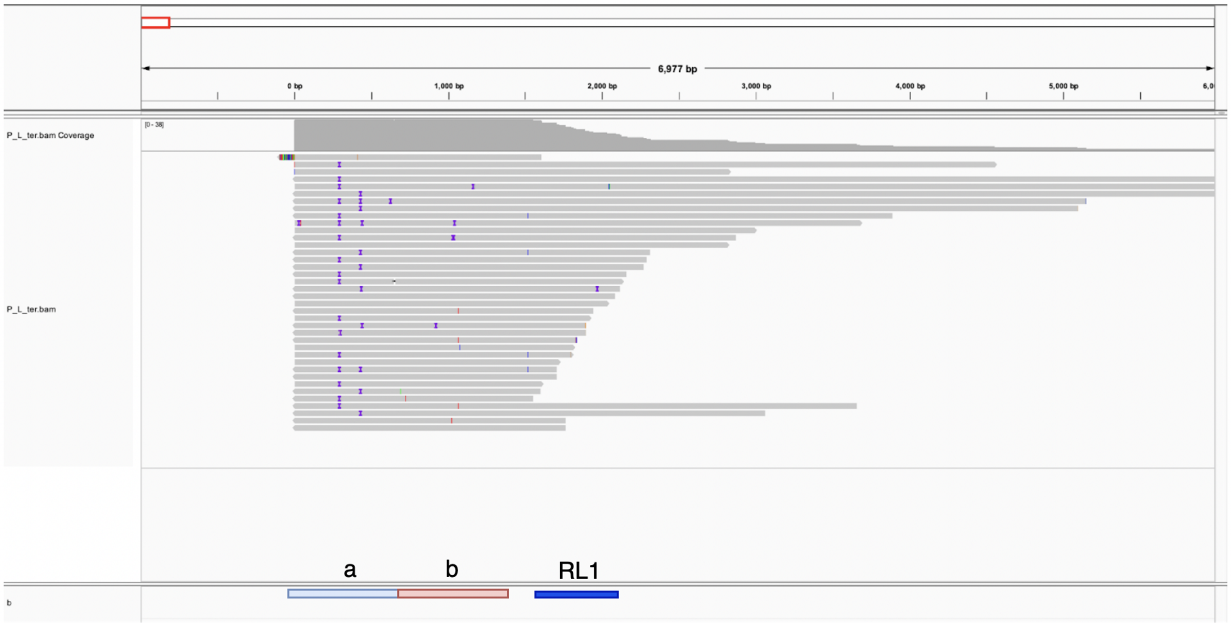
*position is reported relative to the sequence itself not the whole genomic coordinates.

** Length of the *a sequence = 704bp and b sequence = 600bp*

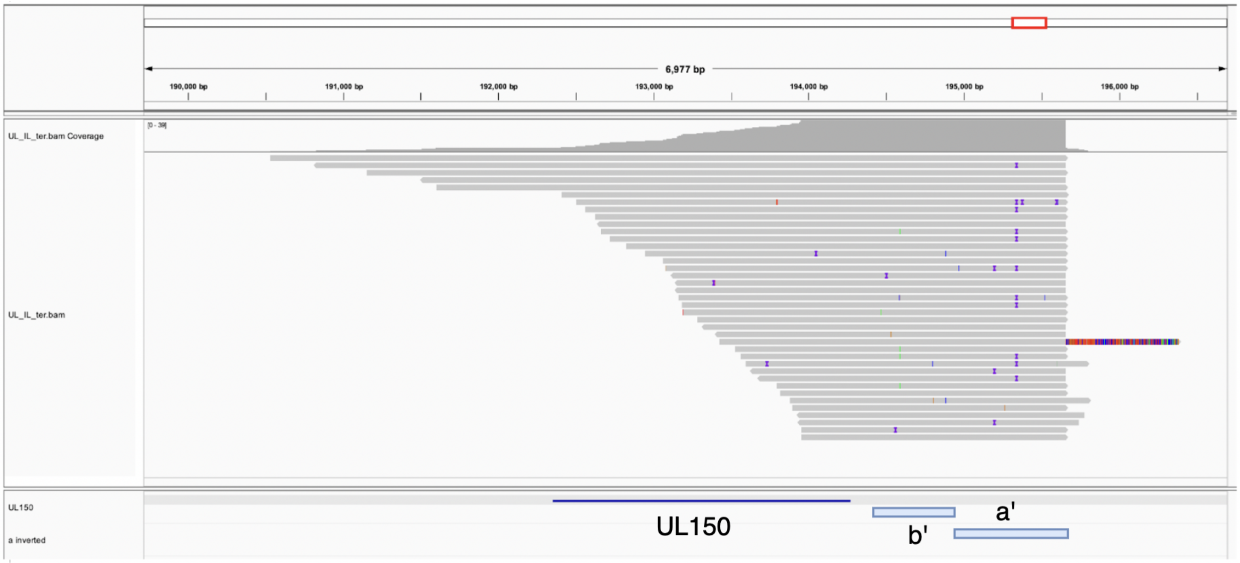
3.4) Terminal Ends By Definition Plots

The following alignment plot shows the reads based on the definition provided in section 2 (Figure 13) for S3. The S1 terminal ends are not included in this report for the inability to infer IS or the prototype S ends.

