Transcriptomic analysis of a human and an insect DNA virus using an integrated sequencing approach

Ph.D. Thesis

Ádám Fülöp

Department of Medical Biology

Doctoral School of Interdisciplinary Medicine

Faculty of Medicine

University of Szeged

Supervisor: Prof. Dr. Zsolt Boldogkői Ph.D., DSc

Szeged

- 2024 -

1. List of publications:

1.1 Publications directly related to the subject of the thesis

I. **Fülöp Á**, Torma G, Moldován N, Szenthe K, Bánáti F, Almsarrhad IAA, Csabai Z, Tombácz D, Minárovits J, Boldogkői Z. Integrative profiling of Epstein-Barr virus transcriptome using a multiplatform approach. Virol J. 2022 Jan 6;19(1):7. doi: 10.1186/s12985-021-01734-6. **IF:4.8**

II. Torma G, Tombácz D, Moldován N, **Fülöp Á**, Prazsák I, Csabai Z, Snyder M, Boldogkői Z. Dual isoform sequencing reveals complex transcriptomic and epitranscriptomic landscapes of a prototype baculovirus. Sci Rep. 2022 Jan 25;12(1):1291. doi: 10.1038/s41598-022-05457-8. **IF:4.6**

1.2 Other related publications

III. Prazsák I, Csabai Z, Torma G, Papp H, Földes F, Kemenesi G, Jakab F, Gulyás G, **Fülöp Á**, Megyeri K, Dénes B, Boldogkői Z, Tombácz D. Transcriptome dataset of six human pathogen RNA viruses generated by nanopore sequencing. Data Brief. 2022 Jun 18;43:108386. doi: 10.1016/j.dib.2022.108386. **IF:1.2**

IV. Kakuk B, Dörmő Á, Csabai Z, Kemenesi G, Holoubek J, Růžek D, Prazsák I, Dani VÉ, Dénes B, Torma G, Jakab F, Tóth GE, Földes FV, Zana B, Lanszki Z, Harangozó Á, **Fülöp Á**, Gulyás G, Mizik M, Kiss AA, Tombácz D, Boldogkői Z. In-depth Temporal Transcriptome Profiling of Monkeypox and Host Cells using Nanopore Sequencing. Sci Data. 2023 May 9;10(1):262. doi: 10.1038/s41597-023-02149-4. **IF:9.8**

V. Tombácz D, Torma G, Gulyás G, **Fülöp Á**, Dörmő Á, Prazsák I, Csabai Z, Mizik M, Hornyák Á, Zádori Z, Kakuk B, Boldogkői Z. Hybrid sequencing discloses unique aspects of the transcriptomic architecture in equid alphaherpesvirus 1. Heliyon. 2023 Jun 28;9(7):e17716. doi: 10.1016/j.heliyon.2023.e17716. **IF:4**

VI. Torma G, Tombácz D, Csabai Z, Almsarrhad IAA, Nagy GÁ, Kakuk B, Gulyás G, Spires LM, Gupta I, **Fülöp Á**, Dörmő Á, Prazsák I, Mizik M, Dani VÉ, Csányi V, Harangozó Á, Zádori Z, Toth Z, Boldogkői Z. Identification of herpesvirus transcripts from genomic regions around the replication origins. Sci Rep. 2023 Sep 29;13(1):16395. doi: 10.1038/s41598-023-43344-y. **IF:4.6**

Cumulative IF:29

2. Table of contents

3. Introduction

Cells are characterized by a set of RNA molecules expressed at a given time, called the transcriptome. These include both protein-coding and non-coding RNAs, as well as transcripts with splice and alternative initiator and terminator sites (K-H Liang, 2013). Understanding them is important for studying functional elements of the genome (Wang et al., 2009). The main goal of transcriptomics is to map and quantify the characteristics of RNA molecules.

3.1 Sequencing Technologies

RNA sequencing has become a common and ubiquitous tool for analyzing quantitative changes in gene expression between experimental groups (differential gene expression or DGE) (Young et al., 2012) or in longitudinal sampling of tissues and microorganisms (Hubbard et al., 2013). The first form of RNA sequencing was used in 1977 when Fredrick Sanger developed the chain-end method, which is the first-generation sequencing platform (Adams et al., 1995, 1991). This platform had many technical difficulties due to sequencing length limits and low throughput. In this process, radioactive or fluorescently labeled dideoxynucleotides are used to stop the synthesis of the DNA molecule, resulting in shorter DNA fragments that are separated by capillary electrophoresis at the end of the process. An automated version of this method was developed in 1986 (Heather and Chain, 2016). This method was based on the chemical cleavage or degradation of molecules. Walter Gilbert and Allan Maxam developed a chemical degradation technique to sequence the DNA of the bacteriophage Phix174 (Satam et al., 2023). In the experiment, the DNA was labeled with radioactive phosphate at the 5' end, and then the bases were removed from purines and pyrimidines by various chemical treatments. Subsequently, the phosphodiester bond was cleaved with piperidine to produce fragments of different sizes, which could later be separated by gel electrophoresis. In the 1990s, The Human Genome Project was launched with the goal of sequencing the entire human genome, creating a significant demand for high-throughput technologies. Tag-based techniques were used to enable higher throughput, more accurate isoform detection, and quantitative analysis, such as sequential analysis of gene expression (SAGE) (Velculescu et al., 1995) or cap analysis gene expression (CAGE) (Shiraki et al., 2003). These methods are still used to some extent by the virology community on modern sequencing platforms (Djavadian et al., 2018; Wyler et al., 2017)

Innovations in microfluidics and nanotechnology have ushered in an era of next-generation sequencing platforms. New generation sequencing (NGS) or second generation sequencing platforms offer a key advantage over classical Sanger sequencing, as they do not require bacterial cloning and electrophoretic separation. They provide high throughput and can deliver whole-genome information, simultaneously sequencing millions of cDNA molecules in parallel, reducing the overall time and cost of sequencing, and vastly increasing the amount of information output. In contrast to first-generation platforms, second-generation platforms are sensitive to the expression level of splice isoforms and can be used to discover novel genes and non-coding RNAs (Wang et al., 2009). In 2005, Roche's model 454 was the first, a synthesis-based bioluminescence method (Metzker, 2010; van Dijk et al., 2018). In this reaction, dNTPs are cyclically added, and pyrophosphate released upon incorporation is detected. Roche 454 had a significant advantage in long reads (~1 kb) but a disadvantage in low coverage. Other technology-based developments have been initiated, resulting in the release of the Ion PGM platform in 2010 (Liu et al., 2012), which works on the basis that when a polymerase incorporates a nucleotide into DNA, a proton is released, causing a pH change that can be detected. This does not require fluorescence or optical detection, so the platform can be operated at a lower cost. Illumina was launched in 2007 and is currently the most widely used platform in genomics (Turnbull et al., 2018; Weimer, 2017). Its operation is Sequencing by Synthesis (SBS) based, where the molecule to be sequenced is hybridized to oligonucleotides on the flow cell, which are amplified multiple times by bridge amplification to form clusters (Metzker, 2010). The sequencing process uses fluorescently labeled dNTPs, whose fluorescence is detected by a camera upon incorporation. A major advantage of Illumina sequencers is the large number of reads, approximately 30-100 million (van Dijk et al., 2014).

Second-generation technologies are not suitable for the detection of long RNA molecules and their isoforms (Byrne et al., 2019). Currently, two large sequencing platforms, PacBio and ONT MinION, are the most widely used for the detection of long RNA molecules while retaining high throughput. PacBio is based on nanosensor technology (Eid et al., 2009). The template molecule to be sequenced is ligated to a hairpin-shaped adapter (SMRTbell), creating a circular molecule. This molecule is loaded onto an SMRTcell, which has many thousands of picoliter-sized holes (ZMVs) (Levene et al., 2003). A DNA polymerase is attached to the bottom of these ZMWs, which, during synthesis, incorporates fluorescently labeled nucleotides of different colors according to the four nucleotides. These are detected by an optical detector when the labeled phosphorylated end is detached, and a light signal is obtained (Rhoads and Au, 2015). The advantage of this technology is that a molecule can be sequenced multiple times, as the template is a circular molecule. The system does not require amplification of the starting material and is able to accommodate native DNA (Coupland et al., 2012). PacBio has released two platforms in recent years, RSII and Sequel (Satam et al., 2023; Travers et al., 2010).

A new competitor in the sequencing industry is Oxford Nanopore Technologies. Their sequencing approach is based on a completely new approach using protein nanopores embedded in a synthetic membrane (Lu et al., 2016). These pores are alpha-hemolysates embedded in a lipid bilayer (Schneider and Dekker, 2012). The pores are arranged in a flow cell, with 2048 pores divided into four groups of 512 (Lu et al., 2016). During cDNA library construction, a motor protein is ligated to DNA, or in the case of direct RNA sequencing, to the cDNA-RNA hybrid (Ayub et al., 2013), aiding in the unwinding and translocation of the double-stranded molecule through the pore (Satam et al., 2023; van Dijk et al., 2018) As the molecule passes through the pore, the sensors detect changes in ionic current corresponding to the characteristics of each nucleotide passing through. This information provides the signal used for base search. The specifics of platforms mentioned in this section can be found in **Table 1**.

Table 1. Comparison of sequencing platforms.

	Sequencing Method	**Detection Method**	**Read Length**	**Advantage**	**Disadvantage**
Roche 454	Pyrosequencing, cleavage of released pyrophosphate	Light	700	Average read length	High homopolymer error rate
Ion Torrent	Ion semiconductor sequencing	pH	200	Rapid runs	High homopolymer error rate
Illumina MiSeq	Sequencing by synthesis	Light	2*300	Low error rates, run cost	Short-read length
PacBio Sequel	Polymerase incorporating colored NTPs	Light	up to 15,000	Low error rare, long-read length	High cost
ONT MinION	Molecule traverses pore	Current	up to 1M	Long-read length, low cost	Relatively high error rate

3.2 Native RNA sequencing

Two major problems can arise during transcriptome sequencing, leading to the detection of false isoforms (Cocquet et al., 2006). One is false priming, where the oligod(T) primer used for reverse transcription binds to an adenine-rich region, resulting in a truncated cDNA with a defective 3' end (Balázs et al., 2019). Another significant problem is the temple shift (TS), when the reverse transcriptase (RT) enzyme jumps homology-dependently to another template strand and the strand shift phenomenon can occur between separate cDNA strands. This leads to the generation of transcripts containing false introns with non-canonical splice sites. The potency of TS is enhanced when the concentration of templates is high, homologous sequences are long or the temperature of reverse transcription is low (Odelberg et al., 1995).

To avoid these two defects, direct RNA sequencing developed by nanopore technology, which directly sequences RNA, provides an efficient alternative (Garalde et al., 2018). This platform has several significant benefits, as it avoids errors from PCR amplification and reverse transcription, and it is possible to detect 5-methylcytosine and 6-methyladenine modifications. The quality of the resulting reads is poorer than that of MinION cDNA-seq, but this is not a problem for isoform detection if a well annotated reference genome is available. A major drawback of the technology is that it cannot accurately detect RNA approximately 30 nucleotides from the 5' end of the RNA (Soneson et al., 2019; Tombácz et al., 2020) This problem arises because the ratchet molecule releases the RNA strand a few nucleotides (15-30) before the 5' end and it passes through the pore, resulting in a rapid sliding of the molecule through the pore, resulting in a weak signal that is not possible to base-called. Since the sequencing rate is six times slower compared to cDNA-Seq, another current major drawback of direct RNA sequencing is its low throughput (Garalde et al., 2018).

3.3 Transcriptional overlaps and Virus Genetic Regulations

Viral transcripts can overlap with each other (Boldogköi, 2012; Boldogkői et al., 2019b). Based on the position of the genes, these overlaps can be convergent (tail-to-tail), divergent (head-to-head) and parallel (head-to-head). These overlaps can be 'soft' when only some reads overlap or 'hard' overlaps where all reads overlap. All these overlaps may play a potential role in gene regulation, meaning that they may regulate and synchronize the kinetics of viral genes through the physical interaction of transcriptional machinery. This hypothesis is called the Transcription Interference Network (TIN).

Another interesting phenomenon may be generated by replication-associated RNAs transcribed near replication origins. Several such molecules have previously been detected in herpesviruses using Long-Read Sequencing (LRS) (Boldogkői et al., 2019a; Torma et al., 2023). These RNAs are thought to regulate the initiation of replication and the orientation of the replication fork. DNA replication and transcription processes are likely to generate genome-wide interference, where these two processes tightly regulate each other at the genomic level. A particularly interesting aspect and evidence for the existence of this process may be that the genes surrounding the replication origins (Ori) are regulatory genes involved in the initiation of replication and transcription. In BK viruses, it was observed that these replication-associated RNAs (raRNAs) bind to sense and antisense DNA strands and prevent viral replication (Tikhanovich et al., 2011). In another study published on Epstein-Barr virus (EBV), it was shown that an RNA named BHLF1, which partially overlaps with the origin of replication, is able to bind to DNA and therefore inhibit replication (Rennekamp and Lieberman, 2011).

3.4 Viral transcriptome analysis with long-read sequencing

Our group has analyzed several human and animal viral transcriptomes, among them: Human herpes viruses: Herpes simplex virus type 1 (HSV-1), Epstein-Barr virus (EBV), Human betaherpesvirus 5 (HCMV), Kaposi’s sarcoma-associated herpesvirus (KSHV) (Fülöp et al., 2022; Kakuk et al., 2021b; Prazsák et al., 2023; Tombácz et al., 2020). Animal herpes viruses: Pseudorabies virus (PRV), Bovine alphaherpesvirus-1 (BoHV-1), Equine herpesvirus-1 (EHV-1) (Moldován et al., 2020; Tombácz et al., 2023; Torma et al., 2021a). Other animal pathogenic viruses: African Swine Fever Virus (ASFV), Autographa californica nucleopolyhedrovirus (AcMNPV), Vesicular Stomatitis Indiana (VSV), Vaccinia virus (Kakuk et al., 2021a; Tombácz et al., 2021; Torma et al., 2022, 2021b).

3.5 Epstein–Barr virus, a human gammaherpesvirus

The Epstein-Barr virus (EBV, human gammaherpesvirus 4) is a member of the subfamily Gammaherpesvirinae within the family Herpesviridae (Davison et al., 2009). EBV is predominantly transmitted by saliva and is widespread in human populations (Rickinson et al., 2007). EBV plays a role in the pathogenesis of Burkitt's lymphoma and other lymphomas, and it is also involved in the development of nasopharyngeal carcinoma and a subset of gastric carcinomas (Shannon-Lowe and Rickinson, 2019; Young et al., 2016). EBV is classified as a group 1 carcinogen in humans (de Martel et al., 2020). Moreover, EBV reactivation is considered a major cause of long COVID symptoms (Gold et al., 2021). Primary EBV infection in early childhood is typically mild or asymptomatic. Later, however, it can cause infectious mononucleosis (IM, glandular fever), a lymphoproliferative disease accompanied by pharyngitis, tonsillitis, fever and lymphadenopathy. In the majority of cases, IM is a self-restrictive disease due to a strong T cell response to proliferating B cells infected with EBV and expressing viral antigens (Meckiff et al., 2019). Although B cells are the main target of EBV, the virus also replicates in oropharyngeal epithelial cells. Both B cells and epithelial cells are able to produce new EBV particles that carry a linear, double-stranded viral genome.

Following primary infection, the virus creates a lifelong latency in memory B cells (Thorley-Lawson, 2015). Only a limited number of viral genes from the circular, episomal, chromatinized EBV genome are expressed in latently infected cells and can be classified into three gene expression programs (types I, II and III) (Price and Luftig, 2015). It is characterized by the limited expression of six EBV-encoded nuclear antigens (EBNA-1, -2, -LP (leader protein), -3A, -3B and -3C), two latent membrane proteins (LMP-1 and -2), two small non-coding RNAs (EBER1 and -2) and a set of viral microRNAs (miRNAs). All EBNA proteins are expressed from a common transcript that spans most of the EBV genome through alternative promoters, splicing and polyadenylation (Bodescot et al., 1987; Rogers et al., 1990). EBNA-1, EBERs and BamHI a rightward transcripts (miR-BARTs) are all expressed in three latency types (Price and Luftig, 2015). During cell division, EBNA-1 is necessary for the maintenance of the episomal viral genome in infected cells (Deakyne et al., 2017; Hodin et al., 2013; Kanda et al., 2013). EBNA-2, together with all other latency genes, is expressed in type III latency. It transactivates the expression of LMP-1 and has a positive effect on EBNA-2 activity together with EBNA-LP and a negative effect on EBNA-2 activity together with EBNA-3 (Kempkes and Ling, 2015). Expressed in both type II and type III latency, LMP-1 and -2 are transmembrane proteins that promote the proliferation and B cell function of latently infected and transformed cells by modulating host signaling pathways (Kang and Kieff, 2015; Vrazo et al., 2012). Non-coding EBERs and BART-derived viral miRNAs also promote latency by modulating host expression at the post-transcriptional level (Cai et al., 2006; Lee et al., 2016; Marquitz et al., 2015). The complete absence of latent gene expression is sometimes regarded as "0" (zero) latency (Thompson and Kurzrock, 2004). EBV-associated lymphoma and carcinoma cells and in vitro immortalized lymphoblastoid cell lines (LCL) may also carry a latent EBV genome in addition to memory B cells. The epigenetic machinery of the host cell interacts with viral episomes and the activity of latent EBV promoters is regulated by epigenetic signals deposited by host cell enzymes on transcriptional control sequences of the viral genome

(Takacs et al., 2010). In latently infected cells, viruses are replicated once per cell cycle by the host DNA synthesis machinery, with viral episomes binding to the nuclear matrix at oriP, the latent origin of EBV replication (Hammerschmidt and Sugden, 2013). Different signals can disrupt latency and induce EBV lytic reactivation both in vitro and in vivo (Kenney and Mertz, 2014; Li et al., 2018).

Induction of EBV lytic replication results in a change in the limited latent expression pattern of EBV genes by sequential transcription of immediate early (IE), early (E) and late (L) EBV genes. IE genes, BZLF1 and BRLF1, are transactivator proteins of early genes that turn on the transcription of early genes (Liu and Speck, 2003; Schaeffner et al., 2019). The E gene products contain, among others, the core set of lytic replication proteins shared by Herpesvirus species (Rennekamp and Lieberman, 2011). Lytic EBV DNA synthesis happens in the replication compartments inside the host cell nucleus (Nagaraju et al., 2019). The exponential amplification of the viral genome is initiated at one of two copies of oriLyt, the lytic replication origin of EBV DNA synthesis, in contrast to the replication of latent episomes (Hammerschmidt and Sugden, 2013). It is hypothesized that during productive replication, the EBV IE and E genes are transcribed from chromatinized templates, while the unchromatinized, unmethylated templates are used to transcribe the L genes encoding the virion's structural proteins (Chakravorty et al., 2019). Late RNA transcription of EBV is promoted by the virion's preinitiation complex (Djavadian et al., 2018). After the synthesis of the virus' structural proteins, epigenetically naive, unmethylated, linear dsDNA molecules are packaged into EBV virions (Chakravorty et al., 2019; Woellmer and Hammerschmidt, 2013). These linear genomes undergo chromatization, circularization and epigenetic modification in the newly infected host cells.

