Herpesvirus Transcriptomics as a Module of DNA Replication-Global Transcription Dynamics

Ph.D. Thesis

Islam Almsarrhad M.Sc.

University of Szeged

Faculty of Medicine

Department of Medical Biology

Doctoral School of Multidisciplinary Medicine

Supervisors: Prof. Dr. Zsolt Boldogkői Ph.D., Ds.C. and Dr. Dóra Tombácz Ph.D.

Szeged

2023

Publications directly related to the subject of the thesis

Gábor Torma, Dóra Tombácz, Islam A.A. Almsarrhad, Zsolt Csabai, Gergely Ármin Nagy, Balázs Kakuk, Gábor Gulyás, Lauren McKenzie Spires, Ishaan Gupta, Ádám Fülöp, Ákos Dörmő, István Prazsák, Máté Mizik, Virág Éva Dani, Viktor Csányi, Zoltán Zádori, Zsolt Toth, Zsolt Boldogkői: Identification of Herpesvirus Transcripts from Genomic Regions around the Replication Origins**. Scientific Reports,** 2023 Sep 29;13(1):16395. doi: 10.1038/s41598-023-43344-y.

IF: 4.6

Ádám Fülöp, Gábor Torma, Norbert Moldován, Kálmán Szenthe, Ferenc Bánáti, Islam A. A. Almsarrhad, Zsolt Csabai, Dóra Tombácz, János Minárovits and Zsolt Boldogkői: Integrative profiling of Epstein–Barr virus transcriptome using a multiplatform approach. **Virology Journal** 19:7.(2022) PMID: 34991630. PMC8740505. DOI: 10.1186/s12985-021-01734-6

IF: 5.913

Abbreviations

asRNA: antisense RNA
BoHV-1: Bovine alphaherpesvirus 1
CDS: coding sequence
DBP: DNA-binding protein
dcDNA-Seq: direct cDNA sequencing
DNP: DNA polymerase
dRNA-Seq: direct RNA sequencing
E: early
EBV: Epstein-Barr virus
EHV-1: Equid alphaherpesvirus
ES: unique short
HCMV: Human cytomegalovirus
HHV-6: Human herpesvirus 6
HSV-1: Herpes simplex virus 1
ICP: infected cell polypeptide
IE: immediate-early
IR: inverted repeat
IRL: internal repeat of UL region
IRS: internal repeat of US region
L/ST: L/S junction-spanning transcript
L: late
LAT: latency-associated transcript
LLT: long latency transcript
lncRNA: long noncoding RNA
LRS: long-read sequencing
miRNA: micro RNA
ncRNA: non-coding RNA
ONT: Oxford Nanopore Technologies
ORC: origin recognition complex
ORF: open reading frame
Ori: replication origin
PacBio: Pacific Biosciences
PRV: Pseudorabies virus
raRNA: replication origin-associated RNA
RNP: RNA polymerase
SRS: short-read sequencing
SVV: Simian varicella virus
TES: transcript end site
TF: transcription factor
TI: transcript isoform
TO: transcriptional overlap
TR: transcription regulator
TRL: terminal repeat of UL region
TRS: terminal repeat of US region
TSS: transcript start site
UL: unique long
UTR: untranslated region
VZV: Varicella-zoster virus
αHV: alphaherpesvirus
βHV: betaherpesvirus
γHV: gammaherpesvirus

Table of contents

1. INTRODUCTION

1.1 Regulation of herpesvirus transcription

The lytic transcription of herpesviruses follows a sequential order, which is divided into three distinct temporal phases: immediate-early (IE), early (E), and late (L)[1]. Cellular and viral factors complexly regulate the expression of herpesvirus genes during productive infection. Gene expression of the early infection process in herpes simplex virus type 1 (HSV-1), a representative member of alphaherpesviruses (αHVs), is regulated by four immediate early (IE) proteins. The essential transcription regulator of HSV-1 is the *ICP4* viral protein [encoded by rs1 (*icp4*)], which attracts cellular factors that influence transcription (TFs; e.g., TFIID) to viral promoters to enhance (or sometimes repress) the initiation of transcription[2]. *ICP22* (encoded by *us1*) has been demonstrated to enhance the transcription extension of viral RNAs[3]. The *us1* gene of HSV-1 is found in a single copy in the unique short (US) region of the genome while its promoter is duplicated in the inverted repeat (IR) region (the other IR copy regulates the expression of the *ul12* gene). In the Varicellovirus genus, the *us1* gene is translocated to the IR region where it is duplicated. *ICP0* (encoded by *rl2* [*icp0*)], strictly speaking, is not a TF because it does not attach to DNA or other TFs. This viral protein can increase viral gene expression by affecting pre-chromatin interactions before histones are bound to the viral DNA[4]. *ICP27* (encoded by *ul54*), like the other viral proteins mentioned above, has multiple functions. It is involved in engaging the RNA polymerase (RNP) to viral promoters[5], and also in regulating gene expression and DNA synthesis after transcription[6].

Other viruses that belong to the same group as HSV-1 (αHVs) have a similar way of controlling how they break down cells, during the lytic cycle, by transcription regulation. The main difference is how fast they make some of their genes (*rl2*, *us1*, and *ul54* orthologs) work: these genes work faster in pseudorabies virus (PRV) and equid alpha herpesvirus type 1 (EHV-1) than in HSV-1; the advancement is their expression during the early stage (E) through the evolution. Also, the *icp0* gene was structurally and functionally simplified in these viruses. The other subfamilies, Beta herpesviruses (βHVs) and Gamma herpesviruses (γHVs), have a similar way of controlling the genome-wide viral transcription. In human cytomegalovirus (HCMV), the prototype member of βHVs, the major IE genes (*ie1* and *ie2*) regulate global viral transcription[7]. In Epstein-Barr (EBV), which is a representative member among γHVs, two IE proteins (*BZLF1* and *BRLF1*) control how the next group of viral genes work by transactivating the E genes- transcription[8].

Another scenario employed by herpesviruses involves the establishment of latency, during which most of the viral genetic material remains inactive in terms of transcription, with only a limited number of specific viral RNAs being actively expressed. During the latent phase of HSV-1 infection, the only viral gene that is highly active is the latency-associated transcript (*LAT*)[9]. This non-coding RNA (ncRNA) represses the lytic gene expression by blocking the activity of *icp4*[10] and facilitating heterochromatin formation on the HSV-1 genome[11]. *LAT* has many potential coding regions, but none of them seem to code any proteins[12,13]. There are also other long non-coding RNAs (lncRNAs) that do not code for proteins and are expressed during latency, such as the long-latency transcript (*LLT*; overlapping both the *icp0* and *icp4* genes)[14], and the L/S junction-spanning transcripts (L/STs; overlapping the *icp34.5* and *icp4* genes[15]).

Members of the non-coding *NOIR*-1 transcript family, described in αHVs [PRV[16], varicella-zoster virus (VZV)[17] and EHV-1[18]] are 3′-coterminal with the *LLT* transcripts and are expressed during the lytic cycle. Another non-coding RNA called *NOIR*-2, transcribed in reverse orientation to *NOIR*-1, is only found in PRV. PRV also has a ncRNA called *ELIE* in the lytic cycle[19]. *ELIE* partially overlaps the long isoform of HSV-1 *L/ST*.

1.2 DNA replication

DNA replication varies significantly among life’s domains, but they also have many common features. Prokaryotic genomes have a single site where DNA synthesis begins (called replication origin; Ori) that is determined by consensus sequences[20], while eukaryotic genomes usually have tens of thousands of Oris that are defined by their chromatin structure[21,22]. Viruses have one or a few Oris, which are specified by a combination of structural properties and sequence specificity (typically AT-rich regions) of the particular DNA segment[23]. The replication of eukaryotic genomes starts with a binding of the origin recognition complex (ORC) to the Ori[24]. The function of ORC is to serve as a platform for the assembly of the replisome, which consists of a wide range of proteins such as DNA helicase, DNA polymerase (DNP), topoisomerase, primase, DNA gyrase, single-stranded DNA-binding protein (ssDBP), RNase H, DNA ligase, and telomerase enzymes.

Several proteins that are essential for DNA replication are encoded by herpesviruses. For instance, an origin-binding protein (OBP) (*ul9*), an ssDBP (*ul29*), two DNPs (*ul30* and *ul42*), and three helicase/primase enzymes (*ul5*, *ul8*, and *ul52*) are coded by HSV-1[25,26]. Some viral factors that play roles in nucleotide metabolism [ribonucleotide reductase (*ul39/40*), thymidine kinase (*ul23*), uracil glycosylase (*ul12*), deoxyuridine triphosphatase (*ul50*), alkaline nuclease (*ul12*)] allowing herpesvirus replication in non-dividing cells[27].

HSV-1 has three Oris: one in the unique long (UL) region (OriL) and two in the inverted repeats (IRs) that flank the US region (OriS) (**See Figure 1**). OriL is located between *ul29* and *ul30*, two E genes that are essential for DNA replication. The 2 copies OriS are surrounded by IE genes: *icp4* and *us1* in the internal repeat of the US region (IRS) and *icp4* and *us12* in the terminal repeat of the US region (TRS). Certain members of Varicelloviruses, such as VZV and Bovine alphaherpesvirus 1 (BoHV-1), do not possess an OriL. In contrast, in other Varicelloviruses like PRV and EHV-1, the location of OriL has shifted to the non-coding region situated between the *ul21* and *ul22* gene pair. HCMV, on the other hand, has only one replication origin known as OriLyt, which is found at a semi-orthologous position, adjacent to *ul57*, which is homologous to HSV-1 *ul29*. Human gammaherpesviruses, including EBV and Kaposi's sarcoma herpesvirus (KSHV), feature two lytic origins (OriLyt-L and OriLyt-R), along with a latent replication origin referred to as OriP in EBV and terminal repeat in KSHV[28–31]. The EBV OriP consists of two primary elements: the dyad symmetry (DS) and the family of repeats (FR). Both of these components contain multiple binding sites for the EBNA-1 protein. When EBNA1 binds to the DS, it acts as an origin, facilitating the recruitment of ORC[32]. During DNA replication in its lytic phase, seven EBV replication proteins are essential. OriLyt is bound by *BZLF1*, which possesses the capability to bind numerous viral replication proteins, thereby triggering the onset of the lytic phase. The *BALF5* protein encodes the catalytic subunit of DNP and interacts with the helicase-primase complex. *BALF2* serves as the single-strand DNA binding protein (ssDBP). *BMRF1* is a DNP accessory subunit that can act as a coactivator for *BZLF1* and forms a complex with *BALF5*, creating the DNP holoenzyme[33].

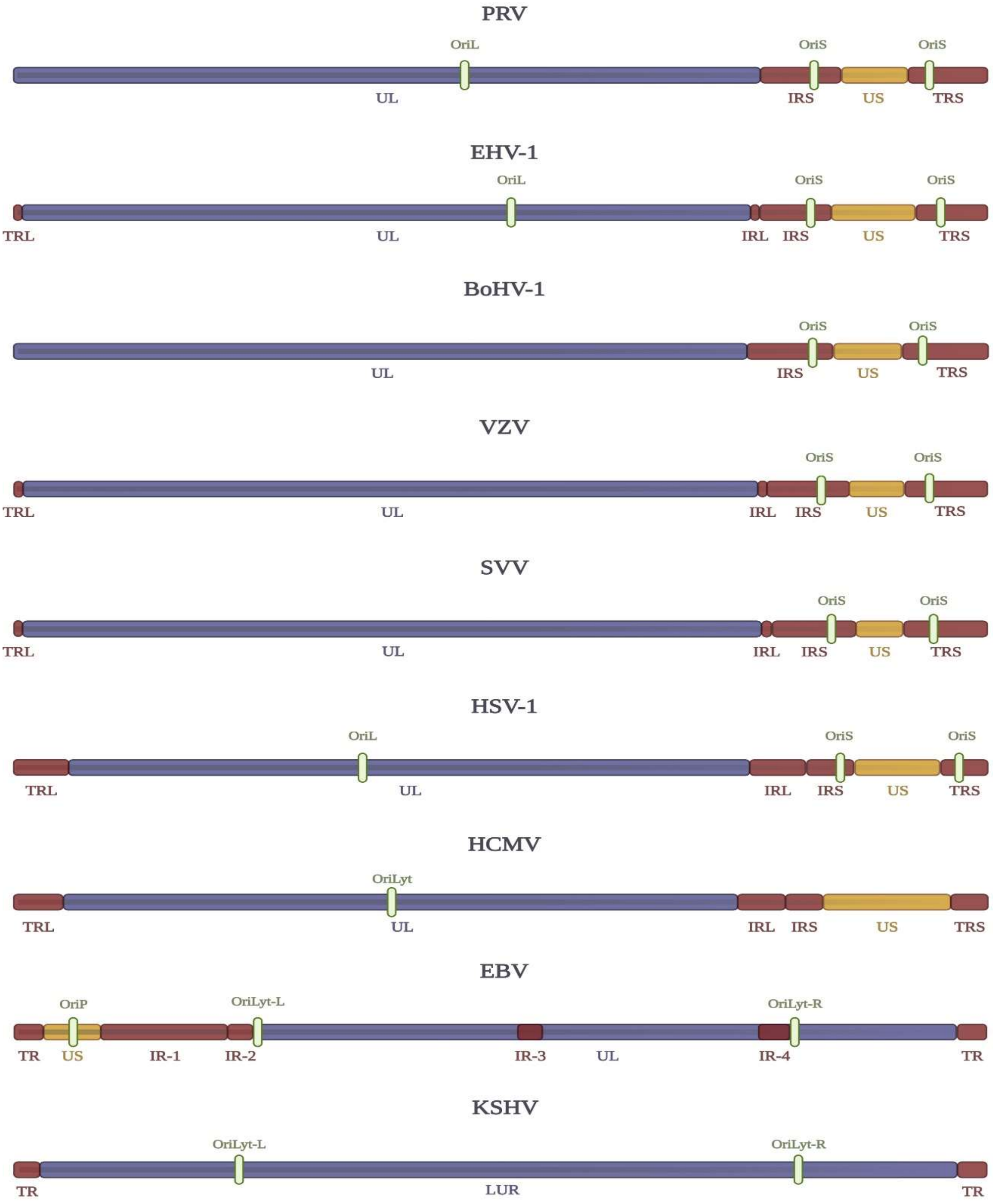


Figure 1. Replication origin positions on herpesvirus genomes: This figure shows the locations of inverted repeats and replication origins in the herpesvirus genomes examined in this study. Abbreviations Ul: unique long, US: unique short, IRS: internal repeat of US region, TRS: terminal repeat of US region, IRL: internal repeat of UL region, TRL: terminal repeat of UL region.

1.3 Overlapping viral transcripts

Recent research studies[34] have revealed that all genes of the herpesvirus exhibit various forms of transcriptional overlaps (TOs), which include divergent (where they are oriented head-to-head), convergent (tail-to-tail), and parallel (tail-to-head) configurations. Tandem genes form parallel-overlapping multigenic, 3′-coterminal transcripts, representing the archetypal genomic organization of herpesviruses. Moreover, many viral genes express 5′-truncated transcripts with different transcription start sites (TSSs) but the same transcription end site (TES), containing 'nested' open reading frames (ORFs) that encode N-terminally-truncated polypeptides[35,36]. Most co-located divergent genes produce 'hard' TOs where the canonical transcripts overlap each other. However, in a few cases, only the long transcript isoforms (TIs), create head-to-head TOs ('soft' TOs), not the canonical transcript. Convergently oriented genes form 'soft' TOs through transcriptional read-through, but only a few 'hard' TOs can be observed (e.g. in αHVs, the *ul7/8*, *ul30/31*, and *ul50/51* gene pairs[37]).

1.4 Non-coding RNAs regulating DNA replication

Some non-coding transcripts, such as short ncRNAs (sncRNAs), (e.g., miRNAs[38]), and long ncRNAs (lncRNAs), have been shown to play critical roles in the regulation of DNA replication[39]. For example, a particular kind of lncRNAs originating from sequences close to the Oris has been identified in all three life domains and viruses in the last ten years. Recent research indicates that about 72% of mammalian ORC1s are linked to active promoters, with over half regulated by ncRNAs[40]. Replication RNAs have several modes for controlling DNA replication. These include the regulation of RNA primer synthesis through hybridization with DNA sequences[9] or the formation of hybrids with mRNAs. This latter process initiates their degradation by RNase H, thereby inhibiting the translation of replication proteins. Additionally, these transcripts can help recruit ORC to the Ori[41].

1.5 Replication origin-associated herpesvirus transcripts

Replication-origin-associated RNAs (raRNAs) have been identified in all three subfamilies of herpesviruses. These transcripts have been previously characterized in βHVs and γHVs but were mostly overlooked in αHVs until recently. One example of a non-coding raRNA in HCMV is RNA4.9, which originates from the OriLyt and has multiple functions in regulating viral DNA replication[42] and gene expression[43]. RNA4.9 can form DNA: RNA hybrids and modulate the level of ssDBP encoded by *ul57*. RNA4.9 may also have other roles in cis and trans, such as repressing the major IE promoter during latency[44]. This discovery indicated that HCMV possesses a distinct method of managing replication, for which there's no proven evidence. Another vital non-coding raRNA crucial for HCMV DNA replication is the SRT (smallest replicator transcript), which is also located at the OriLyt. Two more raRNAs (vRNA-1 and vRNA-2) overlapping the OriLyt have been described in HCMV[45,46]. Rennekamp and Lieberman[47] reported that a ncRNA designated *BHLF1* forms an RNA: DNA hybrid at the OriLyt region of EBV. Additionally, a bidirectional promoter[48] and a highly structured RNA were identified within this region of EBV[44]. The role of this subsequent transcript is to aid the viral EBNA1 and HMGA1a proteins in attracting ORC[49]. Two co-terminal lncRNAs with the same end near the OriS of HSV-1 were also described earlier[50].

Long-read sequencing (LRS) techniques have enhanced transcript research and aided in the identification of new viral transcripts and their TIs, such as splice, TSS, and TES variants. These studies have detected several lncRNAs close to the OriS and OriL regions of αHVs[34,51–55]. However, their precise function remains unknown, so labeling them as 'replication RNAs' would suggest a specific involvement in DNA replication. Moreover, the genes around the Oris produce TIs with long 5′-untranslated regions (5′ UTRs – TSS isoforms), or 3′ UTRs (TES isoforms) that overlap with the replication origin[54,55].

1.6 Epstein-Barr virus

The Epstein-Barr virus (EBV), also known as human gamma herpesvirus 4, belongs to the Gammaherpesvirinae subfamily in the Herpesviridae family[56]. EBV has the ability to cause cancer and is implicated in the development of Burkitt's lymphoma, other lymphomas, nasopharyngeal carcinoma, and certain gastric carcinomas[57,58]. It is classified as a Group 1 carcinogenic agent in humans[59]. Two trans-activator proteins called *BZLF1* and *BRLF1*, produced by the IE genes, play a role in activating the transcription of early genes[59,60]. Lytic EBV DNA synthesis occurs in replication compartments within the host cell nuclei[61]. Unlike the replication of latent episomes, the replication of the viral genome during the lytic cycle starts at one of the two copies of OriLyt, the lytic replication origin of EBV DNA synthesis, leading to unlicensed and exponential amplification[62]. EBV late RNA transcription is aided by the viral preinitiation complex[63]. Initially, it was believed that all viral genes encoded on the approximately 170 kb EBV genome were actively transcribed during the lytic cycle[64–67]. However, recent studies have shown a more complex pattern of viral gene expression during the disruption of EBV latency in various cell lines. It has been discovered that lytic cycle transcription is bidirectional, and many newly identified transcribed regions do not code for proteins[67–71]. This suggests that hundreds of viral lncRNAs may be generated during productive EBV replication. Moreover, new splicing occurrences add to the variety of the EBV transcriptome that is expressed during active replication. [72].

The EBV transcriptome has previously been analyzed using both Illumina-based short read-sequencing and Pacific Biosciences RS II-based long-read sequencing technologies. Since the various sequencing methods have distinct strengths and limitations, the use of multiplatform approaches has proven to be valuable. LRS is more efficient than short-read sequencing (SRS) for determining 5- and 3-UTR isoforms, splice variants, the long RNA molecules, including the polygenic transcripts, as well as the overlapping and embedded transcripts[17,34,73,74]. Compare to ONT platform, the major limitation of PacBio and Illumina approaches are that they are inefficient in reading nucleic acid sequences within the range of 200–800 nucleotides. In this work, we analyzed the EBV lytic transcriptome using the Oxford Nanopore Technologies (ONT) MinION sequencing platform, which is suitable to provide a complete picture on the viral transcriptomic architecture[54,75–77]. The aim of this study is to provide a more complete picture on the transcriptomic architecture of EBV.

2. Objective

In this study, I will illustrate the potential of long-read sequencing technologies for analyzing viral transcriptomes, using herpesviruses as a model system. Herpesviruses are known to generate various transcripts that are located near or overlap with replication origins and adjacent genes associated with transcription or replication, many of which have established or potential regulatory functions. In our research, we utilized both newly generated datasets, which were specifically investigated by our group, as well as previously published long-read sequencing (LRS) and short-read sequencing data to discover additional transcripts located close to Oris in nine herpesviruses belonging to all three subfamilies (alpha, beta, and gamma). While these will serve as the main focal points of my study, for the purpose of meeting the requirements for a Ph.D. demonstration, I will incorporate some aspects of the Epstein-Barr virus (EBV) research in which I was involved, as it is already integrated into this study.

3. RESULTS

3.1 Multiplatform sequencing for characterization of viral transcripts

This study used novel and existing data from sequencing the transcripts of nine herpes viruses that infect humans and animals. Five human and four veterinary pathogenic herpesviruses, which were as follows: six alpha herpesviruses: [a Simplexvirus: HSV-1[78] and five Varicelloviruses: PRV[79]; VZV [77]; BoHV-1[53]; EHV-1[18]; and simian varicella virus (SVV[80]]; as well as a βHV [HCMV[81]]; and two γHVs [EBV[70,82] and Kaposi's sarcoma-associated herpesvirus (KSHV)]. The new sequencing methods were: dcDNA-Seq on ONT MinION device for PRV, EHV-1, and KSHV, dRNA-Seq for EHV-1 and KSHV, and amplified cDNA sequencing for VZV also on ONT device, and SRS on Illumina platform for EHV-1. To identify the TSS regions in EHV-1 and KSHV, we used Cap Analysis of Gene Expression (CAGE) sequencing, CAGE-Seq, on the Illumina platform **(Figure 2)**. We also enriched the capped transcripts by using a Terminator enzyme-based method for both dcDNA-Seq and dRNA-Seq. In addition to the oligo(dT) primer-based RT used for these techniques, we also used random hexamer priming for VZV sequencing. Moreover, we reanalyzed previous herpesvirus transcriptome data from our group or others that were generated by various methods **(Table 1)**: SRS on different Illumina platforms[51,83–86], and LRS on ONT MinION[18], PacBio - RSII and Sequel[17,70], and LoopSeq[53] using a wide range of library preparation techniques and CAGE-Seq for VZV [80] and EBV[70].

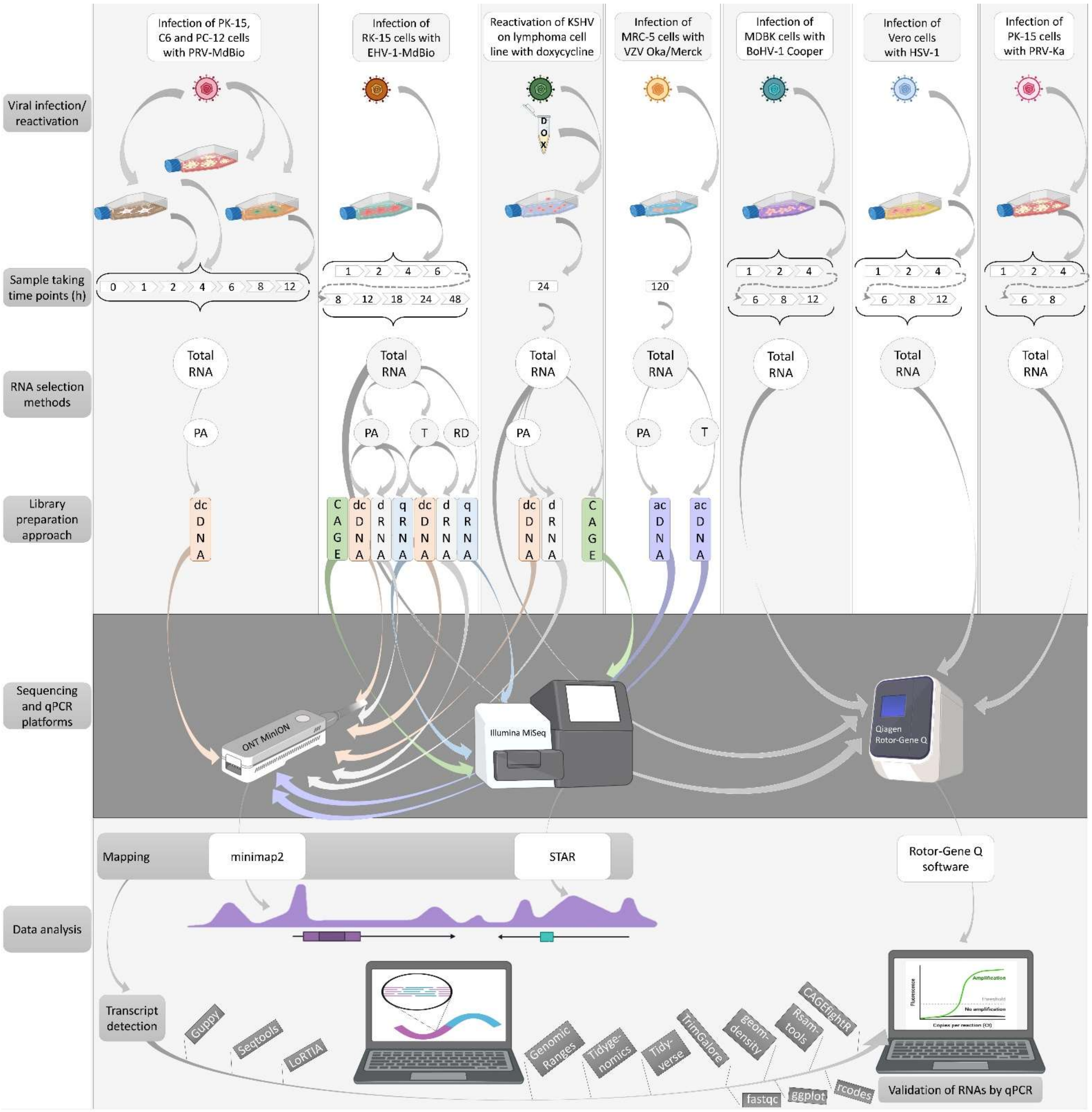

Infection of PK-15, C6 and PC-12 cells with PRV-MdBio
Infection of RK-15 cells with EHV-1-MdBio
Reactivation of KSHV on lymphoma cell line with doxycycline
Infection of MRC-5 cells with VZV Oka/Merck
Infection of MDBK cells with BoHV-1 Cooper
Infection of Vero cells with HSV-1
Infection of PK-15 cells with PRV-Ka
Viral infection/ reactivation
DOX
Sample taking time points (h)
0 1 2 4 6 8 12
1 2 4 6 8 12 18 24 48
24
120
1 2 4 6 8 12
1 2 4 6 8 12
1 2 4 6 8
RNA selection methods
Total RNA
PA
T
RD
Library preparation approach
dcDNA
CAGE
dRNA
qRNA
acDNA
Sequencing and qPCR platforms
ONT MinION
Illumina MiSeq
Qiagen Rotor-Gene Q
Mapping
minimap2
STAR
Rotor-Gene Q software
Data analysis
Transcript detection
Guppy
Seqtools
LoRTIA
GenomicRanges
Tidygenomics
Tidyverse
TrimGalore
geom-density
Rsamtools
CAGEfightR
fastqc
ggplot
rcodes
Amplification
Threshold
No amplification
Fluorescence
Copies per reaction (Ct)
Validation of RNAs by qPCR

Figure 2. Workflow: Here, we outline the techniques employed to generate new sequencing data, including infection of cells with various viruses, library preparation, sequencing and bioinformatics. The qRT-PCR validation workflow for several transcripts is also depicted. Abbreviations: PA: polyadenylated RNAs; T: Terminator-handled samples; RD: ribodepleted RNAs; dcDNA: direct cDNA-seq; CAGE: Cap Analysis Gene Expression-Seq; dRNA: direct RNA-Seq; qRNA: short-read RNA-Seq (library generated by qRNA-seq kit); acDNA: amplified cDNA-Seq.

Table 1. Techniques and datasets from earlier publications used in this study. This table shows the annotated transcripts mapping to the genomic loci examined in this study.