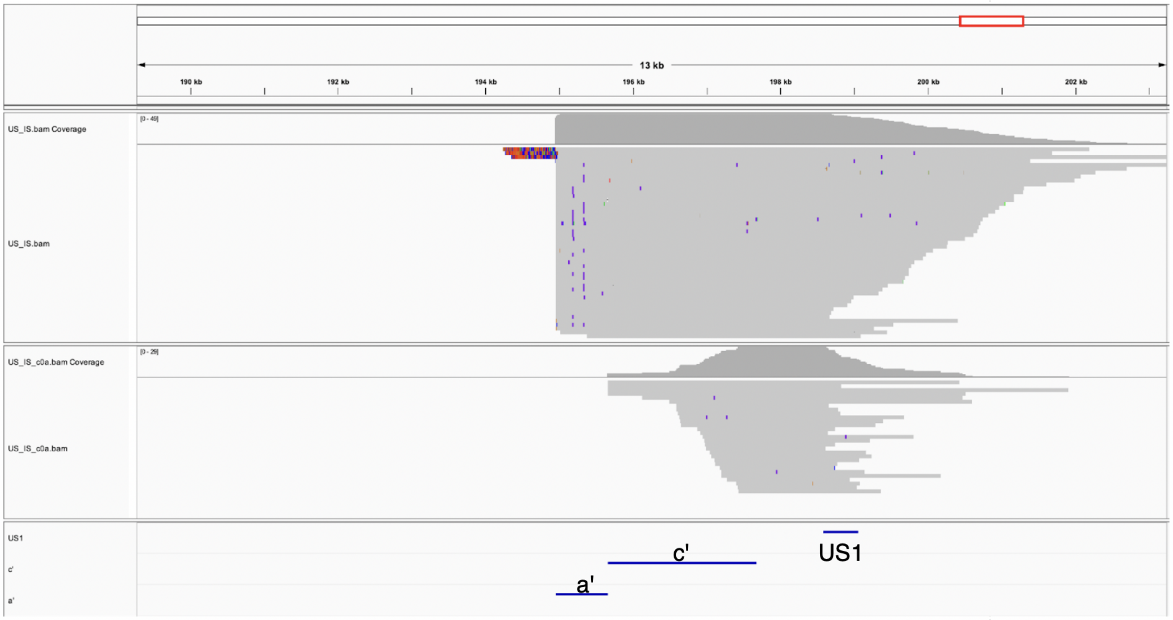
Terminals of S3 isoforms



S3.dPlot.1) P terminal reads. One read (colored top) had parts of the *c sequence*. Two sites of poly(G) tract insertions of 1,2,3 nucleotides within the *a sequence*.

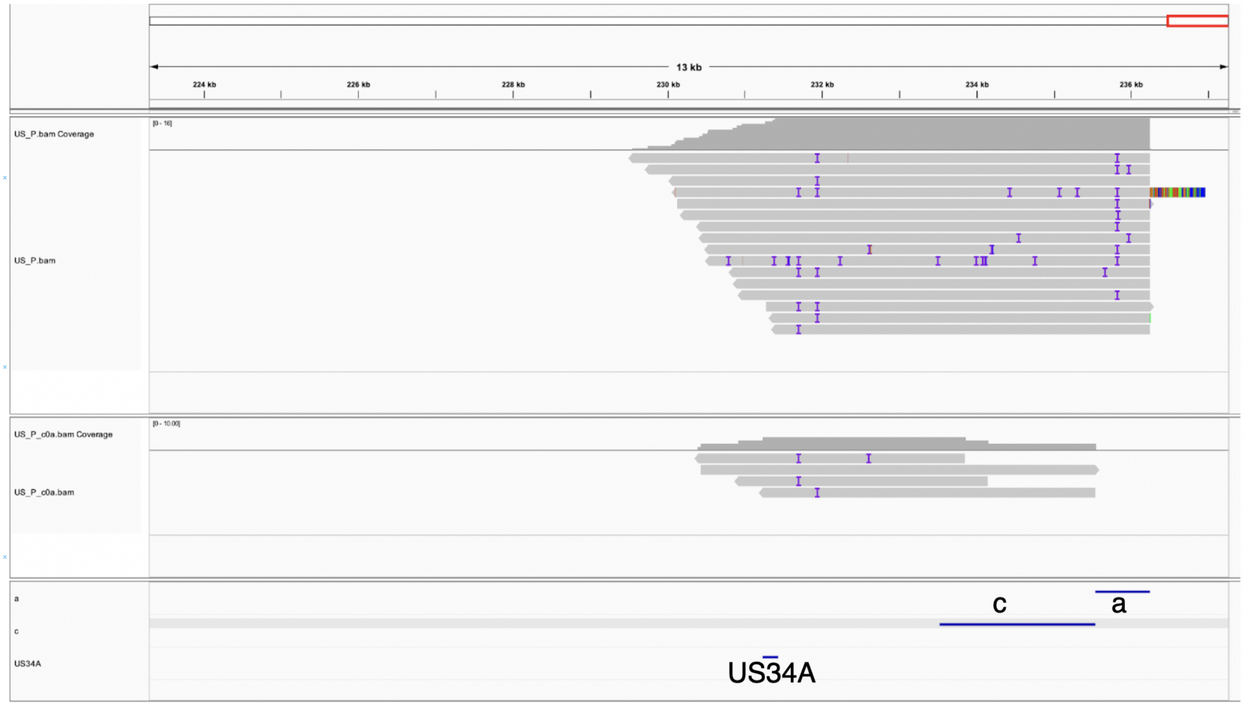


S3.dPlot.2) IL terminal reads. One read (colored top) had parts of the *a sequence and* 5 reads contained parts of the *c* sequence.