Initial studies have shown that all viral genes mapped in the approximately 170 kb EBV genome are actively transcribed during the lytic cycle and can encode more than 100 different gene products, of which 69 EBV-encoded proteins have recently been identified by proteomic analysis (Arvey et al., 2012; Dresang et al., 2011; Ersing et al., 2017; Yuan et al., 2006). The genome structure of the virus is shown in Figure 1. However, recent studies of the viral transcriptome have revealed a more complex pattern of viral gene expression after EBV latency disruption in different cell lines. It has been shown that transcription of the lytic cycle is bidirectional and that several newly identified transcribed regions do not encode proteins (Cao et al., 2015; Majerciak et al., 2018; O’Grady et al., 2016, 2014). These data indicate that hundreds of viral long non-coding RNAs (lncRNAs) may be produced during productive EBV replication. Moreover, novel splicing events further increase the diversity of the EBV transcriptome expressed during productive replication (Concha et al., 2012)

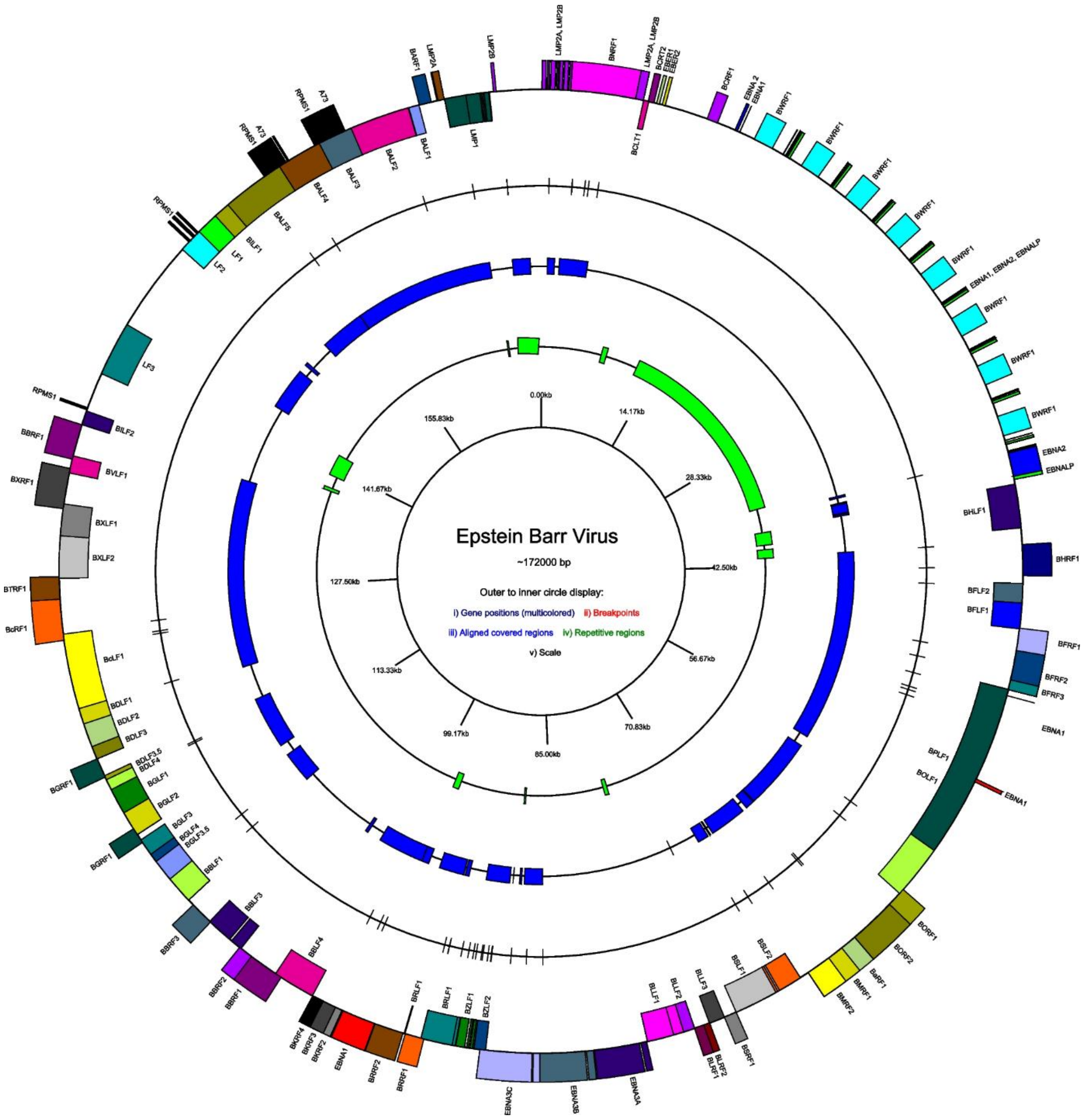


Figure 1 EBV genome map with positions of recombination breakpoints. From outer to inner, circles display genomic positions for (i) gene positions, (ii) breakpoints, (iii) aligned covered regions, (iv) repetitive regions, and (v) scale. Genes are color-coded based on the gene exons. Genes on the outside are transcribed clockwise and the inner are counterclockwise. This figure was drawn by GenomeVx (Agwati et al., 2022)

3.6 AcMNPV, a baculovirus

The Autographa californica multiple nucleopolyhedrovirus (AcMNPV) is an insect virus of the family Baculoviridae (Blissard and Rohrmann, 1990). It is utilized as a biopesticide in agriculture, for protein production in research and industry, and as a gene delivery system for mammalian cell transduction (A. Kost et al., 2010; Hu, 2006, 2005; Kenoutis et al., 2006;

Kost et al., 2005). The developed recombinant SARS-CoV-2 antibody nanoparticle vaccine is based on this virus (Tian et al., 2021). Insects become infected through their digestive tract when they consume vegetation contaminated with the virus inclusion bodies, which resistant protein structures that carry the virus. After ingestion, the virus binds to and enters the host endothelial cells, from where it infects the entire insect body by budding (George F Rohrmann., 2008a). During infection, two different forms of the virion are produced: occlusion-derived viruses, surrounded by an envelope containing viral proteins that ensure survival even in harsh environments such as the midgut of insects, and budding viruses, which have an envelope and some proteins facilitating their systemic spread in the near-neutral environment of insect tissue (Volkman et al., 1976). The 134 kbp long, double-stranded circular viral DNA contains 150 closely spaced open reading frames (ORFs) (Ayres et al., 1994). As shown by our group and others, the proximity of ORFs causes overlap between several transcripts of AcMNPV (Chen et al., 2013; Moldován et al., 2018a). The AcMNPV genes are expressed in three phases: early (E), late (L), and very late (VL) (George F Rohrmann., 2008b). Early transcription (0-6 hours post-infection, (p.i.)) produces transcriptional activators (Guarino and Summers, 1986) and the molecular machinery for DNA replication (Kool et al. 1994). E genes are transcribed by host RNA polymerase II, which recognizes TATA promoter elements located upstream of the transcription start site (TSS) or at the initiator element of the arthropod CAGT (Kogan et al., 1995), however, some E genes lack a canonical initiator sequence or any recognizable promoter motif (Lu and Carstens, 1993). After a transient early/late phase, some E genes cease to be expressed, while others are transcribed throughout the infection cycle, presumably due to the presence of early and late promoters and/or initiators (Kovacs et al., 1991). The L phase starts at the beginning of genome replication (6-18 h p.i.). The viral RNA polymerase (RNP) transcribes the L and VL genes, which recognize the consensus late initiator sequence (TAAG) on DNA and start synthesizing RNAs from the second nucleotide of the motif (Chen et al., 2013; Garrity et al., 1997). VL gene expression (18-72 h p.i.) is characterized by the synthesis of occlusion body proteins: polyhedrin and p10, as well as transcription factors such as very late expression factor 1 (VLF-1). The VL genes contain an A/T-rich region (Ooi et al., 1989), referred to as the "burst sequence" after their late initiation sequence (LIS) recognized by VLF-1, which promotes the high expression of these VL genes (McLachlin and Miller, 1994).

Capping of viral RNAs is performed by both host and viral proteins: LEF-4 RNA exhibits 5′-triphosphatase and guanyltransferase functions (Li and Guarino, 2008), while MTase (encoded by Ac69) methylates guanosine in the Cap structure (Wu and Guarino, 2003).

The majority of AcMNPV transcripts contain a canonical polyadenylation signal (PAS) upstream of their transcription end site (TES). The PASs are detected by the cleavage and polyadenylation apparatus of the host, which nick the transcripts in their 3′-UTR region and perform non-templated adenine addition. Viral (RNP) has also been shown (Jin and Guarino, 2000) to initiate catalysis of poly(A)-tail formation after transcription of uracil-rich regions, which may lead to alternative terminations in late transcript species (Moldován et al., 2018a). The genome of the virus is shown in Figure 2.

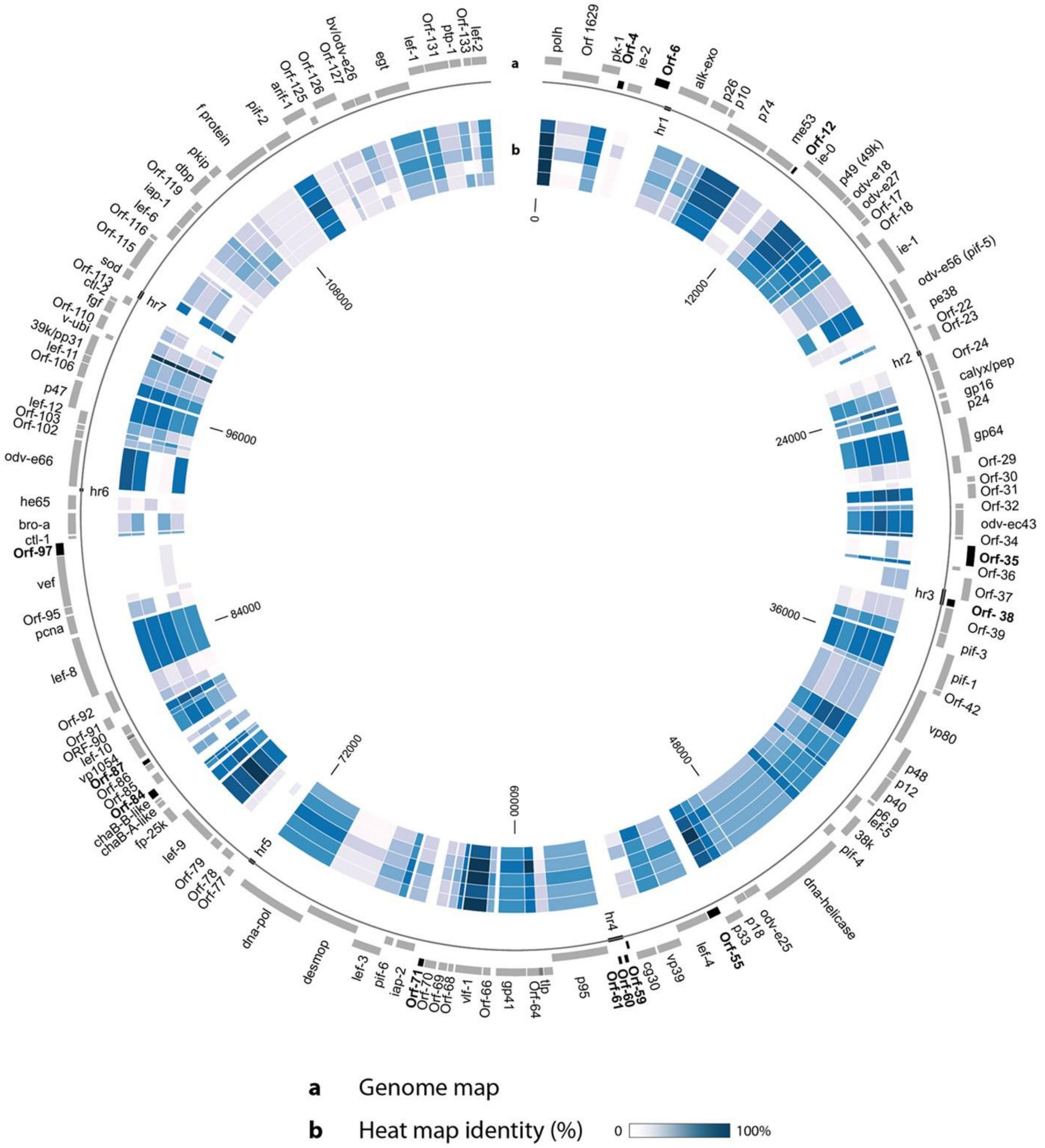


Figure 2 a) Genome map showing all the 134 ORFs. The unique ORFs are represented in black. The outer track contains forward orientation ORFs and the inner track contains reverse orientation ORFs. Hrs are shown on the line below the genome. (b) Heat map identity of the genomes of the species AcMNPV, ThorNPV, MaviMNPV, DekiNPV and AnpeNPV (from the outside to the inside) compared to ortholog ORFs from LoobMNPV. The darker the blue, the higher is the correlated ORF identity (Aragão-Silva et al., 2016).

4. Aims

The aim of this work is to detect full-length RNAs and to create a close-to-complete transcriptome atlas of AcMNPV and EBV.

In the case of EBV, previous studies based on Illumina short-read sequencing (SRS) (Djavadian et al., 2018; O'Grady et al., 2016, 2014; Peng et al., 2019) and Pacific Biosciences (PacBio) RS II sequencing (O'Grady et al., 2016) have identified a large number of transcripts. However, SRS is not optimal for detecting transcriptome complexity (Liu et al., 2012; Steijger et al., 2013), and RS II sequencing also has a limitation in detecting transcripts within a certain size range (Balázs et al., 2017). In this work, we analyze the EBV lytic transcriptome using the MinION sequencing platform from Oxford Nanopore Technologies (ONT), which is capable of providing a complete picture of the transcriptomic architecture of the virus (Moldován et al., 2017, 2018a; Prazsák et al., 2018; Tombácz et al., 2019) and integrate previous results with our data.

For AcMNPV the structure of the transcriptome has already been described in a study using Illumina SRS (Chen et al., 2013) and in our work (Moldován et al., 2018a) using third-generation long-read cDNA and direct RNA sequencing. Other studies have focused on characterizing transcriptional dynamics using microarrays (Smith, 2007), real-time PCR analysis (Jiang et al., 2006) and Illumina SRS (Chen et al., 2013). The techniques used in these gene expression analyses are not suitable for addressing the structural complexity of the baculovirus transcriptome. The aims of this work are to update the AcMNPV transcriptome using a dual LRS approach and to detect RNA methylation and editing by ONT sequencing.

For each virus, we aimed at six things: Identification of the 5' end of mRNAs. Determination of the 3' end of the mRNAs. Detection of promoter elements (TATA box, CAAT box, GC box) and polyadenylation signals. Linking annotated TSS and TES positions to transcripts. Categorization and abundance determination of annotated transcript isoforms, polycistronic RNAs, ncRNAs, antisense RNAs and 5' truncated RNAs. Detection of transcriptional overlaps

5. Materials and methods

5.1 Cells and viruses

Each step of this section has been carried out according to the guidelines and regulations for virus propagation and decontamination.

5.1.1 EBV

Close to saturation, Akata cells were diluted one-toone with RPMI-1640 medium supplemented with 10% FCS and pens/strep 24 h before induction. Cells were washed and resuspended to 106 cell/ml in RPMI solution supplemented with Goat anti-human IgG (Jackson, 109-001-003, 17 µg/ml final concentration), or in normal RPMI serving as controls in 7–7 T25 cell culture flasks (O'Grady et al., 2014). 100 µl cell suspensions were aspirated at time points of 10 min, 90 min, 4, 12, 24, 48 and 72 h after resuspension for RNA isolation to verify the success of induction of Epstein–Barr Virus transcription. Applying real-time PCR the activity of the BZLF and GP350 genes were monitored normalized to reference genes. The remaining cells were pelleted and stored at − 70 °C.

5.1.2 AcMNPV

AcMNPV expressing lacZ gene (βgal-AcMNPV) was propagated on the Sf9 cell line (both kindly provided by Ernő Duda Jr., Solvo Biotechnology, Hungary). Cells were cultivated in 200 mL of GIBCO Sf-900 II SFM insect cell medium (Thermo Fisher Scientific) in a Corning spinner flask (Merck) at 70 rpm and 26 °C, and they were infected with a viral titer of 2 multiplicity of infection (MOI = plaque-forming units per cell). A 5 mL sample was measured and centrifuged at 2000 rpm at 4 °C at nine consecutive time points after inoculation (5 min, 1 h, 2 h, 4 h, 6 h, 16 h, 24 h, 48 h, and 72 h), followed by washing with PBS and centrifuged again. Pellets were stored at − 80 °C until use.

5.2 RNA isolation and library preparation

Total RNA was purified from the cells using the NucleoSpin RNA Kit (Macherey–Nagel). For EBV total RNA samples were split in two. Polyadenylated RNAs were isolated from half of the total RNA. Ribodepletion, was carried out to remove ribosomal RNA from the other half of total RNAs using Epicentre Ribo-Zero Magnetic Kit. In case of AcMNPV thirty-five µg of total RNA was pipetted in separate from every time point. Polyadenylated RNAs were isolated samples using the Oligotex mRNA Mini Kit (Qiagen). The concentrations of RNA samples were determined using Qubit 4 (Thermo Fisher Scientific). The RNA BR Assay Kit (Thermo Fisher Scientific) was used for the quantification of total RNAs while the Qubit

RNA HS Assay (Thermo Fisher Scientific) Kit was applied for the measurement of polyadenylated and ribodepleted samples. The RNA quality was measured with a TapeStation 4150(Agilent). RNAs were stored at − 80 °C until use.

5.2.1 Bisulfite conversion.

RNA bisulfite conversion was carried out for the detection of the methylation frequency (5mC) of AcMNPV transcripts. An RNA mixture containing equal amount of RNA from each examined time point and the EZ RNA Methylation Kit (Zymo Research) were used for this experiment. Bisulfite conversion was carried out according to the Kit's manual. In short, RNA was mixed with the RNA Conversion Reagent, and then, they were incubated in a PCR cycler (Veriti, Applied Biosystems). The mixture was cooled down which was followed by an in-column desulfonation using the Zymo-Spin IC Column. First, the RNA Binding Buffer (part of the Kit), then the RNA sample, and finally, 100% ethanol were loaded to the column, then after a brief mixing, the sample was centrifuged at 13,000×g for 30 s. RNA Wash Buffer was added to the column, which was followed by a centrifugation Afterwards, desulfonation was carried out with the addition of the Kit's Desulfonation Buffer to the column. The sample was incubated at room temperature, which was followed by centrifugation The column was washed twice using the RNA Wash Buffer. Finally, the bisulfite-converted RNA was eluted in 15 µL of DNase/RNase-free water. The RNA was stored at − 80 °C until further usage.

5.2.2 Sequencing libraries for PacBio Sequel platform

The cDNAs were produced from the Poly(A) + RNA samples. For this, the SMARTer PCR cDNA Synthesis Kit (Clontech) and the 'Isoform Sequencing (Iso-Seq) protocol without size selection (PacBio) were used. The reverse transcription (RT) reactions were primed by using the oligo(dT) from the SMARTer Kit. The cDNA sample was amplified using KAPA HiFi Enzyme (Kapa Biosystems).

Amplified cDNA sample was used to produce PacBio SMRTbell templates for sequencing on the Sequel platform. using PacBio Template Prep Kit. The concentration of PacBio SMRTbell template was measured using Qubit fluorometer and Qubit dsDNA HS Assay Kit. SMRTbell libraries were annealed to the sequencing primer v3 and bound to Sequel DNA polymerase 2.0 for sequencing using the Sequel Binding Kit 2.0 (PacBio, 100-862-200), and then, the library-polymerase complex was bound to MagBeads using the PacBio's MagBead Binding Kit. The amount of the primer for the annealing and the polymerase for the binding were determined by the PacBio IsoSeq Binding Calculator (Sample Setup Module, PacBio SMRT Link) the MagBead-bound complex was loaded onto the Sequel SMRT Cell 1

M v2. One SMRT Cell was run on the Sequel sequencer. The consensus reads (ROIs) were created using SMRT Link5.0.1.9585.

5.2.3 Poly(A) selected cDNA sequencing libraries for the ONT MinION platform

Amplified cDNA sequencing. Amplified cDNA libraries were prepared from the purified polyA(+) RNAs using ONT Ligation Kit 1D (EBV: SQKL-SK109, AcMNPV: SQK-LSK108). End repair was carried out on Cap-selected and barcoded samples using NEBNext End repair/dA-tailing Module (New England Biolabs). The libraries were barcoded using 1D PCR Barcoding (96) Kit (ONT) following the manufacturer’s instructions. Between each step the samples were purified using Agencourt AMPure XP magnetic beads (Beckman Coulter). The concentration of the libraries was determined by Qubit 4 Fluorometer. Barcoded libraries were loaded on a MinION Flow Cell. For EBV amplified cDNA library was also generated from ribodepleted RNA and custom-made random primers. The consecutive steps were the same as described above.

Non-amplified cDNA sequencing. ONT Direct cDNA Sequencing Kit (SQK-DCS109) was used for the generation of amplification-free libraries. PolyA-selected RNA sample was used for the synthesis of the first cDNA strand using Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific) RNase Cocktail Enzyme Mix (Thermo Fisher Scientific) was used for the removal of RNAs from the single stranded cDNA molecules. The synthesis of the second cDNA strand was performed using LongAmp Taq Master Mix (New England Biolabs). cDNA ends were repaired using NEBNext Ultra II End Repair/dA-Tailing Module. The cDNA ends were repaired using NEBNext Ultra II End Repair/dA-Tailing Module (New England Biolabs). Libraries were barcoded using ONT Native Barcoding Expansion Kit (EXP-NBD104), then the ligation of the sequencing adapter was carried out using NEB Quick T4 DNA Ligase. All conditions were set according to the SQK-DCS109 manufacturer’s protocol.

5.2.4 Cap selection followed by cDNA sequencing on the ONT MinION platform

Cap-selection was carried out with the aim to validate the 5-ends of the transcripts. For this, we used the TelopPrime Full-Length cDNA Amplification Kit of Lexogen. The starting material was a total RNA mixture. The cDNA generation was carried out according to the recommendations from the manual of Lexogen. Detailed protocol can be found in our earlier published data paper (Boldogkői et al., 2018). The cDNA sample was used to produce a sequencing library with the ONT Ligation Sequencing Kit 1D (SQK-LSK108) following the last steps (end-repair and 1D adapter ligation) of ONT's 1D Strand switching cDNA by ligation method. Sequencing was performed on ONT Flow Cells.

5.2.5 Direct RNA library for the ONT MinION platform

The Direct RNA Sequencing (DRS) protocol from ONT was used to generate non-amplified sequencing libraries. For this, we used a mixture of total RNA samples from all time point. The poly(A)+ RNA fraction from this mixture was isolated using the Qiagen's Oligotex protocol, as described above. For the DRS library preparation was mixed with the ONT's oligo(dT)-adapter (ONT Direct RNA Sequencing Kit; SQK-RNA001) and with T4 DNA ligase (New England BioLabs). After incubation the first strand cDNA synthesis carried using the SuperScript III Reverse Transcriptase enzyme (Life Technologies), according to the DRS protocol (Boldogkői et al., 2018). Libraries were sequenced on a MinION Flow Cell. Life Technologies). The cDNAs and the sequencing ready cDNA libraries were washed using AMPure XP beads (Agencourt, Beckman Coulter) after every enzymatic reaction. The samples for dRNA sequencing were handled with RNase OUT-treated (Life Technologies) AMPure XP beads. The library concentrations were measured with Qubit 2.0 and Qubit dsDNA HS Assay Kit

5.2.6 cDNA-PCR sequencing on the ONT MinION platform

A cDNA library from the bisulfite-converted sample was generated for MinION sequencing by using the cDNA-PCR Sequencing Kit (SQK-PCS109), as follows: the primer (VNP) and dNTP (both from the Kit) were mixed with bisulfite-converted RNA. This protocol was followed by the addition of RT buffer, RNaseOUT, nuclease-free water, and strand-switching primer to the sample. Maxima H Minus Reverse Transcriptase (Thermo Scientific) was measured into the RT mix. For the amplification of the first-strand cDNA sample, LongAmp Taq Master Mix (New England Biolabs), cDNA Primer (cPRM, ONT Kit) and Nuclease-free water (Invitrogen) were added. The PCR product was treated with exonuclease (NEB). AMPure XP Beads were used for purification, and the clean samples were eluted. The concentrations of the libraries were detected by Qubit 4.0 and Qubit 1× dsDNA HS Assay Kit, and then, they were loaded to a MinION Flow Cell.