Virus	Sequencing approach	Library	Reference	Data availability
BoHV-1	LRS ONT	direct RNA	Moldován et al., 2020	ENA: PRJEB33511
	LRS ONT	direct cDNA		
	LRS ONT	amplified cDNA		
	Synthetic LRS Illumina	LoopSeq		
	LRS ONT	direct cDNA	Tombácz et al., 2022b	ENA: PRJEB33511
EBV	LRS PacBio RSII	amplified cDNA	O'Grady et al., 2016	GSE: GSE79337
	SRS Illumina	amplified cDNA		
		CAGE		
	LRS ONT	direct cDNA	Fülöp et al., 2022	ENA: PRJEB38992
		amplified cDNA		
HCMV	LRS PacBio RSII	amplified cDNA	Balázs et al., 2017	ENA: PRJEB22072
	LRS PacBio Sequel	amplified cDNA	Balázs et al., 2018	ENA: PRJEB25680
	LRS ONT	amplified cDNA		
	LRS ONT	direct RNA		
	LRS ONT	CAP-selected	Kakuk et al., 2021	ENA: PRJEB25680
HSV-1	SRS Illumina	amplified cDNA	Rutkowski et al., 2015	GEO: GSE59717
		amplified cDNA	Pheasant et al., 2018	SRA: PRJNA505045
		amplified cDNA	Tang et al., 2019	SRA: PRJNA482043, PRJNA483305, and PRJNA533478
		amplified cDNA	Whisnant et al., 2020	GEO: GSE128324
	LRS PacBio Sequel	amplified cDNA	Boldogkői et al., 2018	ENA: PRJEB25433
	LRS ONT	amplified cDNA		
		direct RNA		
	LRS ONT	direct RNA	Depledge et al., 2019	ENA: PRJEB27861
	LRS PacBio RSII	amplified cDNA	Tombácz et al., 2018	GEO: GSE97785
PRV	SRS Illumina	amplified cDNA	Oláh et al., 2015;	ENA: PRJEB9526
	LRS PacBio RSII	direct cDNA	Tombácz et al., 2016	ENA: PRJEB12867
		amplified cDNA		
	LRS PacBio Sequel	amplified cDNA	Tombácz et al., 2018a	ENA: PRJEB24593
	LRS ONT	direct RNA		
	LRS ONT	amplified cDNA		
	LRS ONT	amplified Cap-selected		
	LRS ONT	Terminator-handled amplified cDNA	Torma et al., 2021	ENA: ERP106430 and ERP019579
	LRS ONT	direct cDNA		
	LRS ONT	direct RNA		
SVV	LRS ONT	direct RNA	Braspenning et al., 2021	ENA: PRJEB42868
	SRS Illumina	amplified cDNA		
VZV	LRS ONT	amplified cDNA	Prazsák et al., 2018	ENA: PRJEB25401
		Targeted		
	LRS ONT	amplified Cap-selected	Tombácz et al., 2018b	ENA: PRJEB25401
	SRS Illumina	amplified cDNA	Braspenning et al., 2020	ENA: PRJEB38829.
	SRS Illumina	CAGE		
	LRS ONT	direct RNA		

We annotated the exact locations of the canonical viral RNA transcripts and their TIs in the genome, including TSSs, TESs, and splice variants, in the first part of this study. We validated or adjusted past annotations by combining and reassessing the sequencing datasets. Additionally, we identified cis-regulatory elements for numerous viral RNA transcripts under investigation. Notably, in every instance, promoter elements (TATA boxes) were found within the Ori sequences. We also endeavored to capture an almost comprehensive view of the TOs complexity in the genomic areas we scrutinized. In viruses lacking dRNA-Seq and/or CAGE data, we employed more rigorous standards for transcript annotation, leading to reduced transcript diversity in these viruses. We determined the relative transcript abundances in viruses where enough data were available. Additionally, we used a multi-time-point real-time RT-PCR (RT-PCR) analysis to monitor the expression kinetics of the three most important lncRNAs in PRV.

To confirm the lncRNAs and longer mRNA variants of PRV, BoHV-1, EHV-1, HSV-1, and KSHV, we used RT-PCR. We also used native RNA sequencing to check the cDNA sequencing results. We validated the KSHV and EHV-1 TSSs by ONT sequencing with CAGE-Seq. We acknowledge that the library preparation and the LRS methods may miss or underestimate the expression of long transcripts (>5kb) due to their size-biasing effect. However, our RT-PCR results showed that many of these RNA molecules are expressed at a low level. Due to varying names for orthologous genes across the αHVs, we adopt the HSV-1 nomenclature to enhance consistency in comparison. **Table 2** lists the orthologous gene names.

Table 2. Correspondence of orthologous transcripts.

Gene products	HSV-1	PRV	EHV-1	VZV/SVV	BoHV-1
ICP34.5	RL1	NA	NA	NA	NA
ICP0	RL2	EP0	ORF 63	ORF61	BICP0
UL21	UL21	UL21	ORF 40	ORF38	UL21
gH	UL22	UL22	ORF 39	ORF37	UL22
ICP36	UL23	UL23	ORF 38	ORF36	UL23
UL25	UL25	UL25	ORF 36	ORF34	UL25
ICP8	UL29	UL29	ORF 31	ORF29	UL29
DNA polymerase	UL30	UL30	ORF 30	ORF28	UL30

UL31	UL31	UL31	ORF 29	ORF27	UL31
ICP27	UL54	UL54	ORF 5	ORF4	UL54
ICP4	RS1	IE180	ORF 64	ORF62	BICP4
ICP22	US1	US1	ORF 65	ORF63	BICP22
US10	US10	NA	ORF 66	ORF64/69	US10
US12	US12	NA	NA	NA	NA

3.2 raRNAs: transcripts proximal to replication origins

Transcripts that overlap or map proximal to the Oris include lncRNAs, as well as long 5′ and 3′ UTR isoforms of mRNAs (**Table 3, Figures 3 and 4**).

Table 3. Transcripts with putative regulatory functions. The numbers represent the count of transcripts identified at a specific time point, treatment, etc. for a transcript. **Study Accession: #1** ERR2112243**, #2** ERR2112244**, #3** ERR2112245**, #4** ERR2112247**, #5** ERR2112248**, #6** ERR2112249**, #7** ERR2112250**.**

Start	**Stop**	**Exon composition**	**Strand**	**Cap**	**cDNA**	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**RSII**	**Sequel**
87,121	90,828	87,121-90,828	-	0	0	0	0	0	0	0	0	0	0	0
91,410	91,785	91,410-91,785	+	0	0	0	0	0	0	0	0	0	0	0
91,746	92,120	91,746-92,120	-	0	0	0	0	0	0	0	0	0	0	0
92,891	93,375	92,891-93,375	-	0	0	0	0	0	0	0	0	0	0	0
93,405	94,695	93,405-94,695	-	0	0	0	0	0	0	0	0	0	0	0
93,420	98,453	93,420-98,453	+	11	0	1	0	2	5	1	2	10	10	1,249
94,662	95,313	94,662-95,313	-	0	0	0	0	0	0	0	0	0	0	0
95,858	96,269	95,858-96,269	-	0	0	0	0	0	0	0	0	0	0	0
96,452	96,752	96,452-96,752	-	0	0	0	0	0	0	0	0	0	0	0
98,446	101,778	98,446-100,409; 100,527-101,778	-	0	0	0	0	0	0	0	0	0	0	1
98,446	100,993	98,446-100,409; 100,527-100,993	-	0	0	0	0	0	2	0	0	0	2	2
98,446	100,501	98,446-100,501	-	1	0	0	0	0	0	0	0	0	0	2
98,446	100,593	98,446-100,593	-	0	0	0	0	0	0	0	0	0	0	2
98,446	100,635	98,446-100,635	-	0	0	0	0	0	0	0	0	0	0	3
98,446	100,993	98,446-100,993	-	0	0	2	0	0	0	6	0	5	13	49

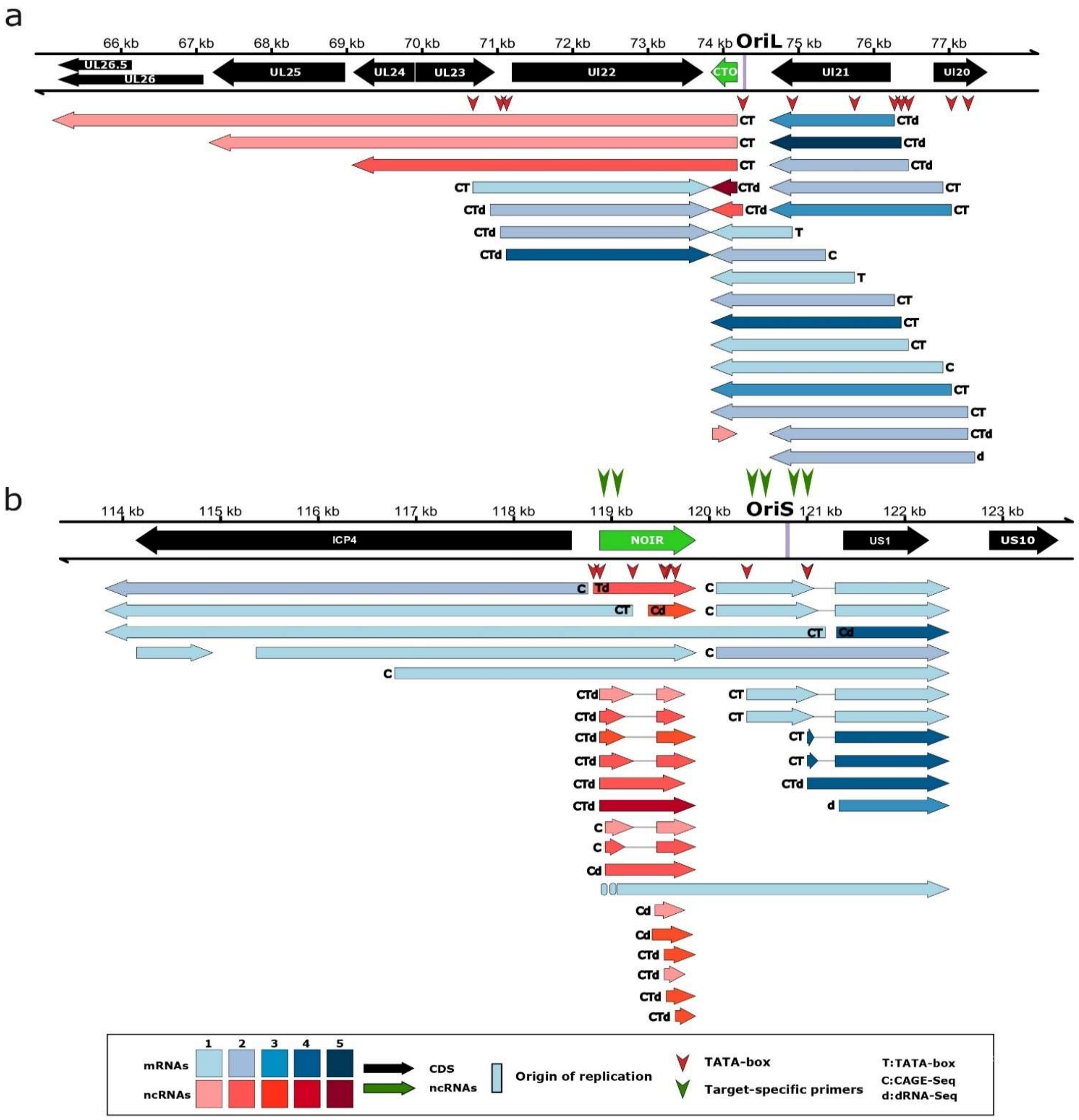


Figure 3. Ori-proximal transcripts of EHV-1: This figure illustrates the transcripts encoded by the OriL- (a) and OriS- (b) proximal regions of Equid alphaherpesvirus 1. For the library preparation, both polyA-selected and ribo-depleted samples were used. However, in both cases, the RNAs were reverse transcribed to cDNA using an oligo(dT) primer. All the putative transcripts were identified by LoRTIA software using dcDNA datasets unless otherwise stated. Protein-coding genes are marked with black arrows, non-coding genes with green arrows, mRNAs with blue arrows, and ncRNAs with red arrows. For better comparability, we use the names for the genes applied in HSV-1 terminology. The relative transcript abundance is indicated by shading. Shades represents relative abundance: 1: 1-9 reads, 2: 10-49 reads, 3: 50-199 reads, 4: 200-999 reads, 5: >1000 reads. CAGE-Seq was performed (transcripts detected by this technique are marked with a 'C').

Transcripts with a proximal TATA box are labeled by a ‘T’ letter, and the vertical red arrows (▼) show TATA box positions on the genome. Transcripts also detected by dRNA-Seq are marked with a ‘d’ letter at upstream positions. Introns are represented by horizontal lines. The position of PCR primers are indicated by vertical green arrows (▼).

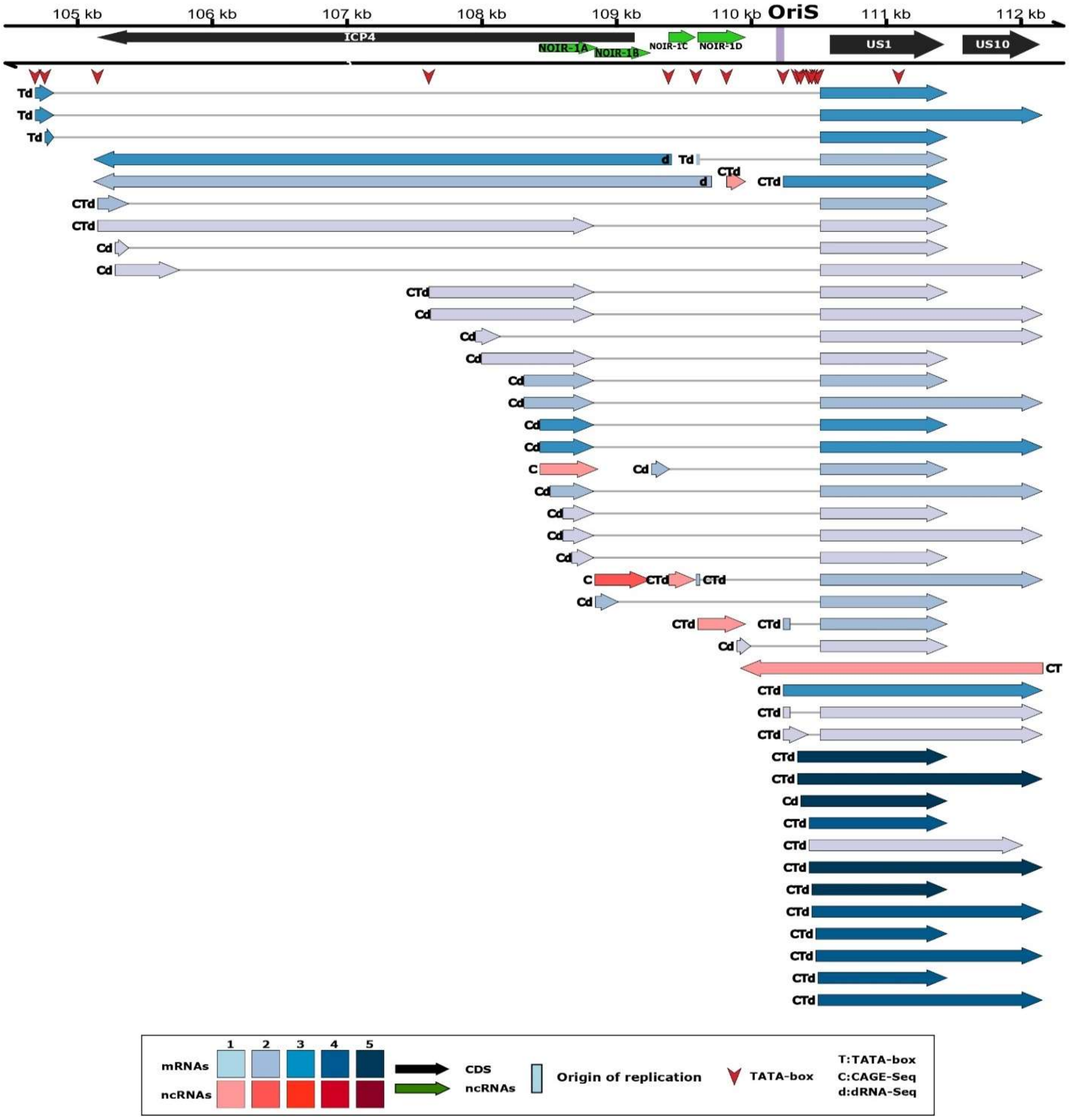


Figure 4. OriL-proximal transcripts of VZV: This figure shows the transcripts encoded by the OriS-proximal region of the Varicella-zoster virus. All putative transcripts were identified by LoRTIA software using dcDNA datasets unless otherwise stated. Protein-coding genes are labeled with black

arrows, non-coding genes with green arrows, mRNAs with blue arrows, and ncRNAs with red arrows. For better comparability, we have adopted the naming convention used for HSV-1 genes. Relative transcript abundance is indicated by shading. Shades represent relative abundance: 1: 1-9 reads, 2: 10-49 reads, 3: 50-199 reads, 4: 200-999 reads, 5: >1000 reads. Transcripts with a proximal TATA box are marked with a 'T' letter, and those detected by dRNA-Seq are marked with a 'd' letter at the upstream positions. Vertical red arrows (▼) show the positions of TATA boxes on the genome. Transcripts detected by CAGE-Seq are marked with a 'C' letter. Horizontal lines represent introns.

3.2.A Alphaherpesviruses – OriS

In this part of the study, we discovered novel transcripts and TIs near the OriSs of α-HVs. The annotated transcripts, including lncRNAs like BoHV-1's OriS-RNA, HSV-1's OriS-RNA1, the *NOIR*-1 transcripts (in PRV, EHV-1, VZV, and SVV), and *NOIR*-2 transcript (in PRV), are shown in **Figures 3-9**. We found that in all examined αHVs, the very long 5′ TIs of transcription regulator genes (*us1* and *icp4*) of BoHV-1, EHV-1, HSV-1, and SVV overlap the OriS. We cannot rule out an opportunity that this is the case for other αHVs also overlapping the OriS, but they might have been missed due to their long size and low abundance. In the HSV-1, we detected a very complex splicing pattern of *US1* transcripts in αHVs. We also identified novel lncRNAs oriented antisense to the HSV-1 OriS-RNA1.

The *NOIR-1* family members exhibit a unique arrangement concerning the *icp4* gene. The standard forms, canonical versions, of these RNAs don't overlap any sequence with *icp4*, whereas the longer *NOIR-1* variants partially overlap with this important TR gene. In viruses like EHV-1, VZV, SVV, and likely PRV, an extended Transcription Initiation (TI) site from *US1* originates from the promoter of the *noir-1* gene. In the case of SVV, we identified an elongated Transcription Start Site (TSS) variant of *NOIR-1* that overlaps with the regular *ICP4* transcript. Both this variant and the typical *NOIR-1* transcript overlap with the OriS region. *NOIR-1* is expressed moderately, while *NOIR-2* has very low levels of expression. In the genomic area of VZV, we identified five long non-coding RNAs (lncRNAs) with distinct TSSs and TESs, labeling them as *NOIR-1A*, -1B, -1C, -1D, and -1E. We observed TIs with TSSs that closely align with the TATA boxes within the OriSs in all six αHVs, indicating the functionality of these promoter elements.

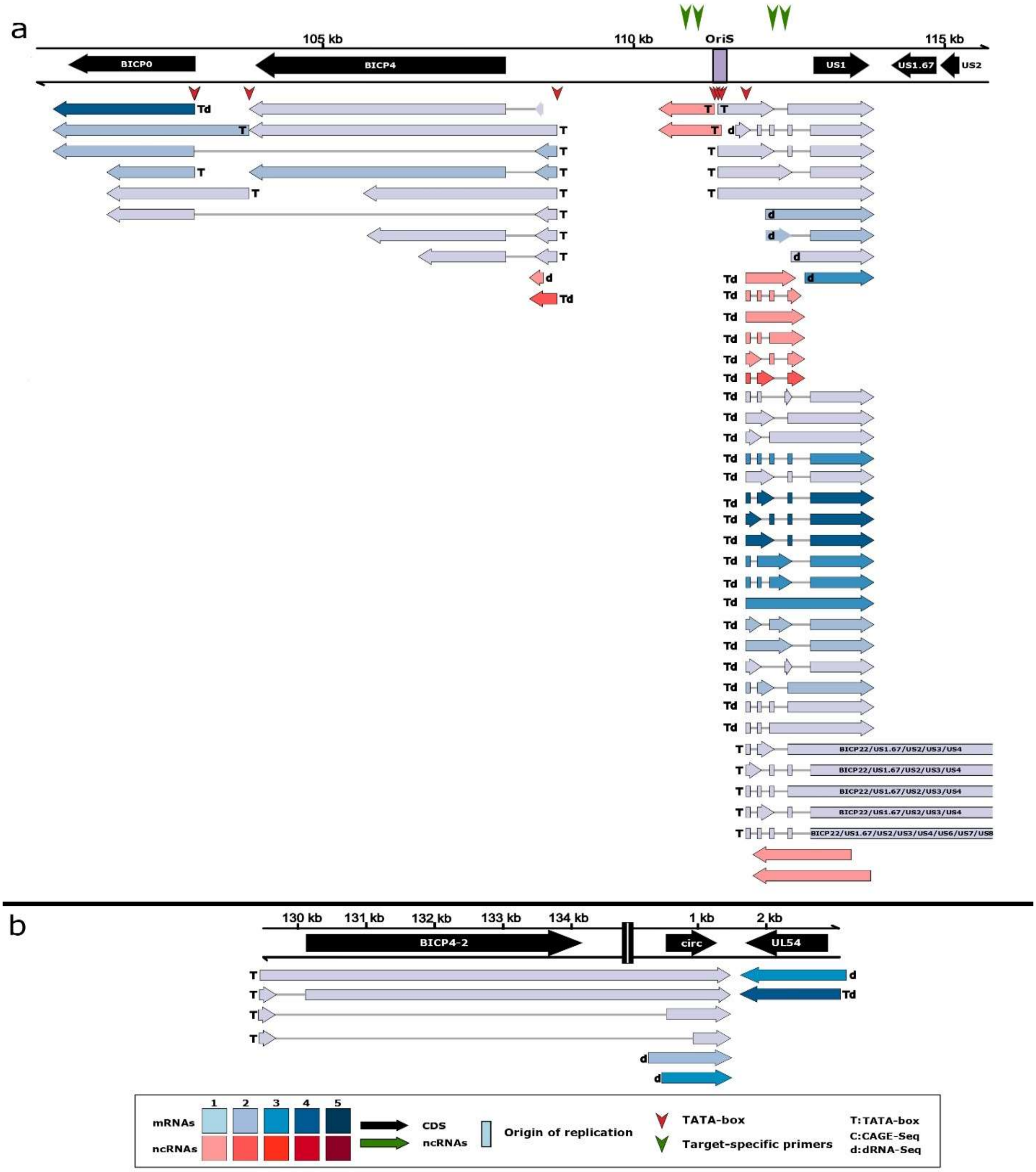


Figure 5. Ori-proximal transcripts of BoHV-1: This image displays the transcripts encoded by the OriS-proximal region of Bovine alphaherpesvirus 1 (IRS: a.; TRS: b.). All potential transcripts were identified by LoRTIA software using dcDNA datasets unless indicated otherwise. Protein-coding genes are labeled with black arrows, non-coding genes with green arrows, mRNAs with blue arrows, and ncRNAs with red arrows. For better comparability, we use the names of the genes applied in HSV-1 terminology. Relative transcript abundance is indicated by shading. Shades represent relative abundance: 1: 1-9 reads, 2: 10-49 reads, 3: 50-199 reads, 4: 200-999 reads, 5: >1000 reads. The coverage

of dRNA-Seq data in BoHV-1 is relatively low therefore, no detection of certain RNAs with this technique does not necessarily mean low reliability. The presence of a proximal TATA box is marked with a ‘T’ letter, and those that were also detected by dRNA-Seq are marked with a ‘d’ letter at the upstream positions. Vertical red arrows indicate the positions of TATA boxes on the genome. Horizontal lines indicate introns.

3.2.B Alphaherpesviruses – OriL

Tombácz and colleagues[17] have reported a group of transcripts called *CTO* that share the same 3′ ends (3′-coterminal). They found three types of *CTO*: *CTO*-S, which is short; *CTO-M*, which starts near the poly(A) signal of the *ul21* gene; and *CTO-L*, which is a transcriptional read-through TI (3′ UTR variant) encoded by the *ul21* gene. We note that the long 3′ UTR isoforms with unique TES are extremely rare in αHV mRNAs. The list of this family has been updated later[55] and a more detailed update is published in this current report. *CTO* transcripts were also found in EHV-1 (**Figures 4, 6**), but not in other herpesviruses with the annotated transcriptome. CTO-S is very abundant in both PRV and EHV-1. We observed a tail-to-tail (convergent) transcriptional overlap (TO) between the 3′ UTR isoforms of *CTO-S* and *UL22* transcripts and identified very long read-through *CTO* transcripts in both viruses. We studied two PRV strains: one from the lab, the laboratory-strain Kaplan, (PRV-Ka[87]), and one from the field (strain MdBio: PRV-MdBio[88]). In HSV-1, both members of the divergent *ul29-ul30* gene pair generate long 5′ UTR variants that overlap the OriL. No lncRNA was detected near the HSV-1 OriL.

3.2.C Betaherpesviruses

We examined the HCMV transcripts near the OriLyt **(Figure 9)** and found that *RNA4.9*[43], a major lncRNA, starts from the OriLyt. We also confirmed the presence of *ul59*[89], *SRT*[45] and *vRNA-2* (one of two *vRNA*s)[46] using our earlier dataset[36]. We found two longer versions of *UL58* lncRNA and a shorter version of *UL59* lncRNA).

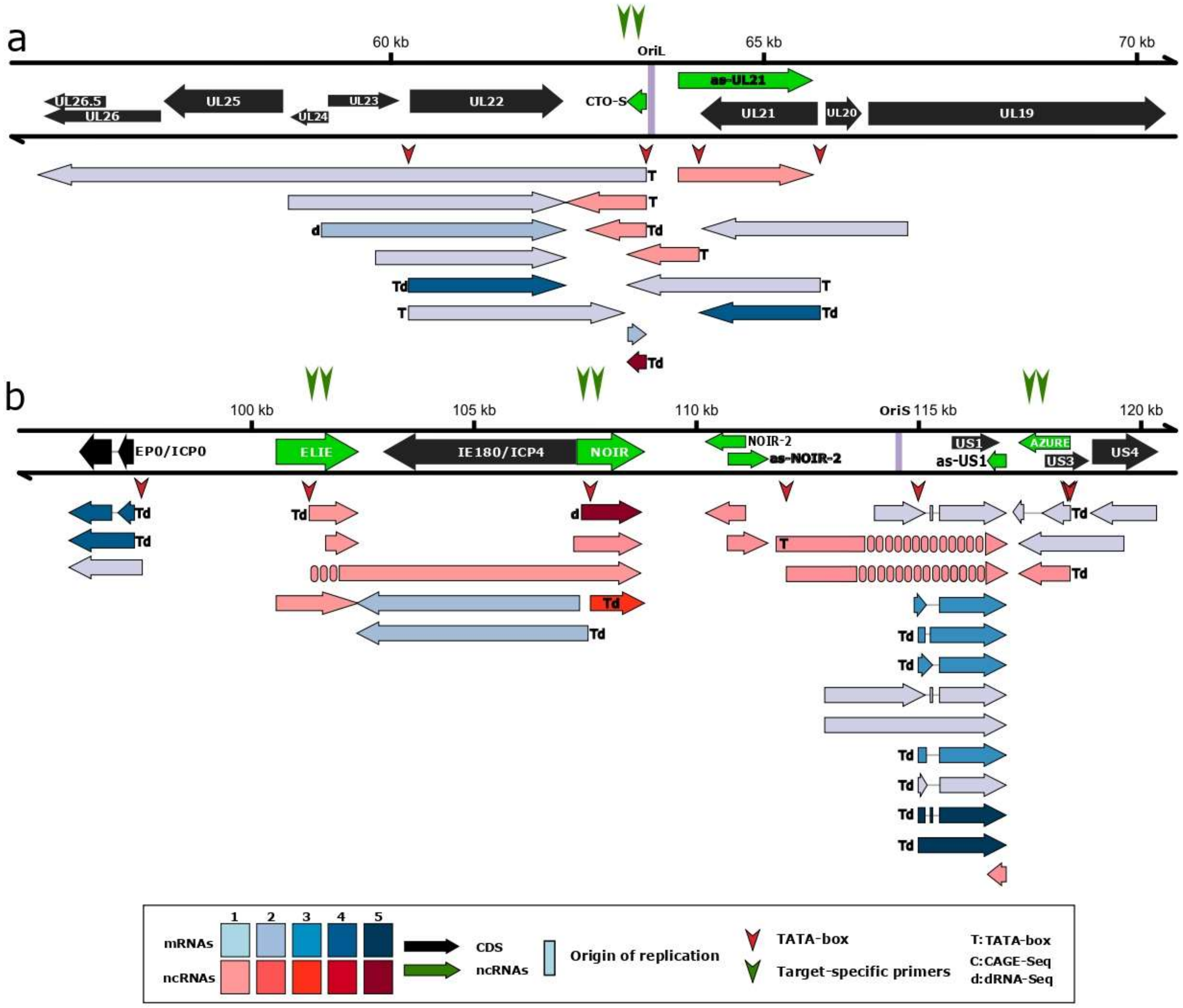


Figure 6. Ori-proximal transcripts of PRV: This image displays the transcripts encoded by the OriL- (**a**) and OriS-proximal regions (**b**) of Pseudorabies virus. All putative transcripts were identified by LoRTIA software using dcDNA datasets unless indicated otherwise. Protein-coding genes are marked with black arrows, non-coding genes with green arrows, mRNAs with blue arrows, and ncRNAs with red arrows. For better comparability, we have adopted the naming convention used for HSV-1 genes. Relative transcript abundance is indicated by shading. Shades represent relative abundance 1: 1-9 reads, 2: 10-49 reads, 3: 50-199 reads, 4: 200-999 reads, 5: >1000 reads. Transcripts with a proximal TATA box are marked by a 'T' letter, and the vertical red arrows indicate the positions of TATA boxes on the genome. Those transcripts that were also detected by dRNA-Seq are marked with a 'd' letter at the upstream positions. Introns are indicated by horizontal lines.