S3.dPlot.3) IS isomer terminal reads. Reads by definition which includes *a and c sequence* (top panel) shows one read containing an extra copy of the *a sequence* (top colored)*,* a total of 3 reads containing parts of the *b sequence,* and a single read that had a deletion within the *c sequence.*

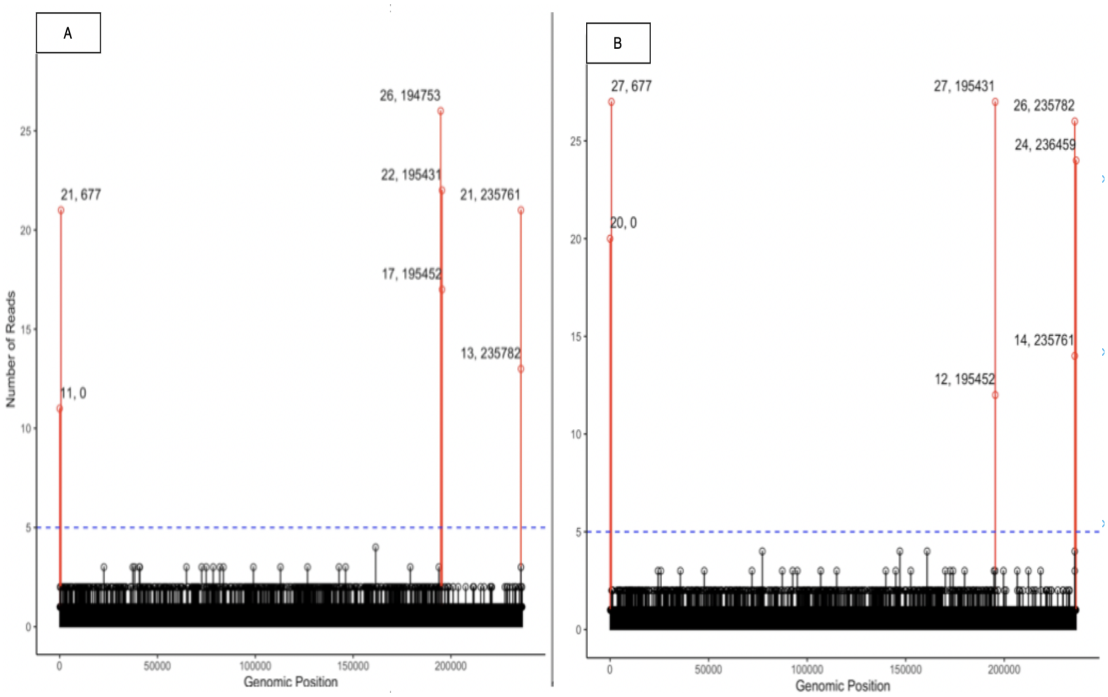
Reads by definition which exclude the *a sequence* (bottom panel) shows random termination of reads.

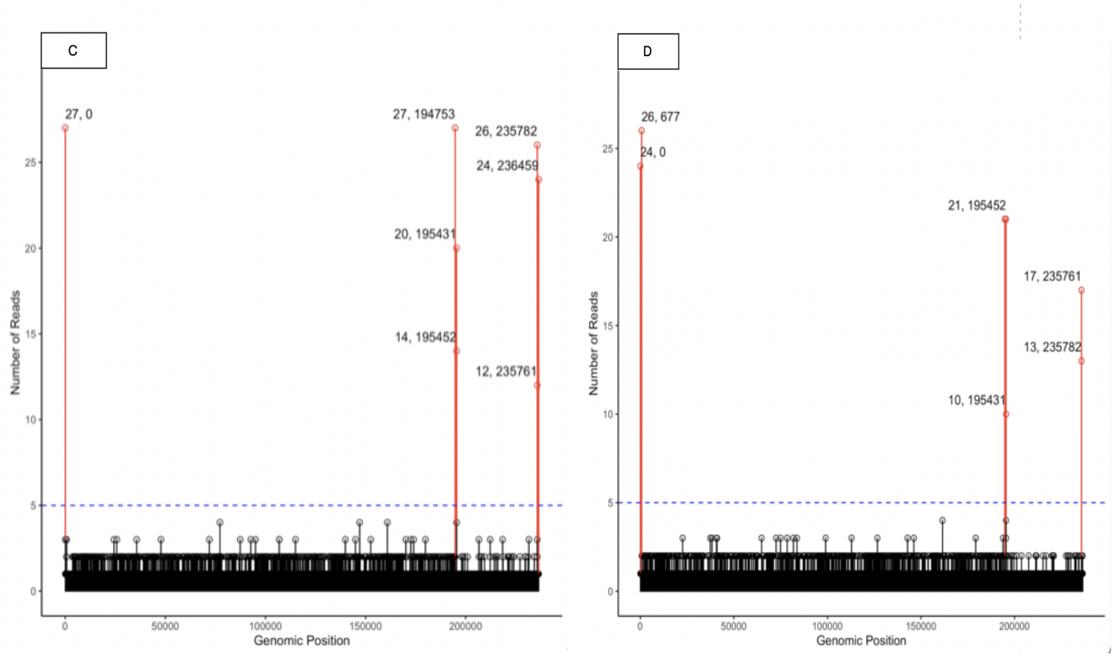


S3.dPlot.4) P isomer S segment terminal reads. Reads by definition which includes *a and c sequence* (top panel) shows one read containing an extra copy of the *a sequence* and only one of the two Poly(G) variants found on the *a sequence* elsewhere in the genome. Reads by definition which exclude the *a sequence* (bottom panel) shows only four reads that include US34A ORF and part of the *c sequence.*