5.3 Read processing and analysis

MinION data were base-called and demultiplexed using Guppy base caller (EBV: v. 3.3.3, AcMNPV: v.2.11) with –qscore_filtering turned on. Reads with a Q-score larger than 7 were mapped to the circularized viral genome (NCBI nucleotide accession: EBV: KC207813.1, AcMNPV: KM667940.1) using the Minimap2 software (Li, 2018). Adapter sequences and poly(A) tails were preserved on reads to determine 5′ and 3′ ends and the orientation of the transcripts. For EBV previously published (Cao et al., 2015; Lin et al., 2013, 2010; O'Grady et al., 2016, 2014; Ungerleider et al., 2018) CAGE-Seq, PA-seq, Illumina and PacBio RSII data were retrieved for TSS, TES, intron and transcript validation. Annotation of the TSSs, TESs, and introns was performed using the LoRTIA software suite (https://github.com/ zsolt-balazs/LoRTIA) with specific settings for each sample type. We used SeqTools, our in-house scripts for the generation of the descriptive quality statistics of reads (Read-Statistics) and for the analysis of promoters (MotifFinder), which are available on GitHub: https://github.com/moldovannorbert/seqtools. In this study, the LoRTIA (v.0.9.9) pipeline developed in our laboratory was used for the identification of transcripts and transcript isoforms. Briefly, sequencing adapters and the homopolymer A sequences were checked by the LoRTIA software for the detection of TSS and TES, respectively. For the elimination of false transcript ends, the putative TSSs and TESs were tested against the Poisson distribution (using Bonferroni correction). Introns were identified by applying the following criteria: they have one of the three most frequent splice consensus sequences (GT/AG, GC/AG, AT/AC), and their frequency exceed 1‰ compared to the local coverage

The raw Illumina reads were trimmed with the Trimgalore. (https://www.bioinformatics.babraham.ac.uk/projects/ trim_ galore/) The above-mentioned EBV reference genome was indexed using STAR aligner v2.7.3a (Dobin et al., 2013) using the following settings: –genomeSAindexNbases 7, followed by the mapping of the reads with default options. STAR software was also used to detect introns from the SRS samples. Bam files obtained from CAGE-seq were converted to BigWig format to detect 5′ end coverage. The CAGEfightR (R/Bioconductor) package (Thodberg et al., 2019) was used to determine TSS positions. The TSS clusters within a 10 nucleotides window were termed identical. Clusters with a "minimum pooled value" (--pooledcutoff) of 0.1 and below were excluded from the further analysis. Then, the cluster positions with a score of 25 or lower were filtered out. The same approach was used for the TES identification, with the following exception: "minimum pooled value" was set to 10.

For transcript isoform annotation, TSSs and TESs were selected for EBV: were accepted as real if presented in either two of our techniques, or for TSSs in one of our techniques and one in either CAGE-Seq or PacBio results, for TESs in one of our techniques and either PA-Seq or the PacBio results. Likewise, putative introns were accepted as real if they were present in two of our techniques, or in one of our techniques and the PacBio results. and for AcMNPV: that were present in at least two samples, while introns were selected if they were present in at least two samples and if their orientation matched the orientation of reads in which they were present, as the LoRTIA software is blind for the orientation of the reads when looking for introns. Transcript isoforms were annotated for each sample using these features and the Transcript Annotator module of LoRTIA.

A read was considered a transcript isoform if it started in the ± 5 nt vicinity of a TSS and if it ended in the ±5 nt vicinity of a TES. Transcripts enclosing the same ORFs as a previously annotated transcript but starting upstream of its TSS were denoted longer (L) 5′-UTR isoforms, while those starting downstream, shorter (S) 5′-UTR isoforms. Transcripts with the same ORFs as a previously annotated transcript but ending upstream or downstream of its TES were denoted transcript isoforms with alternative termination (AT). Transcripts with longer 5′ or 3′-UTRs overlapping multiple ORFs in the same orientation were considered polygenic. If a TSS of a novel transcript isoform was positioned downstream of a previously described ORF's AUG, with an alternative in-frame start codon downstream from the TSS, the isoform was considered putative protein coding transcript, while those without a 5′-truncated ORF were considered 5′-truncated (TR) non-coding transcripts. Both of these transcript species are conterminal with their previously annotated isoforms. If a transcript isoform started in the same TSS as a previously described protein coding transcript, but its TES was located upstream of the stop codon of previously described ORF, the novel transcript was denoted as non-coding (NC). Transcripts in the opposite orientation of an annotated transcript were named non-coding antisense (AS) transcripts. Very long transcripts overlapping multiple ORFs in different orientation were denoted as complex (C) transcripts. Any other transcript configuration not containing a previously annotated ORF was denoted as NC. Very long unique or low-abundance reads which could not be detected using LoRTIA were evaluated and annotated manually. These reads were also accepted as putative transcript isoforms if they were longer than any other overlapping RNA molecule.

In case of EBV We used the conventional terminology for naming the EBV transcriptome (O'Grady et al., 2016). Novel transcript isoforms were named after the most abundant previously annotated transcript of a gene.

5.3.1 Detection of RNA modifications

To detect the modifications in the RNA nucleotides, we base-called the raw fast5 files of our previously published direct RNA sequencing dataset deposited in the European

Nucleotide Archive under sample accession SAMEA10458962. Modification detection was performed by Tombo software suite43 (v.1.3.1.).

5.3.2 Coding potential estimation

In order to estimate coding potential of the transcripts with previously undetected ORFs, we extracted the transcript sequences from the reference genome and used the Coding-Potential Assessment Tool (CPAT) (Wang et al., 2013) with default settings.

6. Results

6.1. Sequencing and mapping statistics

For AcMNPV, the PacBio Sequel and ONT MinION LRS platforms were used in this study to characterize the structure of the transcriptome and epitranscriptome. Sequel sequencing resulted in a total of 47 880 circular consensus sequences (CCS), of which 25 371 were aligned to the viral genome and 23 884 to the insect host genome (Sf9 cells). The total number of reads was less than the sum of the reads mapped to both genomes, as the chimera reads generated during library construction were mapped to both genomes. The Cap-selected samples resulted in a total of 1 830 476 reads, of which 198 516 mapped to the AcMNPV genome and 1 631 960 to the host genome, while the non-Cap-selected samples resulted in 1 119 716 reads, of which 290 039 mapped to the virus and 760 533 to the host genome. Sequencing resulted in longer average mapped read lengths than ONT, whereas Cap-selected and non-Cap-selected ONT reads exhibited similar mapped read lengths. The difference in average read length between the two platforms can be explained by the step used to mitigate the loading bias of the PacBio sequencers during Sequel library construction, which results in the loss of short cDNAs.

The lytic EBV transcriptome was analyzed using our new amplified and unamplified ONT sequencing dataset, as well as transcriptomics data generated by others using PacBio RSII and Illumina platforms. ONT and PacBio data were used to identify full-length RNA molecules, while Illumina CAGE-Seq and Poly(A)-Seq data were used to validate TSSs, TESs and splice sites. The data obtained from this multi-platform approach were integrated to detect novel EBV transcripts and to validate already described RNA molecules. Libraries were generated from eight consecutive lytic time points. Due to low coverage, especially at early time points, kinetic analysis was not feasible from this data set. A total of 22 358 unamplified and 54 271 amplified reads were mapped to the viral genome, with an average length of 838.66 nucleotides (nts) and 1098.43 nts for mapped reads, respectively. The number of reads obtained with other techniques was PacBio: 104 469, Illumina Cage-Seq: 3 344 162 and Illumina polyA-Seq: 93 817 061. We also generated random hexamer primer-loaded amplified libraries from collected samples and sequenced them on the MinION platform. Read statistics for each our library are shown in Table 2 below.

Table 2. **Detailed read statistics of the sequencing libraries.** The following abbreviations were used: dRNA: direct RNA sequencing; cDNA: sequencing of amplified cDNAs; dcDNA: sequencing of non-amplified cDNAs

		Read count		**Read length (nt) (Mean±SD)**		**Mismatch %**	**Insertion %**	**Deletion %**
		Raw	**Mapped**	**Raw**	**Mapped**			
AcMNPV	**PacBio**	**47,880**	**25,371**	**2790±1550**	**1514±742**	**0.2%**	**2%**	**0.4%**
	ONT dRNA	**66,003**	**2,710**	**669±51**	**615±380**	**4%**	**3%**	**8%**
	ONT Cap	**6,862,026**	**198,516**	**747±531**	**555±276**	**4%**	**3%**	**4%**
	ONT non-Cap	**1,119,716**	**290,039**	**844±588**	**563±416**	**4%**	**5%**	**5%**
	ONT bisulfite conversion	**7,077,229**	**125,448**	**296±257**	**268±135**	**14%**	**1%**	**3%**
EBV	**ONT cDNA**	**7,826,210**	**54271**	**1359±1130**	**1144±791**	**5%**	**4%**	**6%**
	ONT dcDNA	**3,042,936**	**22358**	**1188±872**	**874±732**	**3%**	**6%**	**4%**

6.2 Transcriptional start, end sites and introns

For AcMNPV, the criterion for accepting TSSs and TESs as true transcript ends is if they were detected in two amplified ONT samples and in another technique that was either sequel or Cap-sequencing. A more stringent criterion was applied for non-coding and 5'-truncated transcripts: confirmation was required in two amplified ONT samples, along with cap-selection. After screening, a total of 311 TSSs and 261 TESs were obtained, as well as 13 splice junctions. TATA boxes were identified for 60 TSSs. The average distance of the TATA boxes from the TSS was 32 nts. Twenty-two GC boxes were identified and their average distance from TSSs was 66 nts. The average distance of the 15 CAAT boxes identified from the TSSs was 108 nts. Canonical CAGT initiators were present in only 6% of TSSs, while TAAG initiators were found in 61% of cases, and non-TAAG initiators were observed in 33% of cases (Figure 3a).

Canonical PAS were found in approximately 80% of TESs at an average distance of 27.23 nts upstream of the TES. Consistent with previous results describing polyadenylation signals in arthropods (Calvo et al., 2009), the environment of viral TESs ± 50 nts was characterized by A/U-rich sequences with increased adenine content immediately upstream of the cleavage site. Interestingly, sequences containing PAS showed a slight increase in adenine between - 26 and - 12 nts upstream of the TES, whereas this phenomenon was not observed in sequences without PAS (Figure 3b).

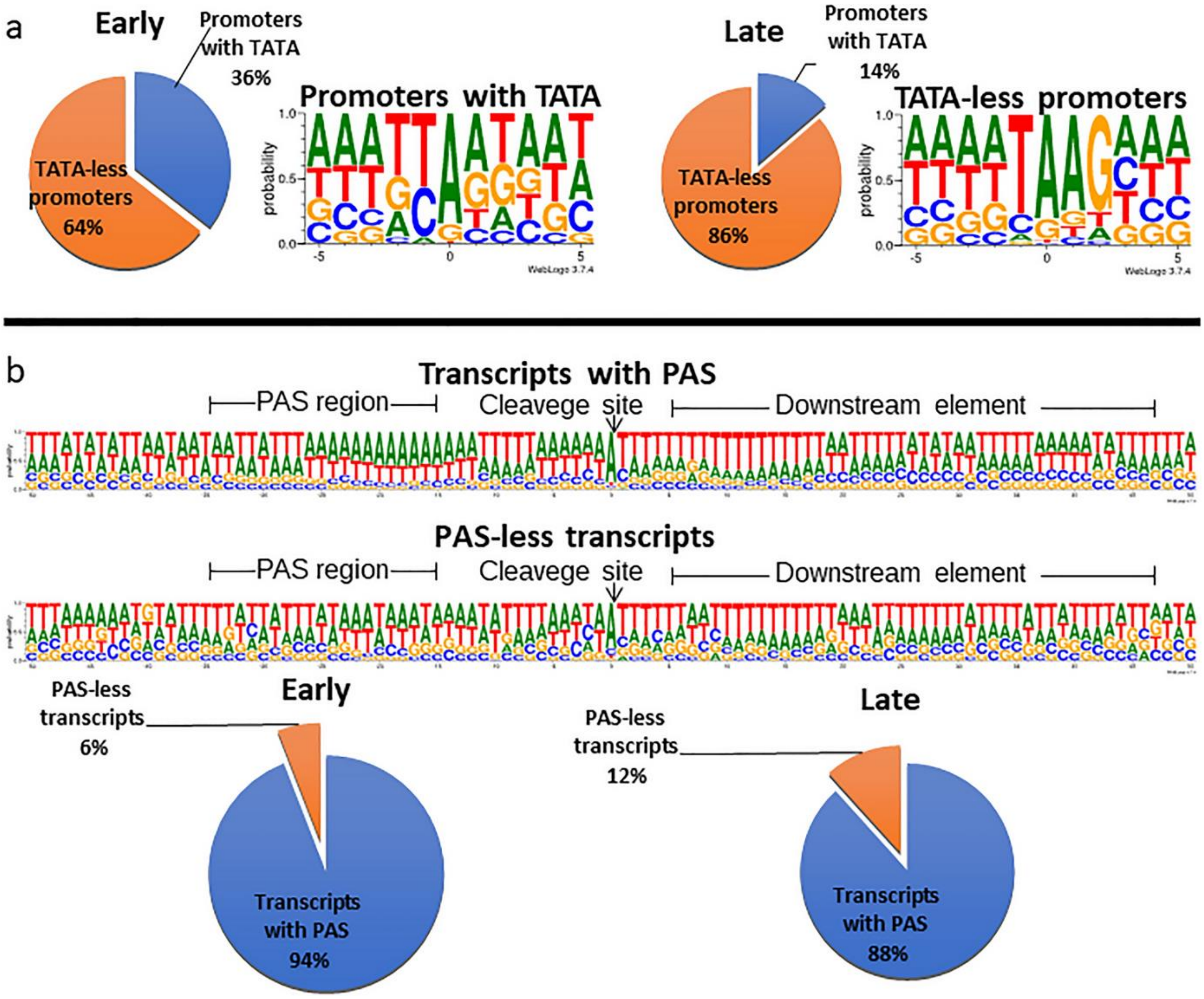


Figure 3. Utilization of TATA box and polyadenylation signal (PAS) in early and late viral transcripts. (a) The pie charts show the percentage of TATA (blue) and non-TATA (orange) promoters for early and late time points. The pie chart was created with Microsoft Excel 2021 software. (b) The weblogo shows the probability of the occurrence of TSSs and nucleotides in their genomic environment. The weblogo shows the poly(A) signal in the downstream element and the probability of occurrence of nucleotides in their vicinity. Pie charts for early and late time points show the percentages of transcripts with poly(A) signal (blue) and without Poly(A) signal (orange). The weblogo image is generated using weblogo 3.0. The pie chart was created with Microsoft Excel 2021 software.

A total of 398 putative TSS were detected in EBV. CAGE-Seq, ONT-MinION and PacBio datasets were used to validate the TSSs. These were accepted if they were present in at least two of our techniques or in one of our techniques and either the CAGE-Seq or PacBio dataset. This rigorous filtering resulted in a total of 322 TSSs, of which 145 were novel (Figure 4a). We identified all TSSs that were also detected by CAGE-Seq; however, 4.66% of the TSSs (14 out of 322) were not detected by CAGE-Seq. We identified upstream TATA boxes in 20% of the TSSs, with an average distance of -31.43 nts. Nucleotide composition analysis of these start sites revealed a G-rich initiator region (Figure 4c). Sixty-two GC boxes were identified with an average distance of 64.70 nt from the TSSs. The 17 CAAT boxes identified had an average distance of 110.23 nts from the TSSs. Both GC and CAAT boxes are promoter consensus elements that bind specific transcription factors (SP1 and NF-1, respectively). The consensus sequence of the GC box is GGGCGG, and it is located 100 nucleotides upstream of the TSSs The consensus sequence for the CAAT box is CAATCT, located approximately 75 nucleotides upstream of the TSSs.

A total of 65 putative transcription end sites (TES) were detected using the LoRTIA software package. A TES was accepted if it was present in at least two of our techniques, either in one of our techniques and the PacBio dataset or in the PA-seq dataset. The analysis resulted in the detection of 57 TESs, with 12 being novel (Figure 4b) We identified PAS in 89% of the TESs, with an average distance of -24.51 nts. TESs with PAS showed an A-rich cleavage site and a G/T-rich downstream region. These sequences are similar to the mammalian cleavage and polyadenylation motifs (Tian and Graber, 2012) (Figure 4d). TESs without PAS showed an ACCTC sequence near the cleavage site and a TTATT sequence between positions + 11 and + 15 (Figure 4e). The latter is a variant of the termination signal of RNAPIII (Bogenhagen and Brown, 1981), a gene transcribed by a polymerase specialized in the transcription of rRNAs and tRNAs. However, it has also been described as a terminator of the avian adenovirus CELO VA RNA gene (Liu et al., 2012), suggesting that RNAPIII may perform random transcription of protein-coding genes. To annotate transcript isoforms, we used our validated TSSs, TESs and introns. Using the LoRTIA toolkit, we detected 205 introns and applied the criterion that a putative intron must be present in at least 2 of our techniques or in one of our techniques and either the Illumina or PacBio dataset. The canonical GT/AG splice junction consensus was present in all identified introns.

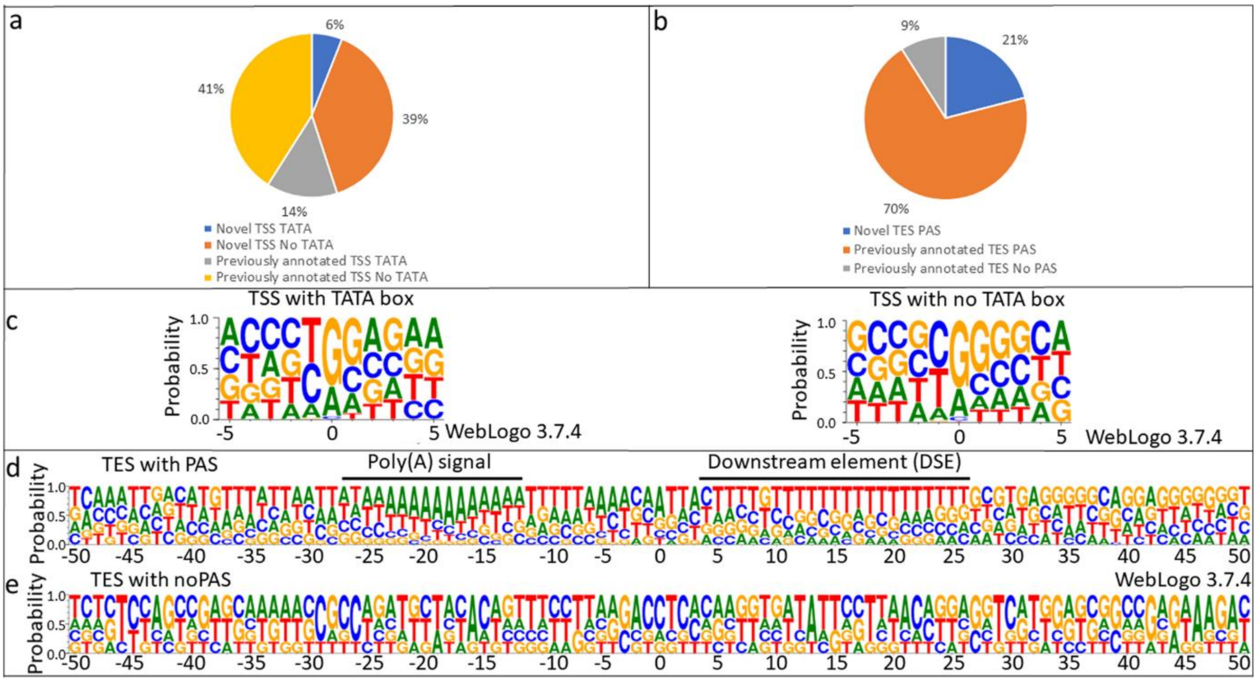


Figure 4. Sequence motifs and proportions of TSSs and TESs. Distribution of TATA and non-TATA promoters (a). Distribution of TESs without a poly(A) signal (PAS) and with a PAS signal (b). Sequences surrounding the (c) TSS contain a G-rich initiator region for TSSs with a TATA box. TSSs lacking a TATA box have G-rich + 1 and + 2 positions. TESs with a PAS have a canonical A-rich cleavage site and a canonical GU-rich downstream element (DSE) (d), while those without a PAS have a C-rich cleavage with no recognizable DSE (e)

6.3. Annotating the viral transcriptome

The strength of the LRS lies in its ability to detect full-length transcripts, leading to the discovery of complex transcriptional landscapes for many viruses (Balázs et al., 2017; Depledge et al., 2019; Moldován et al., 2017, 2018a; Tombácz et al., 2017, 2019).

A total of 875 transcript types were detected in AcMNPV, and the identified transcripts are presented in Table 3 and Figure 5.

Table 3. The number of previously annotated and novel transcripts of AcMNPV.