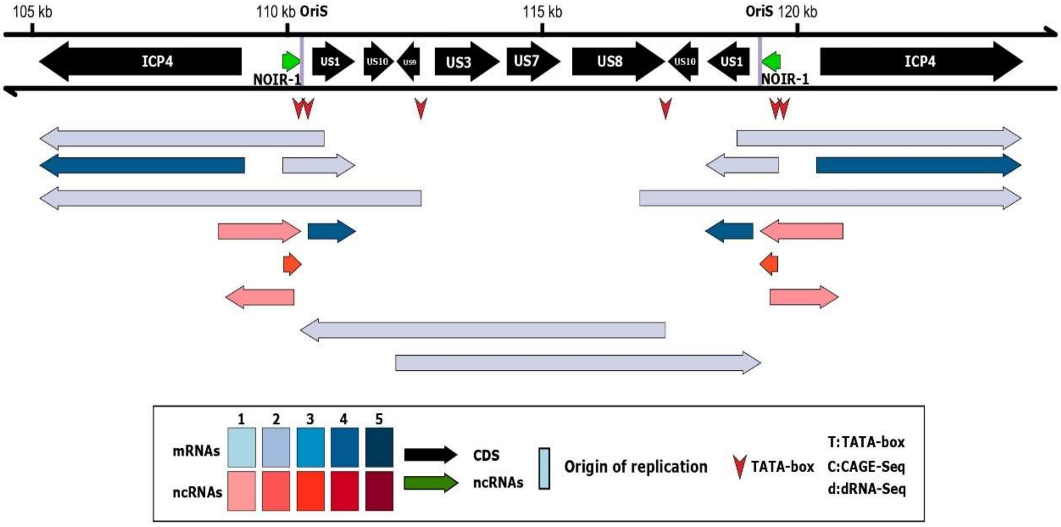


Figure 7. Ori-proximal transcripts of SVV: This figure shows the transcripts encoded by the OriS-proximal region of the Simian varicella virus. All putative transcripts were identified by LoRTIA software using dcDNA datasets unless indicated otherwise. Protein-coding genes are labeled with black arrows, non-coding genes with green arrows, mRNAs with blue arrows, and ncRNAs with red arrows. For better comparability, we employ the HSV-1 naming conventions for the genes. Relative transcript abundance is indicated by shading. Shades represent relative abundance: 1: 1-9 reads, 2: 10-49 reads, 3: 50-199 reads, 4: 200-999 reads, 5: >1000 reads. Vertical red arrows indicate the positions of TATA boxes on the genome.

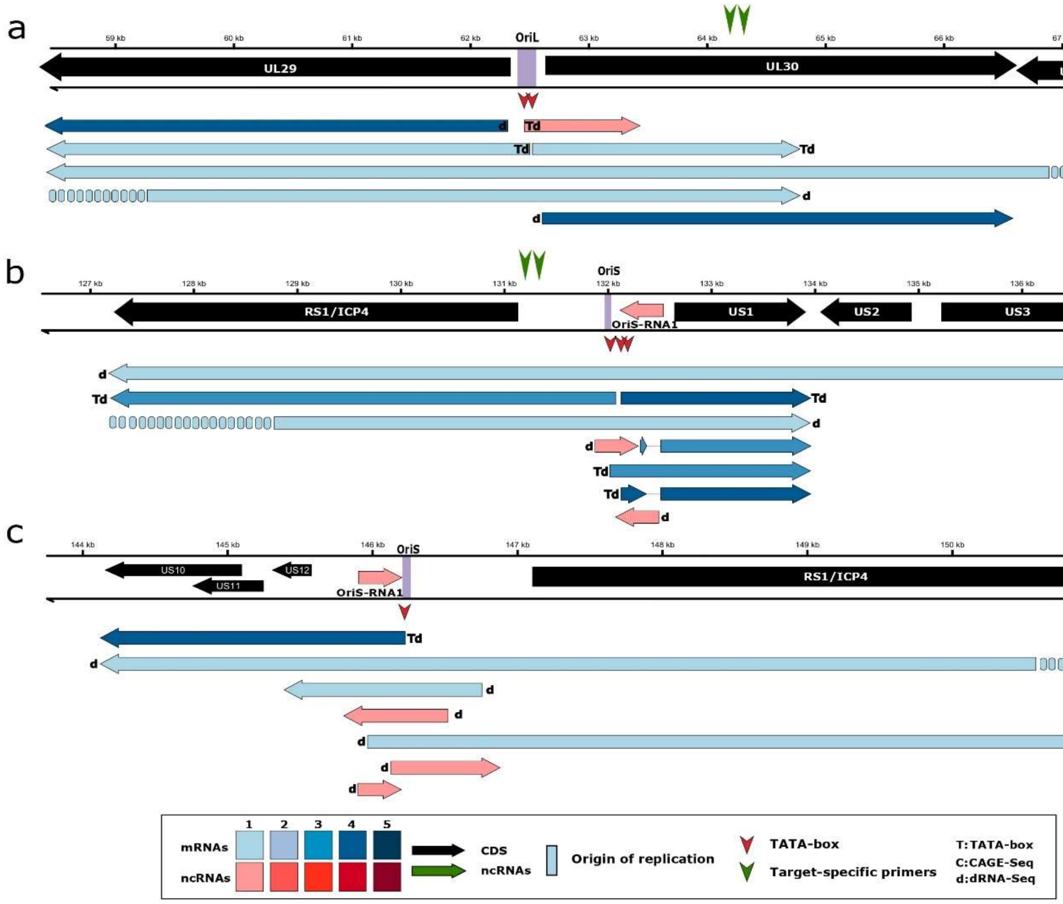


Figure 8. Ori-proximal transcripts of HSV-1: This figure shows the transcripts encoded by the Ori-proximal regions (OriL: a.; OriS of IRS: b.; OriS of TRS: c.) of Herpes simplex virus 1. All the putative transcripts were identified by LoRTIA software using dcDNA datasets unless indicated otherwise. Protein-coding genes are labeled with black arrows, non-coding genes with green arrows, mRNAs with blue arrows, and ncRNAs with red arrows. Relative transcript abundance is indicated by shading. Shades represent relative abundance: 1: 1-9 reads, 2: 10-49 reads, 3: 50-199 reads, 4: 200-999 reads, 5: >1000 reads. Transcripts with a proximal TATA box are marked by a 'T' letter, and vertical red arrows indicate the positions of TATA boxes on the genome. Those transcripts that were also detected by dRNA-Seq are marked with a 'd' letter at the upstream positions. The striped ends of certain arrows (illustration of transcripts) indicate that these terminals have been not or not accurately annotated.

3.2.D Gammaherpesviruses

The long 5′ UTR isoform of EBV *BCRF1* gene overlaps the OriP[90], as shown in **Figure 10a**. Similarly, the long 5′ UTR variants of the *BHRF1* gene overlap OriLyt-L. One of these transcripts is also a spliced form of this gene. The promoter of the *BHLF1* gene is located within the Orilyt-L[47]. We also report novel isoforms of lncRNAs that either have introns that overlap the OriLyt-R or start from the replication origin. We found that many ncRNAs of different lengths associated with OriLyt-L can be made from the same TSS besides the 1.4-kb ncRNA when KSHV is reactivated (**Figure 11**).

The OriLyt-L is surrounded by short genes that code for proteins such as *K4.2, K4.1,* and *K4* on the left and *K5*, *K6*, and *K7* on the right side[91]. Previous studies showed that *K4*, *K4.1,* and *K4.2* are expressed as mono-, bi-, and tri-cistronic mRNAs[92], but our analysis reveals a more complex expression pattern that includes unspliced RNAs of different lengths and spliced RNA variants. We also found that *K5/K6* genes can be expressed not only separately but also through splicing, which produces mRNAs with a first exon of varying length. Importantly, our results agree with previous genomics studies[93–95] but also add to the number of different viral RNA transcripts that can be produced from the OriLyt-L locus, which can potentially increase the coding potential of viral mRNAs. KSHV latency locus is located between *K12* and *LANA* (*ORF73*), which codes for 4 protein-coding latent genes (*K12, K13, ORF72, ORF73*) and 12 pre-miRNAs[95–97]. Here, we detected several lncRNAs that are antisense to the miRNA-coding genomic regions.

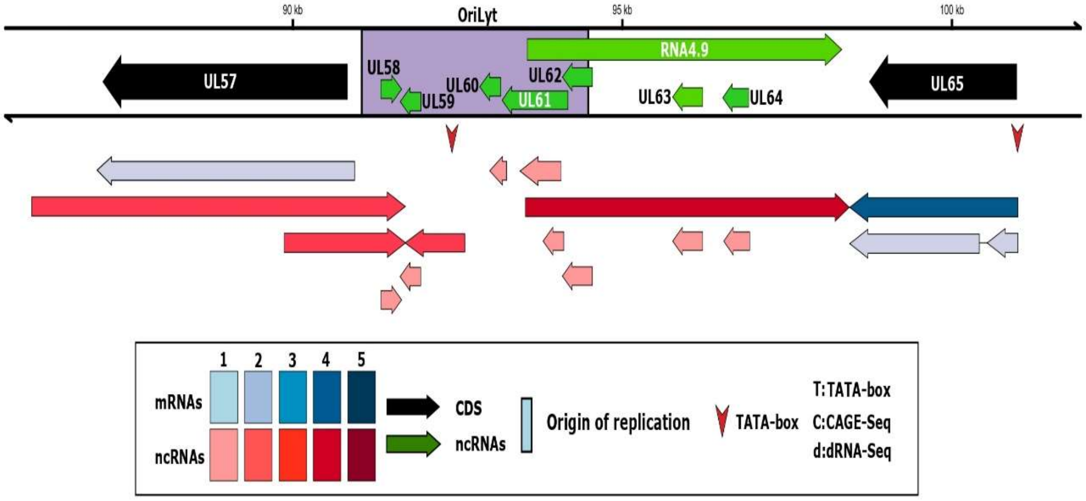


Figure 9. Ori-proximal transcripts of HCMV: This figure shows the transcripts specified by the Ori-proximal region of Human cytomegalovirus. All the putative transcripts were identified by LoRTIA software using dcDNA datasets unless indicated otherwise. Protein-coding genes are labeled with black

arrows, non-coding genes with green arrows, mRNAs with blue arrows, and ncRNAs with red arrows. Relative transcript abundance is indicated by shading. Shades represent relative abundance: 1: 1-9 reads, 2: 10-49 reads, 3: 50-199 reads, 4: 200-999 reads, 5: >1000 reads. Vertical red arrows indicate the positions of TATA boxes on the genome.

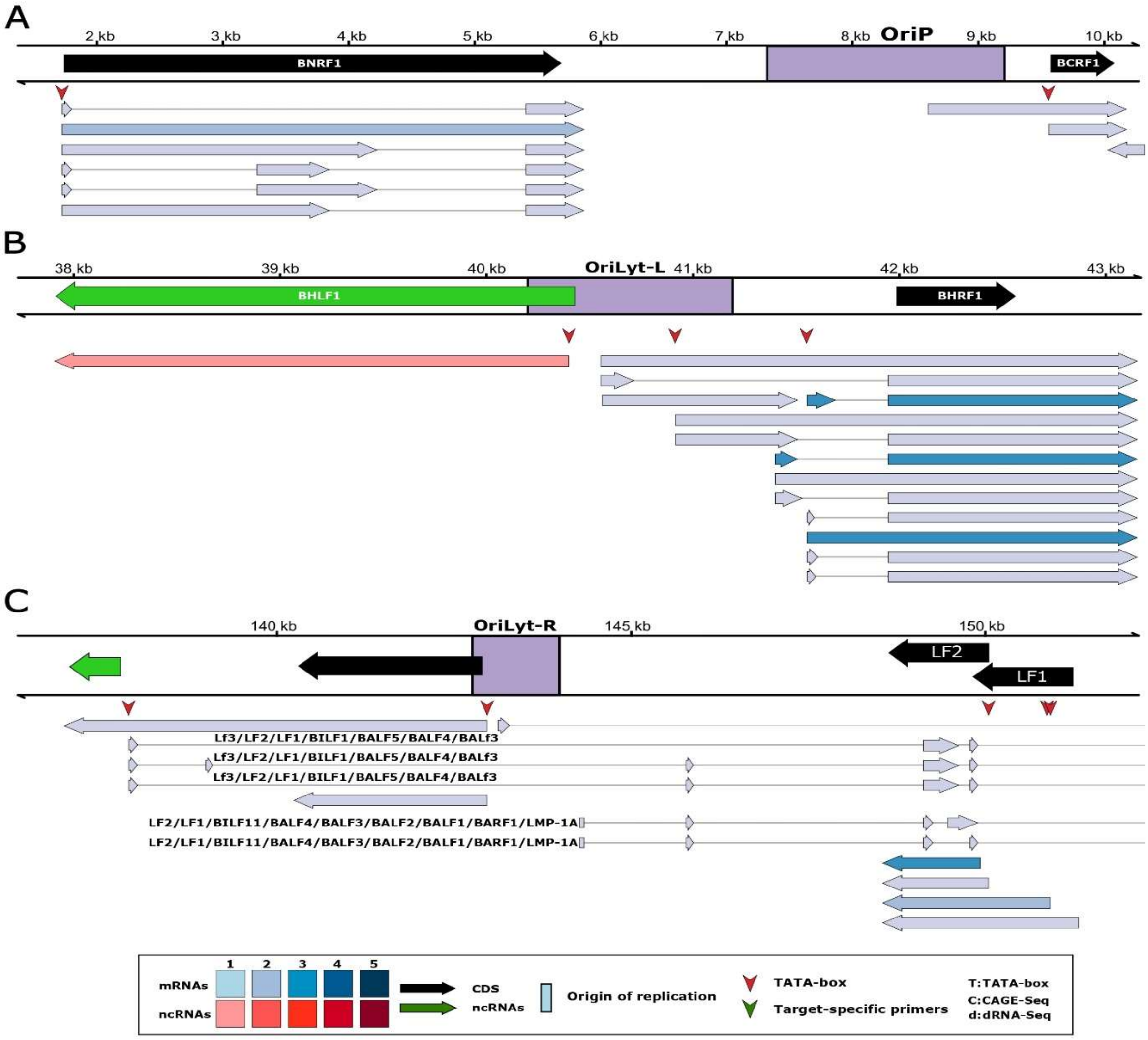


Figure 10. Ori-proximal transcripts of EBV: This figure shows the transcripts specified by the Ori-proximal regions of Epstein-Barr virus (OriP: **a.**; Orilyt-L: **b.**; OriLyt-R: **c.**). All putative transcripts were identified by LoRTIA software using dcDNA datasets unless indicated otherwise. Protein-coding genes are labeled with black arrows, non-coding genes with green arrows, mRNAs with blue arrows, and ncRNAs with red arrows. Relative transcript abundance is indicated by shading. Shades represent relative abundance: 1: 1-9 reads, 2: 10-49 reads, 3: 50-199 reads, 4: 200-999 reads, 5: >1000 reads. Vertical red arrows indicate the positions of TATA boxes on the genome. In the case of very long transcripts, the names of the genes overlapped by these transcripts are enlisted. Introns are indicated by

horizontal lines. All TSS of EBV transcripts have been validated by CAGE analysis, therefore this information is not indicated at the figure.

3.3 RNA isoforms of transcription regulator genes

The *us1* gene produces very long 5′ UTR variants, which were found to form head-to-head TOs with the *ICP4* transcripts in EHV-1 (**Figure 3**), VZV (**Figure 4**), and HSV-1 (**Figure 8**). In HSV-1, the 5′ UTR isoforms of *US10-12* polycistronic transcripts can form a divergent TO with the *icp4* gene. In HSV-1, it was observed that the 5′ UTR isoforms of *US10-12* polycistronic transcripts establish a divergent TO with the *icp4* gene. Furthermore, HSV-1 showed the presence of an extended 5′ UTR in the *ICP4* gene, which overlaps with the *us1* gene. In BoHV-1, a different isoform of the *icp4* gene located in the 3′ UTR was reported to form a parallel TO with the downstream *icp0* gene. However, in this case, the *icp4* ORF is removed from the transcript, resulting in a hybrid RNA that contains the complete *icp0* gene and a section of the 5′ UTR from the *ICP4* transcript[98]. Additionally, *ICP4* transcription initiation sites were found to produce similar hybrid and two-gene transcripts with the BoHV-1 *CIRC* RNA, which is in the nearby genomic region in the circular or concatenated viral genome.

3.4 Non-coding RNAs mapping near the transcription regulator genes

In this study, we identified antisense RNAs (asRNAs) that overlap with the *us1* gene in both BoHV-1 and PRV. *ELIE* was previously distinguished in PRV, but then again, we identified a transcript with an analogous genomic location in EHV-1 (**Figure 3**). *ELIE* is located between the *icp4* and *ep0* genes. It shares one of its TIs with the *NOIR-1* transcripts. We found a similar transcript in EHV-1. We also found an asRNA in EHV-1, named *as64*, that has the same orientation as PRV *ELIE*, but within the *icp4* gene. An HSV-1 transcript that starts at the 3′ end of the *icp4* gene and ends at the *us1* gene is also reported in this study. Moreover, we identified a TSS of a long 5′ UTR variant of the VZV *us1* gene, which is located downstream of *icp4* gene, at the same position as the TSS of PRV *ELIE*. *AZURE* is another lncRNA in PRV located in a reverse direction to the *us1* gene.

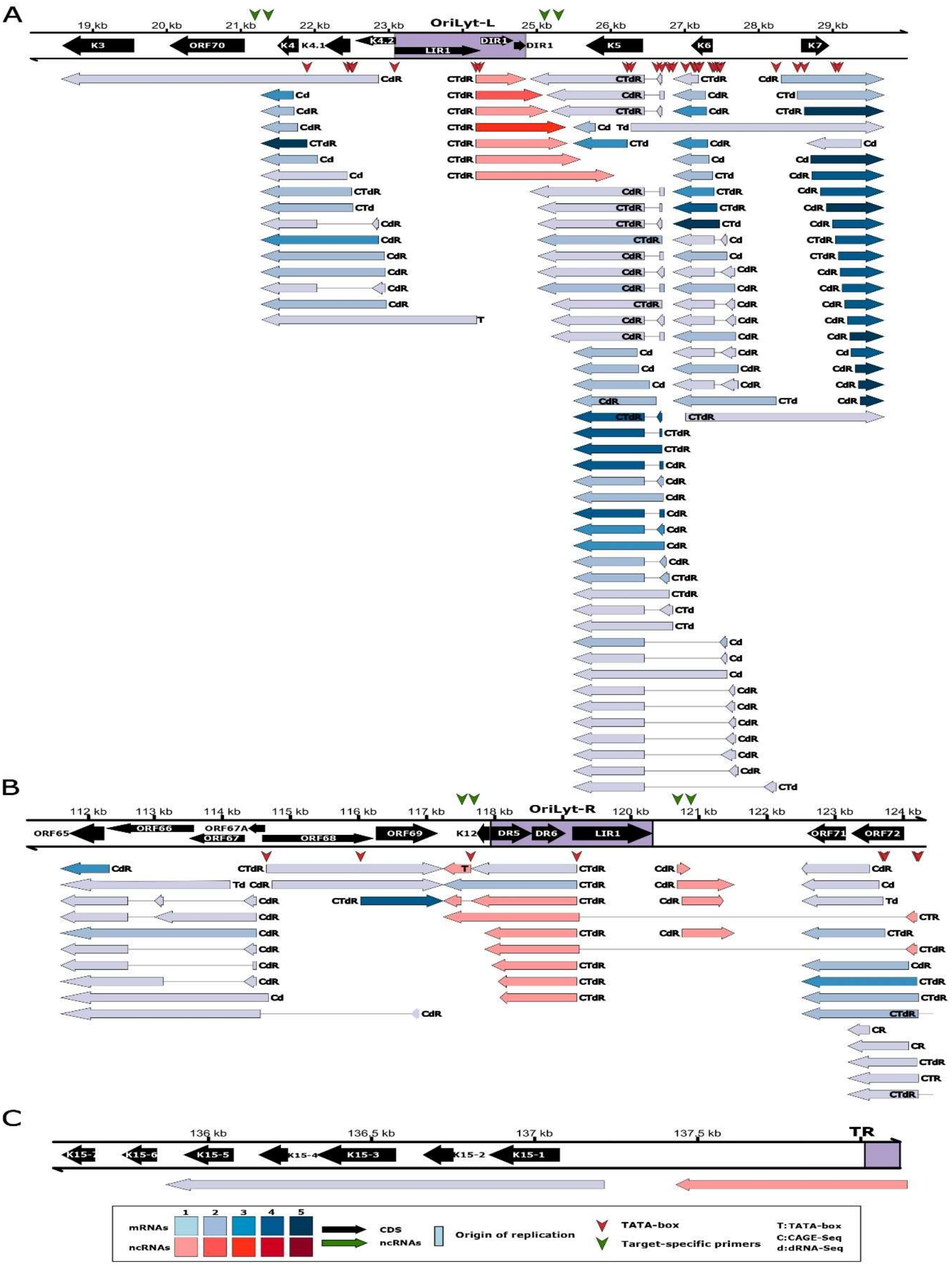


Figure 11. Ori-proximal transcripts of KSHV: This image illustrates the transcripts specified by the Ori-proximal regions of Kaposi's sarcoma-associated herpesvirus (Orilyt-L: a; OriLyt-R: b; TR: c). All putative transcripts were identified by LoRTIA software using dcDNA datasets unless otherwise stated.

Protein-coding genes are marked with black arrows, non-coding genes with green arrows, mRNAs with blue arrows, and ncRNAs with red arrows. Relative transcript abundance is indicated by shading. Shades represent relative abundance: 1: 1-9 reads, 2: 10-49 reads, 3: 50-199 reads, 4: 200-999 reads, 5: >1000 reads. Transcripts with a proximal TATA box are marked by a 'T' letter, and those that were also detected by dRNA-Seq are marked with a 'd' letter at the upstream positions. The vertical red arrows (▼)indicate the positions of TATA boxes on the genome. Transcripts detected by CAGE-Seq are marked by a 'C' letter. RAMPAGE data were also available for KSHV (transcripts detected by this technique are marked by an 'R' letter). Introns are represented by horizontal lines. The position of PCR primers are indicated by vertical green arrows (▼).

3.5 Transcriptional overlaps of replication genes

The long 5′ UTR isoforms of the *ul29-ul30* genes that are situated in divergent orientations in Simplex viruses overlap not only the OriL but also part of each other (**Figure 8**). Interestingly, both genes code for proteins that control DNA replication. In HCMV (*ul57*) and human herpesvirus type 6 (HHV-6) (*ul42*), the *ul29* orthologs are adjacent to the OriLyt. Intriguingly, in αHVs, three 'hard' TOs between gene pairs are present, and one of the partners in these is always a gene playing a role in viral replication. These gene pairs are: *ul30/ul31, ul6-7/ul8-9, ul50/ul51* (*ul30*: DNP; *ul8*: DNA helicase; *ul9*: OBP; *ul50*: deoxyuridine triphosphatase).

3.6 Defining the TSS patterns of examined genomic regions

Identifying the transcription start sites (TSS) of RNA molecules poses a significant challenge in the field of transcriptome research. In our study, we tackled this challenge by employing various long-read sequencing (LRS) and short-read sequencing (SRS) methods. Specifically, we performed CAGE-Seq experiments for EHV-1 and KSHV, as illustrated in **Figure 12**). Additionally, we incorporated CAGE data obtained from other sources for VZV[99] and EBV[70] in our analysis, as depicted in **Figures 3,4, and 13**. To validate our findings, we compared the TSS results from our KSHV CAGE-Seq with the RAMPAGE-Seq results reported by other researchers[100]. Our analysis revealed a total of 199 KSHV transcripts, with 192 confirmed by CAGE and 159 by RAMPAGE. It is worth noting that all the TSSs identified by RAMPAGE were also detected by CAGE.

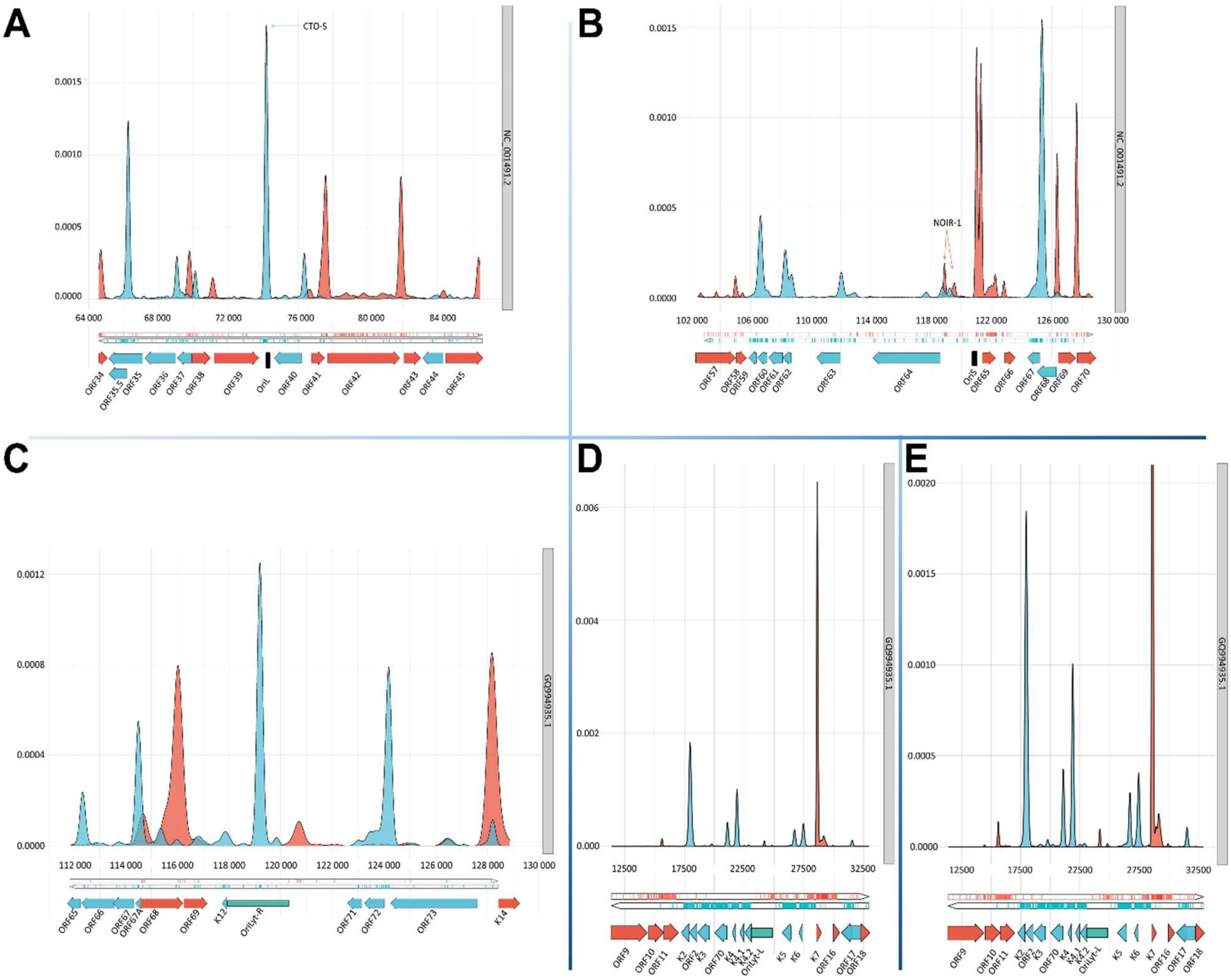


Figure 12. TSS Distribution in examined genomic regions determined by CAGE-Seq: TSS distributions are illustrated in the following genomic regions of EHV-1 and KSHV: a. EHV-1 OriL; b. EHV-1 OriS of the IRS; c. KSHV OriLyt-R; d. KSHV OriLyt-L; e. KSHV OriLyt-L. A higher resolution is used for the better visibility of the low-abundance TSSs. CTO-S transcript is highly expressed (a), whereas NOIR-1 is a group of relatively low-expressed transcripts (b). Smoothed density plots of the 5′ ends in the CAGE data. The y-axis shows the probability estimation of the 5′ ends using a probability density function (details in the Materials and Methods section). Coding sequence annotations for the respective genomes (displayed with the accession number on the right) are visualized in the lower part. Positive strand coverage and the coding sequence annotation are shown in red, while in KSHV they are depicted in blue in the negative strand. The Ori regions in EHV-1 are shown in black, whereas in KSHV they are depicted in green. In the latter case, an accompanying white box displays the 20-nt binding site for the DNA replication origin-binding protein.