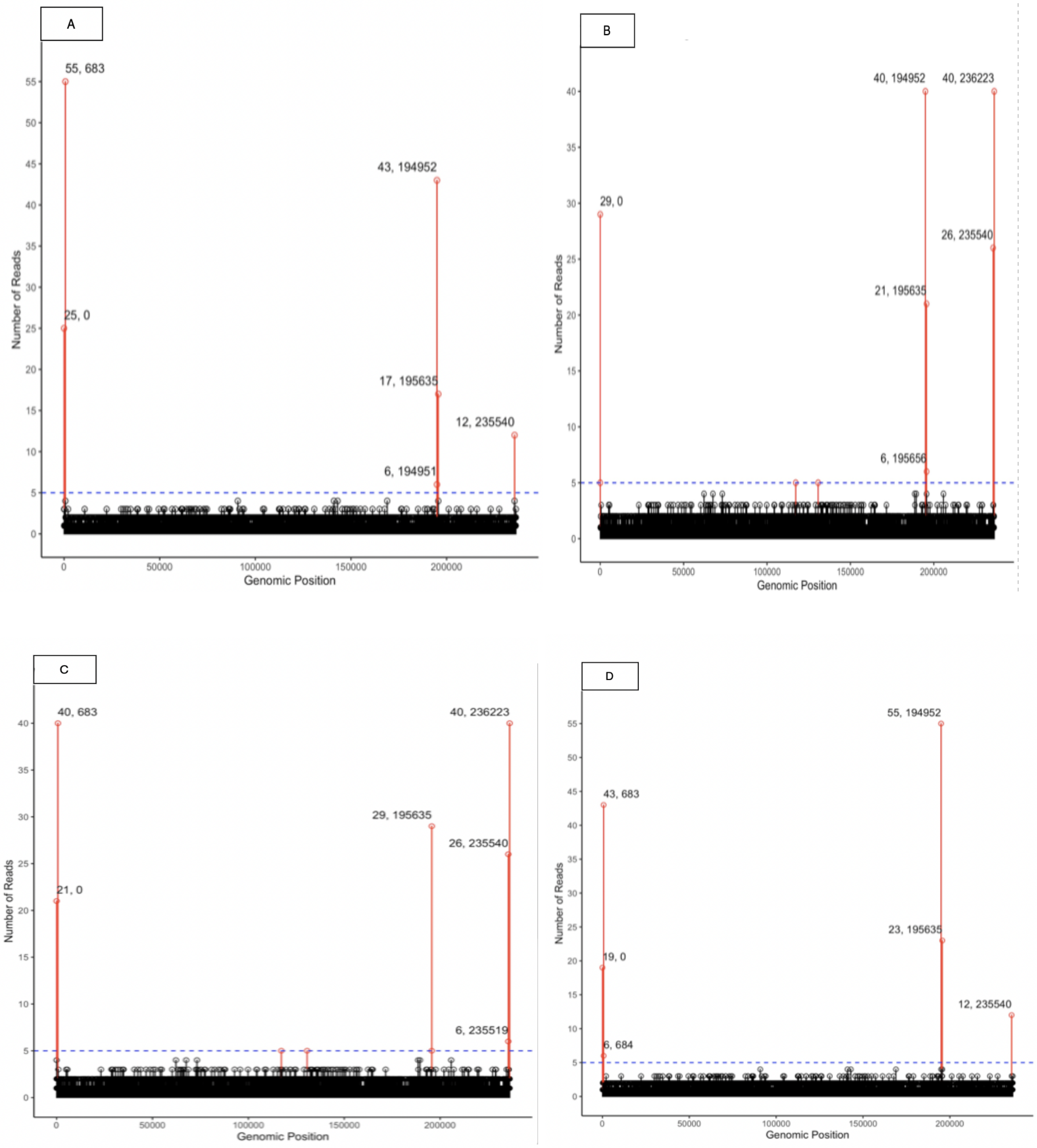
3.5) Mapping of Reads Ends

The output of *stat_cuts.py* which implement the logic described in Figure 14 are reported below. Each peak represents the number of reads-end mapped at a given position with respect to the strain genome.





S1.plot.sc) Histogram of read ends mapped against S1 genomic sequence: mapping of the left and right termini of the original (A/B) and reverse complement (C/D) orientation respectively of CCS reads against S1 WGS. Each point represents (the number of reads mapped, genomic coordinates). The genomic coordinates reflect the starting position of the alignment.



S3.plot.sc) Histogram of reads ends mapped against S3 genomic sequence: mapping of the left and right terminal of the original (A/B) and reverse complement (C/D) orientation respectively of CCS reads against S1 WGS. Each point represents (the number of reads mapped, genomic coordinates). The genomic coordinates reflect the starting position of the alignment.