Transcript types	Number
Previously annotated transcripts	116
Novel monocistronic transcripts	14
5′-UTR isoforms	164
3′-UTR isoforms	174
5′-UTR isoforms with alternative termination	166
Polygenic transcripts	45
Complex transcripts	54
Non-coding transcripts	78
… of which antisense transcripts	23
Novel putative protein-coding	41

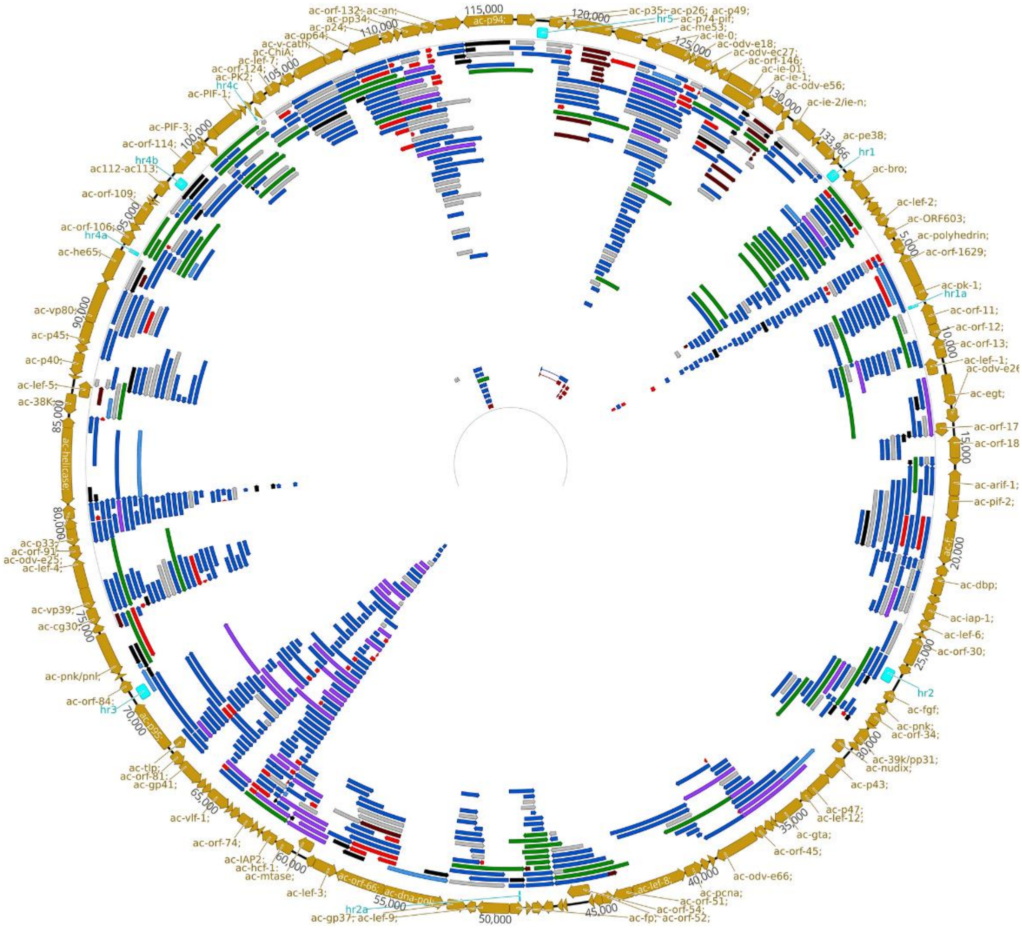


Figure 5. AcMNPV transcripts and isoforms transcribed along the circular genome. Color code. brown arrows: ORFs; aqua rectangles: replication origins; grey: formerly annotated transcripts; light blue: novel monocistronic transcripts; purple: novel polygenic transcripts; green: complex transcripts; red: non-coding transcripts; black: 5′-truncated transcripts; dark blue: TSS and TES isoforms. The figure was created with the Geneious 2021.2.2 software (https://www.geneious.com)

In EBV after the filtering we annotated in a total of 351 polyadenylated transcripts (Table 4). We compared the transcripts reported by O'Grady et al (O’Grady et al., 2016) with those annotated in our work. We found that 108 transcripts were identified in both studies, with O'Grady et al detecting 185 transcripts and our study identifying 241 transcripts. while 185 transcripts were detected by only O’Grady and colleagues and 241 transcripts in only our study. The discrepancy between the two studies is explained by two reasons: the average read-length were higher in the PacBio (1176 nts) than in the ONT (1009 nts) data, and that we applied very strict criteria for accepting reads as true transcripts. Some transcripts are represented by only a single read, which is below the threshold of detection of LoRTIA. Because of their low abundance their TSS is uncertain, hence we denote these as putative transcript isoforms. Watanabe and colleagues (Watanabe et al., 2015) analyzed the impact of two uORFs upstream of BGLF3.5 ORF on the translation of BGLF4, a protein kinase involved in replication and nuclear regress (Gresburg et al., 2007), and found that point mutations in the two disruption of uORFs (duORFs) had no effect on protein levels of BGLF4. The most abundant transcript in the BGLF3.5-BGLF4 cluster is BGLT16, a bicistronic mRNA consisting of a wild-type uAUG upstream (Figure 6).

Table 4. The number of previously annotated and novel EBV transcripts

Transcript types	Previously annotated	Novel
Monocistronic transcripts	7	15
5′-UTR isoforms	23	104
Multigenic transcripts	14	47
Isoforms with alternative termination	1	7
Complex transcripts	2	6
Splice isoform	31	42
Non-splice isoforms	2	5
Putative protein-coding transcripts	25	47
Non-coding transcripts …	7	21
… of which antisense transcripts	1	1

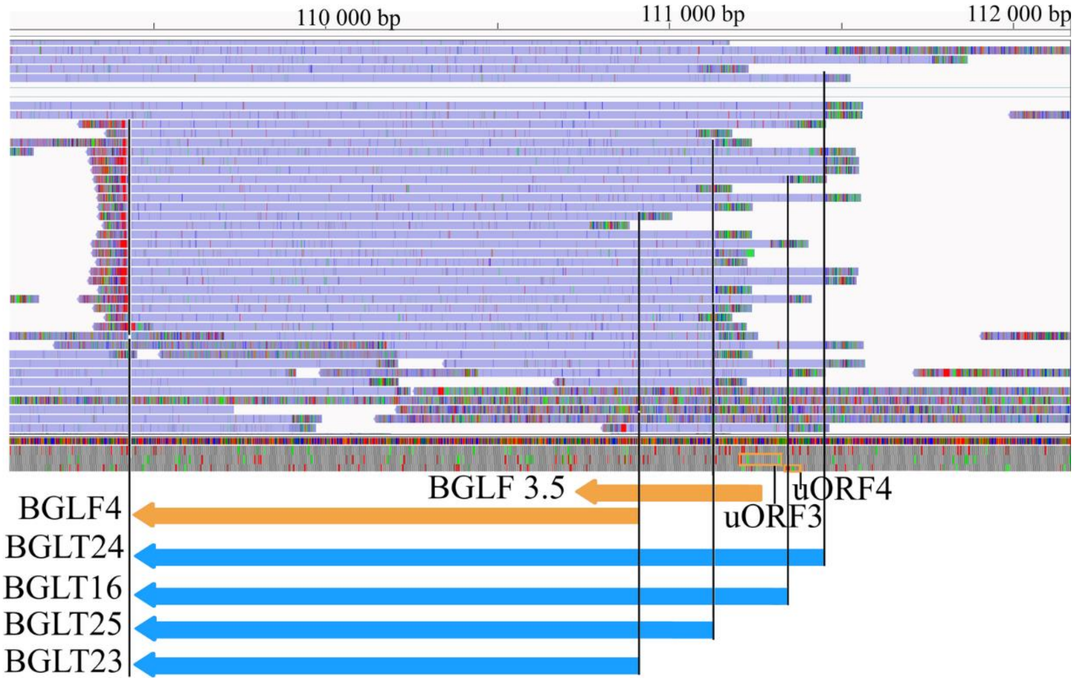


Figure 6. Transcripts overlapping BGLF4 and BGLF3.5 ORFs and their uORFs. Yellow arrows indicate the BGLF4 and BGLF3.5 genes, while the blue arrows show the transcripts encoded by these genes (BGLT24, BGLT16, BGLT25, BGLT23). Two upstream ORFs (uORF3 and uORF4) are also indicated. Purple bars indicate the reads obtained by the ONT MinION sequencing. Rectangular lines show the transcript ends.

6.3.1 Monocistronic transcripts

Due to challenges faced by SRS in assembling genomic regions with complex transcriptional overlap patterns, precise transcript annotation is lacking for many AcMNPV genes. Through LRS, we annotated 14 novel monocistronic transcript species with single-nucleotide accuracy. Canonical TATA boxes were observed upstream of the TSSs of ORF85 and ORF112-113. These transcripts originated from a non-TAAG-initiator element (Inr) and contained canonical PAS upstream of their TES. Transcripts encoding DNA polymerase were initiated on a canonical arthropod initiator (GCATA), while the helicase was initiated on a similar but non-canonical sequence (GCAATA). Both DNAPOL and HEL contained a canonical PAS upstream of their TES. Nine of the transcripts (ORF1629, P47, ORF72, ORF84, 38K, ORF108, PP34, P49 and ORF154) were initiated on TAAG-Inr, PP34 was initiated upstream of TSS 87 nts with a canonical CAAT sequence (CCA ATC), and five of these transcripts had a PAS.

In EBV, we report the detection of 22 monocistronic transcripts, with 15 being novel monocistronic. We identified non-spliced BNRT10, BHLF1, BORF2 and BGLT18 transcripts, of which only spliced variants were previously detected (O’Grady et al., 2016; Concha et al., 2012). We have also discovered 10 new monocistronic transcripts with full-length ORFs, of which only shorter isoforms with incomplete ORFs (lacking in-frame AUG) have been previously described (O’Grady et al., 2016). No transcripts have been recorded in the genomic region of BFRF3, despite studies showing that this region is transcriptionally active (Batisse et al., 2005; Han et al., 2009). We identified BFRT3 with a novel transcript that completely overlaps the ORF of BFRF3 and ends in a new terminus.

6.3.2 TSS and TES isoforms

In this part of the study, 330 transcript isoforms were found to have longer or shorter 5′-UTRs than previously annotated transcripts encoded by the same genes in AcMNPV. Less than half of them (32.06%) showed an E/L initiation region shift compared to the Inr of their previously annotated transcript. Among the TSS transcript isoforms, 7.06% were found to be TAAG-Inr driven, and they were previously annotated as non-TAAG-Inr driven but encoded by the same gene. However, non-TAAG-Inr isoforms were identified in 25% of the transcripts previously annotated as TAAG-Inr. This phenomenon suggests that many AcMNPV genes are transcribed by both the host and the viral RNP, resulting in altered 5′-UTR lengths. In addition to the genes described in our previous work (Chen et al., 2013), we detected 57 genes that are transcribed by both host and viral RNP. The polymorphism in 5′-UTR length is likely to have biological significance, although we cannot exclude the possibility that it is merely transcriptional noise. In many cases, longer 5′-UTRs harbor upstream ORFs (uORFs), which have been shown to alter protein coding sequence translation through processes such as ribosome rearrangement, ribosome arrest or dissociation, and ribosome bypass (Calvo et al., 2009; Kronstad et al., 2013). We identified 75 gene products that contain at least one uORF.

In total, 340 novel TES isoforms were identified in this work, 76.35% of which contained canonical PAS upstream of their 5′-end. The phenomenon of non-templated adenine addition by viral RNP has been described previously (Jin and Guarino, 2000). This in vitro study also suggested the presence of a T-rich termination signal for this enzyme and non-templated thymine addition prior to adenine incorporation. Consistent with this work, we found that 51.85% of the 3′-UTR isoforms with LIS terminated near ± 3 nts from the TES) the T-rich region. This finding is in contrast to the 22.51% of 3′-UTR isoforms with non-TAAG-Inr. However, we could not confirm the presence of non-templated thymines upstream of the poly(A) tail. The average 5′-UTR length was 153.06 nts (σ = 270.438) and the average 3′-UTR length was 529.09 nts (σ = 729.266), both measured from the first ORF overlapped by the transcript. The difference was significant for transcript and 3′-UTR length, suggesting that the viral RNP tends to produce longer RNA molecules.

A previous LRS study showed a high diversity of TSSs and TESs for several EBV genes (O’Grady et al., 2016). Here, we identified 104 novel 5′-UTR isoforms, with 47 having longer and 57 having shorter 5′-UTRs. CAGE-Seq data analysis validated 98% of the longer TSSs and 92.98% of the shorter TSSs among the investigated isoforms. 5′-UTRs can regulate translation through their secondary structures (Leppek et al., 2018) and upstream AUGs (uAUGs) or upstream ORFs (uORFs) (Calvo et al., 2009; Kronstad et al., 2013). Watanabe and co-workers (Watanabe et al., 2015) analyzed the effect of two upstream uORFs (duORFs) in the BGLF3.5 ORF on the translation of BGLF4, a protein kinase involved in replication and nuclear regression (Gershburg et al., 2007), and found that point mutations in two interruptions of uORFs (duORFs) had no effect on the protein level of BGLF4.

The most abundant transcript of the BGLF3.5-BGLF4 cluster is BGLT16, a bicistronic mRNA that consists of wild-type uAUGs upstream (Figure 6). We detected two short 5′-UTR isoforms of this transcript (BGLT23 and BGLT25), which exclusively carry the BGLF4 gene. This RNA molecule lacks the duORFs mutated by Watanabe and colleagues (Watanabe et al., 2015). Furthermore, the longer 5′-UTR isoform of BGLT16, BGLT24, contained additional wild-type uAUGs and uORFs before the point mutations introduced by Watanabe and colleagues (Figure 6).

Our transcriptomic analysis revealed 7 isoforms with alternative polyadenylation sites, 4 of which are novel. Interestingly, all of them are located in the same 5 kb region, BZLT42, BZLT43, BZLT44 and BZLT50 are 3′-UTR isoforms of BZLF2, while BELT6, BELT8 and BELT9 are isoforms of the BELT1 transcript. As a consequence, a convergent overlap of 10 nt in length is formed between BZLT44, BELT8, BELT9 and BERT3 (Figure 7).

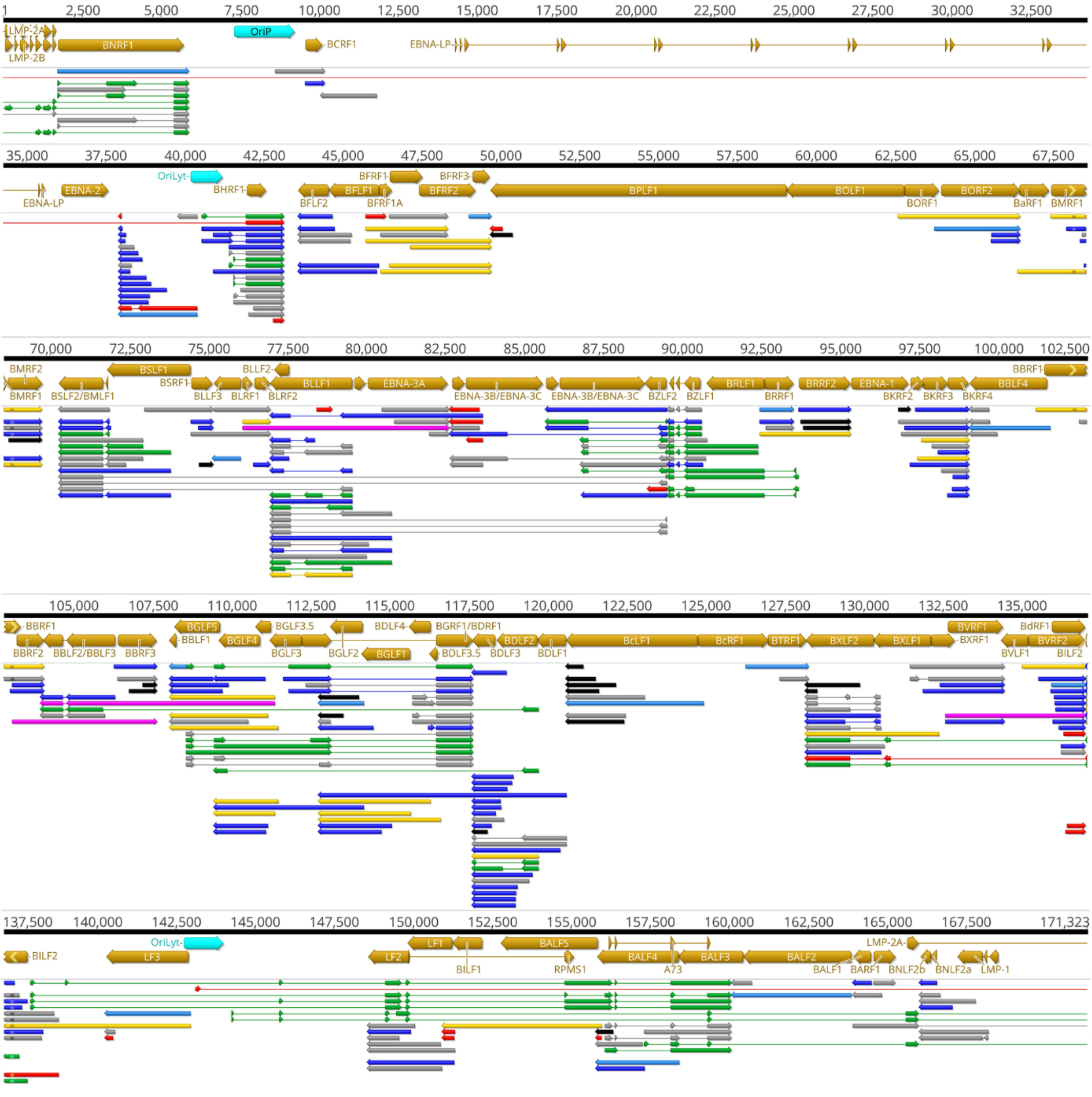


Figure 7. The lytic transcriptome of EBV. Previously detected (O'Grady et al., 2016) and novel transcript isoforms are shown on the genome of strain Akata (NCBI accession: KC207813.1). The EBV transcriptome color code: brown arrows: ORFs; aqua rectangles: replication origins; grey: formerly annotated transcripts; light blue: novel monocistronic transcripts; purple: novel polycistronic transcripts; pink: complex transcripts; red: non-coding

transcripts; black: 5′-truncated transcripts; dark blue: TSS and TES isoforms. Transcripts generated by the junction region of the circular EBV genome are indicated by asterisk at the genome ends.

6.3.3 Splice isoforms

Chen et al (Chen et al., 2013) previously reported twelve introns in AcMNPV with a frequency of more than 1%. We detected five additional introns. Twelve of the introns detected in this study contained the canonical GT/AG splice junction, while one contained the less common GC/AG. Chen et al. associated a spliced antisense transcript with ORF115. We detected 2 RNAs with similar positions (ORF117-L-SP-1 and ORF117-L-SP-2), and the introns of ORF117-L-SP-1 were identical to the previously annotated transcript. ORF117-L-SP-2 had the same acceptor site position, but its donor site was located 85 nts downstream of the previously annotated genomic site. We were unable to annotate the TSS of these transcripts accurately, but our data indicate that it was located upstream of the TSS of ORF117-L-1. The splicing of ORF117-L-SP-2 led to frame shifting within the previously annotated ORF and the creation of a new 246 nt long ORF upstream of the original ATG.

In EBV using the LoRTIA toolkit, 205 introns were detected. Reverse transcription and PCR can create gaps in cDNAs through template switching (TS) events, leading to potential errors in intron annotation. The LoRTIA software package can eliminate these artifacts by detecting the absence of splice junction consensus or the presence of repeat regions favorable for TS. The identification of putative introns required their presence in at least 2 of our techniques or in one of our techniques and either the Illumina or PacBio dataset. The canonical GT/AG splice junction consensus was present in all identified introns.

6.3.4 Polycistronic and complex transcripts

In the transcriptome of the AcMNPV virus, several long RNA molecules (more than 760 nts) were detected, consisting of polycistronic and complex transcripts. Polycistronic transcriptome species were defined as those containing only tandem ORFs, while complex transcripts were defined as multigenic transcripts that contained at least one ORF in an orientation opposite to the others. In this study, we identified 241 polycistronic transcript species with at least two ORFs. The main initiator motif for these long transcripts was LIS, as 81.74% of them started with a TAAG sequence. Additionally, we found 79 complex transcript species, of which 21 were transcript isoforms, and the rest were mapped to unique genomic locations. The longest complex transcript, P10-74-ME53-C-1, had only one sense and two anti-parallel ORFs, while ORF51-52-53-LEF-10-ORF54-55-56-C-1 exhibited the highest number of ORFs (6 sense and 1 anti-parallel ORF).

Multigenic transcripts are present in abundance in all the major DNA viruses studied, including herpesviruses (Balázs et al., 2017; Moldován et al., 2017, 2018a; Prazsák et al., 2018; Tombácz et al., 2016, 2019). Previous studies have detected multigenic mRNAs in EBV using both SRS (Majerciak et al., 2018) and LRS (O'Grady et al., 2016) techniques. In this study, we identified forty-seven multigenic transcripts, 27 of which were novel. Our findings indicate that essentially all lytic gene clusters with ORFs of the same orientation overlap with at least one multigenic transcript isoform. Additionally, we discovered 4 novel complex transcripts out of a total of 6 complex transcripts (BLRT8, BBRT18, BGLT29, and BVRT9) containing genes of opposite polarity. An overview of the transcripts discovered in this study is illustrated in Figure 7.

6.3.5 Putative mRNAs

In AcMNPV, we identified 41 putative novel genes that generated 5′-truncated versions of canonical mRNAs, containing shorter in-frame ORFs. Among these, 19 transcripts were initiated with a TAAG sequence. Nine of them had previously annotated isoforms (EGT, DNAPOL, HCF1, PNK/PNL, HEL, HE65, 94K, IE1, and IE2) that were not initiated with TAAG-Inr. This observation suggests that early genes were partially transcribed by the viral RNP at late time points. Interestingly, among the previously annotated transcripts, eleven (AC-BRO, POLH, ORF19, PP31, ORF66, ORF84, ODV-E25, BV/ODV-C42, ORF117, CHIT, ODV-EC27) that started from a TAAG sequence had 5′-truncated isoforms that did not start on TAAG-Inrs. This implies that the 5′-truncated isoforms of some late genes are transcribed by the cellular RNP. All these transcripts were present in the Cap-selected samples.

We detected several transcripts in EBV with truncated 5'-ends, where the TES was identical to the host mRNA. These short RNA molecules lack a canonical ORF but contain downstream in-frame AUGs and thus may encode N-terminally truncated proteins (Boldogkői et al., 2019b). We report 72 such RNA molecules, nineteen of which are novel. Seventy-two TSSs of these transcripts were confirmed by the CAGE-Seq dataset. In addition to alternative transcription initiation within the gene, alternative splicing can also produce transcripts with altered coding potential when splicing occurs within the ORF. In this analysis, 42 novel splicing isoforms and 5 previously annotated spliced transcripts were detected as non-spliced variants. Nineteen transcripts contained introns within the ORFs. Of these, we identified 9 frame-shifting, 2 nonsense termination (by intron retention leading to an early stop codon), 4 ORFs containing deleted amino acids (in-frame deletion) and 4 intergenic terminations. These latter transcripts contain the regular AUG and a new stop codon at the intergenic position. The coding potential of these transcripts was assessed using the Coding-Potential Assessment Tool (CPAT) with default settings (Wang et al., 2013). CPAT uses four parameters to estimate the coding potential. The CPAT analysis revealed that all but 9 of the 5 truncated isoforms and the ORFs of the 4 splice isoforms with intergenic termination have coding potential. Therefore, we consider these latter transcripts as non-coding RNA (ncRNA). To investigate the homology of proteins encoded by alternative spliced transcripts, the translation of the modified ORFs was queried in the NCBI non-redundant protein database using protein BLAST. BNRT11 and BNRT12 have the first 21 amino acids of the BNRF1 ORF, but end with a stop codon immediately after the first splice acceptor position. For BHLF2, splicing results in frame shifting. The first 77.38% of the ORF is identical to the ORF of BHLF1, while the amino acids following the splice acceptor position show no similarity with other proteins in the database. The splice acceptor position of BSLT12, BSLT18 and BSLT21 is different from the splice acceptor of the major isoform (BSLT13). This results in altered amino acids downstream of the acceptor position, which are not identical to other proteins in the database. BZLT46 and BZLT48 encode the first 75 amino acids of the BZLF2 ORF, while the following amino acids and the stop codon are spliced from the transcript. Thus, the modified ORF continues and terminates in the second exon of these transcripts, and the amino acids following the acceptor position show no similarity with the proteins in the database. The second splice donor position of BZLT39 and BZLT40 is different from that found in the major isoform (BZLF1), resulting in frame shifting, with 7 amino acids and a stop codon following the corresponding splice acceptor. In BZLT51, the second exon of BZLF1 has been deleted, it is 35 amino acids short, but the deletion does not cause frame shifting. In BART17, the splice isoform of the BART transcript retains the first and third introns. The first intron contains an in-frame stop codon. The resulting modified protein shows partial homology with the first exon of the a73 ORF. The putative proteins BZLT47 and BZLT49 and BRLT8 do not show homology with any other NCBI entry.