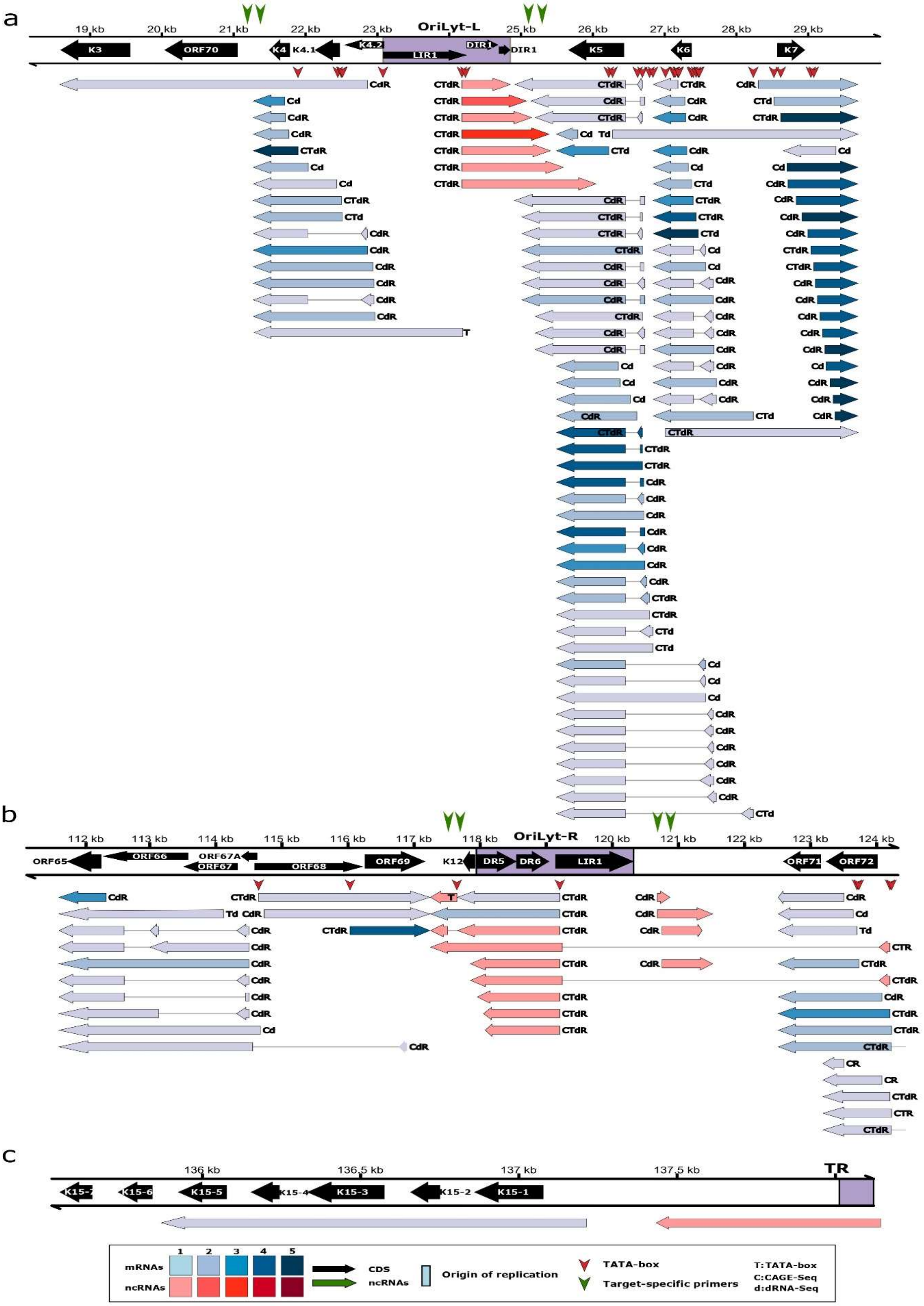

a
19 kb
20 kb
21 kb
22 kb
23 kb
OriLyt-L
25 kb
26 kb
27 kb
28 kb
29 kb
K3
ORF70
K4
K4.1
K4.2
LIR1
DIR1
DIR1
K5
K6
K7
CdR
Cd
CTdR
CTd
Td
T
b
112 kb
113 kb
114 kb
115 kb
116 kb
117 kb
118 kb
OriLyt-R
120 kb
121 kb
122 kb
123 kb
124 kb
ORF65
ORF66
ORF67A
ORF67
ORF68
ORF69
K12
DR5
DR6
LIR1
ORF71
ORF72
CdR
Td
Cd
CTdR
T
CTR
CR
c
136 kb
136.5 kb
137 kb
137.5 kb
TR
K15-7
K15-6
K15-5
K15-4
K15-3
K15-2
K15-1
mRNAs
ncRNAs
1
2
3
4
5
CDS
ncRNAs
Origin of replication
TATA-box
Target-specific primers
T:TATA-box
C:CAGE-Seq
d:dRNA-Seq

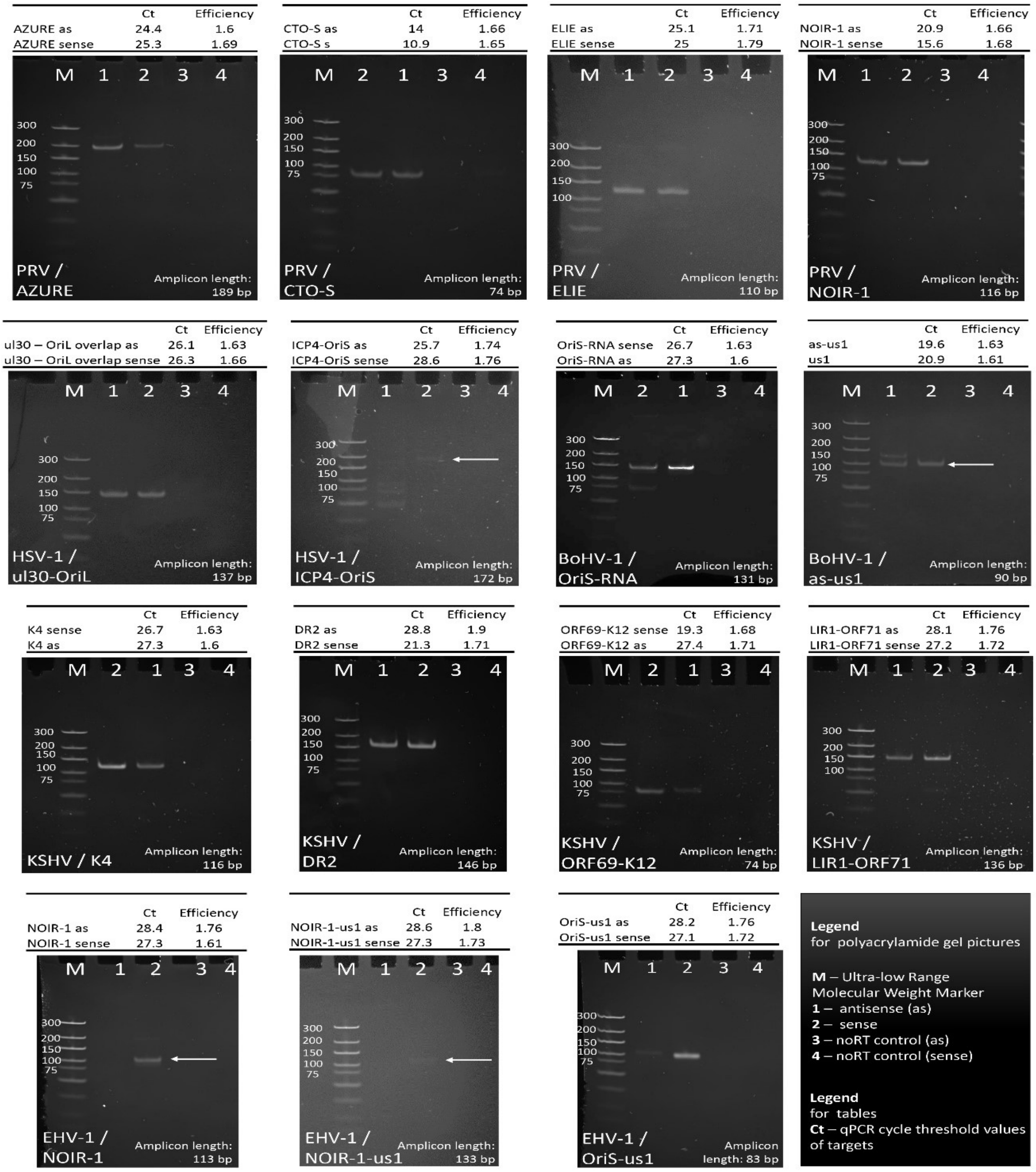
Figure 13. Ori-proximal transcripts of KSHV: This image illustrates the transcripts specified by the Ori-proximal regions of Kaposi's sarcoma-associated herpesvirus (Orilyt-L: a; OriLyt-R: b; TR: c). All putative transcripts were identified by LoRTIA software using dcDNA datasets unless otherwise stated. Protein-coding genes are marked with black arrows, non-coding genes with green arrows, mRNAs with blue arrows, and ncRNAs with red arrows. Relative transcript abundance is indicated by shading. Shades represent relative abundance: 1: 1-9 reads, 2: 10-49 reads, 3: 50-199 reads, 4: 200-999 reads, 5: >1000 reads. Transcripts with a proximal TATA box are marked by a 'T' letter, and those that were also detected by dRNA-Seq are marked with a 'd' letter at the upstream positions. The vertical red arrows (▼) indicate the positions of TATA boxes on the genome. Transcripts detected by CAGE-Seq are marked by a 'C' letter. RAMPAGE data were also available for KSHV (transcripts detected by this technique are marked by an 'R' letter). Introns are represented by horizontal lines. The position of PCR primers are indicated by vertical green arrows (▼).

3.7 Transcript validation using qRT-PCR

Using qRT-PCR (**Figure 14**), 15 transcripts of the viruses (PRV, BoHV-1, EHV-1, HSV-1, and KSHV) were confirmed. The TR genes that overlap the Ori and each other have long TSS isoforms that are expressed at a low level, as undoubtedly indicated by the Ct values.

3.8 Transcription kinetics of three lncRNAs of PRV

In this study, temporal sequence analyses (**Figure 15)** were performed on three PRV lncRNAs (*CTO-S, NOIR-1,* and *AZURE*), early transcripts (*EP0, UL50,* and *UL51*), and late-2 transcripts (*UL19, UL25, UL47,* and *UL48*) using both untreated and phosphor-amino acid (PAA)-treated samples. PAA is used to block DNA replication. analysis was conducted in triplicates using dcDNA-Seq and observed that PAA treatment reduced overall transcript levels; therefore, we adjusted the number of transcripts relative to the total viral RNAs. The effect of the treatment was measured using the PAA/UT ratio, where a value less than 1 indicated late-2 expression kinetics. qRT-PCR analysis was also performed at 12 hours post-infection, which confirmed their dcDNA-Seq findings. Overall, it is evident that PAA treatment exerted the most pronounced effect on L2 genes, manifesting in a notable reduction in their relative abundance within the overall viral reads when compared to untreated specimens. Conversely, early transcripts manifested elevated ratios, especially during later stages, in the PAA-treated samples. While CTO-S exhibited an obvious early kinetic characteristic, the remaining two lncRNAs displayed complex expression trends. Notably, AZURE is characterized by low expression levels, which might explain its irregular expression dynamics.



	Ct	Efficiency
AZURE as	24.4	1.6
AZURE sense	25.3	1.69

	Ct	Efficiency
CTO-S as	14	1.66
CTO-S s	10.9	1.65

	Ct	Efficiency
ELIE as	25.1	1.71
ELIE sense	25	1.79

	Ct	Efficiency
NOIR-1 as	20.9	1.66
NOIR-1 sense	15.6	1.68

	Ct	Efficiency
ul30 – OriL overlap as	26.1	1.63
ul30 – OriL overlap sense	26.3	1.66

	Ct	Efficiency
ICP4-OriS as	25.7	1.74
ICP4-OriS sense	28.6	1.76

	Ct	Efficiency
OriS-RNA sense	26.7	1.63
OriS-RNA as	27.3	1.6

	Ct	Efficiency
as-us1	19.6	1.63
us1	20.9	1.61

	Ct	Efficiency
K4 sense	26.7	1.63
K4 as	27.3	1.6

	Ct	Efficiency
DR2 as	28.8	1.9
DR2 sense	21.3	1.71

	Ct	Efficiency
ORF69-K12 sense	19.3	1.68
ORF69-K12 as	27.4	1.71

	Ct	Efficiency
LIR1-ORF71 as	28.1	1.76
LIR1-ORF71 sense	27.2	1.72

	Ct	Efficiency
NOIR-1 as	28.4	1.76
NOIR-1 sense	27.3	1.61

	Ct	Efficiency
NOIR-1-us1 as	28.6	1.8
NOIR-1-us1 sense	27.3	1.73

	Ct	Efficiency
OriS-us1 as	28.2	1.76
OriS-us1 sense	27.1	1.72

Legend
for polyacrylamide gel pictures

M – Ultra-low Range Molecular Weight Marker
1 – antisense (as)
2 – sense
3 – noRT control (as)
4 – noRT control (sense)

Legend
for tables
Ct – qPCR cycle threshold values of targets

Figure 13. Validation of Ori-adjacent transcripts using qRT-PCR: This figure presents the validation of key Ori-adjacent transcripts using qRT-PCR and gel electrophoresis. The virus and the transcript names are indicated at the appropriate panels. We examined the transcriptional activity of both DNA strands and used no-RT controls for each transcript. The lanes in every panel are as follows. M: molecular weight marker; 1: antisense transcripts; 2: sense transcripts (mRNA, or the canonical lncRNA); 3: no-RT for the

antisense transcripts; 4. no-RT for the sense transcripts. We also indicated the amplicon lengths the Ct and efficiency values, which allow transcript quantity estimation.

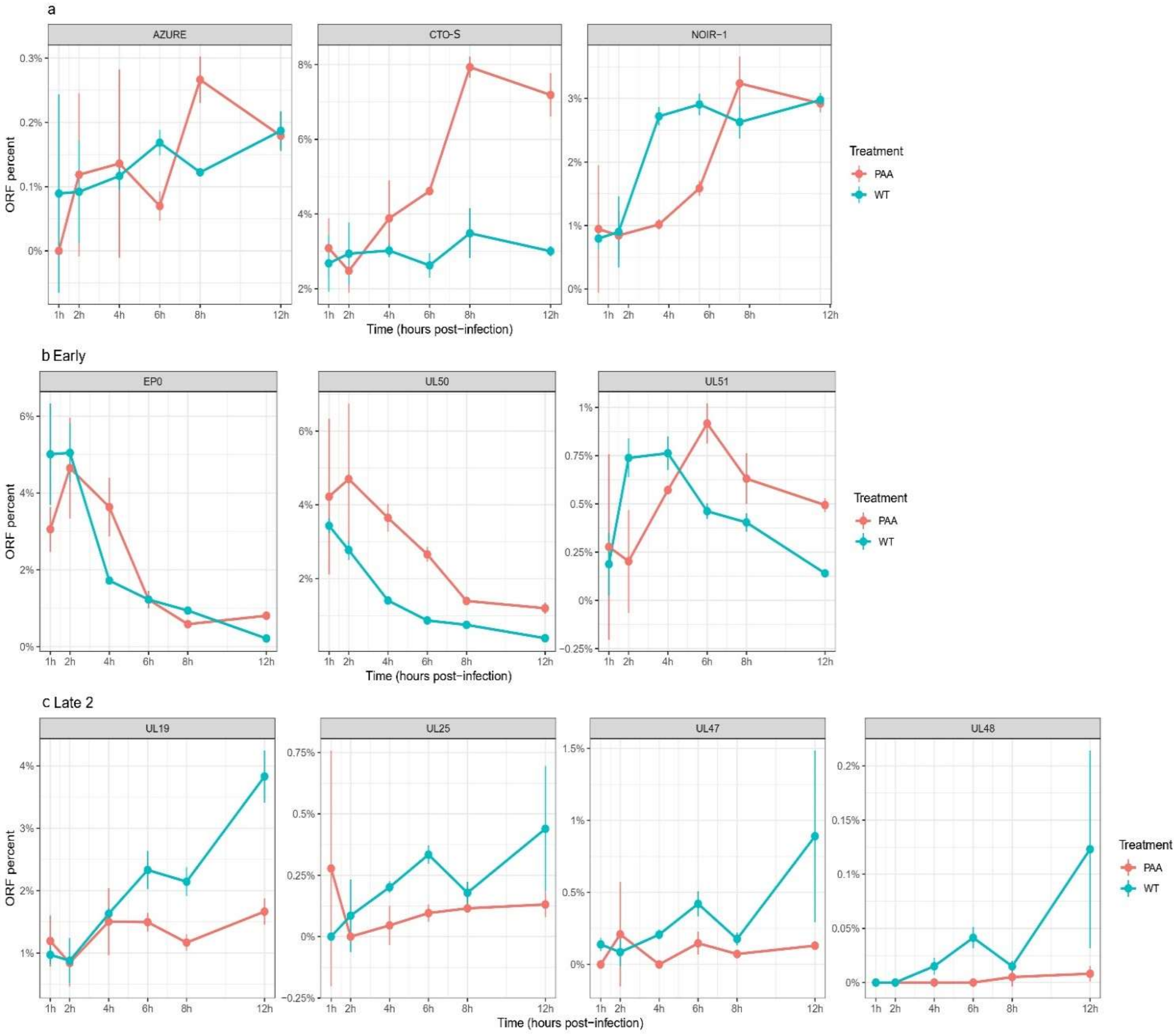


Figure 15. Expression kinetics of PRV transcripts in untreated and in PAA-treated samples

a. Three lncRNAs: AZURE, CTO-S, NOIR-1

b. Three early transcripts: EP0, UL30, UL50, UL51

c. Four late (L2) transcripts: UL19, UL25, UL47, UL48

The ratios of every time points were calculated by normalizing the read count of a specific transcript against the read counts of total viral transcripts in both the untreated and PAA-treated samples. The influence of PAA-treatment was most pronounced on the L2 genes, as their expression is contingent upon DNA replication, a process inhibited by PAA. CTO-S clearly displayed early expression kinetics, whereas other two lncRNAs demonstrated complex expression dynamics.

3.9 Epstein-Barr virus (EBV):

3.9.A Multiplatform profiling of the EBV transcriptome

In this study, we examined the lytic EBV transcriptome using a combination of novel amplified and non-amplified ONT sequencing data, as well as transcriptomic data generated by others using PacBio RSII[70] and Illumina platforms Identification of new viral genes and transcript isoforms during[69–72,101,102]. 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. By integrating data from multiple platforms, we aimed to detect new EBV transcripts and confirm previously described RNA molecules using our LoRTIA program for annotation and filtering out spurious transcripts. We produced cDNA libraries from eight sequential lytic stages, employing both oligo(dT)-primed amplified and non-amplified techniques. Yet, the sparse coverage, especially during the initial stages, made kinetic analysis unviable with this data collection. A total of 22,358 non-amplified and 54,271 amplified reads were mapped to the viral genome, with average mapped read lengths of 838.66 nts and 1098.43 nts, respectively. Other techniques yielded the following read counts: PacBio: 104,469, Illumina Cage-Seq: 3,344,162, and Illumina polyA-Seq: 93,817,061. Additionally, we generated a random hexamer-primed amplified library from pooled samples and sequenced them using the MinION platform.

3.9.B Transcripts ends and alternative ends

This study detected a total of 398 putative TSSs. CAGE-Seq, ONT-MinION, and PacBio datasets were used[70] for the validation of our TSSs. A TSS was accepted if it was present in at least two of our techniques or in one of our techniques and either in the CAGE-Seq or in the PacBio dataset. This strict filtering gives rise to a total of 322 TSS of which 145 are novel (**Figure 16a**). The identified TSSs matched all of which were detected by CAGE-Seq, moreover, 4.66% (14 out of 322) of the TSSs were undetected by CAGE-Seq. Upstream TATA boxes were identified for 20% of the TSSs at an average distance of −31.43 nts (σ=3.31). The nucleotide composition analysis of these start sites revealed a G-rich initiator region (**Figure 16c**). Sixty-two GC boxes were identified with a 64.70 nt average distance from the TSSs. The average distance of the identified 17 CAAT boxes from the TSSs is

110.23 nts. Both the GC and CAAT boxes are promoter consensus elements, which bind specific transcription factors (SP1 and NF-1, respectively). The GC box consensus sequence is as follows: GGGCGG. This sequence is located within 100 nucleotides upstream of the TSSs. The CAAT box consensus sequence is CAA TCT, which is located ~75 nucleotides upstream of the TSSs.

A total of 65 potential TESs was identified using the LoRTIA. A TES was considered valid if it appeared in at least two of our techniques or in either one of our techniques and in the PacBio dataset or the PA-seq dataset[70,72,75,102]. This analysis led to the discovery of 57 TESs, out of which 12 were previously unknown (**Figure 16b**). For 89% of the TESs, polyadenylation signals (PASs) were found, located at an average distance of −24.51 nts (σ=6.99). TESs with a PAS displayed an A-rich cleavage site and a G/T-rich downstream region, bearing resemblance to the mammalian cleavage and polyadenylation motifs[103] (**Figure 16d**).

Numerous transcripts are represented by just a single read, which is beneath detection limit of *LoRTIA*. Owing to their scant presence, the precise transcription starts site (TSS) of these transcripts remains ambiguous, prompting us to categorize them as potential transcript variants. Altogether, we pinpointed 52 potential transcripts, with 33 extending beyond any other overlapping transcripts. However, we believe that with higher data coverage, a much larger population of these transcripts would be revealed. In a previous study using long-read sequencing (LRS), a wide range of TSSs and transcription end sites (TESs) for several EBV genes were disclosed[70]. In this study, we discovered 104 novel 5′-UTR isoforms, with 47 having longer and 57 having shorter 5′-UTRs. The CAGE-Seq data analysis confirmed 98% of our longer TSS isoforms and 92.98% of the shorter TSS isoforms.

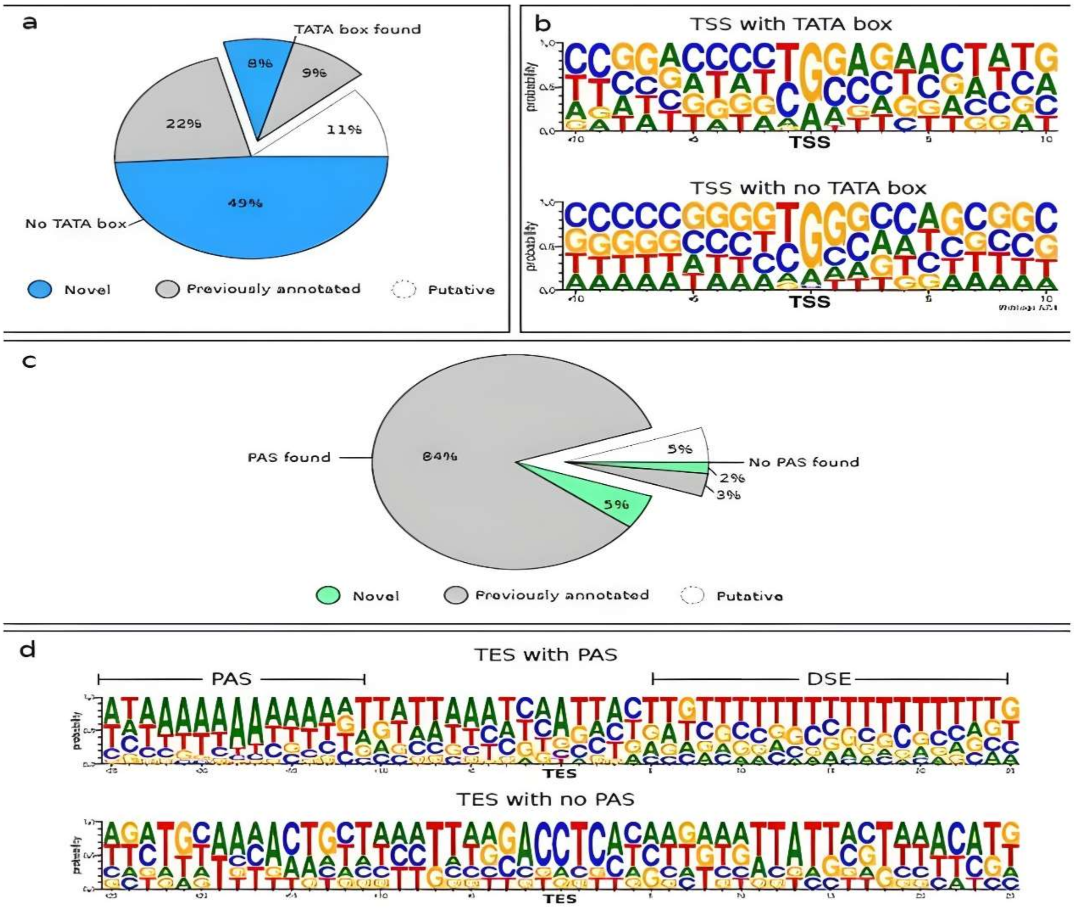


Figure 16. TSS and TES proportions and sequence motifs. The proportion of previously annotated, novel and putative (a) TSSs and (c) TESs with the extracted slices represent the proportion of TSSs or TESs with TATA box or PAS respectively, while the rest of the pie chart represent the proportion of TSSs or TESs without a TATA box or PAS. The sequence surrounding the (b) TSS shows a G-rich initiator region for TSSs with a TATA box. TSSs lacking a TATA box have a G-rich +1 and +2 position. (d) TESs with a PAS have a canonical A-rich cleavage site and a canonical GU-rich downstream element (DSE), while those without a PAS showed a C-rich cleavage with no recognizable DSE.

The 5′-UTRs possess the ability to regulate translation through various mechanisms, including secondary structures[104], upstream AUGs (uAUGs), or upstream ORFs (uORFs)[105,106]. In a study by Watanabe *et al.*[107], they investigated the impact of two uORFs upstream of the *BGLF3.5* ORF on the translation of *BGLF4*, a protein kinase that plays a role in replication and nuclear regress[108]. Surprisingly, they found that point mutations in these two upstream ORFs (uORFs) did not affect the protein levels of *BGLF4*.

Within the *BGLF3.5-BGLF4* cluster, the most abundant transcript is *BGLT16*, which is a bicistronic mRNA containing a wild-type uAUG upstream (**Figure 17**). However, during our analysis, we identified two short 5′-UTR isoforms of this transcript (*BGLT23* and *BGLT25*) that only carry the *BGLF4* gene. These RNA molecules lack the uORFs that were mutated by Watanabe and colleagues[107]. Additionally, we discovered *BGLT24*, a longer 5′-UTR isoform of *BGLT16*, which contains additional wild-type uAUGs and uORFs located upstream of the point mutations introduced by Watanabe and coworkers (**Figure 17**).

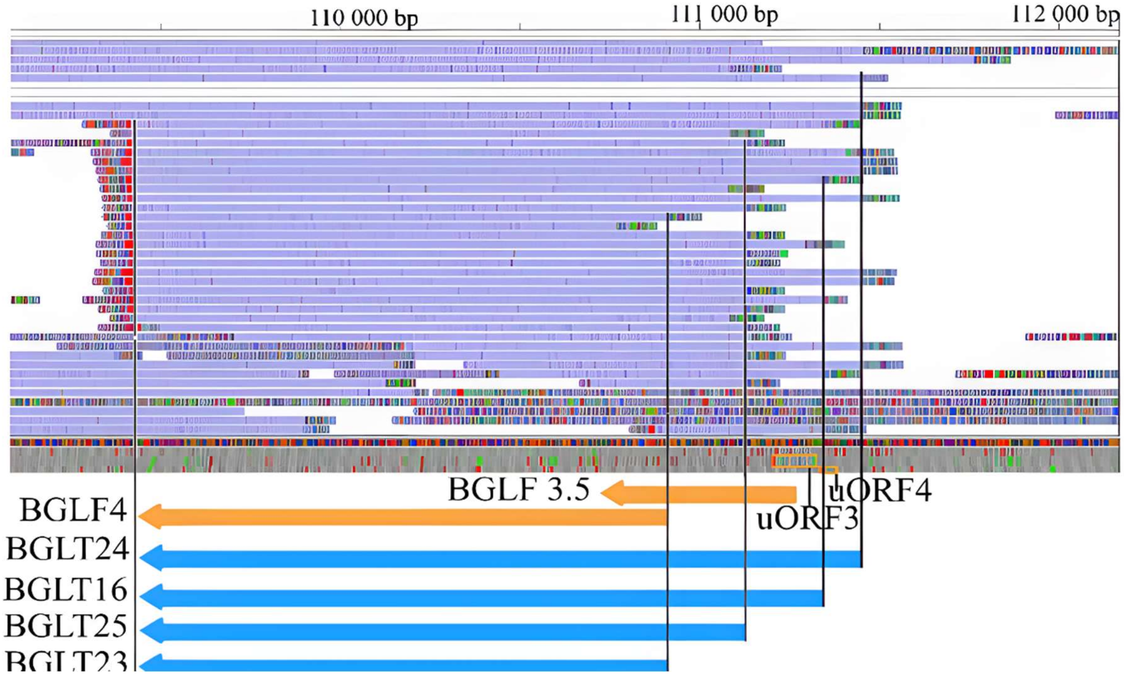


Figure 17. Transcripts overlapping *BGLF4* and *BGLF3.5* ORFs and their uORFs. Yellow arrows indicate the *BGLF4* and *BGLF3.5* genes, while the blue arrows show the transcripts encoded by these genes (*BGLT24, BGLT16, BGLT25*, *BGLT23*). Two upstream ORFs (*uORF3* and *uORF4*) are also indicated. Purple bars indicate the reads obtained by the ONT MinION sequencing. Rectangular lines show the transcript ends.

