Mapping the Temporal Transcriptomic Signature of a Viral Pathogen through CAGE and Nanopore sequencing

Dóra Tombácz, Balázs Kakuk, Gábor Torma, Ádám Fülöp, Ákos Dörmő, Gábor Gulyás, Zsolt Csabai, Zsolt Boldogkői\*

Department of Medical Biology, Albert Szent-Györgyi Medical School, University of Szeged, Szeged, Hungary

**\*** Corresponding author: [boldogkoi.zsolt@med.u-szeged.hu](mailto:boldogkoi.zsolt@med.u-szeged.hu)

*Emails*:

DT: [tombacz.dora@med.u-szeged.hu](mailto:tombacz.dora@med.u-szeged.hu)

BK: [kakuk.balazs@med.u-szeged.hu](mailto:kakuk.balazs@med.u-szeged.hu)

GT: [torma.gabor@med.u-szeged.hu](mailto:torma.gabor@med.u-szeged.hu)

ÁF: [fulop.adam@med.u-szeged.hu](mailto:fulop.adam@med.u-szeged.hu)

AD: [dormo.akos@med.u-szeged.hu](mailto:dormo.akos@med.u-szeged.hu)

GG: [gulyas.gabor@med.u-szeged.hu](mailto:gulyas.gabor@med.u-szeged.hu)

ZC: [csabai.zsolt@med.u-szeged.hu](mailto:csabai.zsolt@med.u-szeged.hu)

ZB: [boldogkoi.zsolt@med.u-szeged.hu](mailto:boldogkoi.zsolt@med.u-szeged.hu)

*Keywords*: Equid herpesvirus 1, EHV-1, transcriptome, long-read sequencing, nanopore sequencing, direct cDNA-sequencing, CAGE-Seq, kinetic classes

**Abstract**

**Introduction**

Equid alphaherpesvirus 1 (EHV-1), a veterinary pathogen belonging to the Varicellovirus genus, is responsible for significant economic losses in the global equine sector. This research involved timescale gene expression profiling and transcriptional reannotation of this herpesvirus.

**Methods**

We employed cap analysis gene expression sequencing (CAGE-Seq) on Illumina platform to determine the transcript start sites alongside long-read direct cDNA sequencing (dcDNA-Seq) on Oxford Nanopore Technology platform to detect full-length viral transcripts. Samples were collected at nine distinct stages of the viral lifecycle, with triplicates taken at each stage. We also applied protein synthesis inhibition to determine the immediate-early gene expression of the virus. Earlier data on native RNA sequencing was also utilized to validate the results.

**Results**

The time-course analysis of viral transcript expression using long-read dcDNA-Seq enabled the characterization of these transcripts based on their kinetic behavior throughout the replication cycle. Furthermore, the study involved a comprehensive reannotation of the EHV-1 transcriptome. CAGE-Seq helped identify the transcription start sites and promoter regions, while dcDNA-Seq provided a more accurate approach to capturing full-length transcripts and isoform diversity. Using an integrated approach, we identified and validated close to 200 novel transcripts, refining the EHV-1 transcriptome annotation and revealing new insights into viral gene regulation. Together, these methods allowed for a more detailed and accurate mapping of the EHV-1 transcriptome, uncovering previously unknown transcripts and refining the existing annotations.

**Conclusions**

The shifting patterns in transcript isoforms and overlaps suggest a sophisticated regulatory network that enables EHV-1 to precisely modulate gene expression throughout its replication cycle. The presence of multiple isoforms per gene indicates that the virus can adapt to different stages of infection by producing a variety of transcripts. This likely enhances its genomic efficiency and allows it to respond more effectively to the host's environment.

**Introduction**

Equid alphaherpesvirus 1 (EHV-1), also referred to as *Varicellovirus equidalpha1* [1] commonly presents with symptoms such as upper respiratory tract disease, spontaneous abortion in pregnant mares, neonatal death, and life-threatening myeloencephalopathy [2]. EHV-1 contains an approximately 150 kilobase pair double-stranded DNA genome. This genome is organized into two segments designated as unique short (US) and unique long (UL) regions, both surrounded by inverted repeats (IRs) [3,4]. The complete viral genome contains 80 open reading frames (ORFs) [5] among which five genes (ORF1, 2, 67, 71, and 75) are absent in other alphaherpesviruses with annotated genomes [6]. Given that four ORFs are situated in the IR region, the EHV-1 genome comprises a total of 76 unique protein-coding genes. We note that the term 'ORF' is used to denote the entire genes in EHV-1, not just the protein-coding parts. Similar to other alphaherpesviruses, EHV-1 can either productively infect cells or enter a latent state in specific peripheral neurons [7].

EHV-1 genes fall into three categories: immediate-early (IE), early (E), and late (L). The regulation of these genes is governed by viral transcription factors, specifically ORF5, 12, 63, 64, and 65 [8,9]. Notably, EHV-1 has a sole IE gene [10], ORF64, which bears homology to *rs1* gene of herpes simplex virus type 1 (HSV-1) coding for the ICP4 protein. Most of the viral E genes encode enzymes required for DNA replication, whereas the L genes specify structural proteins found in the virion, such as capsid and spike proteins. Late genes are further subdivided into leaky late (L1) and true late (L2), depending on their reliance on DNA replication for their expression [10].