6.3.6 Novel non-coding transcripts

In AcMNPV, 101 novel transcript isoforms were identified that did not contain previously annotated ORFs, two-thirds of which were longer than 200 nts, representing long non-coding RNAs (lncRNAs), while one-third fell within the size range of short non-coding RNAs (sncRNAs). We identified 41 sense ncRNAs, all of which overlapped with a canonical transcript but lacked stop codons or stop codons and some of the 5′-UTR. Only 10.3% of the ncRNAs had a canonical TATA promoter before their TSS, while 70.5% were initiated with a

TAAG initiator, which may indicate their late transcription. Twenty-three antisense RNAs (asRNAs) were found to be driven by their own promoter. These asRNAs were encoded by complementary DNA strands of 11 genes. ORF-60-AS-1 was the only asRNA whose promoter contained a TATA box, while 86% contained a TAAG initiator sequence. All ncRNAs were also present in the Cap-selected samples.

In case of EBV transcripts without an ORF longer than 10 amino acids were classified as non-coding. In this part of the study, 2 short ncRNAs and 19 lncRNAs were detected. Among the lncRNAs, 14 are 5′-truncated, while three (BFRT14, BLRT9 and BZLT45) are 3′-truncated isoforms of previously annotated RNAs. BLRT9, an lncRNA, starts at the same position as BLRT5 but ends 490 nts downstream, a result confirmed by both our analysis and Illumina PA-Seq. BLRT9 overlaps the BZTL and BELT regions in antisense orientation.

6.3.7 Transcriptional overlaps

The AcMNPV genome contains 37 convergent gene pairs. Our LRS analysis revealed that all convergent gene pairs exhibited transcriptional readthroughs. Among these, only three pairs exclusively overlapped in their 3'-UTRs, while the remaining pairs exhibited overlaps in their ORFs. Out of the 34 gene pairs, 32 showed divergent transcriptional overlaps, and 84 demonstrated parallel overlaps in 87 gene pairs. It is assumed that with higher data coverage, overlaps would be detected in all transcripts.

All three forms of transcriptional overlap were detected in EBV RNAs. These can be formed between transcripts of adjacent genes, such as BDRF1 and BILF2, or between long multigene and monocistronic transcripts, such as BBRT18, a bicistronic transcript that overlaps with isoforms of BBTR16 and BBRT14, both in the same orientation, and with isoforms of BBRT18 and BGLT29 in opposite orientations. Several long splicing transcripts also span multiple genes. BDLT30, for example, starts upstream of BDLF2 and spans the transcripts of 12 genes in the same orientation as BDLF2 and the transcripts of 3 genes in reverse orientation. Although transcriptional overlaps are a common phenomenon in EBV, the intergenic regions between convergent BHRF1 and BHLF1 showed very low levels of overlap. The intergenic region of BMRF2 and BSLF2/BMLF1 was found to be devoid of transcriptional activity. However, a higher overall transcript coverage with low levels of activity was detected in this region.

6.3.8 Ori-associated transcripts

Homologous repeat (hr) regions are located at multiple genomic positions in AcMNPV, believed to contain replication origins (Oris). Our LRS approach detected overlapping transcriptional activity at all nine hr sequences. However, LoRTIA did not identify transcripts for hr5. Nevertheless, we were able to detect reads without both exact TSSs and TESs. In total, 55 transcript species were identified in hr regions, with 50 containing TAAG initiator sequences. Of these RNAs, 15 were polygenic, 32 were TSS variants, 3 were TES isoforms, and 8 were monocistronic transcripts. The majority of overlapping transcripts (12) were transcribed at the genomic junction (hr1), including 7 complex transcripts, 4 TSS isoforms, and 1 monocistronic RNA.

Eukaryotic replication origins are generally associated with coding and non-coding transcripts (Hangauer et al., 2013; Sequeira-Mendes et al., 2009), with overlapping transcripts previously detected in alpha- (Boldogkői et al., 2019a; Tombácz et al., 2016), beta- (Gatherer et al., 2011; Tai-Schmiedel et al., 2018) and gammaherpesviruses (Wang et al., 2006). The EBV genome has two lytic (Ori-Lyt) and one latent (OriP) replication origins. The left OriLyt has been shown to overlap with splice isoforms of BWRT and BCRT, while BHLT2 is initiated within this Ori (O'Grady et al., 2016). The genomic region containing OriP also shows transcriptional activity: several different ncRNA TSSs are located within Ori (Cao et al., 2015) and a long 5'-UTR transcript isoform of the BCRF1 gene is BCRT3. We detected nine novel isoforms of Ori-associated RNAs, all of which were initiated within one of the lytic replication origins. BHLF1 and 2 transcripts are encoded by the bhlf-1 gene. The BHRT15, 16, 17, 21, and 22 transcripts are splice and 5'-UTR isoforms encoded by the brf1 gene (Figure 8a). The LF3 transcript starts in the right OriLyt region. We annotated LF3 and found four novel spliced transcripts (RPMS2, RPMS3, RPMS4 and RPMS5) that completely overlap the Ori region, and BILT44 and BIRT21 transcripts whose 5'-UTR regions overlap the replication origin (Figure 8b).

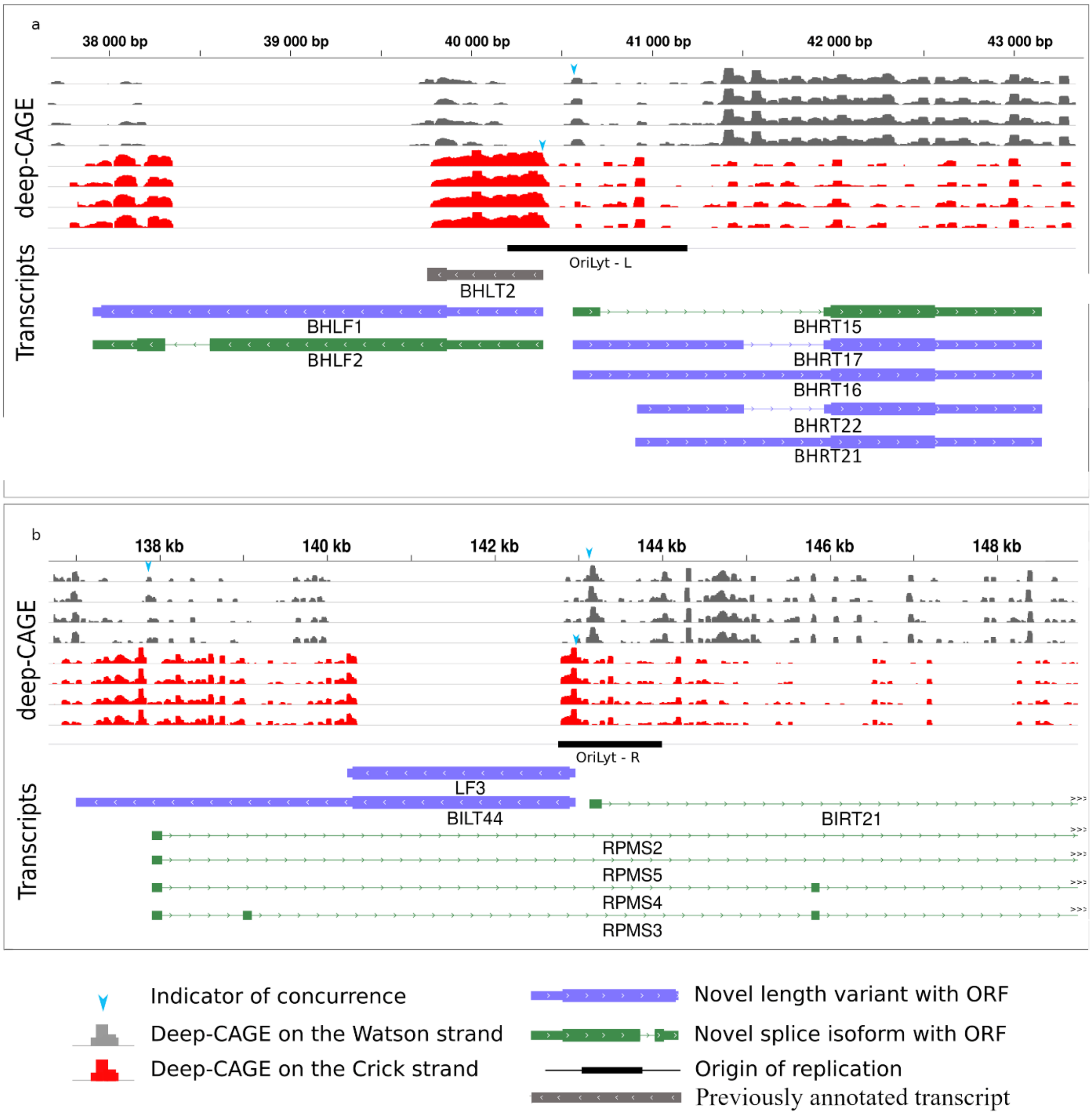


Figure 8. Novel Ori-overlapping RNAs. **a** Three novel length and two novel splice transcript isoforms are initiated in the left lytic Ori region. The CAGE-Seq data from O 'Grady et al. (16) is represented by the histograms with gray reads mapping to the positive, while red are reads mapping to the negative strand. Light blue arrow heads show matching between the CAGE-Seq data and our TSS data. **b** Two novel length and five novel splice isoforms initiated or overlapping the right lytic Ori region

6.4. RNA modifications

We used dRNA-Seq and bisulfite conversion data to detect methylated nucleotides in AcMNPV transcripts.

Tombo analysis. Tombo is a software package used to identify modified nucleotides from nanopore sequencing data. To reduce false positives in the dRNA-Seq sample, we filtered out transcripts with coverage less than 30 and modified fraction less than 30%. No significant correlation was found between coverage and the number of methylated nucleotides in the raw fraction (Figure 9a). Using the Tombo software, we identified a possible methylation consensus sequence (UUAC*CG) (the modified letter C is marked with an asterisk), which indicated a corresponding distribution of log-likelihood ratios (Figure 9b). Our bisulfite conversion experiment confirmed the methylation of this consensus sequence. A deviation from the canonical C-sites was also clearly detected (Figure 9c). After identification of potential false positive sites, 325 putative 5-mC methylation positions were obtained in 12 viral genes (ac-39k, ac-bro, ac-ctl, ac-odve25, ac-orf-58, ac-orf-73, ac-orf-74, ac-orf-75, ac-p40, ac-p6. 9, ac-polyhedryn and ac-vp39). Deviations from canonical C-sites were also clearly detectable (Figure 9c).

Bisulfite conversion analysis. In addition to the Tombo analysis of dRNA-Seq data, bisulfite conversion experiments were performed. While a low read count (2710 viral reads) was obtained in dRNA-Seq, a much higher read count (125 448 reads) was generated by bisulfite sequencing. Furthermore, positive controls (unconverted samples) were used in the latter method. With bisulfite sequencing, 234 of the 325 methylated positions (identified by Tombo analysis) were confirmed. To reduce false positives, a coverage of 25 was set as the threshold for bisulphate analysis. In total, 7897 putative methylation positions were identified in 99 gene transcripts. 31 potential cytosine positions were detected in the 3′-UTR of the ac-Orf-12 transcript, all of which were untranslated and therefore methylated. Overall, 88% of the potential methylation positions (positions with a coverage of at least 25) tested were located in coding regions and 21% in UTRs.

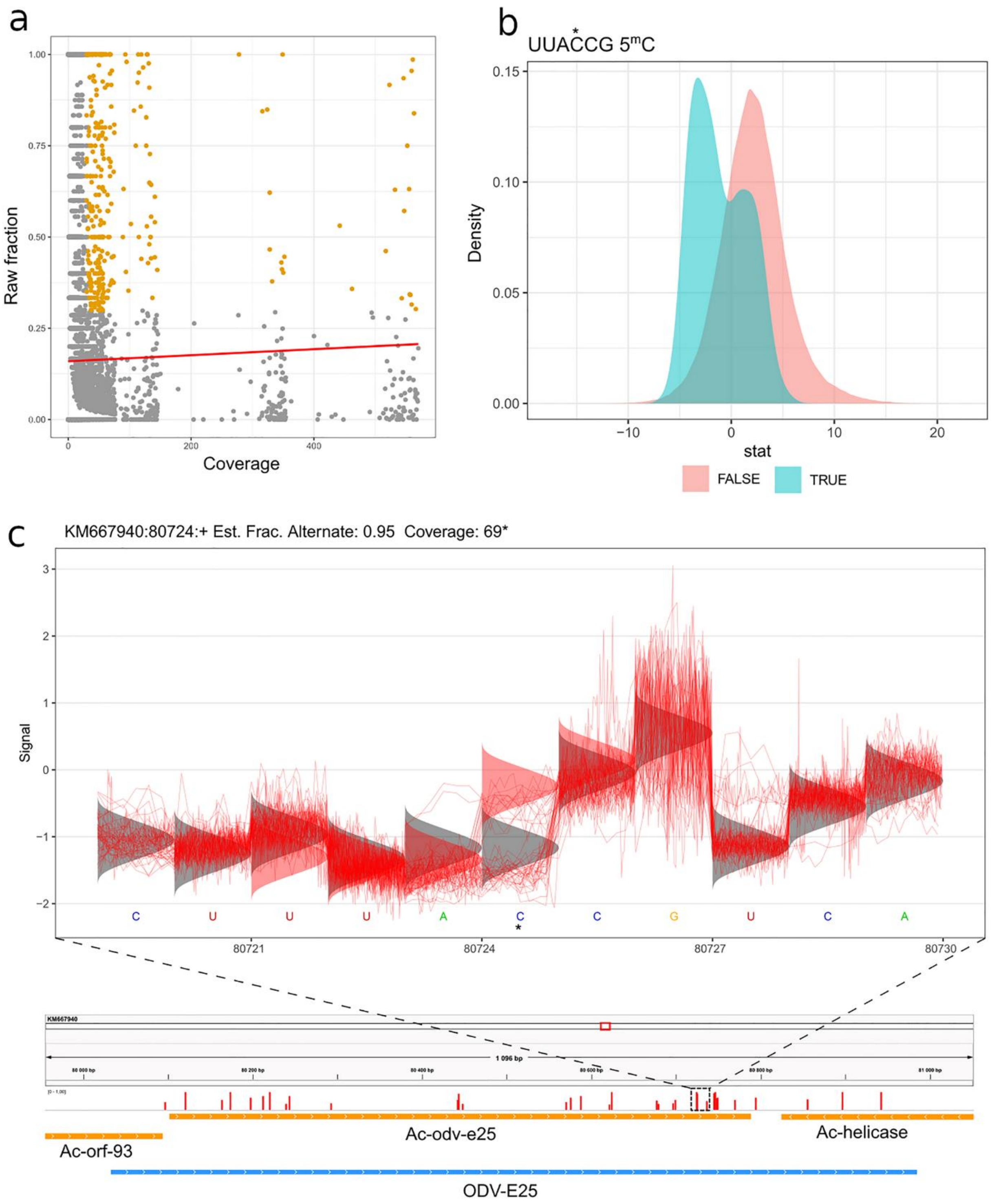
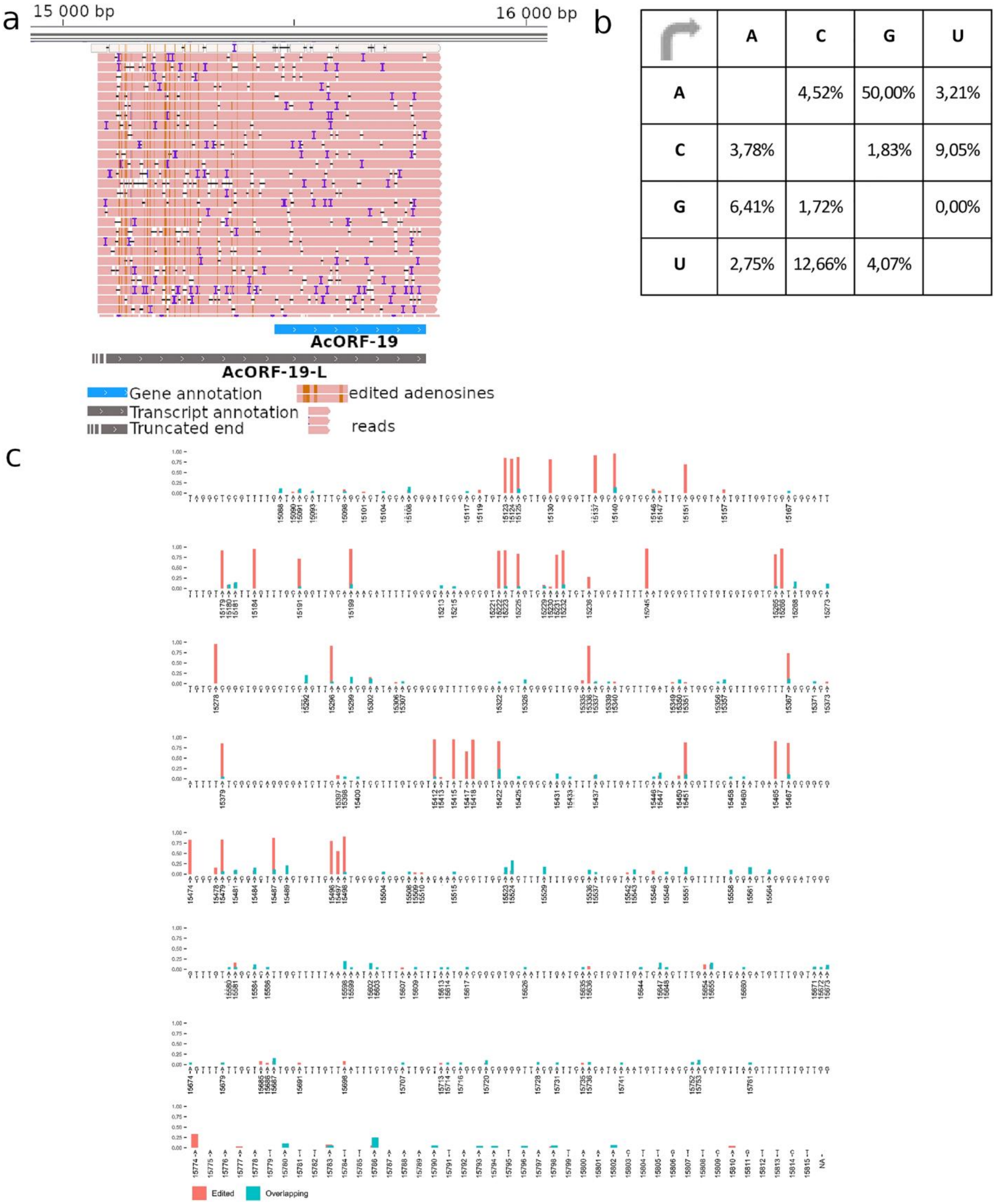


Figure 9. 5-mC methylation of AcMNPV transcripts. (**a**) No significant correlation was observed between the coverage and the number of methylated nucleotides in the raw fraction. Yellow dots indicate positions designated for further analysis. The plot was made with the ggplot2 r package (Wickham et al., 2019). (**b**) Test statistics of UUAC*CG sequence (methylated positions are labelled by asterisks). This panel plots a distribution of test statistics for motif-matching and non-motif-matching sites for all of the provided motives. The plot was made with the ggplot2 r package (Wickham et al., 2019). (**c**) A putative methylated cytosine of the readthrough part of ODV-E25 transcript (5-mC labelled by asterisk). The red curves indicate the electric signals, while the densities indicate the alternative signal levels. The plot was made with the Tombo 1.5.1 software (Matthews et al., 2016).

RNA hyper-editing from A to I. Reads of ORF19-L showed a high frequency of A to I (read A to G by sequencing) substitution, which was not present in the overlapping reads. We found that for ORF19, 50% of all substitutions were A to G (Figure 5a, b), which was significantly higher than the 16.9% of overlapping transcripts in the same region ($p < 0.0001$, one-sided Fisher's exact test) (Figure 5c). A substitution threshold of 16.9% was set to distinguish potentially edited bases from noise due to sequencing inaccuracy. Our results showed that 18% of the total adenines in ORF19-L showed high level ($\overline{x} = 0.839$, $\sigma = 0.153$), while 4% of the adenines in the overlapping reads showed low level A-G editing ($\overline{x} = 0.224$, $\sigma = 0.051$). To identify the presence of a possible editing motif recognized by ADAR, we calculated the frequency of bases in ± 5 nts surrounding the edited A. It has been previously shown that a G-enriched adjacency and an upstream U stabilize the RNA-ADAR complex in mammalian cells (Matthews et al., 2016). A significantly higher Us frequency ($\chi 2(1, N = 79{,}455) = 79{,}338.023$, $p < 0.01$) was observed immediately upstream of the edited base, whereas the frequency of Gs was only slightly higher downstream of the edited base at the + 5 position ($\chi 2(1, N = 79{,}454) = 79{,}340.021$, $p < 0.05$).