Our examination revealed 7 variations of isoforms featuring different polyadenylation sites, with 4 of them being novel. Interestingly, all of these isoforms are situated within the same 5 kb region. Specifically, *BZLT42, BZLT43, BZLT44*, and *BZLT50* represent 3'-UTR isoforms of the *BZLF2* gene, while *BELT6*, *BELT8*, and *BELT9* correspond to isoforms of the

BELT1 transcripts (as depicted in **Figure 18.a**). Consequently, these isoforms, namely *BZLT44*, *BELT8*, *BELT9*, and *BERT3*, exhibit a convergent overlap that spans 10 nucleotides.

3.9.C Novel monocistronic mRNAs with canonical ORFs

In this part of our work, we present the identification of 15 new monocistronic transcripts. Among these, we found unspliced versions of *BNRT10*, *BHLF1*, *BORF2*, and *BGLT18* transcripts, which were previously only known in their spliced forms[70,72] (**Figure 18.c**). Additionally, we discovered ten monocistronic transcripts that contain complete open reading frames (ORFs), whereas previous descriptions only mentioned shorter isoforms with incomplete ORFs lacking an in-frame AUG codon[70].

The genomic region of *BFRF3* has not been annotated with these transcripts yet, despite studies showing its transcriptional activity [109,110]. Among our findings, we identified *BFRT3*, which fully overlaps with the *BFRF3* ORF and has a novel terminus (**Figure 18.b**).

3.9.D Splice junctions and introns

Reverse transcription and PCR have the potential to create gaps in cDNAs due to template-switching (TS) events, resulting in incorrect intron annotation. The *LoRTIA* software suite can effectively address this issue by identifying the absence of splice junction consensuses or the presence of repeat regions that promote template-switching. Using the *LoRTIA*, we identified a total of 205 introns. Our criteria for a putative intron required it to be present in at least two of our techniques or in one of our techniques and either in the Illumina or in the PacBio dataset. Moreover, every identified intron exhibited a canonical GT/AG splice junction consensus.

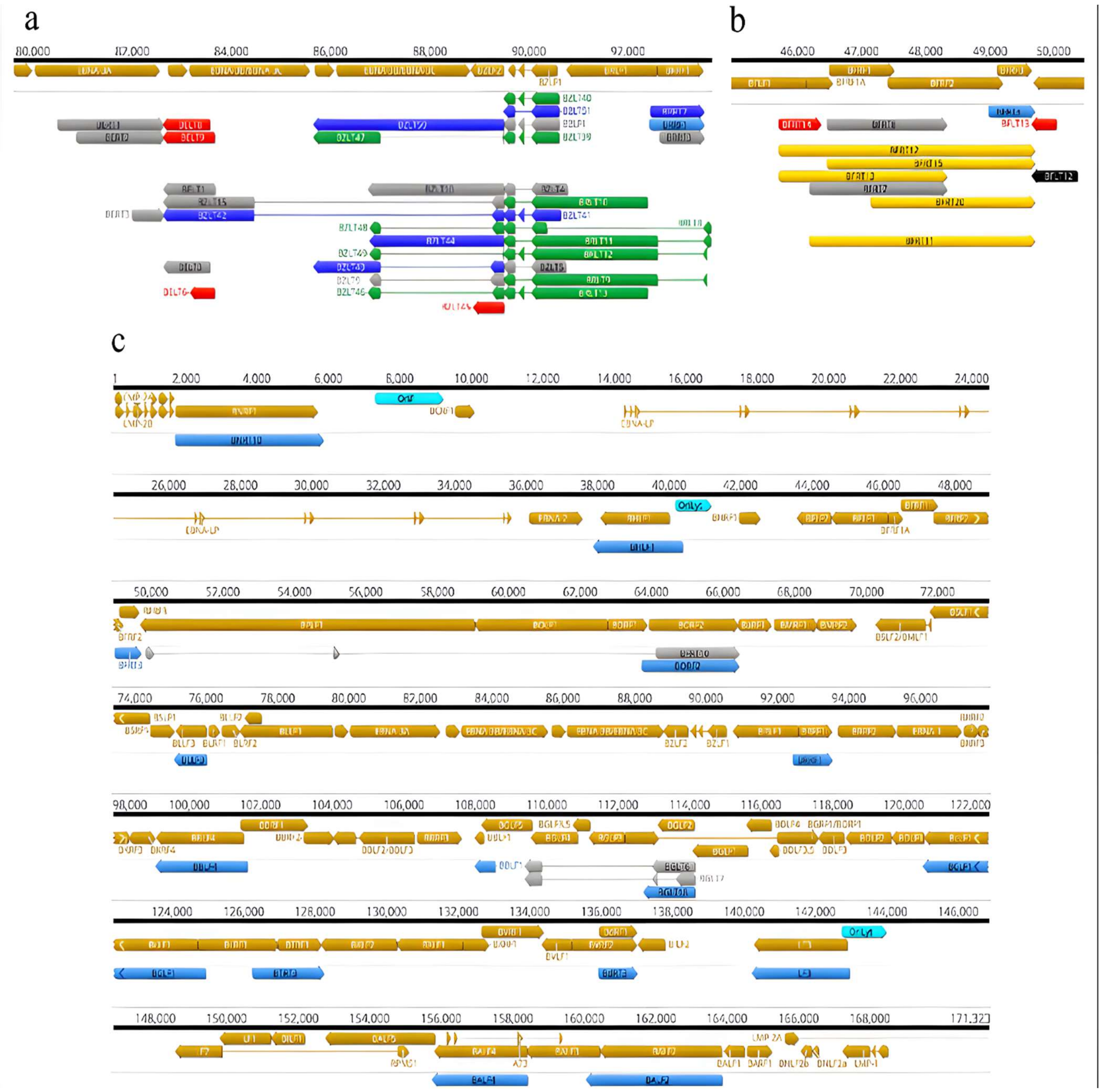


Figure 18. Novel alternative polyadenylation sites and monocistronic mRNA-s with canonical ORFs. Seven annotated transcripts with alternative polyadenylation sites. b Genomic region of *BFRF3* where previously transcripts were not annotated, we identified 1 novel monocistronic and 4 polycistronic transcript. c Novel monocistronic mRNAs where only spliced versions have previously detected. Color codes: brown arrows: ORFs; aqua rectangles: replication origins; grey: formerly annotated transcripts; light blue: novel monocistronic transcripts; yellow: novel polycistronic transcripts; red: non-coding transcripts; black: 5′ -truncated transcripts; dark blue: TSS and TES isoforms

3.9.E mRNAs with altered coding potential

We identified multiple transcripts with truncated 5′-ends that share the same transcription end sites (TESs) as the host mRNAs. These shorter RNA molecules lack the typical open reading frame (canonical) but contain downstream in-frame AUGs, suggesting the potential to encode N-terminally truncated proteins[111]. In our findings, we present a total of 72 such RNA molecules, out of which 19 are newly discovered. The transcription starts sites (TSSs) of these 72 transcripts were verified using the CAGE-Seq dataset (**Figure 19**.**a**).

Additionally, beyond alternative transcription initiation activities within a gene, alternative splicing might yield transcripts with altered coding capacities if the splicing takes place within the ORF. In our study, we identified 42 new splice variants and 5 non-spliced forms of earlier noted spliced transcripts. Among these transcripts, 19 contained introns within the ORFs. We identified 9 instances of frame-shifting, 2 occurrences of nonsense terminations (caused by intron retention leading to premature stop codons), 4 ORFs with deleted amino acids (in-frame deletions) see **Figure 19**.**b**, and 4 intergenic terminations (**Figure 19**.**c**). The intergenic termination transcripts contain the regular AUG start codon and a new stop codon at an intergenic position.

The coding potential of transcripts was assessed using the Coding-Potential Assessment Tool (CPAT) with its default settings[112]. CPAT uses four parameters to estimate the coding potential: maximum open reading frame (ORF) length, ORF coverage, Fickett score (based on codon usage and nucleotide composition), and hexamer score (+coding, +non-coding) that distinguishes coding from non-coding sequences. Sensitivity measures false negatives, specificity evaluates false positives, and accuracy combines sensitivity and specificity. CPAT's performance was validated on known coding and non-coding transcript isoforms, yielding a sensitivity of 0.87, specificity of 0.93, and accuracy of 0.89, indicating the suitability of the default parameters for the dataset. Subsequently, the coding probability of 5′-truncated, alternatively spliced, and unspliced transcript isoforms was calculated. Based on the CPAT analysis, it was found that 9 of the 5′-truncated isoforms and all isoforms, except the ORFs of the 4 splice isoforms with intergenic termination, potentially have coding capacity. As a result, these latter transcripts are classified as non-coding RNAs (ncRNAs). To investigate the homology of proteins encoded by alternatively spliced transcripts, the translation of the altered ORFs was compared to the NCBI non-redundant protein database using protein BLAST.

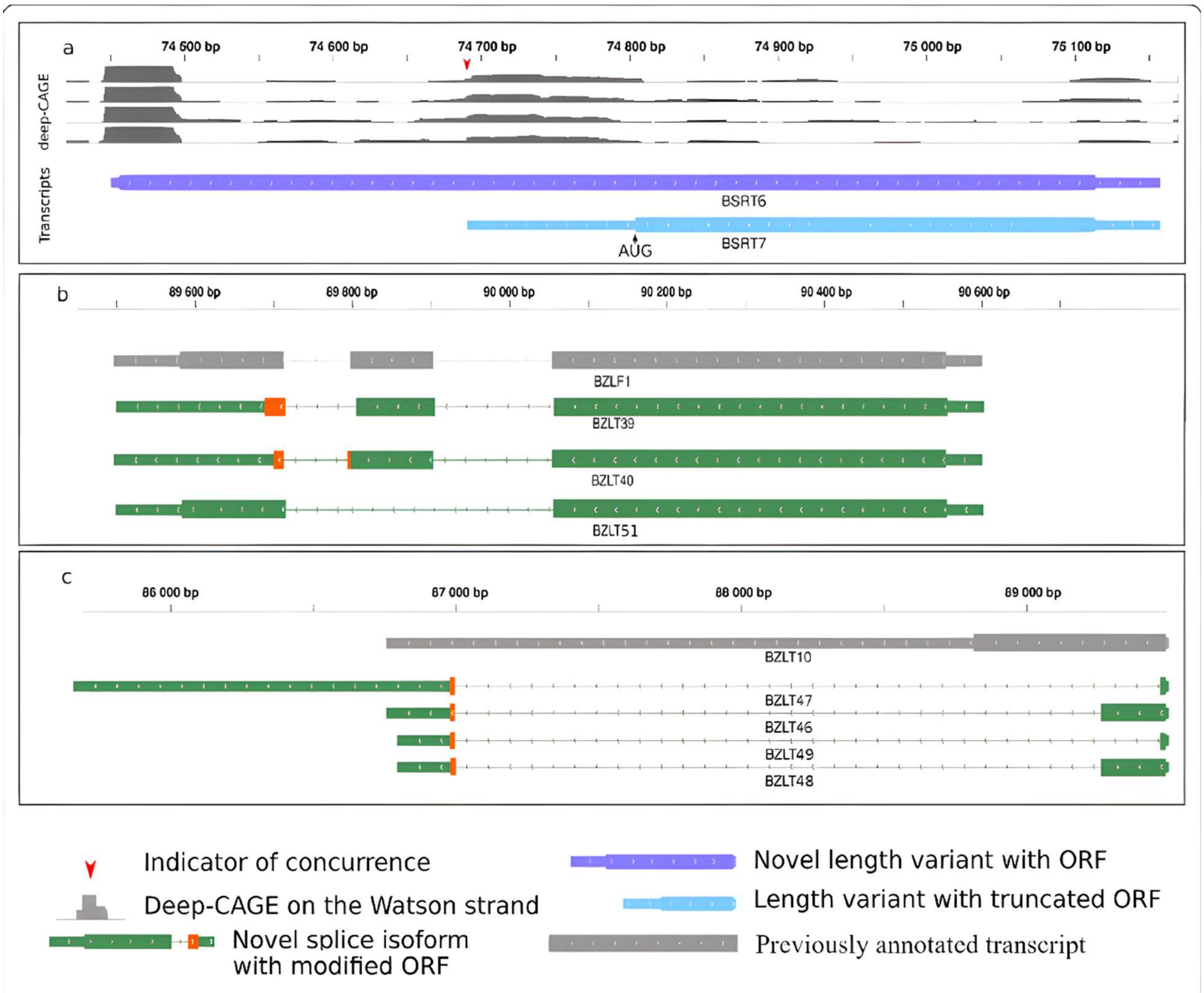


Figure 19. Transcripts with altered coding potential. Transcript isoforms with TSSs located within the canonical ORFs of the host genes in many cases contain in-frame truncated ORFs. The gray histogram illustrates the CAGE-Seq data generated by O 'Grady and colleagues[68]. b Alternative splicing of the *BZLF1* transcripts results in nonsense termination of the ORFs highlighted by the orange wide rectangles and in deletion of the exon in *BZLT51* transcript that does not cause frameshift but an ORF with deleted nucleotides. c Splicing of the unspliced *BZLT10* leads to the deletion of a large portion of the ORF and results in an intergenic stop codon (first exon shown by green, while the second exon shown by the orange wide rectangles). This figure shows examples, and not a summary of all data.

For *BNRT11* and *BNRT12*, the first 21 amino acids of the *BNRF1* ORF are present, but they end with a stop codon immediately after the first splice acceptor position. In the case of *BHLF2*, splicing leads to frameshifting, where the first 77.38% of the ORF matches the *BHLF1* ORF, while the amino acids following the splice acceptor position do not match any proteins in the database.

For *BSLT12*, *BSLT18*, and *BSLT21*, the splice acceptor position differs from the main isoform (*BSLT13*), resulting in altered amino acids following the acceptor position, which has no similarity to any other proteins in the database. *BZLT46* and *BZLT48* encode the first 75 amino acids of the *BZLF2* ORF, and the following amino acids and the stop codon are spliced from the transcript, showing no similarity to proteins in the database.

In *BZLT39* and *BZLT40*, the second splice donor position differs from the main isoform (*BZLF1*), leading to frameshifting, resulting in 7 amino acids and a stop codon following the corresponding splice acceptor. *BZLT51* lacks the second exon of *BZLF1*, causing a 35 amino acid shortening, but without any frameshifting.

3.9.F Non-coding transcripts

Transcripts that do not contain an open reading frame (ORF) longer than 10 amino acids were classified as non-coding in this segment of the study. During this phase of the research, we identified two sncRNAs that are shorter than 200 nucleotides and 19 lncRNAs that are longer than 200 nucleotides. Among the lncRNAs, fourteen of them are 5′-truncated, while three of these lncRNAs (*BFRT14*, *BLRT9*, and *BZLT45*) represent 3′-truncated variants of previously known RNAs. Specifically, *BLRT9*, which is one of the lncRNAs, begins at the same position as *BLRT5* but is terminated at 490 nucleotides downstream. This was confirmed both by our analysis and the Illumina PA-Seq. Furthermore, *BLRT9* overlaps with the *BZTL* and *BELT* regions in the antisense orientation.

3.9.G Replication origin-associated transcripts

Eukaryotic replication origins are typically linked to both coding and non-coding transcripts[113,114]. Previous studies have shown the existence of Ori-overlapping transcripts in alpha-[17,77], beta-[114,115], and gammaherpesviruses[116]. The Epstein-Barr virus (EBV) genome contains two lytic origins of replication (OriLyt) and latent (OriP). The left OriLyt has been observed to overlap with splice isoforms of *BWRT* and *BCRT*, while *BHLT2* initiates within this Ori[70].

Additionally, the genomic region containing OriP exhibits transcriptional activity, with several transcription start sites (TSSs) of various non-coding RNAs located within the Ori[69]. Moreover, there is a long 5′-UTR transcript isoform of the *BCRF1* gene, known as *BCRT3*, which is associated with this region. During our research, we discovered nine new 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 bhlf1gene. *BHRT15, 16, 17, 22*, and *22* transcripts are splice and 5′-UTR isoforms and are encoded by the *brf1* gene (**Figure 20**.**a**). The *LF3* transcript starts in the right OriLyt region. We annotated *LF3* and identified four new spliced transcripts (*RPMS2, RPMS3, RPMS4*, and *RPMS5*) that completely overlap with the Ori region. Additionally, we found *BILT44* and *BIRT21* transcripts, of which the 5′-UTR regions overlap with the replication origin (**Figure 20**.**b**).

3.9.H Transcriptional overlaps

We have identified all types of transcriptional overlaps within EBV RNAs: divergent (head-to-head), convergent (tail-to-tail), and tandem (parallel, tail-to-head) overlaps. These overlaps occur between transcripts of neighboring genes, such as *BDRF1* and *BILF2*, as well as long multigenic and monocistronic transcripts. One example of the latter is *BBRT18*, a bicistronic transcript that overlaps with isoforms of both *BBTR16* and *BBRT14* in the same orientation, and also with *BBRT18* and the isoforms of *BGLT29* in the opposite orientation. Additionally, we found several long-spliced transcripts that overlap with multiple genes. For instance, *BDLT30* initiates upstream of *BDLF2* and overlaps with the transcript of 12 genes in the same orientation as *BDLF2*, and the transcript of 3 genes in the reverse orientation.

Though transcriptional overlaps are common in EBV, we observed significantly low levels of overlapping transcripts in the intergenic regions between the convergent *BHRF1* and *BHLF1*. Moreover, there was no transcriptional activity detected in the intergenic region of

BMRF2 and *BSLF2*/*BMLF1*. However, it's worth noting that a higher overall transcript coverage might reveal low-level activity in this region.

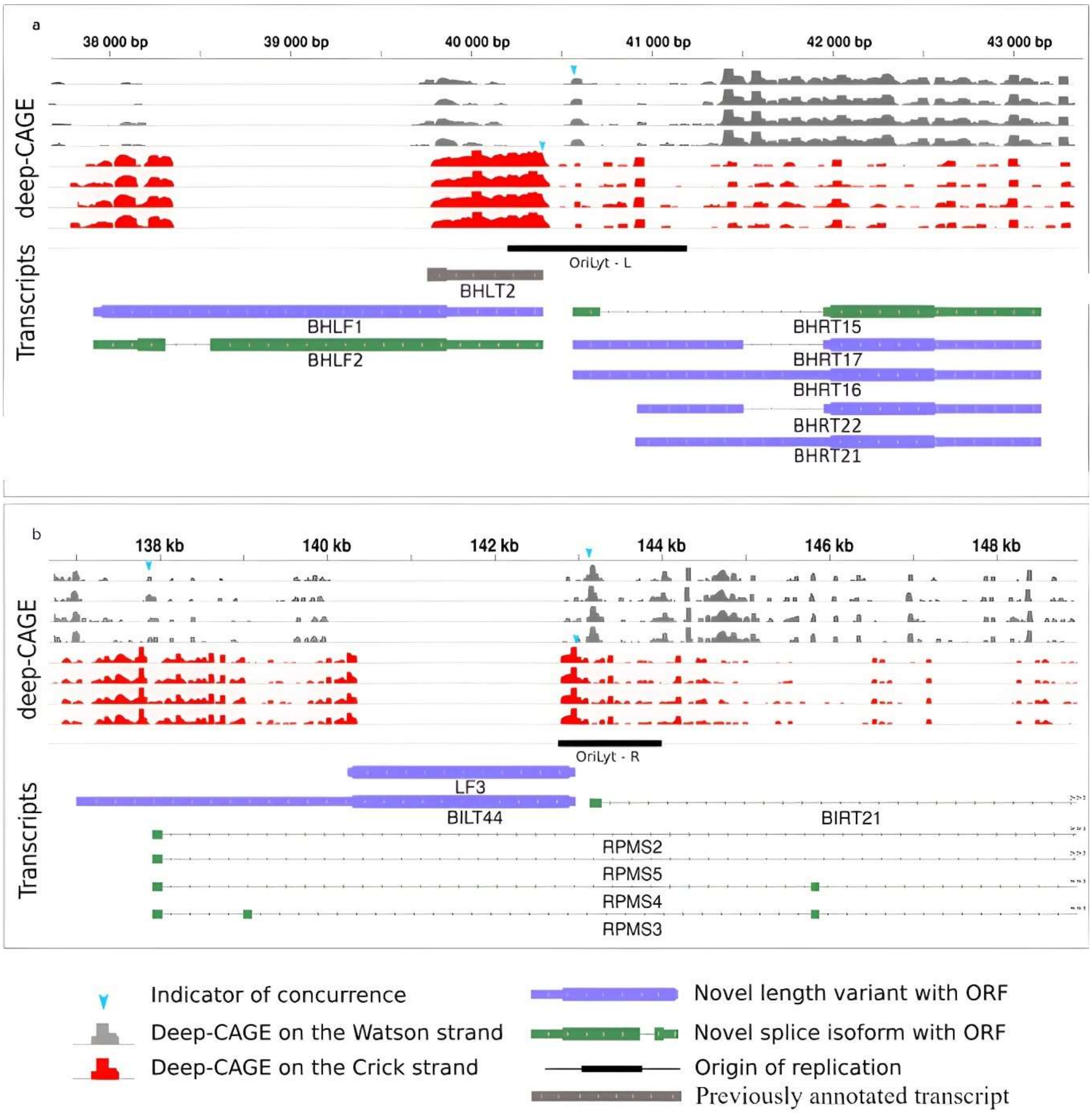


Figure 20. Novel Ori-overlapping RNAs. a Three novel length and two novel splice transcript isoforms are initiated in the left lytic Ori region. The CAGE-Seq data from O 'Grady *et al.* [68] is represented by the histograms with gray reads mapping to the positive, while red are reads mapping to the negative 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.

4. DISCUSSION

With LRS technologies, we can now identify and accurately annotate transcripts and RNA isoforms, such as those with length and splice variants. These technologies include PacBio's Single Molecule, Real-Time (SMRT) sequencing[70,99], ONT's nanopore sequencing [77], and Loop Genomics' LoopSeq synthetic LRS (which uses Illumina platform)[53]. They have been used alone or together with each other or with SRS to mark viral transcripts in herpesviruses from all three subfamilies (HSV-1[117,118]; VZV[99,119]; PRV[17]; BoHV-1[120]; HCMV[81]; and EBV[70,82]).

We and other researchers have found out that different viruses have a hidden and complex network of genes that overlap with each other when they are transcribed into RNA molecules[17,23,70,118,121,122]. It has been shown that the RNA molecules encoded by closely spaced genes overlap each other in a parallel, divergent, or convergent manner. This phenomenon implies an interaction between the transcription machinery at the TOs throughout the entire viral genome[123]. We and others have previously demonstrated that in several viruses, the replication origins overlap with specific lncRNAs and with long 5′ or 3′ UTR isoforms of mRNA[124]. We note that many of these long versions of mRNAs, 5′ UTR isoforms, may not code for proteins, because their start codons are far away from their TSSs. The only exceptions may be those transcripts whose large parts of the 5′ UTR are spliced out. Functional analyses discovered how some replication RNAs regulate DNA synthesis by forming RNA: DNA hybrids in numerous viruses[115].

Our effort is to review the structure and function of RNA isoforms that overlap with or are close to the origin of replication (Ori) of herpesviruses. These RNAs potentially control how the virus copies and expresses its genes. We focus more on αHVs because they are less studied. We used novel and previously published sequencing data to find novel transcripts and fix errors in previous annotations. We found complex interactions between transcripts encoded by replication and transcription regulatory genes and from long non-coding RNAs (lncRNAs) near them. We identified promoter consensus elements in the Oris of all herpesviruses we examined. Even though Oris have AT-rich regions that may look like TATA boxes, we distinguished the corresponding transcripts with TSSs being proximal to these sequences in all cases. The terminology of the lncRNAs we reported are not very consistent, unlike protein-coding genes, because they are not well conserved across different viruses. It is also possible that some lncRNAs with the same term (e.g., *NOIR-1*) have different origins. However, the CTO and *NOIR-1* transcript families are apparently orthologous in PRV and EHV-1.

The co-temporal activity of DNA replication and transcription within the same genomic region can create interference when they restrict each other[125]. These clashes seem more significant when they go in opposite directions (convergent) than when they go in the same direction (co-directional)[126]. Some molecular ways reduce the conflict between the RNP and the replication fork[127]. Nevertheless, our current study suggests the existence of an in-built, intrinsic regulatory system based on an interaction between the two apparatuses for controlling the initiation of both replication and transcription cooperatively. This system is thought to depend on the clash and competition between the transcribing RNP and the replisome, as well as the assembly of pre-replication and transcription initiation complexes by the ongoing DNA and RNA syntheses (**Figure 21-23**).

In general, αHVs' OriS is highly conserved, and is always located in the same location in the genome: it is positioned in the US repeats and enclosed by the major TF genes, *icp4,* and *us1.* In this work, we verify that the TF genes code for transcript variants that overlap not only the OriS but likewise the transcripts of the opposite TF gene and/or specific proximal lncRNAs. These TOs might aid these genes to interact more with each other – besides the TF/promoter interaction, which might happen through RNA: DNA or probably by RNA: RNA hybridization, and also the interference between the transcription machineries. Consequently, the US-IR region of αHVs seems to work as a 'super regulatory center', where both DNA replication and global transcription are cooperatively regulated by a complex system that affects each other at different levels. In addition, this genomic segment governs the transition between the lytic cycle and latency, as well as the maintenance of these processes. Hence, this region is functionally the most complex genomic locus of αHVs, encoding lncRNAs such as intergenic transcripts and asRNAs, long TIs of mRNAs, and an intricate TO pattern of local transcripts. Besides the lytic transcripts, several latent lncRNAs (*LAT, LLT, L/ST*) and miRNAs are also expressed from this genomic region[128].

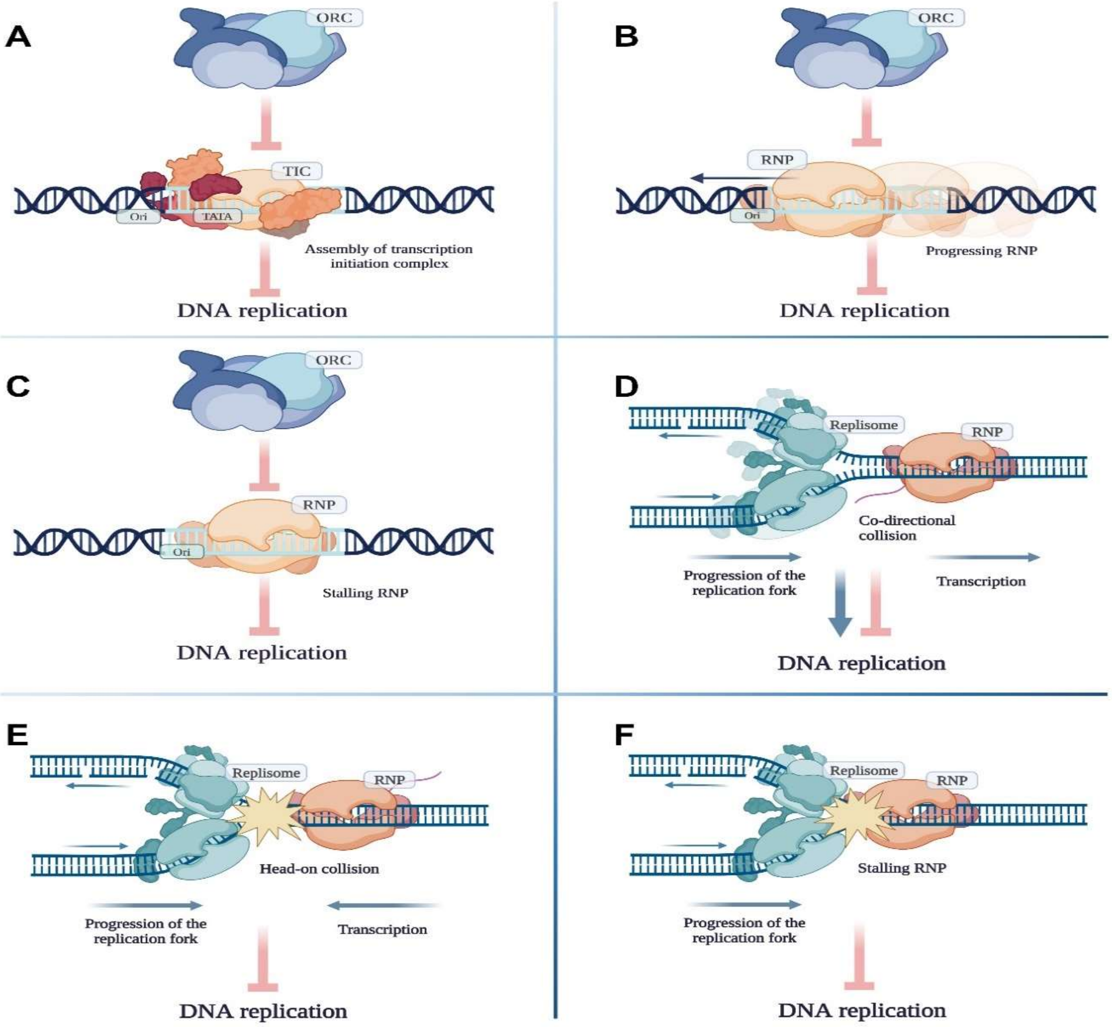


Figure 21. Potential effects of transcription on the DNA replication: A. Recruitment of transcription initiation complex on promoters located within the replication origin inhibits assembly of ORC and replisome. B. RNA polymerase II passage across the replication origin inhibits of ORC and replisome assembly. C. RNA polymerase II stalling on the replication origin inhibits ORC and replisome assembly. D. Co-directional movement of replication and transcription machineries can slow down or speed up replication fork progression. Transcription can facilitate DNA replication by pre-opening the two DNA strands. E. Head-on collision of replication and transcription machineries inhibits both processes. F. RNA polymerase II stalling inhibits replication fork progression.