Discussion

Defining Human Cytomegalovirus terminal repeats has been challenging using second-generation sequencing due to the technology's limitation of spanning over the wide and complex regions of short tandem repeats of the virus that can reach up to 10.7 kb in length. In addition to sequencing limitations, the presence of four genomic isoforms in any viral sample hinders an accurate computational reconstruction of each isoform. These challenges, in turn, overshadow the discovery of structural variants of the repeating region. Such limitations keep the repeats functionally ambiguous in HCMV. Here, I showed a pipeline that aimed to provide a bioinformatical pipeline to reconstruct, analize, and infer a better definition of HCMV terminal repeats. Findings of this research confirmed previous speculation of the nature of HCMV genomic ends in the sense of accurate cleavage mechanism at terminal ending with an *a* sequence. In addition, this research findings supported previous reports of predominantly presence of one *a* sequence at left termini in both P and IL isoforms and in the junction. The findings also supported an equimolar prevalence of the L segment in both viral strains. Furthermore, data analysis indicates the presence of an additional copy of the *a* sequence at the right termini in Ig-KG-H2P14S strain, selective variation of within *a sequence* on the right termini when compared to the sequence elsewhere.

The reconstruction pipeline generates multiple sets of subread-containing files called filtered read files (FRFs, figure 12). Each FRF has been aligned to a different segment of its respective genome to reduce the number of non-relevant reads for downstream analysis and to narrow the scope down enough to isolate reads belonging to one isoform over another. The pipeline generates 19 files with increased compoundness (tables 3, 4, and 5). FRFs a, b, and c were used as a baseline for further filtering steps and to call inner variants reported in table 6. Both the *a* and *b sequences* showed a degree of inner variation within them whereas the *c sequence* seems to be more resilient to inner variations in both HCMV strains S1 and S3. Concerning the *a sequence*, abrupt insertion of 1, 2, and 3 guanine bases within poly(g) tracts (and similar indels of cytosine within poly(c) within the *a' sequence*) was present in both S1 and S3. In addition, there is strong evidence that there is a different poly(G)/Poly(C) selective internal variation within *a sequences* of S3. Reads of the left terminal *a sequence* in the P isoform contained two poly(G) variants at different positions while reads of the left terminal *a sequence* in the IL isoform contained only one poly(G) variation (S3.dPlots 1 and 2). Similarly, evidence of one poly(G) variation within reads belonging to the P isoform right termini were found in contrast to two poly(G) variation in the counterpart region in IS isoform (S3.dPlot 3 and 4). Such inner variation of the *a* sequence suggests that during recombination events each unique *a* sequence remains adjacent to a specific unique end on either the L and S arms.

Compounded FRFs (table 4) were used to count the number of reads that contain more than one repeating sequence. Most notably and for S3, the number of reads that had both a and b sequences (n=337) within them was about seven times the number of reads that contained all three repeating sequences (n=48). Such observation should not be misunderstood as an issue of sequencing. Rather, it is a direct issue with alignment prediction by scoring. In other words, The higher number of reads in the ab file contains reads belonging to the terminal end of the IL genome (reads with no c sequence) that were included due to the high scoring during the mapping process. Evidently, excluding reads that contained the c sequence in (ab_no_c, n= 289) are missing 48 reads mapped to the abc file. In addition, the ca_no_b file contains reads that map to the terminal or junction region of the viral genome of IS or P isoform respectively.

End-stage FRFS, on the other hand, are files that include reads aligned to at least one repeating sequence and a unique sequence of L or S segments (table 5). Visualization of each of these FRFs against their respective whole genomic sequences reported in S1.plots 1 and 2 [b] and S3.plots 1 and 2. For S1, the same set of reads was mapped to both counterparts of the P and IL terminal and junctional regions (S1.plot.1.a and S1.plot.1.b). Such observation suggests an equimolar presence of the P and IL isoforms of virus S1. Analysis of S termini was not possible for S1 due to the duplication of the three genes that elongated the *c* repeats. As a result, the number of reads long enough to include multiple repeats was lacking. For S3, a similar number of reads mapped to both counterparts of the P and IL terminal and junctional regions (S3.plot.1.a and S3.plot.1.b), indicating an equimolar presence of the P and IL isoforms. On the other hand, mapping of reads belonging to the S segment of both isoforms showed a preference for the IS isoform (S3.plot.2.a and S3.plot.2.b) as the number of reads belonging to IS termini was about two-fold the number of reads from P isoform right termini (table 5). S3.plot.1.a shows the alignment of reads that are mapped to the left end of the P isoform. The alignment shows the expected sets of reads to map the genomic regions: UL_P_ter, UL_IL_jun, and shows the sets of all reads that contain RL1 for reference. This is a strong suggestion that both isoforms, P and IL, have close abundance within the S1 sample. Both files had two reads with an extra copy of the a sequence and a single read that contained a small part of the c sequence. This wasn't the case in S3 where the two sets contained different numbers of reads. Such differences, although minimal (n=5 reads), suggests a possibility of strain-dependent isoform disproportional prevalence. This is supported by a similar quantitative difference of reads found in the US ends-containing FRFs of S3 (table 5, S3.plot.2).

Evidence of a different ratio of L to IL and S to IS ends was found in both samples. This is a direct indication of the non-equimolar presence of the isoforms. Concerning S1, the reads count supports that the S segment of P is more prevalent than it is for the IS isoform (supplementary table 1). Reads counts of S3, on the other hand, support that the IS is more prevalent than the P isoform (table 5). This is also supported by the alignment of the right end of both IS (S3.dPlot3) and P (S3.dPlot4). Such observations suggest a strain-specific preference for one isoform over another. These observations can be verified using Southern hybridization to detect unique restriction fragments of each isoform.