	A	C	G	U
A		4,52%	50,00%	3,21%
C	3,78%		1,83%	9,05%
G	6,41%	1,72%		0,00%
U	2,75%	12,66%	4,07%	

Figure 10. Hyper-editing of AcORF-19 transcript. (**a**) A- > I hyper-edited AcORF-19 reads. The image shows the AcORF19-L transcript reads in pink. The brown lines on the reads indicate base modifications. These modifications are located in the 5′-UTR of the canonical transcript. The reads were visualized with IGV 2.11.3 software (Thorvaldsdottir et al., 2013). (**b**) Substitution matrix of AcORF-19 reads. Reference nucleotides are found on the left side of the substitution matrix, while the right side of the matrix contains the percentage of nucleotides. It can be seen that that more than 50% of them are G letter. The substitution matrix was created with Microsoft Excel 2021 software. (**c**) The sites and frequency of A- > G substitution on AcORF-19 transcript

(indicated on the genomic region encoding this transcript). High-frequency substitutions indicate A->G editing events, whereas low-frequency substitutions indicate sequencing errors. Red color indicates the high, whereas blue color shows the low frequency of editing. The plot was made with the ggplot2 r package(Wickham et al., 2019).

7. Discussion

Standard next-generation sequencing techniques are limited by short-read length, as fragmented sequences must be re-assembled computationally, leading to a loss of valuable information on the transcriptome. Long-read sequencing (LRS) is particularly useful for analyzing nested and alternatively spliced transcripts. In this study, we applied two LRS techniques, the SMRT Sequel platform from PacBio and the MinION platform from ONT, for profiling the AcMNPV transcriptome. Amplified and direct RNA sequencing were performed on the ONT platform. Earlier studies have annotated an incomplete lytic transcriptome of EBV. Our multiplatform, integrative approach allowed us to obtain a more complete picture of the transcriptomic architecture of this important human pathogen. For the Epstein–Barr virus lytic transcriptome, we applied the MinION platform using amplified and non-amplified cDNA libraries. Transcript annotation utilized our dataset, as well as SRS and LRS data published previously by others (Cao et al., 2015; Lin et al., 2013, 2010; O'Grady et al., 2016, 2014; Ungerleider et al., 2018). Altogether, we identified 876 novel transcript species, including mRNAs, ncRNAs, mono- and polygenic transcript species, transcript isoforms, and novel splice sites in AcMNPV. A total of 241 novel lytic EBV transcripts were detected, and 110 previously detected transcripts were confirmed.

A recent study (Donovan-Banfield et al., 2020) on human adenovirus type 5, a linear dsDNA virus with medium genome size, disclosed a huge plasticity in intron and TES usage. The authors speculate that this flexibility in viral RNA synthesis can lead to selection advantages and thus fuel viral evolution.

AcMNPV offers unique support for the existence of these variable UTR isoforms by the presence of a LIS located at the TSS. The same TSSs and TESs are used by multiple transcript isoforms of neighboring, or in some cases, distant genes, resulting in polycistronic and complex transcript isoforms. This organization of the transcriptome, especially the intensive usage of the same TSS for multiple transcript isoforms containing varying TESs, is uncommon in herpesviruses (Depledge et al., 2019). However, we observed a somewhat similar pattern of TSS/TES usage in African swine fever virus (ASFV) (Torma et al., 2021b), which is related to insect viruses.

Previous works on herpesviruses (Balázs et al., 2017; Depledge et al., 2019; Moldován et al., 2018b; Prazsák et al., 2018; Tombácz et al., 2016), including the EBV (Majerciak et al., 2018; O'Grady et al., 2016), uncovered a great variety of transcript length isoforms, the function of which is still mostly unclear. TSS isoforms through uORFs, uATGs, and other cis-acting elements of the 5-UTRs are suggested to play an essential role in translational regulation (Geballe and Mocarski, 1988). Additionally, transcripts with alternative termination may have different turnover times (Mayr and Bartel, 2009; Pereira et al., 2017), localization (Macdonald and Struhl, 1988), and altered translation (Martin and Ephrussi, 2009). AcMNPV resembles ASFV and vaccinia virus and differs from herpesviruses in that it exhibits higher heterogeneity in their TESs than TSSs. The alternative use of 3′-UTRs generates long tail-to-tail and tail-to-head transcriptional overlaps. This part of the transcripts may contain cis-regulatory elements, which can bind to regulatory proteins or micro RNAs, thereby controlling the translation and the decay of mRNAs (Matoulkova et al., 2012).

Several multigenic transcripts were detected in each virus. Operons encoding multigenic transcripts represent the basic organization principle of prokaryotic genomes, but they are rare in eukaryotes. The reason for this is that in bacteria, the Shine–Dalgarno sequences allow the translation of every gene in the mRNA (Shine and Dalgarno, 1975). However, in eukaryotes, only the most upstream gene of a multigenic transcript is translated because of the Cap-dependent initiation system. Despite the fact that the viruses of eukaryotes use the same or similar mechanisms as their host organisms, they produce a large variety of multigenic transcripts, the function of which has not yet been described (Boldogkői et al., 2019c). It is hypothesized that transcriptional readthrough in tandem genes (and also on convergent genes) plays a role in a transcription interference-based mechanism (Boldogköi, 2012).

In this study, we detected novel promoters, Inr sequences, and poly(A) sites. In AcMNPV, identified TAAG-Inr motifs bind viral RNP at late and very late phases of the viral life cycle, and non-TAAG-Inr motifs are recognized by both viral and host RNPs at early time points. Our results clearly demonstrate that viral RNP generates longer transcripts than the host RNP. The 3′-cleavage of the viral RNAs and the formation of poly(A) tails are carried out by the polyadenylation machinery of both the host and the virus, although the latter is not well understood.

We detected a relatively large number of short transcripts embedded in larger host genes, containing truncated in-frame ORFs. These transcript types have been described in other viruses, but it turned out that they are more prevalent than earlier believed (Tombácz et al., 2020). Further studies are needed to determine whether these transcripts carry the information of N-terminally truncated polypeptides. If so, this kind of nested transcription significantly increases the coding potential of viruses. In this part of the work, we detected a large number of low-abundance transcript isoforms; nonetheless, their potential functional significance has to be ascertained.

In this work, we report the detection of novel length variants and splice isoforms that may alter the coding potential of several viral genes. Further proteomic studies are needed to conclude the potential significance of these transcripts.

AcMNPV contains 9 AT-rich repetitive sequences (hr regions), which are thought to be replication origins (Pearson et al., 1992; van Oers and Vlak, 2007). However, others have demonstrated that none of them is essential for viral replication (Carstens and Wu, 2007). We detected overlapping transcription from each hr region. They are assumed to play a role in the regulation of replication (Boldogkői et al., 2019a). Our analysis in EBV also revealed novel Ori-overlapping transcripts. Rennekamp and Lieberman showed in their study (Rennekamp and Lieberman, 2011) that the BHLF1 transcript (overlapping the left Ori-lyt) stably binds to its DNA template, and either BHLF1 or the divergent BHRF1 transcript is necessary for the initiation of lytic replication from this Ori. We detected the TSS and TES of BHLF1 with nucleotide precision and the existence of a splice isoform of BHLF1, the BHLF2 transcript. We also identified three isoforms of BHRF1, the BHRT15, the BHRT16, and the BHRT17, with a longer than previously detected Ori-overlapping 5-UTR. The effect of these novel isoforms on viral replication is yet to be evaluated. Our research group has identified several Ori-associated transcripts in various viruses (Boldogköi, 2012; Boldogkői et al., 2019c; Han et al., 2009; Irimia et al., 2014; Kakuk et al., 2021b). We have suggested an interaction between the replication and transcription machineries, which may play a role in the determination of the orientation of the replication fork and the progression of DNA synthesis (Boldogkői et al., 2019a).

Theoretically, the electric signals (squiggles) generated by the nanopore sequencing of native RNAs might be used to identify 5-mC modifications. Currently, only the Tombo package provides support for this, but our results show that this software tends to produce false positive results. We also obtained false negative results with the Combo software, but it is explained by the low dRNA-Seq data coverage. In order to exclude the false results, we also applied the traditional bisulfite conversion method, which converts non-methylated cytosines with 99.5% efficiency. We could validate 70% of the methylation position obtained by Tombo using bisulfate sequencing. With this approach, we detected 5-mC methylation in transcripts of 99 AcMNPV genes and identified the UUA CCG sequence, which is assumed to be a methylation consensus sequence. We found that the majority of methylated positions were located in GC-rich genomic regions. This phenomenon has already been described in mammalian animals

(Yang et al., 2017). Yang and co-workers have demonstrated that 5-mC bases enhance nuclear export via the ALYREF adapter protein in mammalian cells (Yang et al., 2017). Boyne and colleagues have come to the same conclusion.

We detected A to I hyper-editing in the 5′-UTR region of the longer TSS isoform of ORF19 canonical transcript. In cellular organisms, this process plays an important role in innate immunity (Mannion et al., 2014), which is unlikely to be the case in AcMNPV. A-I editing is thought to decrease the affinity of antisense transcripts to the complementary mRNAs through inhibiting the binding of dsRNA nucleases (such as RNase) (Nishikura, 2006). Since cDNA-Seq-based editing detection is still in its infancy, our results need further confirmation.

A genome-wide antisense expression of the EBV genome has already been described using an SRS approach (O'Grady et al., 2014). In this work, we applied an LRS approach that is able to map the transcript ends. According to our results, the majority of antisense transcripts are the results of transcriptional readthroughs between convergent genes or the head-to-head overlap of transcripts encoded by divergently oriented gene pairs. The question as to whether these transcriptional overlaps are functional, or if so, what their significance is, remains unknown. We have suggested the Transcriptional Interference Network hypothesis (Boldogköi, 2012), which claims that one of their functions is to provide a genome-wide gene regulatory mechanism. However, we cannot exclude that at least a part of these transcriptional overlaps represents transcriptional noise without any function. However, the parallel (co-oriented) transcriptional overlaps are not only common but can be considered as a prototypic design of viral genomes. Therefore, we think that other types of overlaps are also functional.

In conclusion, multiplatform approaches are essential in transcriptomic studies because different platforms have distinct advantages and limitations, representing independent techniques vital for validating results obtained by a particular method.

8. Conclusions

This study uses an integrative, multi-technique sequencing approach to gain a more complete understanding of the transcriptomic architecture of AcMNPV and EBV. We have identified several novel transcripts and RNA isoforms, including transcript length and splice variants, as well as novel genes embedded in longer abundance genes containing 5′-truncated in-frame open reading frames potentially encoding N-terminally truncated proteins. We also detected several novel non-coding RNAs and mono- and multigenic transcripts. This work has also identified novel replication origin transcripts. For AcMNPV, we have also detected RNA methylation and RNA hyper-expression in the longer 5′-UTR transcript isoform.

9. Acknowledgements

The study of the EBV transcriptome presented in the thesis was supported by the National Research, Development, and Innovation Office grants FK 128252 to DT and K 128247 to ZB. The APC was covered by the University of Szeged Open Access Fund: 5500. Gábor Torma was supported by the UNKP-21–3-SZTE-105 New National Excellence Program of the Ministry of Human Capacities and study of the AcMNPV transcriptome presented in this thesis was supported by OTKA K 128247 to ZB and OTKA FK 128252 to DT. The APC was covered by the University of Szeged Open Access Fund: 5372.

I would like to express my gratitude to all my colleagues at the Department of Medical Biology for their constant help and support, primarily to my supervisor Prof. Zsolt Boldogkői, who helped with my research, planned the studies, and supervised my work. I also extend thanks to Dr. Dóra Tombácz and Dr. Zsolt Csabai, who conducted the vast majority of the wet lab work for the presented studies, and to Dr. Norbert Moldován and Gábor Torma, who assisted with my analysis work. Additionally, I thank my family and friends for their support.

10. References

A. Kost, T., Patrick Condreay, J., S. Ames, R., 2010. Baculovirus Gene Delivery: A Flexible Assay Development Tool. Curr. Gene Ther. 10, 168–173. https://doi.org/10.2174/156652310791321224

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., Kerlavage, A.R., McCombie, W.R., Venter, J.C., 1991. Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project. Science (80-.). 252, 1651–1656. https://doi.org/10.1126/science.2047873

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D., White, O., 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature 377, 3–174.

Agwati, E.O., Oduor, C.I., Ayieko, C., Ong'echa, J.M., Moormann, A.M., Bailey, J.A., 2022. Profiling genome-wide recombination in Epstein Barr virus reveals type-specific patterns and associations with endemic-Burkitt lymphoma. Virol. J. 19, 208. https://doi.org/10.1186/s12985-022-01942-8

Aragão-Silva, C.W., Andrade, M.S., Ardisson-Araújo, D.M.P., Fernandes, J.E.A., Morgado, F.S., Báo, S.N., Moraes, R.H.P., Wolff, J.L.C., Melo, F.L., Ribeiro, B.M., 2016. The complete genome of a baculovirus isolated from an insect of medical interest: Lonomia obliqua (Lepidoptera: Saturniidae). Sci. Rep. 6, 23127. https://doi.org/10.1038/srep23127

Arvey, A., Tempera, I., Tsai, K., Chen, H.S., Tikhmyanova, N., Klichinsky, M., Leslie, C., Lieberman, P.M., 2012. An atlas of the Epstein-Barr virus transcriptome and epigenome reveals host-virus regulatory interactions. Cell Host Microbe 12, 233–245. https://doi.org/10.1016/j.chom.2012.06.008

Ayres, M.D., Howard, S.C., Kuzio, J., Lopez-Ferber, M., Possee, R.D., 1994. The Complete DNA Sequence of Autographa californica Nuclear Polyhedrosis Virus. Virology 202, 586–605. https://doi.org/10.1006/viro.1994.1380

Balázs, Z., Tombácz, D., Csabai, Z., Moldován, N., Snyder, M., Boldogkői, Z., Boldogkoi, Z., Boldogkői, Z., 2019. Template-switching artifacts resemble alternative polyadenylation.

BMC Genomics 20, 824. https://doi.org/10.1186/s12864-019-6199-7

Balázs, Z., Tombácz, D., Szucs, A., Csabai, Z., Megyeri, K., Petrov, A.N., Snyder, M., Boldogkoi, Z., 2017. Long-Read Sequencing of Human Cytomegalovirus Transcriptome Reveals RNA Isoforms Carrying Distinct Coding Potentials. Sci. Rep. 7. https://doi.org/10.1038/s41598-017-16262-z

Batisse, J., Manet, E., Middeldorp, J., Sergeant, A., Gruffat, H., 2005. Epstein-Barr Virus mRNA Export Factor EB2 Is Essential for Intranuclear Capsid Assembly and Production of gp350. J. Virol. 79, 14102–14111. https://doi.org/10.1128/jvi.79.22.14102-14111.2005

Blissard, G.W., Rohrmann, G.F., 1990. Baculovirus Diversity and Molecular Biology. Annu. Rev. Entomol. 35, 127–155. https://doi.org/10.1146/annurev.en.35.010190.001015

Bodescot, M., Perricaudet, M., Farrell, P.J., 1987. A promoter for the highly spliced EBNA family of RNAs of Epstein-Barr virus. J. Virol. 61, 3424–3430. https://doi.org/10.1128/jvi.61.11.3424-3430.1987

Bogenhagen, D.F., Brown, D.D., 1981. Nucleotide sequences in Xenopus 5S DNA required for transcription termination. Cell 24, 261–270. https://doi.org/10.1016/0092-8674(81)90522-5

Boldogköi, Z., 2012. Transcriptional interference networks coordinate the expression of functionally related genes clustered in the same genomic loci. Front. Genet. 3, 1–17. https://doi.org/10.3389/fgene.2012.00122

Boldogkői, Z., Balázs, Z., Moldován, N., Prazsák, I., Tombácz, D., 2019a. Novel classes of replication-associated transcripts discovered in viruses. RNA Biol. 16, 166–175. https://doi.org/10.1080/15476286.2018.1564468

Boldogkői, Z., Moldován, N., Balázs, Z., Snyder, M., Tombácz, D., 2019b. Long-Read Sequencing – A Powerful Tool in Viral Transcriptome Research. Trends Microbiol. 27, 578–592. https://doi.org/10.1016/j.tim.2019.01.010

Boldogkői, Z., Moldován, N., Szűcs, A., Tombácz, D., 2018. Transcriptome-wide analysis of a baculovirus using nanopore sequencing. Sci. Data 5, 180276. https://doi.org/10.1038/sdata.2018.276

Boldogkői, Z., Tombácz, D., Balázs, Z., 2019c. Interactions between the transcription and replication machineries regulate the RNA and DNA synthesis in the herpesviruses. Virus

Genes. https://doi.org/10.1007/s11262-019-01643-5

Byrne, A., Cole, C., Volden, R., Vollmers, C., 2019. Realizing the potential of full-length transcriptome sequencing. Philos. Trans. R. Soc. B Biol. Sci. 374, 20190097. https://doi.org/10.1098/rstb.2019.0097

Cai, X., Schäfer, A., Lu, S., Bilello, J.P., Desrosiers, R.C., Edwards, R., Raab-Traub, N., Cullen, B.R., 2006. Epstein–Barr Virus MicroRNAs Are Evolutionarily Conserved and Differentially Expressed. PLoS Pathog. 2, e23. https://doi.org/10.1371/journal.ppat.0020023

Calvo, S.E., Pagliarini, D.J., Mootha, V.K., 2009. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc. Natl. Acad. Sci. U. S. A. 106, 7507–7512. https://doi.org/10.1073/pnas.0810916106

Cao, S., Moss, W., O'Grady, T., Concha, M., Strong, M.J., Wang, X., Yu, Y., Baddoo, M., Zhang, K., Fewell, C., Lin, Z., Dong, Y., Flemington, E.K., 2015. New Noncoding Lytic Transcripts Derived from the Epstein-Barr Virus Latency Origin of Replication, oriP , Are Hyperedited, Bind the Paraspeckle Protein, NONO/p54nrb, and Support Viral Lytic Transcription . J. Virol. 89, 7120–7132. https://doi.org/10.1128/jvi.00608-15

Carstens, E.B., Wu, Y., 2007. No single homologous repeat region is essential for DNA replication of the baculovirus Autographa californica multiple nucleopolyhedrovirus. J. Gen. Virol. 88, 114–122. https://doi.org/10.1099/vir.0.82384-0

Chakravorty, A., Sugden, B., Johannsen, E.C., 2019. An Epigenetic Journey: Epstein-Barr Virus Transcribes Chromatinized and Subsequently Unchromatinized Templates during Its Lytic Cycle. J. Virol. 93, 2247–2265. https://doi.org/10.1128/jvi.02247-18

Chen, Y.-R., Zhong, S., Fei, Z., Hashimoto, Y., Xiang, J.Z., Zhang, S., Blissard, G.W., 2013. The Transcriptome of the Baculovirus Autographa californica Multiple Nucleopolyhedrovirus in Trichoplusia ni Cells. J. Virol. 87, 6391–6405. https://doi.org/10.1128/JVI.00194-13

Cocquet, J., Chong, A., Zhang, G., Veitia, R.A., 2006. Reverse transcriptase template switching and false alternative transcripts. Genomics 88, 127–131. https://doi.org/10.1016/j.ygeno.2005.12.013

Concha, M., Wang, X., Cao, S., Baddoo, M., Fewell, C., Lin, Z., Hulme, W., Hedges, D., McBride, J., Flemington, E.K., 2012. Identification of New Viral Genes and Transcript Isoforms during Epstein-Barr Virus Reactivation using RNA-Seq. J. Virol. 86, 1458–1467. https://doi.org/10.1128/jvi.06537-11

Coupland, P., Chandra, T., Quail, M., Reik, W., Swerdlow, H., 2012. Direct sequencing of small genomes on the Pacific Biosciences RS without library preparation. Biotechniques 53, 365–372. https://doi.org/10.2144/000113962

Davison, A.J., Eberle, R., Ehlers, B., Hayward, G.S., McGeoch, D.J., Minson, A.C., Pellett, P.E., Roizman, B., Studdert, M.J., Thiry, E., 2009. The order Herpesvirales. Arch. Virol. 154, 171–177. https://doi.org/10.1007/s00705-008-0278-4

de Martel, C., Georges, D., Bray, F., Ferlay, J., Clifford, G.M., 2020. Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. Lancet Glob. Heal. 8, e180–e190. https://doi.org/10.1016/S2214-109X(19)30488-7

Deakyne, J.S., Malecka, K.A., Messick, T.E., Lieberman, P.M., 2017. Structural and Functional Basis for an EBNA1 Hexameric Ring in Epstein-Barr Virus Episome Maintenance. J. Virol. 91. https://doi.org/10.1128/JVI.01046-17

Depledge, D.P., Srinivas, K.P., Sadaoka, T., Bready, D., Mori, Y., Placantonakis, D.G., Mohr, I., Wilson, A.C., 2019. Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. Nat. Commun. 10, 754. https://doi.org/10.1038/s41467-019-08734-9

Djavadian, R., Hayes, M., Johannsen, E., 2018. CAGE-seq analysis of Epstein-Barr virus lytic gene transcription: 3 kinetic classes from 2 mechanisms. PLOS Pathog. 14, e1007114. https://doi.org/10.1371/journal.ppat.1007114

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Donovan-Banfield, I., Turnell, A.S., Hiscox, J.A., Leppard, K.N., Matthews, D.A., 2020. Deep splicing plasticity of the human adenovirus type 5 transcriptome drives virus evolution. Commun. Biol. 3, 1–14. https://doi.org/10.1038/s42003-020-0849-9

Dresang, L.R., Teuton, J.R., Feng, H., Jacobs, J.M., Camp, D.G., Purvine, S.O., Gritsenko, M.A., Li, Z., Smith, R.D., Sugden, B., Moore, P.S., Chang, Y., 2011. Coupled transcriptome and proteome analysis of human lymphotropic tumor viruses: insights on the detection and discovery of viral genes. BMC Genomics 12, 625. https://doi.org/10.1186/1471-2164-12-625

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., DeWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., Turner, S., 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. Science (80-.). 323, 133–138. https://doi.org/10.1126/science.1162986

Ersing, I., Nobre, L., Wang, L.W., Soday, L., Ma, Y., Paulo, J.A., Narita, Y., Ashbaugh, C.W., Jiang, C., Grayson, N.E., Kieff, E., Gygi, S.P., Weekes, M.P., Gewurz, B.E., 2017. A Temporal Proteomic Map of Epstein-Barr Virus Lytic Replication in B Cells. Cell Rep. 19, 1479–1493. https://doi.org/10.1016/j.celrep.2017.04.062

Fülöp, Á., Torma, G., Moldován, N., Szenthe, K., Bánáti, F., Almsarrhad, I.A.A., Csabai, Z., Tombácz, D., Minárovits, J., Boldogkői, Z., 2022. Integrative profiling of Epstein-Barr virus transcriptome using a multiplatform approach. Virol. J. 19, 7. https://doi.org/10.1186/s12985-021-01734-6

Garalde, D.R., Snell, E.A., Jachimowicz, D., Sipos, B., Lloyd, J.H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., Jordan, M., Ciccone, J., Serra, S., Keenan, J., Martin, S., McNeill, L., Wallace, E.J., Jayasinghe, L., Wright, C., Blasco, J., Young, S., Brocklebank, D., Juul, S., Clarke, J., Heron, A.J., Turner, D.J., 2018. Highly parallel direct RNA sequencing on an array of nanopores. Nat. Methods 15, 201–206. https://doi.org/10.1038/nmeth.4577

Garrity, D.B., Chang, M.-J., Blissard, G.W., 1997. Late Promoter Selection in the Baculovirusgp64 Envelope Fusion ProteinGene. Virology 231, 167–181. https://doi.org/10.1006/viro.1997.8540

Gatherer, D., Seirafian, S., Cunningham, C., Holton, M., Dargan, D.J., Baluchova, K., Hector, R.D., Galbraith, J., Herzyk, P., Wilkinson, G.W.G., Davison, A.J., 2011. High-resolution human cytomegalovirus transcriptome. Proc. Natl. Acad. Sci. U. S. A. 108, 19755–19760. https://doi.org/10.1073/pnas.1115861108

Geballe, A.P., Mocarski, E.S., 1988. Translational control of cytomegalovirus gene expression is mediated by upstream AUG codons. J. Virol. 62, 3334–40.