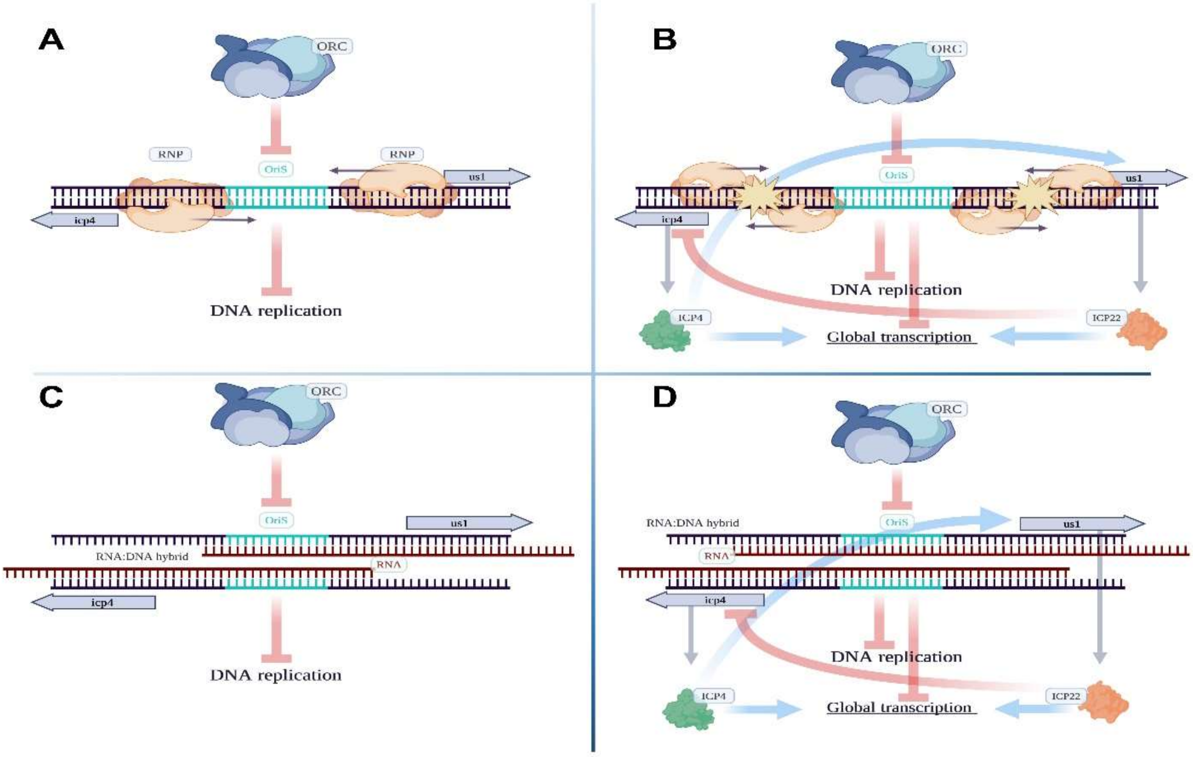


Figure 22. The operation of the Putative Super Regulatory Center: A. Passage of RNA polymerases across the OriS by reading the long 5′ UTR of us1 and *icp4* genes inhibits the assembly of ORC and replisome. Stalling of RNP on OriS leads to the same consequences. B. Reading the long 5′ UTRs of us1 and icp4 genes by RNP inhibits each other expressions which exert an effect on the global transcription. Inhibition of *icp4* expression through interference between RNPs reading the two genes leads a decreased expression in most of the herpesvirus genes, including *us1* gene. Inhibition of *us1* gene expression *ICP22* inhibits icp4 expression and leads a more complex effect on genome-wide gene expression (Takács et al., 2013)[135]. C. Transcriptional overlapping of replication origin may lead to the formation of RNA:DNA hybrids and thereby interfering (inhibiting or perhaps facilitating) the assembly of the replisome. D. Transcriptional overlapping of replication origin and the transcription regulatory genes may lead to the formation of RNA:DNA hybrids and thereby interfering (inhibiting or perhaps facilitating) the assembly of the replisome and also the transcription.

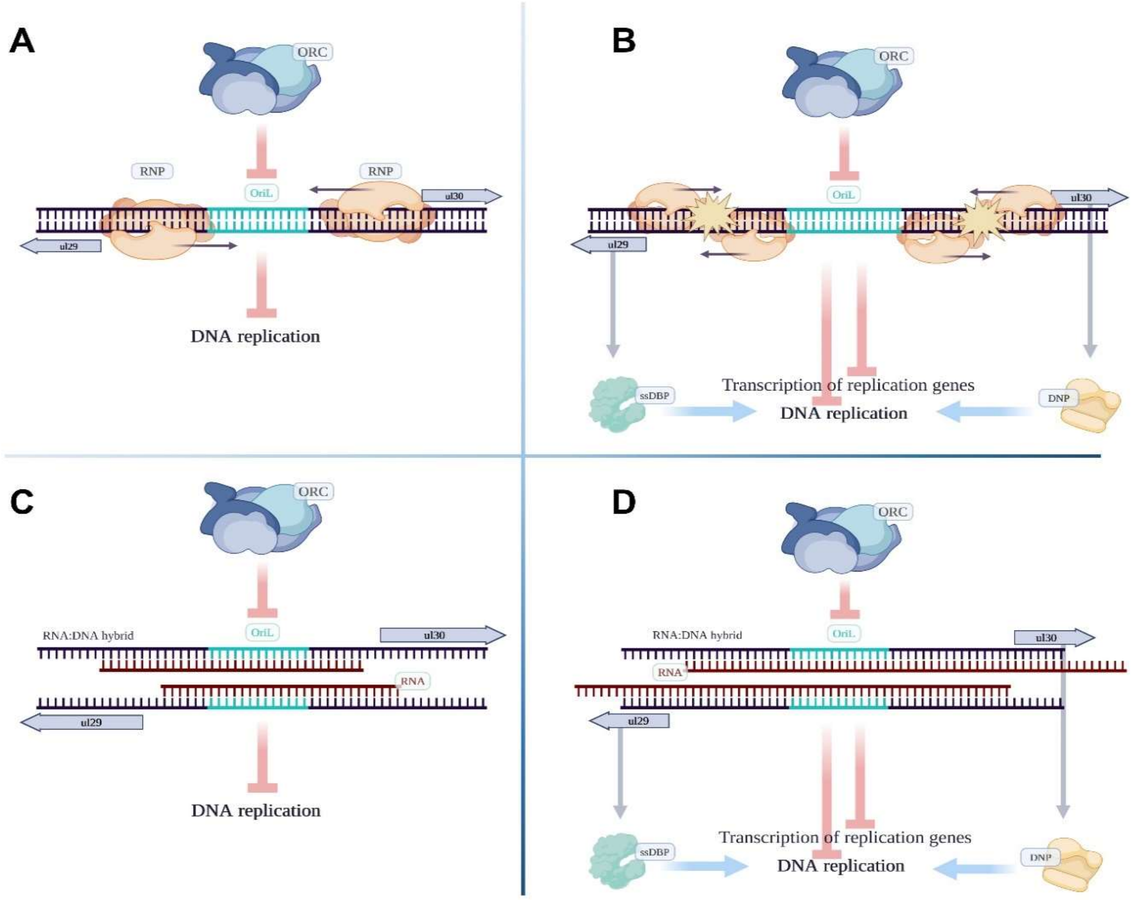


Figure 23. Interference between the replication and transcription apparatuses: A. Passage of RNA polymerases across the OriS by reading the long 5′ UTR of ul29 and ul30 genes inhibits the assembly of ORC and replisome. Stalling of RNP on OriS leads to the same consequences. B. Head-on collision between the two replication genes leads to the inhibition of each other transcription and the initiation replication because of the passage of RNPs across the OriL. The replication is also affected by the inhibition of the generation of replication proteins. C. Formation of RNA:DNA hybrids at the OriL interferes with the initiation of DNA replication. D. Formation of RNA:DNA hybrids at the OriL and the replication genes interferes with both the initiation of DNA replication and the transcription of replication genes. The replication is also affected by the inhibition of the generation of replication proteins.

The US-IR contains the *NOIR-1* transcript family, which is a unique feature of the Varicellovirus genus. These lncRNA molecules have different locations in different viruses. In PRV, the long TIs of *NOIR-1* overlap the *icp4*, but no OriS overlapping *NOIR-1* TES isoforms have been detected in this virus. This could be because *NOIR-2*, which is opposed to *NOIR-**1*** and only found in PRV RNA, blocks *NOIR-1* from extending its transcription to a readthrough. The PRV *NOIR-1* ends at the same point, 3′ co-terminal, as the *LLT* transcripts that are expressed throughout latency. In EHV-1, the *NOIR-1* long TI overlaps with *icp4*, while the long TSS isoform of US1 transcript, driven by the *NOIR-1* promoter, overlaps the OriS. In VZV and SVV, the *NOIR-1* promoters also control the transcription of the long TSS variants of *US1*. Additionally, in VZV, the upstream *NOIR-1* transcripts also overlap the *icp4*. Intriguingly, in SVV, the *NOIR-1* terminates at the OriS, which appears to be a unique solution for the interplay between the two machineries.

Although the *NOIR-1* in BoHV-1 is absent, the virus has a long version of the *US1* transcript that overlaps the OriS. It also has another ncRNA called OriS-RNA that goes in parallel to *icp4* and has its promoter situated on the other side of the OriS. The promoter of this transcript also controls the long TSS isoform of the *ICP4* transcript. *NOIR-1* transcripts may be involved in controlling transcription through overlapping the *icp4* gene, and/or in controlling DNA replication through overlapping the OriS (in cases when they are the upstream part of the *US1* TI), or in controlling both transcription and replication (in cases when they overlap both OriS and *icp4*). The same might also be true for the HSV-1 OriS-*RNA1*, as it also serves as the 5′ UTR part of the long *ICP4* TI.

It's possible that the proximity of the OriL in Simplexviruses to the primary replication genes (*ul29/30*) is not a random occurrence. Our hypothesis suggests that these genes could influence the initiation of replication, not solely through the traditional TF/promoter binding mechanism but also via interactions involving RNA: DNA and/or RNA: RNA hybridization. Additionally, interference between the transcription and replication processes at the overlapping region might be involved. Previously, *CTO* transcripts were identified in PRV[51], and now, we report the presence of orthologous transcripts in EHV-1. Although the canonical *CTO-S* doesn't directly overlap with the OriL, it may impact replication by assisting in separating the two DNA strands, thus determining the replication orientation or through other means. Notably, one of the extended *CTO-S* transcription initiation sites is regulated by a promoter located within the OriL, while another one corresponds to the TES variant of the *ul21* gene.

We've observed an intriguing occurrence in αHVs: three 'hard' convergent transcription overlaps (TOs) are formed, suggesting a significant clash between transcription processes while RNA molecules are being synthesized. What caught our attention is that one of the partners involved in these TOs is consistently an auxiliary replication gene. This frequent involvement of such genes in 'hard' TOs may not be coincidental but could have unknown biological significance. One possible speculation is that the 'hard' TO between the *ul30/ul31* partner genes in Simplexvirus enables the *ul31* gene to produce a long TES isoform that overlaps with the OriL due to transcriptional read-through. In fact, we observed a substantial occurrence of transcriptional read-through from this gene in PRV using RT-PCR. Interestingly, although the OriL is situated between the *ul21* and *ul22* genes in both PRV and EHV-1, a long TES isoform of *ul21* also overlaps with the OriL. Since Simplexviruses do not produce long TES isoforms of *ul21*, and Varicelloviruses do not produce long TES isoforms of *ul31*, it is reasonable to infer that these types of transcriptional overlaps may interfere the initiation of DNA replication.

The HCMV OriLyt is overlapped by long non-coding RNAs (lncRNAs), with *RNA4.9* being previously demonstrated to control DNA replication[42]. We've previously documented new Ori-overlapping lncRNAs for HCMV[36], as well as in our recent report. In the case of EBV, OriLyt-L is situated near the non-coding *BHLF1* gene, a regulator of latency[129], while OriLyt-R is adjacent to *LF2*, which inhibits replication[130]. KSHV *LANA* is crucial as a latency regulator[131]. In our current study, we detail additional transcripts that overlap with the Ori, encompassing lncRNAs and extended 5' UTR variants of both mono- and polycistronic mRNAs.

We measured the presence of exceptionally lengthy 5'UTR variants using RT-PCR and observed that their limited occurrence cannot be solely attributed to sequencing preferences favoring smaller fragments. These transcripts are genuinely expressed at a low rate. While some of them contain full ORFs, it's likely they remain untranslated, given the considerable distance between their transcription and translation initiation sites. Consequently, they may serve as incidental outcomes of transcription without serving as functional protein-coding sequences or even functional RNA molecules.

In summary, even among closely related herpesvirus species, distinct strategies have evolved to create transcriptional overlaps (TOs) at the Oris and TR genes. This highlights the importance of TOs in controlling DNA replication and overall transcription. While the primary TR genes seem to regulate each other and the initiation of DNA replication at the OriS region of αHVs, in Simplexviruses, the principal replication genes at the OriL region may instead govern each other and DNA replication through mechanisms involving TOs. These potential mechanisms offer multiple layers of regulation beyond the traditional interaction of transcription factors with promoters. *ICP4* has been demonstrated to enhance the expression of *icp0*[132] genes by binding to their promoter. Conversely, *ICP22* (the product of *us1*) suppresses the expression of both *icp4* and *icp0*[133]. Additionally, *ICP0* transforms *ICP4* from a repressor into an activator of mRNA synthesis in HSV-1[134]. Mutating the *us1* gene has a differential effect on the transcription kinetics of E and L genes[135].

We believe that the significance of our findings extends far beyond our specific study and offers a more universal insight into how the control of herpes viral replication and transcription has developed over time in tandem. Grasping these intricate relationships among different genes and regulatory components can offer valuable insights into the overall mechanisms that govern herpesvirus replication and gene activity. Delving deeper into these potential interactions and their functional importance could pave the way for novel therapeutic strategies to tackle herpesvirus infections.

For EBV: This study utilized the ONT long-read sequencing platform with both amplified and non-amplified cDNA libraries to profile the Epstein–Barr virus lytic transcriptome. Transcript annotation was carried out using their dataset along with previously published data from other studies[25–27,42,44,45,136,137]. Previous studies had only partially annotated the lytic transcriptome of EBV. By employing a multiplatform, integrative approach, this study achieved a more comprehensive understanding of the transcriptomic architecture of this virus. We identified new transcripts and RNA isoforms and validated previously reported putative transcripts. In total, 241 novel-lytic EBV transcripts were discovered, and 110 previously identified transcripts were confirmed.

A recent study[138] on human adenovirus type 5 revealed significant flexibility in intron (TES) usage, suggesting potential advantages for viral evolution. Previous research on herpesviruses[17,19,77,81,118], including the EBV[70,71], has also shown a wide variety of transcript length isoforms, whose functions remain largely 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[139]. Additionally, transcripts with alternative termination may have different turnover times[140,141], localization[142], and altered translation[143].

In this study, we present the discovery of novel length variants and splice isoforms that could potentially impact the coding potential of several viral genes. To fully understand the significance of these transcripts, further proteomic investigations are required. We observed a relatively large number of short transcripts that are integrated within larger host genes and contain truncated in-frame ORFs. Although such transcript types have been previously described in other viruses, they appear to be more prevalent than previously believed[144].

Additionally, we identified several transcription-start site (TSS) isoforms. Longer transcript variants often include upstream open reading frame (uORF) sequences, which might play a role in translational control. We propose that BGLF4 might bypass the translational interruption described by Watanabe *et al.*[107], either through its shorter isoforms or via a complex interplay of uORFs present in the 5'-UTR of its longer isoform.

Furthermore, we report the identification of numerous multigenic transcripts. While polycistronic transcripts are common in prokaryotic organisms due to Shine-Dalgarno sequences allowing translation of each gene in the polycistronic RNA, such transcripts are rare in eukaryotes, as cap-dependent translation initiation typically leads to the translation of only the most upstream gene in a multigenic transcript. Interestingly, eukaryotic viruses, despite sharing similar translation mechanisms with their host organisms, produce a wide range of multigenic transcripts with yet unknown functions [23].

The EBV genome has previously been found to exhibit genome-wide antisense expression using an SRS approach[68]. In our study, we employed an LRS approach, allowing us to map the transcript ends. Our findings indicate that most antisense transcripts result from transcriptional readthroughs between convergent genes or from the head-to-head overlap of transcripts encoded by divergently-oriented gene pairs. The functional significance of these transcriptional overlaps remains uncertain. We have proposed the Transcriptional Interference Network hypothesis[123], suggesting that these overlaps might serve as a genome-wide gene regulatory mechanism. However, we cannot rule out the possibility that some of these overlaps are merely transcriptional noise without any specific function. Nevertheless, the presence of parallel (co-oriented) transcriptional overlaps, which is a characteristic feature of viral genomes, suggests that other types of overlaps may also have functional roles.

Together, we can conclude that multiplatform approaches are important in transcriptomic studies because the different platforms have distinct advantages and limitations and that they represent independent techniques that are vital for the validation of the results obtained by a particular method.

In our analysis, we discovered new Ori-overlapping transcripts. A previous study by Rennekamp and Lieberman[47] demonstrated that the *BHLF1* transcript, which overlaps the left Ori-lyt, stably binds to its DNA template, and either *BHLF1* or the divergent *BHRF1* transcript is essential for initiating lytic replication from this Ori. In our research, we precisely identified the transcription start site (TSS) and transcription end site (TES) of *BHLF1* and observed a splice isoform of *BHLF1* called the *BHLF2* transcript. Additionally, we identified three isoforms of *BHRF1*, namely *BHRT15*, *BHRT16*, and *BHRT17*, with longer 5'-UTRs overlapping the Ori compared to previously detected isoforms. However, the impact of these novel isoforms on viral replication requires further evaluation.

Our research group has previously identified several Ori-associated transcripts in different viruses[23,36,73,109,123]. We have proposed an interaction between the replication and transcription machinery, which might influence the orientation of the replication fork and DNA synthesis progression [80]. The functional significance of low-abundance transcript variants and multigenic RNA molecules remains uncertain, and it is still unclear if they contribute to the viral proteome or have any specific function. Further investigations are necessary to answer these questions.

In conclusion, multiplatform approaches are crucial in transcriptomic studies because each platform has unique advantages and limitations. These independent techniques are vital for validating results obtained from a particular method.

5. METHODS

5.1 Cells and Viruses

In addition to our novel data, we also used several other datasets for the analyses in this study. The cell types used in this work are listed in **Table 4.**

Table 4. List of the cell types used in this study

Viruses	Cell lines	Time points
EHV-1	Rabbit kidney (RK-13)	1h, 2h, 4h, 6h, 8h, 12h, 18h, 24h, 48 h
HSV-1	African green monkey cells (Chlorocebus sabaeus)	1h, 2h, 4h, 6h, 8h, 10h, 12h, 24 h
PRV	PK-15 porcine kidney epithelial cell line	0h, 1h, 2h, 4h, 6h, 8h, 12 h
VZV	Human primary embryonic lung fibroblast cell line (MRC-5)	Mixed: 1h, 2h, 4h, 6h, 8h, 12 h
SVV	African green monkey (AGM) kidney epithelial BS-C-1 cells, Rhesus macaque kidney epithelial cells LLC-MK2 cells	72h
KSHV	KSHV-positive primary effusion lymphoma cell line iBCBL1-3xFLAG-RTA	24h
EBV	Akata cells	10 min, 90 min, 4, 12, 24, 48, 72 h
HCMV	Human primary embryonic lung fibroblast cell line (MRC-5)	Mixed: 1h, 3h, 6h, 12h, 24h, 72h, 96h, 120h
BoHV-1	Madin–Darby Bovine Kidney	1, 2, 4, 6, 8, 12 h

5.1.A PRV

For the generation of novel transcript data, we employed three immortalized cell lines to propagate the MdBio strain of the pseudorabies virus (PRV-MdBio[88]). The cell lines were: PK-15, which are porcine kidney epithelial cells (ATCC® CCL-33™); C6, which are rat brain tumor cells (ATCC® CCL-107™); and PC-12, which are rat adrenal gland pheochromocytoma that comes from the embryonic origin from the neural crest (ATCC® CRL-1721™). Three replicates with each cell line (triplets) were applied. PK-15 cells were grown in a Dulbecco's modified Eagle medium (DMEM) (Gibco/Thermo Fisher Scientific) with 5% FBS (Gibco/Thermo Fisher Scientific), 80 µg/ml of gentamycin (Gibco/Thermo Fisher Scientific) and 5% CO2 at 37°C. C6 cells were cultivated in another medium (F-12K) (ATCC) with 2.5% FBS, 15% HS (Horse Serum; Sigma-Aldrich), and 5% CO2 at 37°C. We grew the PC-12 cells in a third medium (RPMI-1640) (ATCC) with 5% FBS, 10% HS, and 5% CO2 at 37°C. To make the virus stock solution, the PK-15 cell line was infected with 0.1

multiplicity of infection [MOI = plaque-forming units (pfu)/cell], then was waited until a complete cytopathic effect was detected, then we broke them open by freezing and thawing them three times to release the viruses. All cell lines were infected by a high dose of PRV-MdBio (1 pfu/cell), followed by incubation for 1 h at 37°C to ensure the infection, then viruses were removed, and the cells were washed with phosphate-buffered saline (PBS). Finally, adding fresh medium and incubating the cells for different time points: 0, 1, 2, 4, 6, 8, and 12 h. For later analysis, the medium was thrown away, and stored the infected cells at −80°C

To study the kinetics of infection, we infected PK-15 cells with PRV-Ka at MOI=1. We first synchronized the infection by keeping the cells at 4 ◦C for an hour and then moved them to a 37 ◦C incubator with 5% CO2. Infected cells were collected every half an hour for 8 hours. The culture was rinsed with PBS, removed from the plates, and spun them at 3000 RPM for 5 min at 4 ◦C.

5.1.B EHV-1

This study also used Equid alphaherpesvirus 1 (EHV-1), a virus isolated from the organs of a colt fetus that was aborted in Marócpuszta (Hungary) in the 1980s. The virus (EHV-1-MdBio) was grown in a rabbit kidney (RK-13) epithelial cell line (ECACC 0021715) that covered the whole surface of the culture dish. The experiment was applied in triple technical replicates.

RK-13 was maintained in DMEM (Sigma) with 10% FCS (fetal calf serum: Gibco) and 5% CO2 at 37°C. We made a virus stock by infecting the cells with EHV-1-MdBio at MOI=0.1 then freezing and thawing them three times to release the viruses. For the EHV-1 long-read RNA-seq experiments, RK-13 cells were infected with MOI=4 of the virus, and prepared three technical replicates. The infected RK-15 cells were chilled for 1 h at 4°C, then the viruses were removed, and the cells were washed with PBS. Then new media was supplemented, and the cells were incubated for 1, 2, 4, 6, 8, 12, 18, 24, or 48 h. Finally, the culture medium was taken away and cells were kept at -80°C for later use.

5.1.C KSHV

The KSHV-positive primary effusion lymphoma cell line iBCBL1-3xFLAG-RTA[145] was maintained in RPMI 1640 medium, which was supplemented with 10% Tet System Approved FBS (TaKaRa), penicillin/streptomycin, and 20 µg/ml hygromycin B. KSHV lytic reactivation was induced by treating 1 million of iBCBL1-3xFLAG-RTA cells with 1 µg/ml doxycycline for 24 h. For measuring KSHV gene expression by qRT-PCR, cDNA was generated with iScript cDNA Synthesis kit (Bio-Rad) followed by SYBR green-based real-time quantitative PCR analysis using gene specific primers. The relative viral gene expression was calculated by the delta-delta Ct method where 18S was used for normalization. The sequences of the primers have been reported previously (Toth et al., 2013). The following antibodies were used for immunoblots: anti-FLAG (Sigma F1804), anti-ORF6 (from Dr. Gary S. Hayward, Johns Hopkins University), and anti-Tubulin (Sigma T5326).

5.1.D VZV

A human cell line (MRC-5), embryonic origin lung fibroblast, was used from (ATCC) to propagate the virus. An attenuated form of VZV (OKA/Merck strain) was utilized. Cells were grown in DMEM enhanced with antibiotics, antifungals, and 10% FBS, and then the cells were kept at 37°C with 5% CO2. Infected cells were treated with trypsin and collected when they showed signs of cytopathic effect (in 5 days).

5.1.E HSV-1

To grow HSV-1, immortalized kidney epithelial cells were used (Vero), they were cultured in DMEM with 10% FBS and 100 µl/ml penicillin-streptomycin Mixture (Lonza) and were incubated in a 37°C with 5% CO2. The culture was infected by HSV-1 at a ratio of one virus per cell (MOI=1) and let them interact for an hour. Then viruses were removed, and the cells were rinsed with PBS. A new medium was added followed by incubating the cells for different time periods: 1, 2, 4, 6, 8, or 12 hours then samples were collected from each time point for the next steps.

5.1.F BoHV-1

Madin–Darby Bovine Kidney (MDBK) cell line was used to infect them with the Cooper strain (GenBank Accession # JX898220.1) of BoHV-1.1. The cells were kept at 37°C in a moist incubator with 5% CO2 and grown in DMEM with 5% (v/v) fetal bovine serum, penicillin (100 U/mL), and streptomycin (100 µg/mL). A virus dose of MOI=1 was used to infect the cells. The infection went on until all cells showed signs of damage (cytopathy). The supernatant was collected after 1, 2, 4, 6, 8, and 12 h of infection. To get more viruses out of the cells, three freezing-thawing cycles were applied.

5.2 Inhibition of DNA synthesis using phosphonoacetic acid

Before the infection, the cells were treated with 400 µg/ml PAA (phosphonoacetic acid) for 1 hour at 37°C in the presence of 5% CO_2 to inhibit replication and determine the kinetic class of transcripts. During the experiment, the cells were examined at six different time points (1, 2, 4, 6, 8, 12 hours) using three biological replicates. For the treated samples, we also used a MOI of 10 of the virus for infection. RNA was isolated at 1h, 2h, 4h, 6h, 8h, 12h post-infection time points, and a dcDNA library was prepared according to the library preparation protocol of the direct cDNA Sequencing Kit (SQK-DCS109). The samples were sequenced on Oxford Nanopore MinION flow cells. Subsequently, the data were base-called by Guppy, mapped using minimap2 software, and further analyzed using scripts deposited on GitHub.(https://github.com/Balays/Rlyeh?fbclid=IwAR0HZJNXZjv9YUm_tsJ5J1eT2fKXnkhbJKf7WVoTxX9kvp7fJmdhWQILbjA). Data have been deposited under the project accession number PRJEB64684 into the European Nucleotide Archive (ENA).

5.3 RNA isolation

5.3.A Extraction of total RNA

Total RNA was obtained from cells infected with PRV, EHV-1, VZV, and KSHV using the NucleoSpin® RNA kit (Macherey-Nagel), and following a spin-column protocol. Briefly, cells were broken down by adding a buffer solution (from the kit) that contained chaotropic ions. Nucleic acids then stuck to a silica membrane. To eliminate genomic DNA, DNase I treatment was applied to the samples. The total RNA was subsequently eluted using

RNase-free water. To ensure the removal of any remaining DNA, the TURBO DNA-free™ Kit was employed. Finally, the samples were stored at a temperature of -80 °C.

5.3.B Purification of polyadenylated RNA

For the purification of polyadenylated RNA, the Qiagen Oligotex mRNA Mini Kit was utilized to enrich the mRNAs and other RNAs with polyA-tails from the samples of PRV, EHV-1, and VZV. These enriched RNAs were subsequently used as templates for ONT and Illumina library preparations. The purification process followed the Spin Columns protocol outlined in the kit's manual. Here's a summary of the procedure:

1- The RNA samples were adjusted to a final volume of 250 μL by adding RNase-free water.