Evidence for genomic termini preference to include an *a* sequence seems preferable to terminating with a *c* sequence in S3. This is supported by the reads count (table 5) and alignment view of reads by definition (S3.dPlots 3 and 4). Alignment of the S segments ending with an *a* sequence shows a unified stop at the end of the repeat sequence. However, alignment of the S segments lacking an *a* sequence, from c_no_a files reveals an unexpected termination within the *c* sequence. This is the first evidence that the cryptic pac2 in the *c* sequence is poorly recognized, resulting in inefficient recognition of the cleavage site. In contrast, the cleavage mechanism seems to be highly accurate when a terminus ends with the *a* sequence (supplementary data, figure S1). This is supported by all plots showing the left end of the L segment and the right end of S (except for the alignment of c_no_a files).

Furthermore, alignment plots support the current understanding of having one *a* sequence copy predominantly prevalent at the left termini of both the P and IL isoforms. Few reads supported additional parts of the *a* sequence. However, analysis of the right termini on both IS and S showed evidence of having an extra copy of the *a* sequence in both strains. This observation has not been reported before in the literature. It can be reasoned that due to the sheer low number of reads that supports an additional *a* sequence on both ends (n<5), those reads belong to the junction region and are mismapped to the termini.

To validate the alignment of the repeating sequences on the appropriate location of the genome, mapping the first and last 20bp of each read was performed. The alignment was then quantified by position to reflect the frequency of read termination. The working presumption is that a random termination should occur across the genome, excluding genomic ends of isoforms. Indeed, looking at the S1.plot.sc and S3.plot.sc, we see a spike (n>5 reads) for each genomic region corresponding to an isoform genomic termini.

Here I presented a novel bioinformatics pipeline to reconstruct and analyze the terminal repeats of HCMV with the highest accuracy possible. The analysis supported previous understanding of the structured nature of the virus and revealed evidence of novel findings of the repeating region. This research project took two years to complete, indicating the HCMV genome complexity. I believe that results from this research are of value to the community, for the primary goal of this project was to add to and improve our current understanding of the HCMV genome, hoping to aid the search of antiviral agents.

Supplementary Data

A web link for each novel script is provided below:

S1:

https://github.com/erekevan/HCMV_Analysis_programs_python/blob/master/Terminal_Analyzer.py

S2:

https://github.com/erekevan/HCMV_Analysis_programs_python/blob/master/stats_cuts.py

S3:

https://github.com/erekevan/HCMV_Analysis_programs_python/blob/master/HCMV_Analysis_plots.r

Table S1) End-stage FRFs of S1 read count.

FRF	*Number of Reads* *Out of parent FRF*	*Percentage (total reads)*
UL_P_Ter	*60*	*0.3546 %*
UL_IL_Ter	*62*	*0.3664%*
UL_P_Jun	*62*	*0.3664%*
UL_IL_Jun	*60*	*0.3546 %*
US_P_Jun	*342*	*2.0213 %*
US_IS_Jun	*425*	*2.5118%*
US_P_Ter	*233*	*1.3771 %*
US_P_Ter_no_a	*192*	*1.1348%*
US_IS_Ter	*180*	*1.0638%*
US_IS_Ter_no_a	*72*	*0.4255%*

Samclib was obtained from: https://github.com/tseemann/samclip.git

Pacbio SMRT link was obtained from: https://www.pacb.com/support/software-downloads/

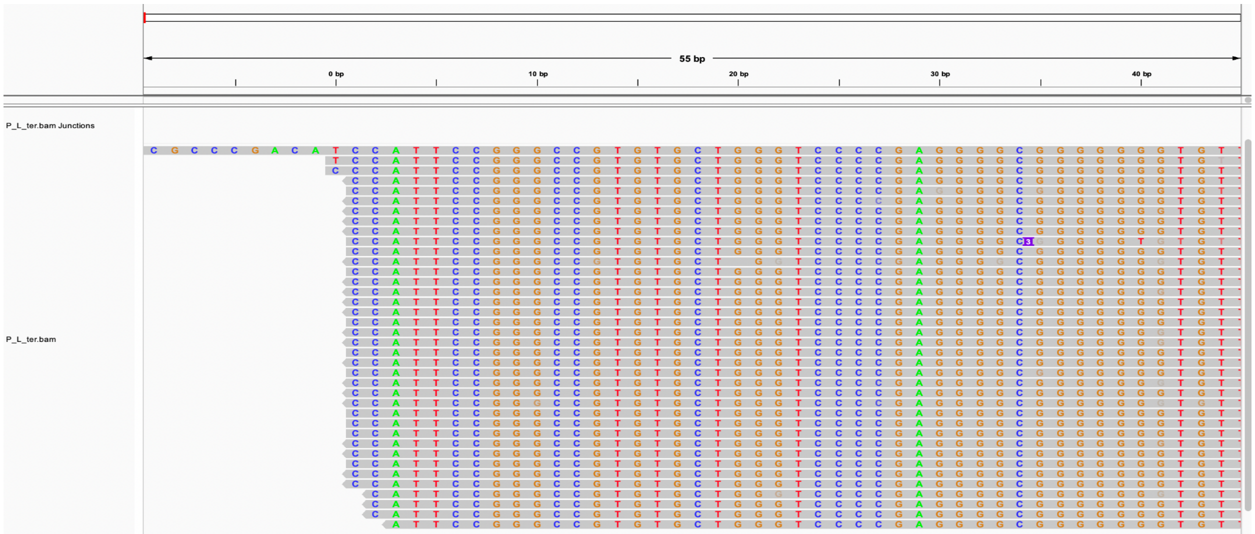

55 bp
0 bp
10 bp
20 bp
30 bp
40 bp
P_L_ter.bam Junctions
P_L_ter.bam
C G C C C G A C A T C C A T T C C G G G C C G T G T G C T G G G T C C C C G A G G G G C G G G G G G G T G T

Figure S1) Left end of the P L segment. Magnified alignment showing a uniform stop at the end of the termini.