George F Rohrmann., 2008a. Baculovirus Molecular Biology, 1 th. ed, Bethesda (MD): National Center for Biotechnology Information (US). Oregon.

George F Rohrmann., 2008b. MOLECULAR BIOLOGY Molecular Biology General. ethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information, Oregon.

Gershburg, E., Raffa, S., Torrisi, M.R., Pagano, J.S., 2007. Epstein-Barr Virus-Encoded Protein Kinase (BGLF4) Is Involved in Production of Infectious Virus. J. Virol. 81, 5407–5412. https://doi.org/10.1128/jvi.02398-06

Gold, J.E., Okyay, R.A., Licht, W.E., Hurley, D.J., 2021. Investigation of long covid prevalence and its relationship to epstein-barr virus reactivation. Pathogens 10, 1–15. https://doi.org/10.3390/pathogens10060763

Guarino, L.A., Summers, M.D., 1986. Functional mapping of a trans-activating gene required for expression of a baculovirus delayed-early gene. J. Virol. 57, 563–571. https://doi.org/10.1128/jvi.57.2.563-571.1986

Hammerschmidt, W., Sugden, B., 2013. Replication of Epstein-Barr Viral DNA. Cold Spring Harb. Perspect. Biol. 5, a013029–a013029. https://doi.org/10.1101/cshperspect.a013029

Han, Z., Verma, D., Hilscher, C., Dittmer, D.P., Swaminathan, S., 2009. General and Target-Specific RNA Binding Properties of Epstein-Barr Virus SM Posttranscriptional Regulatory Protein. J. Virol. 83, 11635–11644. https://doi.org/10.1128/jvi.01483-09

Hangauer, M.J., Vaughn, I.W., McManus, M.T., 2013. Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. PLoS Genet. 9, e1003569. https://doi.org/10.1371/journal.pgen.1003569

Heather, J.M., Chain, B., 2016. The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Hodin, T.L., Najrana, T., Yates, J.L., 2013. Efficient Replication of Epstein-Barr Virus-Derived

Plasmids Requires Tethering by EBNA1 to Host Chromosomes. J. Virol. 87, 13020–13028. https://doi.org/10.1128/JVI.01606-13

Hu, Y., 2006. Baculovirus Vectors for Gene Therapy. pp. 287–320. https://doi.org/10.1016/S0065-3527(06)68008-1

Hu, Y., 2005. Baculovirus as a highly efficient expression vector in insect and mammalian cells. Acta Pharmacol. Sin. 26, 405–416. https://doi.org/10.1111/j.1745-7254.2005.00078.x

Hubbard, K.S., Gut, I.M., Lyman, M.E., McNutt, P.M., 2013. Longitudinal RNA sequencing of the deep transcriptome during neurogenesis of cortical glutamatergic neurons from murine ESCs. F1000Research 2, 35. https://doi.org/10.12688/f1000research.2-35.v1

Irimia, M., Weatheritt, R.J., Ellis, J.D., Parikshak, N.N., Gonatopoulos-Pournatzis, T., Babor, M., Quesnel-Vallières, M., Tapial, J., Raj, B., O'Hanlon, D., Barrios-Rodiles, M., Sternberg, M.J.E., Cordes, S.P., Roth, F.P., Wrana, J.L., Geschwind, D.H., Blencowe, B.J., Irimia M, Weatheritt RJ, Ellis JD, et al., 2014. A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains. Cell 159, 1511–1523. https://doi.org/10.1016/j.cell.2014.11.035

Jiang, S.S., Chang, I.-S., Huang, L.-W., Chen, P.-C., Wen, C.-C., Liu, S.-C., Chien, L.-C., Lin, C.-Y., Hsiung, C.A., Juang, J.-L., 2006. Temporal Transcription Program of Recombinant Autographa californica Multiple Nucleopolyhedrosis Virus. J. Virol. 80, 8989–8999. https://doi.org/10.1128/JVI.01158-06

Jin, J., Guarino, L.A., 2000. 3′-End Formation of Baculovirus Late RNAs. J. Virol. 74, 8930–8937. https://doi.org/10.1128/JVI.74.19.8930-8937.2000

K-H Liang, 2013. Bioinformatics for Biomedical Science and Clinical Applications, 1 st. ed. Woodhead Publishing.

Kakuk, B., Kiss, A.A., Torma, G., Csabai, Z., Prazsák, I., Mizik, M., Megyeri, K., Tombácz, D., Boldogkői, Z., 2021a. Nanopore Assay Reveals Cell-Type-Dependent Gene Expression of Vesicular Stomatitis Indiana Virus and Differential Host Cell Response. Pathog. (Basel, Switzerland) 10, 1196. https://doi.org/10.3390/pathogens10091196

Kakuk, B., Tombácz, D., Balázs, Z., Moldován, N., Csabai, Z., Torma, G., Megyeri, K., Snyder, M., Boldogkői, Z., 2021b. Combined nanopore and single-molecule real-time sequencing survey of human betaherpesvirus 5 transcriptome. Sci. Rep. 11, 14487.

https://doi.org/10.1038/s41598-021-93593-y

Kanda, T., Horikoshi, N., Murata, T., Kawashima, D., Sugimoto, A., Narita, Y., Kurumizaka, H., Tsurumi, T., 2013. Interaction between Basic Residues of Epstein-Barr Virus EBNA1 Protein and Cellular Chromatin Mediates Viral Plasmid Maintenance. J. Biol. Chem. 288, 24189–24199. https://doi.org/10.1074/jbc.M113.491167

Kang, M.-S., Kieff, E., 2015. Epstein–Barr virus latent genes. Exp. Mol. Med. 47, e131–e131. https://doi.org/10.1038/emm.2014.84

Kempkes, B., Ling, P.D., 2015. EBNA2 and Its Coactivator EBNA-LP. pp. 35–59. https://doi.org/10.1007/978-3-319-22834-1_2

Kenney, S.C., Mertz, J.E., 2014. Regulation of the latent-lytic switch in Epstein–Barr virus. Semin. Cancer Biol. 26, 60–68. https://doi.org/10.1016/j.semcancer.2014.01.002

Kenoutis, C., Efrose, R.C., Swevers, L., Lavdas, A.A., Gaitanou, M., Matsas, R., Iatrou, K., 2006. Baculovirus-Mediated Gene Delivery into Mammalian Cells Does Not Alter Their Transcriptional and Differentiating Potential but Is Accompanied by Early Viral Gene Expression. J. Virol. 80, 4135–4146. https://doi.org/10.1128/JVI.80.8.4135-4146.2006

Kogan, P.H., Chen, X., Blissard, G.W., 1995. Overlapping TATA-dependent and TATA-independent early promoter activities in the baculovirus gp64 envelope fusion protein gene. J. Virol. 69, 1452–1461. https://doi.org/10.1128/jvi.69.3.1452-1461.1995

Kost, T.A., Condreay, J.P., Jarvis, D.L., 2005. Baculovirus as versatile vectors for protein expression in insect and mammalian cells. Nat. Biotechnol. 23, 567–575. https://doi.org/10.1038/nbt1095

Kovacs, G.R., Guarino, L.A., Graham, B.L., Summers, M.D., 1991. Identification of spliced baculovirus RNAs expressed late in infection. Virology 185, 633–643. https://doi.org/10.1016/0042-6822(91)90534-I

Kronstad, L.M., Brulois, K.F., Jung, J.U., Glaunsinger, B.A., 2013. Dual Short Upstream Open Reading Frames Control Translation of a Herpesviral Polycistronic mRNA. PLoS Pathog. 9, e1003156. https://doi.org/10.1371/journal.ppat.1003156

Lee, N., Yario, T.A., Gao, J.S., Steitz, J.A., 2016. EBV noncoding RNA EBER2 interacts with host RNA-binding proteins to regulate viral gene expression. Proc. Natl. Acad. Sci. 113, 3221–3226. https://doi.org/10.1073/pnas.1601773113

Leppek, K., Das, R., Barna, M., 2018. Functional 5’ UTR mRNA structures in eukaryotic translation regulation and how to find them. Nat. Rev. Mol. Cell Biol. 19, 158–174. https://doi.org/10.1038/nrm.2017.103

Levene, M.J., Korlach, J., Turner, S.W., Foquet, M., Craighead, H.G., Webb, W.W., 2003. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. Science (80-.). 299, 682–686. https://doi.org/10.1126/science.1079700

Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Hu, J., Luo, X., Bode, A.M., Dong, Z., Cao, Y., 2018. Therapies based on targeting Epstein-Barr virus lytic replication for EBV-associated malignancies. Cancer Sci. 109, 2101–2108. https://doi.org/10.1111/cas.13634

Li, Y., Guarino, L.A., 2008. Roles of LEF-4 and PTP/BVP RNA Triphosphatases in Processing of Baculovirus Late mRNAs. J. Virol. 82, 5573–5583. https://doi.org/10.1128/JVI.00058-08

Lin, Z., Wang, X., Strong, M.J., Concha, M., Baddoo, M., Xu, G., Baribault, C., Fewell, C., Hulme, W., Hedges, D., Taylor, C.M., Flemington, E.K., 2013. Whole-Genome Sequencing of the Akata and Mutu Epstein-Barr Virus Strains. J. Virol. 87, 1172–1182. https://doi.org/10.1128/jvi.02517-12

Lin, Z., Xu, G., Deng, N., Taylor, C., Zhu, D., Flemington, E.K., 2010. Quantitative and Qualitative RNA-Seq-Based Evaluation of Epstein-Barr Virus Transcription in Type I Latency Burkitt’s Lymphoma Cells. J. Virol. 84, 13053–13058. https://doi.org/10.1128/jvi.01521-10

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., Law, M., 2012. Comparison of Next-Generation Sequencing Systems. J. Biomed. Biotechnol. 2012, 1–11. https://doi.org/10.1155/2012/251364

Liu, P., Speck, S.H., 2003. Synergistic autoactivation of the Epstein-Barr virus immediate-early BRLF1 promoter by Rta and Zta. Virology 310, 199–206. https://doi.org/10.1016/S0042-6822(03)00145-4

Lu, A., Carstens, E.B., 1993. Immediate-Early Baculovirus Genes Transactivate the p143 Gene Promoter of Autographa californica Nuclear Polyhedrosis Virus. Virology 195, 710–718.

https://doi.org/10.1006/viro.1993.1422

Lu, H., Giordano, F., Ning, Z., 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics. Proteomics Bioinformatics 14, 265–279. https://doi.org/10.1016/j.gpb.2016.05.004

Macdonald, P.M., Struhl, G., 1988. Cis- acting sequences responsible for anterior localization of bicoid mRNA in Drosophila embryos. Nature 336, 595–598. https://doi.org/10.1038/336595a0

Majerciak, V., Yang, W., Zheng, J., Zhu, J., Zheng, Z.-M., 2018. A Genome-Wide Epstein-Barr Virus Polyadenylation Map and Its Antisense RNA to EBNA. J. Virol. 93. https://doi.org/10.1128/jvi.01593-18

Mannion, N.M., Greenwood, S.M., Young, R., Cox, S., Brindle, J., Read, D., Nellåker, C., Vesely, C., Ponting, C.P., McLaughlin, P.J., Jantsch, M.F., Dorin, J., Adams, I.R., Scadden, A.D.J., Öhman, M., Keegan, L.P., O'Connell, M.A., 2014. The RNA-Editing Enzyme ADAR1 Controls Innate Immune Responses to RNA. Cell Rep. 9, 1482–1494. https://doi.org/10.1016/j.celrep.2014.10.041

Marquitz, A.R., Mathur, A., Edwards, R.H., Raab-Traub, N., 2015. Host Gene Expression Is Regulated by Two Types of Noncoding RNAs Transcribed from the Epstein-Barr Virus BamHI A Rightward Transcript Region. J. Virol. 89, 11256–11268. https://doi.org/10.1128/JVI.01492-15

Martin, K.C., Ephrussi, A., 2009. mRNA Localization: Gene Expression in the Spatial Dimension. Cell. https://doi.org/10.1016/j.cell.2009.01.044

Matoulkova, E., Michalova, E., Vojtesek, B., Hrstka, R., 2012. The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells. RNA Biol. 9, 563–576. https://doi.org/10.4161/rna.20231

Matthews, M.M., Thomas, J.M., Zheng, Y., Tran, K., Phelps, K.J., Scott, A.I., Havel, J., Fisher, A.J., Beal, P.A., 2016. Structures of human ADAR2 bound to dsRNA reveal base-flipping mechanism and basis for site selectivity. Nat. Struct. Mol. Biol. 23, 426–433. https://doi.org/10.1038/nsmb.3203

Mayr, C., Bartel, D.P., 2009. Widespread Shortening of 3′UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. Cell 138, 673–684.

https://doi.org/10.1016/j.cell.2009.06.016

McLachlin, J.R., Miller, L.K., 1994. Identification and characterization of vlf-1, a baculovirus gene involved in very late gene expression. J. Virol. 68, 7746–7756. https://doi.org/10.1128/jvi.68.12.7746-7756.1994

Meckiff, B.J., Ladell, K., McLaren, J.E., Ryan, G.B., Leese, A.M., James, E.A., Price, D.A., Long, H.M., 2019. Primary EBV Infection Induces an Acute Wave of Activated Antigen-Specific Cytotoxic CD4 + T Cells . J. Immunol. 203, 1276–1287. https://doi.org/10.4049/jimmunol.1900377

Metzker, M.L., 2010. Sequencing technologies — the next generation. Nat. Rev. Genet. 11, 31–46. https://doi.org/10.1038/nrg2626

Moldován, N., Balázs, Z., Tombácz, D., Csabai, Z., Szűcs, A., Snyder, M., Boldogkői, Z., 2017. Multi-platform analysis reveals a complex transcriptome architecture of a circovirus. Virus Res. 237, 37–46. https://doi.org/10.1016/j.virusres.2017.05.010

Moldován, N., Tombácz, D., Szucs, A., Csabai, Z., Balázs, Z., Kis, E., Molnár, J., Boldogkoi, Z., Szűcs, A., Csabai, Z., Balázs, Z., Kis, E., Molnár, J., Boldogkői, Z., 2018a. Third-generation Sequencing Reveals Extensive Polycistronism and Transcriptional Overlapping in a Baculovirus. Sci. Rep. 8, 8604. https://doi.org/10.1038/s41598-018-26955-8

Moldován, N., Tombácz, D., Szucs, A., Csabai, Z., Snyder, M., Boldogkoi, Z., 2018b. Multi-platform sequencing approach reveals a novel transcriptome profile in pseudorabies virus. Front. Microbiol. 8, 1–13. https://doi.org/10.3389/fmicb.2017.02708

Moldován, N., Torma, G., Gulyás, G., Hornyák, Á., Zádori, Z., Jefferson, V.A., Csabai, Z., Boldogkői, M., Tombácz, D., Meyer, F., Boldogkői, Z., 2020. Time-course profiling of bovine alphaherpesvirus 1.1 transcriptome using multiplatform sequencing. Sci. Rep. 10, 20496. https://doi.org/10.1038/s41598-020-77520-1

Nagaraju, T., Sugden, A.U., Sugden, B., 2019. Four-dimensional analyses show that replication compartments are clonal factories in which Epstein–Barr viral DNA amplification is coordinated. Proc. Natl. Acad. Sci. U. S. A. 116, 24630–24638. https://doi.org/10.1073/pnas.1913992116

Nishikura, K., 2006. Editor meets silencer: crosstalk between RNA editing and RNA interference. Nat. Rev. Mol. Cell Biol. 7, 919–931. https://doi.org/10.1038/nrm2061

O’Grady, T., Cao, S., Strong, M.J., Concha, M., Wang, X., Splinter Bondurant, S., Adams, M., Baddoo, M., Srivastav, S.K., Lin, Z., Fewell, C., Yin, Q., Flemington, E.K., 2014. Global bidirectional transcription of the Epstein-Barr virus genome during reactivation. J. Virol. 88, 1604–16. https://doi.org/10.1128/JVI.02989-13

O’Grady, T., Wang, X., Höner zu Bentrup, K., Baddoo, M., Concha, M., Flemington, E.K., Höner zu Bentrup, K., Baddoo, M., Concha, M., Flemington, E.K., 2016. Global transcript structure resolution of high gene density genomes through multi-platform data integration. Nucleic Acids Res. 44, e145–e145. https://doi.org/10.1093/nar/gkw629

Odelberg, S.J., Weiss, R.B., Hata, A., White, R., 1995. Template-switching during DNA synthesis by Thermus aquaticus DNA polymerase I. Nucleic Acids Res. 23, 2049–2057. https://doi.org/10.1093/nar/23.11.2049

Ooi, B.G., Rankin, C., Miller, L.K., 1989. Downstream sequences augment transcription from the essential initiation site of a baculovirus polyhedrin gene. J. Mol. Biol. 210, 721–736. https://doi.org/10.1016/0022-2836(89)90105-8

Pearson, M., Bjornson, R., Pearson, G., Rohrmann, G., 1992. The Autographa californica Baculovirus Genome: Evidence for Multiple Replication Origins. Science (80-.). 257, 1382–1384. https://doi.org/10.1126/science.1529337

Peng, R.J., Han, B.W., Cai, Q.Q., Zuo, X.Y., Xia, T., Chen, J.R., Feng, L.N., Lim, J.Q., Chen, S.W., Zeng, M.S., Guo, Y.M., Li, B., Xia, X.J., Xia, Y., Laurensia, Y., Chia, B.K.H., Huang, H.Q., Young, K.H., Lim, S.T., Ong, C.K., Zeng, Y.X., Bei, J.X., 2019. Genomic and transcriptomic landscapes of Epstein-Barr virus in extranodal natural killer T-cell lymphoma. Leukemia 33, 1451–1462. https://doi.org/10.1038/s41375-018-0324-5

Pereira, L.A., Munita, R., González, M.P., Andrés, M.E., 2017. Long 3’UTR of Nurr1 mRNAs is targeted by miRNAs in mesencephalic dopamine neurons. PLoS One 12, 1–15. https://doi.org/10.1371/journal.pone.0188177

Prazsák, I., Moldován, N., Balázs, Z., Tombácz, D., Megyeri, K., Szűcs, A., Csabai, Z., Boldogkői, Z., Szucs, A., Csabai, Z., Boldogkoi, Z., Szűcs, A., Csabai, Z., Boldogkői, Z., 2018. Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus. BMC Genomics 19, 873. https://doi.org/10.1186/s12864-018-5267-8

Prazsák, I., Tombácz, D., Fülöp, Á., Torma, G., Gulyás, G., Dörmő, Á., Kakuk, B., Spires, L.M., Toth, Z., Boldogkői, Z., 2023. KSHV 3.0: A State-of-the-Art Annotation of the Kaposi's Sarcoma-Associated Herpesvirus Transcriptome Using Cross-Platform Sequencing. bioRxiv Prepr. Serv. Biol. 2023.09.21.558842. https://doi.org/10.1101/2023.09.21.558842

Price, A.M., Luftig, M.A., 2015. To Be or Not IIb: A Multi-Step Process for Epstein-Barr Virus Latency Establishment and Consequences for B Cell Tumorigenesis. PLOS Pathog. 11, e1004656. https://doi.org/10.1371/journal.ppat.1004656

Rennekamp, A.J., Lieberman, P.M., 2011. Initiation of Epstein-Barr Virus Lytic Replication Requires Transcription and the Formation of a Stable RNA-DNA Hybrid Molecule at OriLyt. J. Virol. 85, 2837–2850. https://doi.org/10.1128/JVI.02175-10

Rhoads, A., Au, K.F., 2015. PacBio Sequencing and Its Applications. Genomics. Proteomics Bioinformatics 13, 278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Rogers, R.P., Woisetschlaeger, M., Speck, S.H., 1990. Alternative splicing dictates translational start in Epstein-Barr virus transcripts. EMBO J. 9, 2273–2277. https://doi.org/10.1002/j.1460-2075.1990.tb07398.x

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R.P., Banday, S., Mishra, A.K., Das, G., Malonia, S.K., 2023. Next-Generation Sequencing Technology: Current Trends and Advancements. Biology (Basel). 12, 997. https://doi.org/10.3390/biology12070997

Schaeffner, M., Mrozek-Gorska, P., Buschle, A., Woellmer, A., Tagawa, T., Cernilogar, F.M., Schotta, G., Krietenstein, N., Lieleg, C., Korber, P., Hammerschmidt, W., 2019. BZLF1 interacts with chromatin remodelers promoting escape from latent infections with EBV. Life Sci. Alliance 2, e201800108. https://doi.org/10.26508/lsa.201800108

Schneider, G.F., Dekker, C., 2012. DNA sequencing with nanopores. Nat. Biotechnol. 30, 326–328. https://doi.org/10.1038/nbt.2181

Sequeira-Mendes, J., Díaz-Uriarte, R., Apedaile, A., Huntley, D., Brockdorff, N., Gómez, M., 2009. Transcription initiation activity sets replication origin efficiency in mammalian cells. PLoS Genet. 5, 1000446. https://doi.org/10.1371/journal.pgen.1000446

Shannon-Lowe, C., Rickinson, A., 2019. The Global Landscape of EBV-Associated Tumors.