2- To the mixture, 15 μL of Oligotex suspension and 250 μL of OBB buffer (both from the Oligotex kit) were added.

3- The mixture was initially incubated at 70°C for 3 minutes, followed by incubation at 25°C for 10 minutes.

4- The samples were then centrifuged at 14,000×g for 2 minutes, and the supernatants were discarded.

5- To the samples, 400 μL of Oligotex OW2 wash buffer was added, and they were centrifuged in Oligotex spin columns at 14,000×g for 1 minute. This step was repeated once.

6- Finally, the poly(A)+ RNA fraction was eluted from the membrane by adding 60 μL of hot Oligotex elution buffer. To maximize the yield, a second elution step was performed.

For the KSHV total RNA samples, the Lexogen Poly(A) RNA Selection Kit V1.5 was employed to select polyadenylated RNAs. The procedure was as follows:

1- A total of 10 μL of total RNA (5 μg) was denatured at 60°C for 1 minute and then held at 25°C.

2- The RNA samples were mixed with 10 μL of washed beads provided in the kit, and the mixtures were incubated at 25°C for 20 minutes with 1,250 rpm agitation.

3- After incubation, the tubes were placed in a magnetic rack for 5 minutes, and the supernatant was discarded.

4- The beads were then resuspended in Bead Wash Buffer from the Lexogen kit and incubated at 25°C for 5 minutes with 1,250 rpm agitation.

5- The tubes were transferred onto the magnetic rack, and the supernatant was discarded after 5 minutes of incubation. This washing step was repeated twice.

6- Following the second wash, the beads were resuspended in Nuclease-free water from the Lexogen kit and incubated at 70°C for 1 minute.

7- The tubes were again placed on the magnet for 5 minutes, and the supernatant containing the poly(A)+ RNA fraction was transferred to a fresh tube.

5.3.C Removal of rRNA

The Ribo-Zero Magnetic Kit H/M/R (Epicentre/Illumina) was used to enrich mRNAs and remove ribosomal mRNAs. This method also preserves non-polyadenylated RNAs, except for rRNAs. As a startup, 5 µg of EHV-1 total RNA was utilized. The sample was mixed with the Ribo-Zero Reaction Buffer and Ribo-Zero rRNA Removal Solution. The mixture was heated at 68°C for 10 min then a cooling down step at room temperature for 5 min. After that, 225 µl of pre-washed Magnetic beads mixture was added and incubated at room temperature for 5 min and then at 50°C for 5 min. Lastly, rRNA-depleted RNA was separated from the mixture using a magnet and the final purification was further with the Agencourt RNAClean XP Beads (Beckman Coulter) following the Ribo-Zero manual instructions.

5.3.D Enrichment of the 5′ ends of RNAs

Terminator™ 5′-Phosphate-Dependent Exonuclease (Lucigen) was used to enrich the 5′ ends of the transcripts. The process was carried out with a mixture of poly(A)+ RNAs from the EHV-1 samples, which was mixed with Terminator 10X Reaction Buffer A, RiboGuard RNase Inhibitor, and Terminator Exonuclease (1 Unit). The mixture was incubated at 30°C for 60 min, then the reaction was stopped by the addition of 1 µL of 100 mM EDTA (pH 8.0). RNAClean XP beads (Beckman Coulter) were used for final purification.

5.4 Measurement of nucleic acid quality and quantity

5.4.A RNA

The concentration of the RNA was measured with the Qubit Assay Kits (Invitrogen) and the Qubit4 fluorometer. To calculate the quantity of total RNA, RNA BR Assay was used but for poly(A)+ and ribodepleted RNA, RNA High Sensitivity (HS) Assay was applied. For quality checking, the Agilent TapeStation 4150 device and RNA ScreenTape were utilized to check the total RNA samples' quality. RIN scores above 9.6 were used for cDNA production. The RNA quality was assessed with the Agilent 2100 Bioanalyzer (for PacBio sequencing) or Agilent 4150 TapeStation System (for MinION sequencing) and RIN scores above 9.6 were used for cDNA production.

5.4.B cDNA

The Qubit dsDNA HS Assay Kit (Invitrogen) was used to quantify the cDNA samples. For the analysis of Illumina library quality, the Agilent High Sensitivity D1000 ScreenTape was used.

5.5 Long-read sequencing

5.5.A Direct cDNA sequencing

Library preparation was applied on the poly(A)+ fractions of RNAs from PRV, EHV-1, and KSHV as well as from the Terminator-treated EHV-1 samples., without using PCR amplification. For perfect achievement, the ONT's Direct cDNA Sequencing Kit (SQK-DCS109) was utilized following the ONT's protocol. First, RNA molecules were mixed with ONT VN primer and 10 mM dNTPs and heated at 65°C for 5 min. Then, RNaseOUT (Thermo Fisher Scientific), 5x RT Buffer (Thermo Fisher Scientific), and ONT Strand-Switching Primer were added to the mixtures, which were then incubated at 42°C for 2 min. Maxima H Minus Reverse Transcriptase enzyme (Thermo Fisher Scientific) was added to the samples to produce the first cDNA strands. The samples Were kept at 42°C for 90 min and then the reaction was stopped by heating at 85°C for 5 min. RNase Cocktail Enzyme Mix (Thermo Fisher Scientific) was used to remove the RNAs from the RNA:cDNA pairs. This took 10 min at 37°C.

After that, LongAmp Taq Master Mix [New England Biolabs (NEB)] and ONT PR2 Primer were used to make the second cDNA strands. PCR steps were applied: 1 min at 94

°C, 1 min 50°C, then 15 min at 65°C. Then, the end repair and dA-tailing were carried out with the NEBNext End repair /dA-tailing Module (NEB) reagents. These reactions were conducted at 20°C for 5 min and heating the samples at 65°C for 5 min. Next, adapter ligation with the NEB Blunt /TA Ligase Master Mix (NEB) at room temperature for 10 min. The ONT Native Barcoding (12) Kit was utilized to label the libraries, then they were loaded to ONT R9.4.1 SpotON Flow Cells (200 fmol mixture of libraries were loaded to one flow cell. AMPure XP Bead was used after each enzyme step and each sample was eluted in UltraPure™ nuclease-free water (Invitrogen).

5.5.B Amplified Nanopore cDNA sequencing

To map the 5′-end of VZV transcripts more precisely, the first step was random-primer-based RT reactions for library preparation; for this SuperScript IV enzyme (Life Technologies), poly(A)-selected and Terminator-treated RNA samples were used. Following the modified 1D strand switching cDNA by ligation protocol Ligation Sequencing kit (SQK-LSK108; Oxford Nanopore Technologies) to make libraries from the first-strand cDNAs. cDNAs were amplified with KAPA HiFi DNA Polymerase (Kapa Biosystems) and Ligation Sequencing Kit Primer Mix (part of the 1D Kit). To repair the ends of the samples, the NEBNext End repair / dA-tailing Module (New England Biolabs) was used, and then sequencing adapters were ligated, supplied with the kit, and NEB Blunt/TA Ligase Master Mix (New England Biolabs).

5.5.C Direct RNA sequencing (dRNA-seq)

A method called dRNA-seq was used to make libraries from EHV-1 and KSHV samples. This method can help us avoid errors that might happen when we copy RNA into DNA, reverse transcription (RT), or make more copies of DNA (PCR). We wanted to find and confirm novel splice variants as well as 3′ UTR isoforms. For the dRNA-seq experiments, we mixed RNA from different stages of infection from the samples that had a Poly(A)+ and from the Poly(A)+ and Terminator-treated RNAs. The oligo dT-containing T10 adapter for RT priming and the RNA CS for monitoring the sequencing quality (both from the ONT kit) was added to the RNA mix along with NEBNext Quick Ligation Reaction Buffer, and T4 DNA ligase (both from NEB), then let the reaction happen for 10 min at room temperature. next, adding 5X first-strand buffer, DTT (both from Invitrogen), dNTPs (NEB) as well as UltraPure™ DNase/RNase-Free water (Invitrogen). Finally, SuperScript III enzyme (Thermo Fisher Scientific) was added, and let the reaction happen at 50°C for 50 min, followed by heating it up at 70°C for 10 min to denature the enzyme, and subsequently stopping the reaction. RNA adapter (from the ONT kit) was ligated to the RNA:cDNA hybrid sample using the NEBNext Quick Ligation Reaction Buffer and T4 DNA ligase at room temperature for 10 min. The RNAClean XP Beads were used after each additional enzymatic step. Two flow cells were used for dRNA-seq, and 100 fmol from the sample was loaded onto each of them.

5.6 Short-read sequencing

To analyze the transcriptome of the EHV-1 virus using SRS, libraries were prepared from a combination of rRNA-depleted and poly(A)+ enriched samples with the NEXTflex® Rapid Directional qRNA-Seq Kit (PerkinElmer). The NEXTflex® RNA Fragmentation Buffer was used to break down the RNAs enzymatically at 95°C for 10 min. Then, the first strand of cDNA was synthesized. The RNA was mixed with the NEXTflex® First Strand Synthesis Primer and incubated at 65°C for 5 min, followed by drastic cooling. Next, perform the RT reaction with the NEXTflex® Directional First Strand Synthesis Buffer and Rapid Reverse Transcriptase following this protocol: 10 min at 25°C, 50 min at 50°C, and 15 min at 72°C. The second cDNA strands were generated by adding the NEXTflex® Directional Second Strand Synthesis Mix (with dUTPs) at 16°C for 60 min. NEXTflex® Adenylation Mix was used to adenylate the cDNAs at 37°C for 30 min and the reaction was stopped by heating the samples at 70°C for 5 min. Next, Molecular Index Adapters (part of the Kit) were ligated to the sample with the NEXTflex® Ligation Mix at 30°C for 10 min. For the PCR step: first, samples were mixed with the NEXTflex® Uracil DNA Glycosylase and incubated at 37°C for 30 min, then at 98°C for 2 min, and finally cooled on ice. Next, these components were added: PCR Master Mix, qRNA-Seq Universal forward primer, and qRNA-Seq Barcoded Primer (sequence: AACGCCAT; all from the PerkinElmer kit). PCR program: 2 min at 98 °C, then 15 cycles of 30 sec at 98°C, 30 sec at 65°C and 60 sec at 72°C, followed by a final extension of 4 min at 72°C. AMPure XP Beads were used after each enzymatic reaction. The sequencing library was eluted with the NEXTflex® Resuspension buffer and loaded 10 pM from it to the Illumina MiSeq reagent cassette. paired-end RNA sequencing was performed on an Illumina MiSeq sequencer with the MiSeq Reagent Kit v2 (300 cycles).

5.7 Real-time RT-PCR

The aim of using the quantitative qRT-PCR was to measure the kinetic and validate the transcripts. In short, first-strand cDNA was made from the total RNAs of six different viruses (PRV, EHV-1, VZV, HSV-1, BoHV-1, and KSHV) using SuperScript III reverse transcriptase and primers that matched each gene (Integrated DNA Technologies). For each tube, 0.1 μg of total RNA, 2 pmol of primer, dNTP mix (Invitrogen, 10 μM final concentration), 5x First-Strand Buffer, and the RT enzyme (Invitrogen) were mixed in 5μl of final volume and ran the reaction at 55°C for 60 min and then was stopped by heating to 70°C for 15 min. RT product was diluted ten times and used as a template for real-time PCR amplification. ABsolute QPCR SYBR Green Mix (Thermo Fisher Scientific) and a Rotor-Gene Q (Qiagen) cycler for PCR reactions were used, the run conditions. Controls: loading control (28S rRNA as a reference gene), no template (to check for primer-dimer formation), and no RT control (to check for DNA contamination). A mathematical model from Soong *et al.*[147] with some modifications was utilized to calculate the relative expression levels (R): The R values were calculated using the average ECt value of the 6h samples for each gene as a control, which was normalized with the average of the corresponding 28S values (ECt-reference). A special tool was used to get outstanding performance; the Comparative Quantitation module of the Rotor-Gene Q software package that automatically calculates the qPCR efficiency for each sample and sets the cycling threshold values simultaneously.

$$R = \frac{\left(E_{sample6h}\right)^{Ct_{sample6h}}}{\left(E_{sample}\right)^{Ct_{sample}}} : \frac{\left(E_{ref6h}\right)^{Ct_{ref6h}}}{\left(E_{re}\ \right)^{Ct_{ref}}}$$

We used the $2^{-\Delta\Delta Ct}$ method for the comparison of gene expression values of the PAA-treated samples with the untreated samples[148].

5.8 Cap Analysis of Gene Expression (CAGE)

The aim of using the CAGE-Seq method was to study how the TSSs are distributed in some regions of the genomes in EHV-1 and KSHV. The experiment was done in three parallels, by following the instructions of the CAGE™ Preparation Kit (DNAFORM, Japan) CAGE libraries were prepared from 5 µg of total RNA. In short, these steps were done: As a denaturation step, the RNA and an RT primer (random primer mixture from CAGE™ Prep Kit) were heated at 65°C for 5 min to separate the strands. First cDNA strand was synthesized with SuperScript III Reverse Transcriptase (Invitrogen). Enhancer was also utilized, a trehalose/sorbitol mixture (CAGE™ Prep Kit), to ensure the accuracy of the RT enzyme. The reactions were at 25°C for 30 sec and then at 50°C for 60 min.

The Diol group of the Cap at the 5′-end was oxidized (and ribose at the 3′-end) of the RNA with NaIO4 and then attached Biotin (long arm) hydrazine to it. First, the oxidation was carried out with the addition of NaOAc (1M, pH 4.5, CAGE™ Prep Kit) and NaIO4 (250mM, CAGE™ Prep Kit) to the samples and incubated on ice for 45 min in the dark. Then 40% glycerol and Tris-HCl (1M, pH 8.5, CAGE™ Prep Kit) were added to the samples. Next, NaOAc (1M, pH 6.0) and Biotin Hydrazine (10 mM, CAGE™ Prep Kit) were mixed with the samples and left to react at 23°C for 2 hours. This way, the oxidized diol groups were biotinylated.

Capped RNA samples were mixed and bound (Cap-trapping) to the pretreated Streptavidin beads (pretreatment details below) at 37°C 30 min. Then they were incubated on a magnetic rack. The beads were washed with Wash Buffer 1 (twice), then with Wash Buffer 2, and finally with Wash Buffer 3 (both from the CAGE™ Prep Kit). Next, cDNAs were released from the beads: Releasing Buffer was added to the samples and they were incubated at 95°C for 5 min. After a short incubation on a magnetic rack, the supernatant (containing the Capped cDNAs) was transferred to new tubes. RNase I buffer (CAGE™ Prep Kit) was added to the tRNA-Streptavidin bead and they were placed on a magnetic rack. The supernatant was transferred to the tubes containing the cDNAs and they were stored on ice. The samples were treated with an RNase mixture (RNase H and RNase I, both from the CAGE™ Prep Kit) and incubated at 37°C for 15 min. The potential remaining RNA was digested with RNase I. The reaction was performed at 37°C for 30 min. The following steps were performed to prepare the samples for sequencing:

1. Streptavidin beads were coated with tRNA (from CAGE™ Prep Kit), followed by mixing and incubation on ice for 30 min, and then magnetically separated for 3 min. The supernatant was sniped out and they were washed and eluted twice with Wash Buffer 1 (from CAGE™ Prep Kit) and tRNA was also added.

2. The samples were concentrated using a miVac DUO Centrifugal Concentrator (Genevac) and then ligated to barcoded 5′ linkers (from CAGE™ Prep Kit) at 16°C overnight. After a purification step, the miVac DUO was used again in order to concentrate the samples.

3. The 3′ linkers (from CAGE™ Prep Kit) were ligated to the samples at 16°C overnight after pre-heating the linkers and samples at 55°C and 95°C respectively.

4. The linkers were dephosphorylated with Shrimp Alkaline Phosphatase (SAP, from CAGE™ Prep Kit) at 37°C for 30 min and then inactivated at 65°C in 15 min.

5. The samples were treated with USER enzyme to remove the dUTP from the 3′ linker up strand at 37°C for 30 min and then terminated at 95°C in 5 min.

6. The barcoded samples were pooled and concentrated using a miVac DUO device.

7. The second cDNA strands were synthesized with the 2nd primer, DNA polymerase, buffer, and dNTPs (all from CAGE™ Prep Kit) using a thermal cycling protocol; 95°C for 5 min, 55°C for 5 min, and 72°C for 30 min.

8. The sample mixture was digested with Exonuclease I enzyme at 37°C for 30 min and then dried completely using a vacuum concentrator. The sample was resuspended in 10 µl of nuclease-free water.

9. The single-stranded cDNA amount was quantified using Qubit 2.0 and Qubit ssDNA HS Assay Kit.

10. RNAClean XP Beads and AmpureXP Beads were used for purification after various steps as described above.

11. The libraries with different barcodes were sequenced on a MiSeq instrument (Illumina) using v3 (150 cycles) and v2 (300 cycles) chemistries.

12. Qubit 4.0 and 1X dsDNA High Sensitivity (HS) Assay were used to estimate the sample concentration and Tape-Station was used to check the library quality.

5.9 Bioinformatic analyses

5.9.A Illumina CAGE sequencing data analysis

The quality of the reads was assessed by using **fastqc** (https://www.bioinformatics.babraham.ac.uk/projects/fastqc). The reads were then trimmed using **TrimGalore** (https://github.com/FelixKrueger/TrimGalore) with the following settings: -length 151 -q. To map the reads to the KSHV strain TREx reference genome (GQ994935.1), version 2.7.3.a of the **STAR** aligner[149] was employed, with the parameter --*genomeSAindexNbases* set to 8 and other parameters set to their default values. TSSs and TSS clusters were identified using the **CAGEfightR** R package[150], with a minimum pooled value cutoff of 0.1 (pooledcutoff=0.1).

5.9.B ONT sequencing data analysis

For the analysis of ONT sequencing data, the ONT-MinION sequencing reads were basecalled using **Guppy** software (v3.4.5). To ensure high quality, reads with a quality filter of 8 (default) were selected and then mapped to the reference genome using **minimap2**[151] with the following settings: *-ax splice -Y -C5 -cs.* Mapping statistics were computed by using the (**ReadStatistics)** script from **Seqtools** (https://github.com/moldovannorbert/seqtools). The LoRTIA toolkit (alpha version, accessed on 20 August 2019, https://github.com/zsolt-balazs/LoRTIA) was utilized to identify TESs, TSSs, and introns, and subsequently, transcripts were reconstructed based on these identified features.

The LoRTIA workflow was implemented with default settings, which involved the following steps:

1.) For dRNA and dcDNA sequencing, the following parameters were used: −5 TGCCATTAGGCCGGG --five_score 16 --check_in_soft 15 −3 AAAAAAAAAAAAAAA --three_score 16s Poisson–f true.

2.) For o(dT)-primed cDNA reads, the parameters used were: −5 GCTGATATTGCTGGG --five_score 16 --check_in_soft 15 −3 AAAAAAAAAAAAAAA --three_score 16s Poisson–f true.

To search for adapter sequences in the case of EHV-1, the following command was executed: samprocessor.py --five_adapter GCTGATATTGCTGGG --five_score 14 --check_in_soft 15 --three_adapter AAAAAAAAAAAAAAAAAAA --three_score 14 input output. The subsequent step in the workflow involved the annotation of TSS and TES. For TES positions, the wobble value was increased to 20, while the default wobble value was retained for TSS. The 'sam' files were then processed using the following parameters: Stats.py -r genome -f r5 -b 10 and Stats.py -r genome -f l5 -b 10 for TSS detection, and Stats.py -r genome -f r3 -b 20 and Stats.py -r genome -f l3 -b 20 for TES detection. Additionally, Stats.py -r genome -f in was used for intron detection.

Various sequencing techniques were employed in the analysis of BoHV-1, and the LoRTIA workflow was customized accordingly for each technique. For dRNA sequencing, the following parameters were applied: *LoRTIA -5 AGAGTACATGGG --five_score 16 --check_in_soft 15 -3 AAAAAAAAAAAAAAAAAAAAA --three_score 12 -s poisson -f True.* Oligo d(T) cDNA sequencing utilized these parameters: *LoRTIA -5 TGCCATTAGGCCGGGGG --five_score 14 --check_in_soft 15 -3 AAAAAAAAAAAAAAAAAAAAA --three_score 14 -s poisson -f True.* Random primer cDNA sequencing involved the following parameters: *LoRTIA -5 TGCCATTAGGCCGGGGG --five_score 14 --check_in_soft 15 -3 GAAGATAGAGAGCGACA --three_score 14 -s poisson -f True.* Lastly, dcDNA sequencing utilized these parameters: *LoRTIA -5 GCTGATATTATTGCTGGG --five_score 16 --check_in_soft 15 -3 AAAAAAAAAAAAAAAAAAAAAAA --three_score 14 -s poisson -f True.*

A TSS (transcription start site) was considered valid if the adapters were correctly identified, while TESs (transcription end sites) were considered valid if they had polyA tails and no false priming events were detected by LoRTIA. In the case of KSHV, EBV, and EHV-1, TSSs were accepted only if they were supported by at least one dcDNA read, as well as one dRNA or CAGE read. However, for CMV, BoHV-1, PRV, and HSV-1, TSSs were accepted only if they were detected in at least three different samples. When it came to introns, we only considered those identified in dRNA sequencing, as this method is considered the "*Gold Standard*" for identifying alternative splicing variants. Additionally, certain transcripts were manually included if they represented a longer variant of already accepted TSSs. To identify promoter elements around the accepted TSSs, the **MotifFinder** tool from **Seqtools** was utilized.

5.9.C Downstream data analysis and visualization:

Data analysis downstream was performed using R, utilizing packages such as **GenomicRanges**[152], **tidygenomics**[153], and packages from the **tidyverse**[154]. **Gviz**[155] was employed to create **Figures 2-4** and **Supplementary Figures 2-7** (https://github.com/ivanek/Gviz). **Figure 5** was generated using a customized R program. In summary, the '.bam' files were imported into R using **Rsamtools** (https://bioconductor.org/packages/Rsamtools). The 5′ ends were then summed per genomic position, and a density plot was created using **ggplot2's** *geom_density function*, with parameters adjust = 0.025 and default settings for other parameters (including the default Gaussian kernel). The density plot, along with genome annotation, was visualized using a custom plotting function that utilized gggenes (https://github.com/wilkox/gggenes). These scripts can be applied to import other alignments into R and generate similar plots based on their 3′ or 5′ distributions on reference genomes. The scripts are available on GitHub at https://github.com/Balays/R.codes

5.9.D Data availability

Our data used in this study has been deposited to in European Nucleotide Archive (ENA) under the following accessions:

PRJEB24593 (https://www.ebi.ac.uk/ena/browser/view/PRJEB24593), ERP106430 (https://www.ebi.ac.uk/ena/browser/view/PRJEB24593), PRJEB33511 (https://www.ebi.ac.uk/ena/browser/view/PRJEB33511), PRJEB38992 (https://www.ebi.ac.uk/ena/browser/view/PRJEB38992), PRJEB22072 (https://www.ebi.ac.uk/ena/browser/view/PRJEB22072), PRJEB25680 (https://www.ebi.ac.uk/ena/browser/view/PRJEB25680), ERP019579 (https://www.ebi.ac.uk/ena/browser/view/PRJEB17709), PRJEB25401 (https://www.ebi.ac.uk/ena/browser/view/PRJEB25401), PRJEB25433 (https://www.ebi.ac.uk/ena/browser/view/PRJEB25433), PRJEB9526 (https://www.ebi.ac.uk/ena/browser/view/PRJEB9526), PRJEB12867 (https://www.ebi.ac.uk/ena/browser/view/PRJEB12867) and to Gene Expression Omnibus (GEO) under the accession number: GSE97785 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE97785).

We also acquired datasets from other groups, which were downloaded from GEO: GSE79337 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE79337), GSE59717 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59717), GSE128324 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE128324), from Sequence Read Archive: PRJNA505045 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA505045/), PRJNA482043 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA482043), PRJNA483305 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA483305), PRJNA533478 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA533478), and from ENA: PRJEB27861 (https://www.ebi.ac.uk/ena/browser/view/PRJEB27861), PRJEB42868 (https://www.ebi.ac.uk/ena/browser/view/PRJEB42868), PRJEB38829 (https://www.ebi.ac.uk/ena/browser/view/PRJEB38829). https://www.ebi.ac.uk/ena/browser/view/PRJEB64684?fbclid=IwAR2U3rT3i0MTVwb2xYpi9I92DOuuyuvLA8aAtWD48Qip6tw11cxpczt_hLg, PRJEB64684

5.9.E Code availability

LoRTIA: https://github.com/zsolt-balazs/LoRTIA.

R scripts: https://github.com/Balays/Rlyeh
R workflow: https://github.com/Balays/KSHV_RNASeq

6. Conclusions

RNA sequencing has been a pioneering technique in understanding viruses, significantly progressing the acquaintance of viral biology. After characterization of complete genetic sequence of viruses, transcriptomics offers productive visions into their genome structure, organization, and diversity, allowing the identification of viral genes responsible for important functions like replication, transcription, and protein synthesis. By pairwise analysis of viral transcriptomes, RNA sequencing exposes the evolutionary relationship between viruses. Allowing over times genetic-changes track down possible.

The use of long-read sequencing (LRS) techniques has revolutionized transcriptomic research by revealing unexpected transcriptomic complexity in various organisms, including viruses. In this study, we used both newly generated and previously published LRS and short-read sequencing datasets to discover additional Ori-proximal transcripts in nine herpesviruses belonging to all of the three subfamilies (alpha, beta and gamma). We identified novel long non-coding RNAs (lncRNAs), as well as splice and length isoforms of mRNAs and lncRNAs. The analysis revealed an intricate network of transcriptional overlaps, suggesting the existence of a "super regulatory center" that controls both replication and global transcription through multilevel interactions between molecular components

EBV: This research employs a comprehensive and combined sequencing method to gain a deeper understanding of the transcriptomic structure of EBV, a significant human pathogen. By utilizing this approach, we discovered several previously unknown transcripts and RNA isoforms, which include variations in transcript length and splicing patterns. Additionally, we found new genes integrated within longer host genes that possess 5'-truncated in-frame open reading frames, suggesting the possibility of encoding N-terminally truncated proteins. The study also uncovered novel non-coding RNAs, as well as both mono- and multigenic transcripts. This collective data provides a more comprehensive and detailed view of EBV's transcriptomic landscape.

To sum up, RNA sequencing has reformed virology and revolutionized applications for understanding viral biology, diagnosing infections, and studying viral evolution. As RNA sequencing remains to be utilized and upgraded, it will undeniably improve ability of dealing with viral diseases effectively over time.

7. Acknowledgments

This research was supported by the National Research, Development and Innovation Office (NRDIO), Researcher-initiated research projects (Grant numbers: K 128247 and K 142674) to ZB, and by the NRDIO Research projects initiated by young researchers (Grant number: FK 128252) to DT. The work was also supported by NIH grant R01AI132554 to ZT. The APC fee was covered by the University of Szeged, Open Access Fund: 5928. ZC was also by the UNKP-21-4-SZTE-126.

References

1. Whitley, R. J. & Roizman, B. Herpes simplex viruses. in *Clinical Virology: Third Edition* (2022). doi:10.1128/9781555815981.ch19.
2. DeLuca, N. A., McCarthy, A. M. & Schaffer, P. A. Isolation and characterization of deletion mutants of herpes simplex virus type 1 in the gene encoding immediate-early regulatory protein ICP4. *J Virol* **56**, (1985).
3. Fox, H. L., Dembowski, J. A. & DeLuca, N. A. A herpesviral immediate early protein promotes transcription elongation of viral transcripts. *mBio* **8**, (2017).
4. Hagglund, R. & Roizman, B. Role of ICP0 in the Strategy of Conquest of the Host Cell by Herpes Simplex Virus 1. *J Virol* **78**, (2004).
5. Zhou, C. & Knipe, D. M. Association of Herpes Simplex Virus Type 1 ICP8 and ICP27 Proteins with Cellular RNA Polymerase II Holoenzyme. *J Virol* **76**, (2002).
6. Aubert, M. & Blaho, J. A. The Herpes Simplex Virus Type 1 Regulatory Protein ICP27 Is Required for the Prevention of Apoptosis in Infected Human Cells. *J Virol* **73**, (1999).
7. Mark F. Stinski and Jeffery L. Meier. *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis*. (Cambridge University Press, 2007).
8. Liu, P. & Speck, S. H. Synergistic autoactivation of the Epstein-Barr virus immediate-early BRLF1 promoter by Rta and Zta. *Virology* **310**, (2003).
9. Perng, G. C. *et al.* The latency-associated transcript gene of herpes simplex virus type 1 (HSV-1) is required for efficient in vivo spontaneous reactivation of HSV-1 from latency. *J Virol* **68**, (1994).
10. Chen, S.-H. H., Kramer, M. F., Schaffer, P. A., And, ‡ & Coen, D. M. A viral function represses accumulation of transcripts from productive-cycle genes in mouse ganglia latently infected with herpes simplex virus. *J Virol* **71**, 5878–5884 (1997).
11. Wang, Q. Y. *et al.* Herpesviral latency-associated transcript gene promotes assembly of heterochromatin on viral lytic-gene promoters in latent infection. *Proc Natl Acad Sci U S A* **102**, (2005).
12. Boldogköi, Z., Murvai, J. & Fodor, I. G and C accumulation at silent positions of codons produces additional ORFs. *Trends in Genetics* vol. 11 Preprint at https://doi.org/10.1016/S0168-9525(00)89019-8 (1995).
13. Tormanen, K., Allen, S., Mott, K. R. & Ghiasi, H. The Latency-Associated Transcript Inhibits Apoptosis via Downregulation of Components of the Type I Interferon Pathway during Latent Herpes Simplex Virus 1 Ocular Infection. *J Virol* **93**, (2019).