References

- Zuhair, M., Smit, G., Wallis, G., Jabbar, F., Smith, C., Devleesschauwer, B., & Griffiths, P. (2019). Estimation of the worldwide seroprevalence of cytomegalovirus: A systematic review and meta-analysis. *Reviews in medical virology*, *29*(3), e2034.
- Cannon, M. J., Schmid, D. S., & Hyde, T. B. (2010). Review of cytomegalovirus seroprevalence and demographic characteristics associated with infection. Reviews in medical virology, 20(4), 202-213.
- Wang, Y. Q., & Zhao, X. Y. (2020). Human Cytomegalovirus Primary Infection and Reactivation: Insights From Virion-Carried Molecules. Frontiers in microbiology, 11, 1511.
- Stagno, S., Pass, R. F., Cloud, G., Britt, W. J., Henderson, R. E., Walton, P. D., Veren, D. A., Page, F., & Alford, C. A. (1986). Primary cytomegalovirus infection in pregnancy. Incidence, transmission to fetus, and clinical outcome. JAMA, 256(14), 1904–1908
- Adler, S. P. (1988). Molecular epidemiology of cytomegalovirus: viral transmission among children attending a day care center, their parents, and caretakers. The Journal of Pediatrics, 112(3), 366-372
- McVoy, M. A., & Nixon, D. E. (2005). Impact of
- 2-bromo-5,6-dichloro-1-beta-D-ribofuranosyl benzimidazole riboside and inhibitors of DNA, RNA, and protein synthesis on human cytomegalovirus genome maturation. Journal of Virology, 79(17), 11115–11127.
- Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. Human immunology, 82(11), 801–811.
- Gao, M., & Isom, H. C. (1984). Characterization of the guinea pig cytomegalovirus genome by molecular cloning and physical mapping. Journal of Virology, 52(2), 436–447.

- Leung, A. K., Sauve, R. S., & Davies, H. D. (2003). Congenital cytomegalovirus infection. Journal of the National Medical Association, 95(3), 213–218.
- Griffiths, P., Baraniak, I., & Reeves, M. (2015). The pathogenesis of human cytomegalovirus. The Journal of Pathology, 235(2), 288–297.
- Gerna, G., Kabanova, A., & Lilleri, D. (2019). Human Cytomegalovirus Cell Tropism and Host Cell Receptors. Vaccines, 7(3), 70.
- Sinzger, C., Digel, M., & Jahn, G. (2008). Cytomegalovirus cell tropism. Current Topics in Microbiology and Immunology, 325, 63–83
- Simanek, A. M., Dowd, J. B., Pawelec, G., Melzer, D., Dutta, A., & Aiello, A. E. (2011). Seropositivity to cytomegalovirus, inflammation, all-cause and cardiovascular disease-related mortality in the United States. PloS One, 6(2), e16103.
- Kalil, A. C., & Florescu, D. F. (2011). Is cytomegalovirus reactivation increasing the mortality of patients with severe sepsis?. Critical Care (London, England), 15(2), 138.
- Melnick, M., Sedghizadeh, P. P., Allen, C. M., & Jaskoll, T. (2012). Human cytomegalovirus and mucoepidermoid carcinoma of salivary glands: cell-specific localization of active viral and oncogenic signaling proteins is confirmatory of a causal relationship. Experimental and Molecular Pathology, 92(1), 118–125.
- Reeves, M., & Sinclair, J. (2008). Aspects of human cytomegalovirus latency and reactivation. Current Topics in Microbiology and Immunology, 325, 297–313.
- Sijmons, S., Van Ranst, M., & Maes, P. (2014). Genomic and functional characteristics of human cytomegalovirus revealed by next-generation sequencing. Viruses, 6(3), 1049–1072.
- Crough, T., & Khanna, R. (2009). Immunobiology of human cytomegalovirus: from bench to bedside. Clinical Microbiology Reviews, 22(1), 76–98.
- Gugliesi, F., Coscia, A., Griffante, G., Galitska, G., Pasquero, S., Albano, C., & Biolatti, M. (2020). Where do we Stand after Decades of Studying Human Cytomegalovirus?. Microorganisms, 8(5), 685.

- Bale Jr, J. F., O'Neil, M. E., Fowler, S. S., & Murph, J. R. (1993). Analysis of acquired human cytomegalovirus infections by polymerase chain reaction. *Journal of Clinical Microbiology*, *31*(9), 2433-2438.
- Walker, A., Petheram, S. J., Ballard, L., Murph, J. R., Demmler, G. J., & Bale, J. F., Jr (2001). Characterization of human cytomegalovirus strains by analysis of short tandem repeat polymorphisms. *Journal of Clinical Microbiology*, *39*(6), 2219–2226.
- Wilkinson, G. W., Davison, A. J., Tomasec, P., Fielding, C. A., Aicheler, R., Murrell, I., Seirafian, S., Wang, E. C., Weekes, M., Lehner, P. J., Wilkie, G. S., & Stanton, R. J. (2015). Human cytomegalovirus: taking the strain. *Medical Microbiology and Immunology*, *204*(3), 273–284.
- Martí-Carreras, J., & Maes, P. (2019). Human cytomegalovirus genomics and transcriptomics through the lens of next-generation sequencing: revision and future challenges. Virus Genes, 55(2), 138–164.
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008.
- Vanarsdall, A. L., & Johnson, D. C. (2012). Human cytomegalovirus entry into cells. Current Opinion in Virology, 2(1).
- Sauer, A., Wang, J. B., Hahn, G., & McVoy, M. A. (2010). A human cytomegalovirus deleted of internal repeats replicates with near wild type efficiency but fails to undergo genome isomerization. *Virology*, *401*(1), 90–95.
- Dunn, W., Chou, C., Li, H., Hai, R., Patterson, D., Stolc, V., Zhu, H., & Liu, F. (2003). Functional profiling of a human cytomegalovirus genome. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(24), 14223–14228.