Front. Oncol. https://doi.org/10.3389/fonc.2019.00713

Shine, J., Dalgarno, L., 1975. Determinant of cistron specificity in bacterial ribosomes. Nature 254, 34–38. https://doi.org/10.1038/254034a0

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P., Hayashizaki, Y., 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. 100, 15776–15781. https://doi.org/10.1073/pnas.2136655100

Smith, I., 2007. Misleading Messengers? Interpreting Baculovirus Transcriptional Array Profiles. J. Virol. 81, 7819–7821. https://doi.org/10.1128/JVI.00615-07

Soneson, C., Yao, Y., Bratus-Neuenschwander, A., Patrignani, A., Robinson, M.D., Hussain, S., 2019. A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. Nat. Commun. 10, 3359. https://doi.org/10.1038/s41467-019-11272-z

Steijger, T., Abril, J.F., Engström, P.G., Kokocinski, F., Abril, J.F., Akerman, M., Alioto, T., Ambrosini, G., Antonarakis, S.E., Behr, J., Bertone, P., Bohnert, R., Bucher, P., Cloonan, N., Derrien, T., Djebali, S., Du, J., Dudoit, S., Engström, P.G., Gerstein, M., Gingeras, T.R., Gonzalez, D., Grimmond, S.M., Guigó, R., Habegger, L., Harrow, J., Hubbard, T.J., Iseli, C., Jean, G., Kahles, A., Kokocinski, F., Lagarde, J., Leng, J., Lefebvre, G., Lewis, S., Mortazavi, A., Niermann, P., Rätsch, G., Reymond, A., Ribeca, P., Richard, H., Rougemont, J., Rozowsky, J., Sammeth, M., Sboner, A., Schulz, M.H., Searle, S.M.J., Solorzano, N.D., Solovyev, V., Stanke, M., Steijger, T., Stevenson, B.J., Stockinger, H., Valsesia, A., Weese, D., White, S., Wold, B.J., Wu, J., Wu, T.D., Zeller, G., Zerbino, D., Zhang, M.Q., Hubbard, T.J., Guigó, R., Harrow, J., Bertone, P., Bertone, P., 2013. Assessment of transcript reconstruction methods for RNA-seq. Nat. Methods 10, 1177–1184. https://doi.org/10.1038/nmeth.2714

Tai-Schmiedel, J., Karniely, S., Ezra, A., Eliyahu, E., Nachshon, A., Winkler, R., Schwartz, M., Stern-Ginossar, N., 2018. The virally encoded long non-coding RNA4.9 is controlling viral DNA replication, in: International Herpesvirus Workshop 2018. University of British Columbia, Vancouver, p. 2.32.

Takacs, M., Banati, F., Koroknai, A., Segesdi, J., Salamon, D., Wolf, H., Niller, H.H., Minarovits, J., 2010. Epigenetic regulation of latent Epstein–Barr virus promoters. Biochim. Biophys. Acta - Gene Regul. Mech. 1799, 228–235. https://doi.org/10.1016/j.bbagrm.2009.10.005

Thodberg, M., Thieffry, A., Vitting-Seerup, K., Andersson, R., Sandelin, A., 2019. CAGEfightR: analysis of 5′-end data using R/Bioconductor. BMC Bioinformatics 20. https://doi.org/10.1186/s12859-019-3029-5

Thompson, M.P., Kurzrock, R., 2004. Epstein-Barr Virus and Cancer. Clin. Cancer Res. 10, 803–821. https://doi.org/10.1158/1078-0432.CCR-0670-3

Thorley-Lawson, D.A., 2015. EBV persistence-introducing the virus, in: Epstein Barr Virus. Springer International Publishing, pp. 151–209. https://doi.org/10.1007/978-3-319-22822-8_8

Thorvaldsdottir, H., Robinson, J.T., Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. 14, 178–192. https://doi.org/10.1093/bib/bbs017

Tian, B., Graber, J.H., 2012. Signals for pre-mRNA cleavage and polyadenylation. Wiley Interdiscip. Rev. RNA 3, 385–96. https://doi.org/10.1002/wrna.116

Tian, J.-H., Patel, N., Haupt, R., Zhou, H., Weston, S., Hammond, H., Logue, J., Portnoff, A.D., Norton, J., Guebre-Xabier, M., Zhou, B., Jacobson, K., Maciejewski, S., Khatoon, R., Wisniewska, M., Moffitt, W., Kluepfel-Stahl, S., Ekechukwu, B., Papin, J., Boddapati, S., Jason Wong, C., Piedra, P.A., Frieman, M.B., Massare, M.J., Fries, L., Bengtsson, K.L., Stertman, L., Ellingsworth, L., Glenn, G., Smith, G., 2021. SARS-CoV-2 spike glycoprotein vaccine candidate NVX-CoV2373 immunogenicity in baboons and protection in mice. Nat. Commun. 12, 372. https://doi.org/10.1038/s41467-020-20653-8

Tikhanovich, I., Liang, B., Seoighe, C., Folk, W.R., Nasheuer, H.P., 2011. Inhibition of Human BK Polyomavirus Replication by Small Noncoding RNAs. J. Virol. 85, 6930–6940. https://doi.org/10.1128/JVI.00547-11

Tombácz, D., Balázs, Z., Csabai, Z., Moldován, N., Szücs, A., Sharon, D., Snyder, M., Boldogköi, Z., 2017. Characterization of the Dynamic Transcriptome of a Herpesvirus with Long-read Single Molecule Real-Time Sequencing. Sci. Rep. 7, 1–13. https://doi.org/10.1038/srep43751

Tombácz, D., Csabai, Z., Oláh, P., Balázs, Z., Likó, I., Zsigmond, L., Sharon, D., Snyder, M., Boldogkői, Z., Boldogkoi, Z., 2016. Full-Length Isoform Sequencing Reveals Novel Transcripts and Substantial Transcriptional Overlaps in a Herpesvirus. PLoS One 11, e0162868. https://doi.org/10.1371/journal.pone.0162868

Tombácz, D., Moldován, N., Balázs, Z., Gulyás, G., Csabai, Z., Boldogkői, M., Snyder, M., Boldogkői, Z., 2019. Multiple Long-Read Sequencing Survey of Herpes Simplex Virus Dynamic Transcriptome. Front. Genet. 10, 1–20. https://doi.org/10.3389/fgene.2019.00834

Tombácz, D., Prazsák, I., Torma, G., Csabai, Z., Balázs, Z., Moldován, N., Dénes, B., Snyder, M., Boldogkői, Z., 2021. Time-Course Transcriptome Profiling of a Poxvirus Using Long-Read Full-Length Assay. Pathogens 10, 919. https://doi.org/10.3390/pathogens10080919

Tombácz, D., Torma, G., Gulyás, G., Fülöp, Á., Dörmő, Á., Prazsák, I., Csabai, Z., Mizik, M., Hornyák, Á., Zádori, Z., Kakuk, B., Boldogkői, Z., 2023. Hybrid sequencing discloses unique aspects of the transcriptomic architecture in equid alphaherpesvirus 1. Heliyon 9, e17716. https://doi.org/10.1016/j.heliyon.2023.e17716

Tombácz, D., Torma, G., Gulyás, G., Moldován, N., Snyder, M., Boldogkői, Z., 2020. Meta-analytic approach for transcriptome profiling of herpes simplex virus type 1. Sci. Data 7, 223. https://doi.org/10.1038/s41597-020-0558-8

Torma, G., Tombácz, D., Csabai, Z., Almsarrhad, I.A.A., Nagy, G.Á., Kakuk, B., Gulyás, G., Spires, L.M., Gupta, I., Fülöp, Á., Dörmő, Á., Prazsák, I., Mizik, M., Dani, V.É., Csányi, V., Harangozó, Á., Zádori, Z., Toth, Z., Boldogkői, Z., 2023. Identification of herpesvirus transcripts from genomic regions around the replication origins. Sci. Rep. 13, 16395. https://doi.org/10.1038/s41598-023-43344-y

Torma, G., Tombácz, D., Csabai, Z., Göbhardter, D., Deim, Z., Snyder, M., Boldogkői, Z., 2021a. An Integrated Sequencing Approach for Updating the Pseudorabies Virus Transcriptome. Pathogens 10, 242. https://doi.org/10.3390/pathogens10020242

Torma, G., Tombácz, D., Csabai, Z., Moldován, N., Mészáros, I., Zádori, Z., Boldogkői, Z., 2021b. Combined short and long-read sequencing reveals a complex transcriptomic architecture of African swine fever virus. Viruses 13. https://doi.org/10.3390/v13040579

Torma, G., Tombácz, D., Moldován, N., Fülöp, Á., Prazsák, I., Csabai, Z., Snyder, M., Boldogkői, Z., 2022. Dual isoform sequencing reveals complex transcriptomic and epitranscriptomic landscapes of a prototype baculovirus. Sci. Rep. 12, 1291. https://doi.org/10.1038/s41598-022-05457-8

Travers, K.J., Chin, C.-S., Rank, D.R., Eid, J.S., Turner, S.W., 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res. 38, e159–e159. https://doi.org/10.1093/nar/gkq543

Turnbull, C., Scott, R.H., Thomas, E., Jones, L., Murugaesu, N., Pretty, F.B., Halai, D., Baple, E., Craig, C., Hamblin, A., Henderson, S., Patch, C., O'Neill, A., Devereau, A., Smith, K., Martin, A.R., Sosinsky, A., McDonagh, E.M., Sultana, R., Mueller, M., Smedley, D., Toms, A., Dinh, L., Fowler, T., Bale, M., Hubbard, T., Rendon, A., Hill, S., Caulfield, M.J., 2018. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. BMJ k1687. https://doi.org/10.1136/bmj.k1687

Ungerleider, N., Concha, M., Lin, Z., Roberts, C., Wang, X., Cao, S., Baddoo, M., Moss, W.N., Yu, Y., Seddon, M., Lehman, T., Tibbetts, S., Renne, R., Dong, Y., Flemington, E.K., 2018. The Epstein Barr virus circRNAome. PLoS Pathog. 14, 1–27. https://doi.org/10.1371/journal.ppat.1007206

van Dijk, E.L., Auger, H., Jaszczyszyn, Y., Thermes, C., 2014. Ten years of next-generation sequencing technology. Trends Genet. 30, 418–426. https://doi.org/10.1016/j.tig.2014.07.001

van Dijk, E.L., Jaszczyszyn, Y., Naquin, D., Thermes, C., 2018. The Third Revolution in Sequencing Technology. Trends Genet. 34, 666–681. https://doi.org/10.1016/j.tig.2018.05.008

van Oers, M., Vlak, J., 2007. Baculovirus Genomics. Curr. Drug Targets 8, 1051–1068. https://doi.org/10.2174/138945007782151333

Velculescu, V.E., Zhang, L., Vogelstein, B., Kinzler, K.W., 1995. Serial Analysis of Gene Expression. Science (80-.). 270, 484–487. https://doi.org/10.1126/science.270.5235.484

Volkman, L.E., Summers, M.D., Hsieh, C.H., 1976. Occluded and nonoccluded nuclear polyhedrosis virus grown in Trichoplusia ni: comparative neutralization comparative infectivity, and in vitro growth studies. J. Virol. 19, 820–832. https://doi.org/10.1128/jvi.19.3.820-832.1976

Vrazo, A.C., Chauchard, M., Raab-Traub, N., Longnecker, R., 2012. Epstein-Barr Virus

LMP2A Reduces Hyperactivation Induced by LMP1 to Restore Normal B Cell Phenotype in Transgenic Mice. PLoS Pathog. 8, e1002662. https://doi.org/10.1371/journal.ppat.1002662

Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.P., Li, W., 2013. CPAT: Coding-potential assessment tool using an alignment-free logistic regression model. Nucleic Acids Res. 41, 1–7. https://doi.org/10.1093/nar/gkt006

Wang, Y., Tang, Q., Maul, G.G., Yuan, Y., 2006. Kaposi's sarcoma-associated herpesvirus ori-Lyt-dependent DNA replication: dual role of replication and transcription activator. J. Virol. 80, 12171–86. https://doi.org/10.1128/JVI.00990-06

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63. https://doi.org/10.1038/nrg2484

Watanabe, T., Fuse, K., Takano, T., Narita, Y., Goshima, F., Kimura, H., Murata, T., 2015. Roles of Epstein-Barr virus BGLF3.5 gene and two upstream open reading frames in lytic viral replication in HEK293 cells. Virology 483, 44–53. https://doi.org/10.1016/j.virol.2015.04.007

Weimer, B.C., 2017. 100K Pathogen Genome Project. Genome Announc. 5. https://doi.org/10.1128/genomeA.00594-17

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the Tidyverse. J. Open Source Softw. 4. https://doi.org/10.21105/joss.01686

Woellmer, A., Hammerschmidt, W., 2013. Epstein-Barr virus and host cell methylation: Regulation of latency, replication and virus reactivation. Curr. Opin. Virol. 3, 260–265. https://doi.org/10.1016/j.coviro.2013.03.005

Wu, X., Guarino, L.A., 2003. Autographa californica Nucleopolyhedrovirus orf69 Encodes an RNA Cap (Nucleoside-2′- O)-Methyltransferase. J. Virol. 77, 3430–3440. https://doi.org/10.1128/JVI.77.6.3430-3440.2003

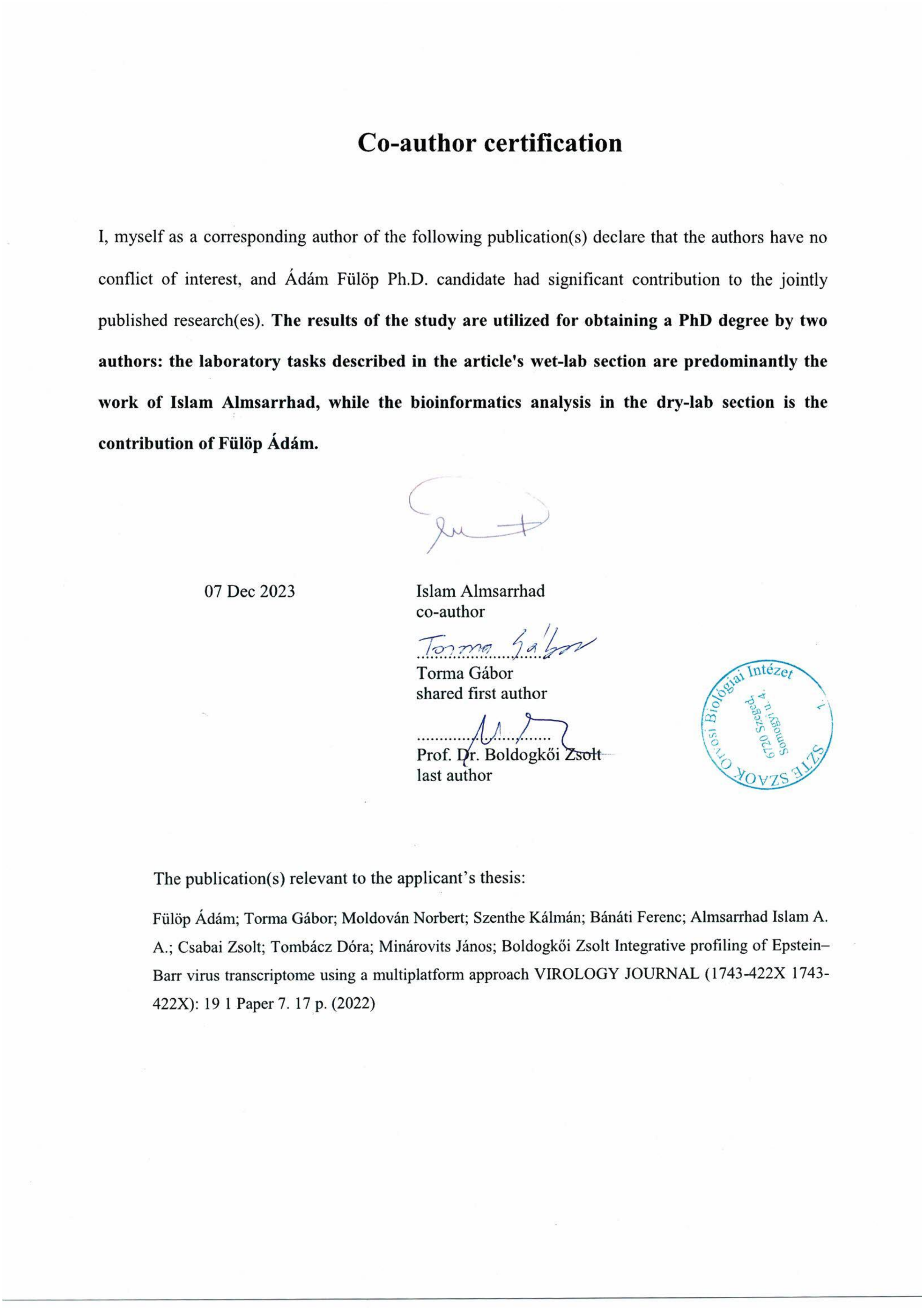
Wyler, E., Menegatti, J., Franke, V., Kocks, C., Boltengagen, A., Hennig, T., Theil, K., Rutkowski, A., Ferrai, C., Baer, L., Kermas, L., Friedel, C., Rajewsky, N., Akalin, A., Dölken, L., Grässer, F., Landthaler, M., 2017. Widespread activation of antisense transcription of the host genome during herpes simplex virus 1 infection. Genome Biol. 18, 209. https://doi.org/10.1186/s13059-017-1329-5

Yang, X., Yang, Y., Sun, B.-F., Chen, Y.-S., Xu, J.-W., Lai, W.-Y., Li, A., Wang, X., Bhattarai, D.P., Xiao, W., Sun, H.-Y., Zhu, Q., Ma, H.-L., Adhikari, S., Sun, M., Hao, Y.-J., Zhang, B., Huang, C.-M., Huang, N., Jiang, G.-B., Zhao, Y.-L., Wang, H.-L., Sun, Y.-P., Yang, Y.-G., 2017. 5-methylcytosine promotes mRNA export — NSUN2 as the methyltransferase and ALYREF as an m5C reader. Cell Res. 27, 606–625. https://doi.org/10.1038/cr.2017.55

Young, L.S., Yap, L.F., Murray, P.G., 2016. Epstein-Barr virus: More than 50 years old and still providing surprises. Nat. Rev. Cancer. https://doi.org/10.1038/nrc.2016.92

Young, M.D., McCarthy, D.J., Wakefield, M.J., Smyth, G.K., Oshlack, A., Robinson, M.D., 2012. Differential Expression for RNA Sequencing (RNA-Seq) Data: Mapping, Summarization, Statistical Analysis, and Experimental Design, in: Bioinformatics for High Throughput Sequencing. Springer New York, New York, NY, pp. 169–190. https://doi.org/10.1007/978-1-4614-0782-9_10

Yuan, J., Cahir-McFarland, E., Zhao, B., Kieff, E., 2006. Virus and Cell RNAs Expressed during Epstein-Barr Virus Replication. J. Virol. 80, 2548–2565. https://doi.org/10.1128/jvi.80.5.2548-2565.2006



Co-author certification

I, myself as a corresponding author of the following publication(s) declare that the authors have no conflict of interest, and Ádám Fülöp Ph.D. candidate had significant contribution to the jointly published research(es). **The results of the study are utilized for obtaining a PhD degree by two authors: the laboratory tasks described in the article's wet-lab section are predominantly the work of Islam Almsarrhad, while the bioinformatics analysis in the dry-lab section is the contribution of Fülöp Ádám.**

07 Dec 2023

Islam Almsarrhad
co-author

Torma Gábor
....
Torma Gábor
shared first author

....
Prof. Dr. Boldogkői Zsolt
last author

SZTE SZAOK Orvosi Biológiai Intézet
1.
6720 Szeged, Somogyi u. 4.

The publication(s) relevant to the applicant's thesis:

Fülöp Ádám; Torma Gábor; Moldován Norbert; Szenthe Kálmán; Bánáti Ferenc; Almsarrhad Islam A. A.; Csabai Zsolt; Tombácz Dóra; Minárovits János; Boldogkői Zsolt Integrative profiling of Epstein–Barr virus transcriptome using a multiplatform approach VIROLOGY JOURNAL (1743-422X 1743-422X): 19 1 Paper 7. 17 p. (2022)

Co-author certification

I, myself as a corresponding author of the following publication(s) declare that the authors have no conflict of interest, and Ádám Fülöp Ph.D. candidate had significant contribution to the jointly published research(es).

07 Dec 2023

....
Dr. Tombácz Dóra
shared first author

Torma Gábor
....
Torma Gábor
shared first author

Moldován
....
Dr. Moldován Norbert
shared first author

....
Prof. Dr. Boldogkői Zsolt
last author

SZTE SZAOK Orvosi Biológiai Intézet
6720 Szeged,
Somogyi u. 4.
1.

The publication(s) relevant to the applicant's thesis:

Torma G, Tombácz D, Moldován N, **Fülöp Á**, Prazsák I, Csabai Z, Snyder M, Boldogkői Z. Dual isoform sequencing reveals complex transcriptomic and epitranscriptomic landscapes of a prototype baculovirus. Sci Rep. 2022 Jan 25;12(1):1291. doi: 10.1038/s41598-022-05457-8 MTMT ID: 32618973