14. Wesley, R. D. & Cheung, A. K. A pseudorabies virus mutant with deletions in the latency and early protein O genes: Replication, virulence, and immunity in neonatal piglets. *Journal of Veterinary Diagnostic Investigation* **8**, (1996).

15. Lee, L. Y. & Schaffer, P. A. A Virus with a Mutation in the ICP4-Binding Site in the L/ST Promoter of Herpes Simplex Virus Type 1, but Not a Virus with a Mutation in Open Reading Frame P, Exhibits Cell-Type-Specific Expression of γ 1 34.5 Transcripts and Latency-Associated Transcripts . *J Virol* **72**, (1998).

16. Tombácz, D. *et al.* Full-length isoform sequencing reveals novel transcripts and substantial transcriptional overlaps in a herpesvirus. *PLoS One* **11**, 1–29 (2016).

17. Tombácz, D. *et al.* Full-length isoform sequencing reveals novel transcripts and substantial transcriptional overlaps in a herpesvirus. *PLoS One* **11**, (2016).

18. Tombácz, D. *et al.* Hybrid Sequencing Reveals Novel Features in the Transcriptomic Organization of Equid Alphaherpesvirus. *SSRN Electronic Journal* (2022) doi:10.2139/ssrn.4141334.

19. Moldován, N. *et al.* Multi-platform sequencing approach reveals a novel transcriptome profile in pseudorabies virus. *Front Microbiol* **8**, (2018).

20. Mackiewicz, P., Zakrzewska-Czerwińska, J., Zawilak, A., Dudek, M. R. & Cebrat, S. Where does bacterial replication start? Rules for predicting the oriC region. *Nucleic Acids Res* **32**, (2004).

21. Vashee, S. *et al.* Sequence-independent DNA binding and replication initiation by the human origin recognition complex. *Genes Dev* **17**, (2003).

22. Erratum: Origins of DNA replication (PLoS Genetics 15:9 (e1008320) DOI: 10.1371/journal.pgen.1008320). *PLoS Genetics* vol. 15 Preprint at https://doi.org/10.1371/journal.pgen.1008556 (2019).

23. Boldogkői, Z., Tombácz, D. & Balázs, Z. Interactions between the transcription and replication machineries regulate the RNA and DNA synthesis in the herpesviruses. *Virus Genes* vol. 55 Preprint at https://doi.org/10.1007/s11262-019-01643-5 (2019).

24. Santocanale, C. & Diffley, J. F. X. ORC- and Cdc6-dependent complexes at active and inactive chromosomal replication origins in Saccharomyces cerevisiae. *EMBO Journal* **15**, (1996).

25. Elias, P. & Lehman, I. R. Interaction of origin binding protein with an origin of replication of herpes simplex virus 1. *Proc Natl Acad Sci U S A* **85**, (1988).

26. Weller, S. K. & Coen, D. M. Herpes simplex viruses: Mechanisms of DNA replication. *Cold Spring Harb Perspect Biol* **4**, (2012).

27. Packard, J. E. & Dembowski, J. A. HSV-1 dna replication—coordinated regulation by viral and cellular factors. *Viruses* vol. 13 Preprint at https://doi.org/10.3390/v13102015 (2021).

28. Hammerschmidt, W. & Sugden, B. Identification and characterization of oriLyt, a lytic origin of DNA replication of Epstein-Barr virus. *Cell* **55**, (1988).

29. Ballestas, M. E., Chatis, P. A. & Kaye, K. M. Efficient persistence of extrachromosomal KSHV DNA mediated by latency- associated nuclear antigen. *Science (1979)* **284**, (1999).

30. Sun, Q. *et al.* Kaposi's sarcoma-associated herpesvirus LANA recruits the DNA polymerase clamp loader to mediate efficient replication and virus persistence. *Proc Natl Acad Sci U S A* **111**, (2014).

31. AuCoin, D. P., Colletti, K. S., Xu, Y., Cei, S. A. & Pari, G. S. Kaposi's sarcoma-associated herpesvirus (human herpesvirus 8) contains two functional lytic origins of DNA replication. *J Virol* **76**, 7890–7896 (2002).

32. Dheekollu, J. *et al.* Cell Cycle-Dependent EBNA1-DNA Cross-Linking Promotes Replication Termination at oriP and Viral Episome Maintenance. *Cell* **184**, 643 (2021).

33. Hammerschmidt, W. & Sugden, B. Replication of Epstein–Barr Viral DNA. *Cold Spring Harb Perspect Biol* **5**, (2013).

34. Boldogkői, Z., Moldován, N., Balázs, Z., Snyder, M. & Tombácz, D. Long-Read Sequencing – A Powerful Tool in Viral Transcriptome Research. *Trends in Microbiology* vol. 27 Preprint at https://doi.org/10.1016/j.tim.2019.01.010 (2019).

35. Ma, Y. *et al.* Human CMV transcripts: An overview. *Future Microbiology* vol. 7 Preprint at https://doi.org/10.2217/fmb.12.32 (2012).

36. Kakuk, B. *et al.* Combined nanopore and single-molecule real-time sequencing survey of human betaherpesvirus 5 transcriptome. *Sci Rep* **11**, (2021).

37. Tombácz, D. *et al.* Transcriptomewide survey of pseudorabies virus using next-A nd third-generation sequencing platforms. *Sci Data* **5**, (2018).

38. Barkley, L. R. & Santocanale, C. MicroRNA-29a regulates the benzo[a]pyrene dihydrodiol epoxide-induced DNA damage response through Cdc7 kinase in lung cancer cells. *Oncogenesis* **2**, (2013).

39. Marchese, F. P. & Huarte, M. A long noncoding RNA in DNA replication and chromosome dynamics. *Cell Cycle* vol. 16 Preprint at https://doi.org/10.1080/15384101.2016.1241604 (2017).

40. Dellino, G. I. *et al.* Genome-wide mapping of human DNA-replication origins: Levels of transcription at ORC1 sites regulate origin selection and replication timing. *Genome Res* **23**, 1 (2013).

41. Tikhanovich, I., Liang, B., Seoighe, C., Folk, W. R. & Nasheuer, H. P. Inhibition of Human BK Polyomavirus Replication by Small Noncoding RNAs. *J Virol* **85**, (2011).

42. Tai-Schmiedel, J. *et al.* Human cytomegalovirus long noncoding RNA4.9 regulates viral DNA replication. (2020) doi:10.1371/journal.ppat.1008390.

43. Gatherer, D. *et al.* High-resolution human cytomegalovirus transcriptome. *Proc Natl Acad Sci U S A* **108**, (2011).

44. Rossetto, C. C., Tarrant-Elorza, M. & Pari, G. S. Cis and Trans Acting Factors Involved in Human Cytomegalovirus Experimental and Natural Latent Infection of CD14 (+) Monocytes and CD34 (+) Cells. *PLoS Pathog* **9**, (2013).

45. Huang, L., Zhu, Y. & Anders, D. G. The variable 3' ends of a human cytomegalovirus oriLyt transcript (SRT) overlap an essential, conserved replicator element. *J Virol* **70**, (1996).

46. Prichard, M. N. *et al.* Identification of Persistent RNA-DNA Hybrid Structures within the Origin of Replication of Human Cytomegalovirus. *J Virol* **72**, (1998).

47. Rennekamp, A. J. & Lieberman, P. M. Initiation of Epstein-Barr Virus Lytic Replication Requires Transcription and the Formation of a Stable RNA-DNA Hybrid Molecule at OriLyt. *J Virol* **85**, (2011).

48. Xu, Y., Cei, S. A., Rodriguez Huete, A., Colletti, K. S. & Pari, G. S. Human Cytomegalovirus DNA Replication Requires Transcriptional Activation via an IE2- and UL84-Responsive Bidirectional Promoter Element within ori Lyt . *J Virol* **78**, (2004).

49. Norseen, J. *et al.* RNA-dependent recruitment of the origin recognition complex. *EMBO Journal* **27**, (2008).

50. Voss, J. H. & Roizman, B. Properties of two 5'-coterminal RNAs transcribed part way and across the S component origin of DNA synthesis of the herpes simplex virus 1 genome. *Proc Natl Acad Sci U S A* **85**, (1988).

51. Tombacz, D. *et al.* Characterization of novel transcripts in pseudorabies virus. *Viruses* **7**, (2015).

52. Tombácz, D. *et al.* Characterization of the Dynamic Transcriptome of a Herpesvirus with Long-read Single Molecule Real-Time Sequencing. *Sci Rep* **7**, (2017).

53. Moldován, N. *et al.* Time-course profiling of bovine alphaherpesvirus 1.1 transcriptome using multiplatform sequencing. *Sci Rep* **10**, (2020).

54. Tombácz, D. *et al.* Article 834 ˝i Z (2019) Multiple Long-Read Sequencing Survey of Herpes Simplex Virus Dynamic Transcriptome. *Front. Genet* **10**, 834 (2019).

55. Torma, G. *et al.* An integrated sequencing approach for updating the pseudorabies virus transcriptome. *Pathogens* **10**, (2021).

56. Davison, A. J. *et al.* The order Herpesvirales. *Archives of Virology* vol. 154 Preprint at https://doi.org/10.1007/s00705-008-0278-4 (2009).

57. Young, L. S., Yap, L. F. & Murray, P. G. Epstein–Barr virus: more than 50 years old and still providing surprises. *Nature Reviews Cancer 2016 16:12* **16**, 789–802 (2016).

58. Rochford, R., Muenz, C., Chabay, P., Rickinson, A. & Shannon-Lowe, C. The Global Landscape of EBV-Associated Tumors. *Frontiers in Oncology | www.frontiersin.org* **1**, 713 (2019).

59. de Martel, C., Georges, D., Bray, F., Ferlay, J. & Clifford, G. M. Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. *Lancet Glob Health* **8**, (2020).

60. Schaeffner, M. *et al.* BZLF1 interacts with chromatin remodelers promoting escape from latent infections with EBV. *Life Sci Alliance* **2**, (2019).

61. Nagaraju, T., Sugden, A. U. & Sugden, B. 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**, (2019).

62. Hammerschmidt, W. & Sugden, B. Replication of Epstein-Barr Viral DNA. doi:10.1101/cshperspect.a013029.

63. Djavadian, R., Hayes, M. & Johannsen, E. CAGE-seq analysis of Epstein-Barr virus lytic gene transcription: 3 kinetic classes from 2 mechanisms. *PLoS Pathog* **14**, (2018).

64. Yuan, J., Cahir-McFarland, E., Zhao, B. & Kieff, E. Virus and Cell RNAs Expressed during Epstein-Barr Virus Replication. *J Virol* **80**, (2006).

65. Dresang, L. R. *et al.* Coupled transcriptome and proteome analysis of human lymphotropic tumor viruses: Insights on the detection and discovery of viral genes. *BMC Genomics* **12**, (2011).

66. Arvey, A. *et al.* Cell Host & Microbe Resource An Atlas of the Epstein-Barr Virus Transcriptome and Epigenome Reveals Host-Virus Regulatory Interactions. doi:10.1016/j.chom.2012.06.008.

67. Ersing, I. *et al.* A Temporal Proteomic Map of Epstein-Barr Virus Lytic Replication in B Cells. *Cell Rep* **19**, (2017).

68. O'Grady, T. *et al.* Global Bidirectional Transcription of the Epstein-Barr Virus Genome during Reactivation. *J Virol* **88**, (2014).

69. Cao, S. *et al.* 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. (2015) doi:10.1128/JVI.00608-15.

70. O'Grady, T. *et al.* Global transcript structure resolution of high gene density genomes through multi-platform data integration. *Nucleic Acids Res* **44**, e145 (2016).

71. Majerciak, V., Yang, W., Zheng, J., Zhu, J. & Zheng, Z.-M. A Genome-Wide Epstein-Barr Virus Polyadenylation Map and Its Antisense RNA to EBNA. doi:10.1128/JVI.01593-18.

72. Concha, M. *et al.* Identification of New Viral Genes and Transcript Isoforms during Epstein-Barr Virus Reactivation using RNA-Seq. (2011) doi:10.1128/JVI.06537-11.

73. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**, (2014).

74. Balázs, Z., Tombácz, D., Szucs, A., Snyder, M. & Boldogkoi, Z. Data Descriptor: Long-read sequencing of the human cytomegalovirus transcriptome with the Pacific Biosciences RSII platform. *Sci Data* **4**, (2017).

75. Moldován, N. *et al.* Third-generation Sequencing Reveals Extensive Polycistronism and Transcriptional Overlapping in a Baculovirus. *Sci Rep* **8**, (2018).

76. Moldován, N. *et al.* Multi-platform analysis reveals a complex transcriptome architecture of a circovirus. *Virus Res* **237**, (2017).

77. Prazsák, I. *et al.* Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus. *BMC Genomics* **19**, (2018).

78. Tombácz, D. *et al.* Article 834 ̋i Z (2019) Multiple Long-Read Sequencing Survey of Herpes Simplex Virus Dynamic Transcriptome. *Front. Genet* **10**, 834 (2019).

79. Oláh, P. *et al.* Characterization of pseudorabies virus transcriptome by Illumina sequencing. *BMC Microbiol* **15**, (2015).

80. Braspenning, S. E. *et al.* The architecture of the simian varicella virus transcriptome. *PLoS Pathog* **17**, (2021).

81. Balázs, Z., Tombácz, D., Szűcs, A., Snyder, M. & Boldogkői, Z. Dual Platform Long-Read RNA-Sequencing Dataset of the Human Cytomegalovirus Lytic Transcriptome. *Front Genet* **9**, (2018).

82. Fülöp, Á. *et al.* Integrative profiling of Epstein–Barr virus transcriptome using a multiplatform approach. *Virol J* **19**, (2022).

83. Rutkowski, A. J. *et al.* Widespread disruption of host transcription termination in HSV-1 infection. *Nat Commun* **6**, (2015).

84. Pheasant, K. *et al.* Nuclear-cytoplasmic compartmentalization of the herpes simplex virus 1 infected cell transcriptome is co-ordinated by the viral endoribonuclease vhs and cofactors to facilitate the translation of late proteins. *PLoS Pathog* **14**, (2018).

85. Tang, S., Patel, A. & Krause, P. R. Hidden regulation of herpes simplex virus 1 pre-mRNA splicing and polyadenylation by virally encoded immediate early gene ICP27. *PLoS Pathog* **15**, (2019).

86. Whisnant, A. W. *et al.* Integrative functional genomics decodes herpes simplex virus 1. *Nat Commun* **11**, (2020).

87. Tombácz, D. *et al.* Strain Kaplan of pseudorabies virus genome sequenced by PacBio single-molecule real-time sequencing technology. *Genome Announc* **2**, (2014).

88. Csabai, Z., Tombácz, D., Deim, Z., Snyder, M. & Boldogkoi, Z. Analysis of the complete genome sequence of a novel, pseudorabies virus strain isolated in Southeast Europe. *Canadian Journal of Infectious Diseases and Medical Microbiology* **2019**, (2019).

89. Hein, M. Y. & Weissman, J. S. Functional single-cell genomics of human cytomegalovirus infection. *Nat Biotechnol* **40**, (2022).

90. Koons, M. D., Van Scoy, S. & Hearing, J. The Replicator of the Epstein-Barr Virus Latent Cycle Origin of DNA Replication, oriP , Is Composed of Multiple Functional Elements . *J Virol* **75**, (2001).

91. Purushothaman, P., Uppal, T. & Verma, S. C. Molecular biology of KSHV lytic reactivation. *Viruses* vol. 7 Preprint at https://doi.org/10.3390/v7010116 (2015).

92. Zhu, F. X., Cusano, T. & Yuan, Y. Identification of the Immediate-Early Transcripts of Kaposi's Sarcoma-Associated Herpesvirus. *J Virol* **73**, (1999).

93. Majerciak, V., Alvarado-Hernandez, B., Lobanov, A., Cam, M. & Zheng, Z. M. Genome-wide regulation of KSHV RNA splicing by viral RNA-binding protein ORF57. *PLoS Pathog* **18**, (2022).

94. Arias, C. *et al.* KSHV 2.0: A Comprehensive Annotation of the Kaposi's Sarcoma-Associated Herpesvirus Genome Using Next-Generation Sequencing Reveals Novel Genomic and Functional Features. *PLoS Pathog* **10**, (2014).

95. Gregory Bruce, A. *et al.* Quantitative analysis of the KSHV transcriptome following primary infection of blood and lymphatic endothelial cells. *Pathogens* **6**, (2017).

96. Schifano, J. M., Corcoran, K., Kelkar, H. & Dittmer, D. P. Expression of the Antisense-to-Latency Transcript Long Noncoding RNA in Kaposi's Sarcoma-Associated Herpesvirus. *J Virol* **91**, (2017).

97. Pearce, M., Matsumura, S. & Wilson, A. C. Transcripts Encoding K12, v-FLIP, v-Cyclin, and the MicroRNA Cluster of Kaposi's Sarcoma-Associated Herpesvirus Originate from a Common Promoter. *J Virol* **79**, (2005).

98. Tombácz, D. *et al.* In-Depth Temporal Transcriptome Profiling of an Alphaherpesvirus Using Nanopore Sequencing. *Viruses* **14**, (2022).

99. Braspenning, S. E. *et al.* Decoding the architecture of the varicella-zoster virus transcriptome. *mBio* **11**, (2020).

100. Ye, X., Zhaoid, Y. & Karijolich, J. The landscape of transcription initiation across latent and lytic KSHV genomes. *PLoS Pathog* **15**, (2019).

101. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods* **10**, (2013).

102. Palla, P., Frau, G., Vargiu, L. & Rodriguez-Tomé, P. QTreds: a flexible LIMS for omics laboratories. *EMBnet J* **18**, (2012).

103. Tian, B. & Graber, J. H. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdisciplinary Reviews: RNA* vol. 3 Preprint at https://doi.org/10.1002/wrna.116 (2012).

104. Leppek, K., Das, R. & Barna, M. Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nature Reviews Molecular Cell Biology* vol. 19 Preprint at https://doi.org/10.1038/nrm.2017.103 (2018).

105. Calvo, S. E., Pagliarini, D. J. & Mootha, V. K. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci U S A* **106**, (2009).

106. Kronstad, L. M., Brulois, K. F., Jung, J. U. & Glaunsinger, B. A. Dual Short Upstream Open Reading Frames Control Translation of a Herpesviral Polycistronic mRNA. *PLoS Pathog* **9**, (2013).

107. Watanabe, T. *et al.* Roles of Epstein-Barr virus BGLF3.5 gene and two upstream open reading frames in lytic viral replication in HEK293 cells. *Virology* **483**, (2015).

108. Gershburg, E., Raffa, S., Torrisi, M. R. & Pagano, J. S. Epstein-Barr Virus-Encoded Protein Kinase (BGLF4) Is Involved in Production of Infectious Virus. *J Virol* **81**, (2007).

109. Han, Z., Verma, D., Hilscher, C., Dittmer, D. P. & Swaminathan, S. General and Target-Specific RNA Binding Properties of Epstein-Barr Virus SM Posttranscriptional Regulatory Protein. *J Virol* **83**, (2009).

110. Batisse, J., Manet, E., Middeldorp, J., Sergeant, A. & Gruffat, H. Epstein-Barr Virus mRNA Export Factor EB2 Is Essential for Intranuclear Capsid Assembly and Production of gp350. *J Virol* **79**, (2005).

111. Crofts, L. A., Hancock, M. S., Morrison, N. A. & Eisman, J. A. Multiple promoters direct the tissue-specific expression of novel N-terminal variant human vitamin D receptor gene transcripts. *Proc Natl Acad Sci U S A* **95**, (1998).

112. Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. doi:10.1093/nar/gkt006.

113. Hangauer, M. J., Vaughn, I. W. & McManus, M. T. Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *PLoS Genet* **9**, (2013).

114. Sequeira-Mendes, J. *et al.* Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS Genet* **5**, (2009).

115. Tai-Schmiedel, J. *et al.* Human cytomegalovirus long noncoding RNA4.9 regulates viral DNA replication. (2020) doi:10.1371/journal.ppat.1008390.

116. Wang, Y., Tang, Q., Maul, G. G. & Yuan, Y. Kaposi's Sarcoma-Associated Herpesvirus ori-Lyt-Dependent DNA Replication: Dual Role of Replication and Transcription Activator. *J Virol* **80**, 12171–12186 (2006).

117. Boldogkői, Z. *et al.* Transcriptomic study of herpes simplex virus type-1 using full-length sequencing techniques. *Sci Data* **5**, (2018).

118. Depledge, D. P. *et al.* Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nat Commun* **10**, (2019).

119. Tombácz, D., Prazsák, I., Moldován, N., Szucs, A. & Boldogkoi, Z. Lytic transcriptome dataset of varicella zoster virus generated by long-read sequencing. *Front Genet* **9**, (2018).

120. Tombácz, D. *et al.* Dynamic Transcriptome Sequencing of Bovine Alphaherpesvirus Type 1 and Host Cells Carried Out by a Multi-Technique Approach. *Front Genet* **12**, (2021).

121. Tombácz, D. *et al.* Dynamic transcriptome profiling dataset of vaccinia virus obtained from long-read sequencing techniques. *Gigascience* **7**, (2018).

122. Torma, G. *et al.* Combined short and long-read sequencing reveals a complex transcriptomic architecture of African swine fever virus. *Viruses* **13**, (2021).

123. Boldogkői, Z. Transcriptional interference networks coordinate the expression of functionally related genes clustered in the same genomic loci. *Front Genet* **3**, (2012).

124. Boldogkői, Z., Balázs, Z., Moldován, N., Prazsák, I. & Tombácz, D. Novel classes of replication-associated transcripts discovered in viruses. *RNA Biology* vol. 16 Preprint at https://doi.org/10.1080/15476286.2018.1564468 (2019).

125. García-Muse, T. & Aguilera, A. Transcription–replication conflicts: how they occur and how they are resolved. *Nature Reviews Molecular Cell Biology 2016 17:9* **17**, 553–563 (2016).

126. Srivatsan, A., Tehranchi, A., MacAlpine, D. M. & Wang, J. D. Co-Orientation of Replication and Transcription Preserves Genome Integrity. *PLoS Genet* **6**, 1000810 (2010).

127. Brambati, A., Colosio, A., Zardoni, L., Galanti, L. & Liberi, G. Replication and transcription on a collision course: Eukaryotic regulation mechanisms and implications for DNA stability. *Frontiers in Genetics* vol. 6 Preprint at https://doi.org/10.3389/fgene.2015.00166 (2015).

128. Umbach, J. L. *et al.* MicroRNAs expressed by herpes simplex virus 1 during latent infection regulate viral mRNAs. *Nature* **454**, (2008).

129. Yetming, K. D. *et al.* The BHLF1 Locus of Epstein-Barr Virus Contributes to Viral Latency and B-Cell Immortalization. *J Virol* **94**, (2020).

130. Calderwood, M. A., Holthaus, A. M. & Johannsen, E. The Epstein-Barr Virus LF2 Protein Inhibits Viral Replication. *J Virol* **82**, (2008).

131. Uppal, T., Banerjee, S., Sun, Z., Verma, S. C. & Robertson, E. S. KSHV LANA—The Master Regulator of KSHV Latency. *Viruses* **6**, (2014).

132. Sampath, P. & DeLuca, N. A. Binding of ICP4, TATA-Binding Protein, and RNA Polymerase II to Herpes Simplex Virus Type 1 Immediate-Early, Early, and Late Promoters in Virus-Infected Cells. *J Virol* **82**, (2008).

133. Guo, L. *et al.* Herpes Simplex Virus 1 ICP22 Inhibits the Transcription of Viral Gene Promoters by Binding to and Blocking the Recruitment of P-TEFb. *PLoS One* **7**, (2012).

134. Liu, M. *et al.* ICP0 antagonizes ICP4-dependent silencing of the herpes simplex virus ICP0 gene. *PLoS One* **5**, (2010).

135. Takács, I. F. *et al.* The ICP22 protein selectively modifies the transcription of different kinetic classes of pseudorabies virus genes. *BMC Mol Biol* **14**, (2013).

136. Blom, J. *et al.* EDGAR: A software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics* **10**, (2009).

137. Pfaff, F. *et al.* Full genome sequence of bovine alphaherpesvirus 2 (BoHV-2). *Arch Virol* **166**, (2021).

138. Donovan-Banfield, I., Turnell, A. S., Hiscox, J. A., Leppard, K. N. & Matthews, D. A. Deep splicing plasticity of the human adenovirus type 5 transcriptome drives virus evolution. *Commun Biol* **3**, (2020).

139. Geballe, A. P. & Mocarski, E. S. Translational control of cytomegalovirus gene expression is mediated by upstream AUG codons. *J Virol* **62**, (1988).

140. Mayr, C. & Bartel, D. P. Widespread Shortening of 3′UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell* **138**, (2009).

141. Pereira, L. A., Munita, R., González, M. P. & Andrés, M. E. Long 3'UTR of Nurr1 mRNAs is targeted by miRNAs in mesencephalic dopamine neurons. *PLoS One* **12**, (2017).

142. Macdonald, P. M. & Struhl, G. Cis- acting sequences responsible for anterior localization of bicoid mRNA in Drosophila embryos. *Nature* **336**, (1988).

143. Martin, K. C. & Ephrussi, A. mRNA Localization: Gene Expression in the Spatial Dimension. *Cell* vol. 136 Preprint at https://doi.org/10.1016/j.cell.2009.01.044 (2009).

144. Tombácz, D. *et al.* Meta-analytic approach for transcriptome profiling of herpes simplex virus type 1. *Sci Data* **7**, (2020).

145. Papp, B. *et al.* Genome-Wide Identification of Direct RTA Targets Reveals Key Host Factors for Kaposi's Sarcoma-Associated Herpesvirus Lytic Reactivation. *J Virol* **93**, (2019).

146. Tombácz, D., Tóth, J. S., Petrovszki, P. & Boldogkoi, Z. Whole-genome analysis of pseudorabies virus gene expression by real-time quantitative RT-PCR assay. *BMC Genomics* **10**, (2009).

147. Bustin, S. A., Gyselman, V. G., Siddiqi, S. & Dorudi, S. Cytokeratin 20 is not a tissue-specific marker for the detection of malignant epithelial cells in the blood of colorectal cancer patients. *Int J Surg Investig* **2**, (2000).

148. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method. *Methods* **25**, (2001).

149. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, (2013).

150. Thodberg, M., Thieffry, A., Vitting-Seerup, K., Andersson, R. & Sandelin, A. CAGEfightR: Analysis of 5′-end data using R/Bioconductor. *BMC Bioinformatics* **20**, (2019).

151. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, (2018).

152. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P. & Carlson, M. Software for Computing and Annotating Genomic Ranges. *PLoS Comput Biol* **9**, 1003118 (2013).

153. *Package 'tidygenomics' Type Package Title Tidy Verbs for Dealing with Genomic Data Frames*. https://github.com/const-ae/tidygenomics (2022).

154. Wickham, H. *et al.* Welcome to the Tidyverse. *J Open Source Softw* **4**, 1686 (2019).

155. Hahne, F. & Ivanek, R. Visualizing genomic data using Gviz and bioconductor. in *Methods in Molecular Biology* vol. 1418 335–351 (Humana Press Inc., 2016).

Co-author certification

I, myself as a corresponding author of the following publication(s) declare that the authors have no conflict of interest, and Islam Almsarrhad 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.**

09 Nov 2023

Fülöp Ádám
……………………………
Fülöp Ádám
first author

Torma Gábor
……………………………
Torma Gábor
shared first author

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

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 Islam Almsarrhad Ph.D. candidate had significant contribution to the jointly published research(es). The results discussed in her thesis were not used and not intended to be used in any other qualification process for obtaining a PhD degree.

09 Nov 2023

Torma ....

Torma Gábor author (shared first authorship)

....

Dr. habil Tombácz Dóra author (shared first authorship)

....

Dr. Csabai Zsolt author (shared first authorship)

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

Torma Gábor (Torma Gábor molekuláris biológia), Tombácz Dóra*; Csabai Zsolt*; **Almsarrhad Islam A. A.*;** Nagy Gergely Ármin; Kakuk Balázs; Gulyás Gábor; Spires Lauren McKenzie; Gupta Ishaan; Fülöp Ádám; Dörmő Ákos; Prazsák István; Mizik Máté, Dani Virág Éva; Csányi Viktor; Harangozó Ákos; Zádori Zoltán; Toth Zsolt; Boldogkői Zsolt Identification of herpesvirus transcripts from genomic regions around the replication origins SCIENTIFIC REPORTS (2045-2322 2045-2322): 13 1 Paper 16395. (2023) MTMT ID 34171230