- Al Qaffas, A., Camiolo, S., Vo, M., Aguiar, A., Ourahmane, A., Sorono, M., Davison, A. J., McVoy, M. A., & Hertel, L. (2021). Genome sequences of human cytomegalovirus strain TB40/E variants propagated in fibroblasts and epithelial cells. *Virology journal*, *18*(1), 112.
- Piret, J., & Boivin, G. (2019). Clinical development of letermovir and maribavir: Overview of human cytomegalovirus drug resistance. *Antiviral Research*, *163*, 91–105.
- Theiß, J., Sung, M. W., Holzenburg, A., & Bogner, E. (2019). Full-length human cytomegalovirus terminase pUL89 adopts a two-domain structure specific for DNA packaging. *PLoS Pathogens*, *15*(12), e1008175.
- Mori, Y., Jinnouchi, F., Takenaka, K., Aoki, T., Kuriyama, T., Kadowaki, M., Odawara, J., Ueno, T., Kohno, K., Harada, T., Yoshimoto, G., Takase, K., Henzan, H., Kato, K., Ito, Y., Kamimura, T., Ohno, Y., Ogawa, R., Eto, T., Nagafuji, K., … Miyamoto, T. (2021). Efficacy of prophylactic letermovir for cytomegalovirus reactivation in hematopoietic cell transplantation: a multicenter real-world data. *Bone Marrow Transplantation*, *56*(4), 853–862.
- Cronn, R., Liston, A., Parks, M., Gernandt, D. S., Shen, R., & Mockler, T. (2008). Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Research, 36(19), e122.
- Segerman B. (2020). The Most Frequently Used Sequencing Technologies and Assembly Methods in Different Time Segments of the Bacterial Surveillance and RefSeq Genome Databases. *Frontiers in Cellular and Infection Microbiology*, *10*, 527102.
- Ewing B, Hillier L, Wendl MC, Green P. (1998): Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8(3):175-185
- Eid, J., et al. (2009) Real-time DNA sequencing from single polymerase molecules. Science, 323(5910), 133–138.
- Kchouk M, Gibrat JF, Elloumi M (2017) Generations of Sequencing Technologies: From First to Next Generation. Biol Med (Aligarh) 9: 395.

- Kulski JK (2016) Next-generation sequencing-An overview of the history, tools, and “Omic” applications, next generation sequencing-advances, applications and challenges. InTech.
- Sutton, T., Clooney, A. G., Ryan, F. J., Ross, R. P., & Hill, C. (2019). Choice of assembly software has a critical impact on virome characterisation. *Microbiome*, *7*(1), 12.
- Al Qaffas, A., Nichols, J., Davison, A. J., Ourahmane, A., Hertel, L., McVoy, M. A., & Camiolo, S. (2021). LoReTTA, a user-friendly tool for assembling viral genomes from PacBio sequence data. *Virus Evolution*, *7*(1), veab042.
- Davison AJ, Eberle R, Ehlers B, Hayward GS, McGeoch DJ, Minson AC, Pellett PE, Roizman B, Studdert MJ, Thiry E. The order Herpesvirales. Arch Virol. 2009;154(1):171-7.
- John D. Shanley (2000). Sexually Transmitted Diseases Vaccines, Prevention and Control. Chapter 10, P 239-257.
- Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34:3094-3100.
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience, 10(2), giab008.
- Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A., & Mesirov, J. P. (2017). Variant Review with the Integrative Genomics Viewer. *Cancer Research*, *77*(21), e31–e34.
- Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., Smirnova, T., Grigoriev, I. V., & Dubchak, I. (2014). The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Research, 42(Database issue), D26–D31.
- Nixon, D. E., & McVoy, M. A. (2002). Terminally repeated sequences on a herpesvirus genome are deleted following circularization but are reconstituted by duplication during cleavage and packaging of concatemeric DNA. Journal of Virology, 76(4), 2009–2013.

- Fulkerson, H. L., Nogalski, M. T., Collins-McMillen, D., & Yurochko, A. D. (2021). Overview of Human Cytomegalovirus Pathogenesis. *Methods in Molecular Biology (Clifton, N.J.)*, *2244*, 1–18.
- Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4.
- Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.
- Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*(6), 461–468.
- RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

VITA

Ahmed Ali Al Qaffas was Born on October, 16, 1978 in Dhahran City, Eastern Province, Saudi Arabia. He graduated from Al-Khobar Highschool, Al Khbar in 2004. He received his Bachelor of Science in Bioinformatics from Virginia Commonwealth University in 2019. He contributed to the field of virology by studying data collected from Human Cytomegalovirus and other microorganisms. He received a Masters of Bioinformatics from Virginia Commonwealth University in 2022.