Long-read sequencing (LRS), developed by Pacific Biosciences (PacBio) through synthesis-based sequencing and by Oxford Nanopore Technologies (ONT) through nanopore sequencing, has become essential in modern transcriptome analysis. The long-read RNA sequencing (lrRNA-Seq) technique is particularly effective in identifying transcription start sites (TSSs), transcription end sites (TESs), splice sites, alternative splicing, embedded transcripts, multigenic RNA molecules, and transcriptional overlap [11]. While lrRNA-Seq platforms can deliver full-length cDNA or native RNA sequences, they come with reduced throughput and a higher incidence of sequencing errors compared to short-read sequencing (SRS) platforms [12–17]. In transcriptomics, inaccuracies in sequencing are not a significant concern when the genomic sequence of a specific organism is well-established. The lrRNA-Seq technology based on ONT is highly suitable for native RNA sequencing. It guarantees correct read orientation and is free from the artifacts generated by reverse transcription and PCR [11,18], and it also facilitates the detection of RNA modifications [19].

Transcriptomic studies in herpesviruses have applied both PacBio and ONT sequencing methods [15,17,20–26]. Temporal dynamics of herpesvirus gene expression has also been examined using SRS [27]. Our previous transcriptome profiling study, which employed a native RNA sequencing approach, identified the canonical EHV-1 transcripts, including mRNAs, non-coding RNAs (ncRNAs) and long multi-gene transcripts [28]. However, direct RNA sequencing (dRNA-Seq) has limitations in accurately identifying TSSs due to 5' truncation caused by motor protein stalling during sequencing. To address this and refine the EHV-1 transcriptome annotation, we integrated cap analysis of gene expression sequencing (CAGE-Seq) with direct cDNA sequencing (dcDNA-Seq). CAGE-Seq provides high-resolution TSS mapping, while dcDNA-Seq captures full-length transcripts without the 5' truncation issues of dRNA-Seq. This combined approach allowed us to validate transcripts previously identified but excluded due to strict filtering criteria and to discover additional transcript isoforms with greater accuracy. Using dcDNA-Seq, we sequenced 27 samples collected at nine time points spanning 1 to 48 hours post-infection, with three replicates taken at each time point. This extensive temporal sampling captured the full dynamics of viral gene expression across the complete infection cycle. By analyzing these time-resolved data, we clustered genes into *de novo* kinetic classes based on their expression dynamics. Additionally, reanalyzing the data with integrated CAGE-Seq refined existing annotations, validated previously excluded transcripts, and identified previously unrecognized transcript isoforms with greater accuracy.

## **Methods**

## **Cells and viruses**

In this study, we utilized the field isolate *equid alphaherpesvirus 1* strain MdBio (EHV-1-MdBio), which was originally isolated from the organs of an aborted colt fetus in the 1980s at Marócpuszta, Hungary, and has been previously described [28]. The virus was propagated in a confluent rabbit kidney (RK-13) epithelial cell line (ECACC: 00021715). Cells were cultivated in DMEM (Sigma), supplemented with 10% fetal calf serum and 80 μg of gentamycin per ml (Gibco) at 37 °C in the presence of 5% CO2. For the preparation of virus stock solution, cells were infected with 0.1 multiplicity of infection [MOI = plaque-forming units (pfu)/cell]. Viral infection was allowed to progress until complete cytopathic effect was observed. As a next step, three successive cycles of freezing and thawing of infected cells were carried out to release of viruses from the cells. For the sequencing experiments, RK-13 cells were infected with 4 MOI of EHV-1-MdBio in three technical replicates. Infected cells were incubated for 1 h at 4 °C, followed by removal of the virus suspension and washing the cells with phosphate-buffered saline. As a next step, new culture medium was added to the infected cells, which were incubated for 1, 2, 4, 6, 8, 12, 18, 24, or 48 h. After the incubation, the culture medium was removed, and the infected cells were frozen at −80 °C until further use.

**Ethics Statement**

Ethics approval is "Not Applicable," as no animal experiments were performed.

**Cycloheximide treatment of cells**

RK-13 cells were grown in DMEM supplemented with 10% fetal bovine serum until they reached 60–70% confluency. The medium was then replaced with 5 mL of serum-free DMEM containing either 20 or 100 μg/mL cycloheximide (CHX). After a 1-hour incubation, this medium was substituted with 2 mL of a 10 MOI virus solution, also containing the same CHX concentrations, and the cells were incubated for either 6 or 8 hours. Post-incubation, the CHX-treated cells were washed once with PBS, scraped off the dish, and centrifuged at 2000 g for 2 minutes. Following the removal of the supernatant, the cell samples were immediately placed on dry ice for future analysis.

## **RNA extraction**

The RNA extraction was conducted using the NucleoSpin RNA kit from Macherey-Nagel. The process began by lysing the cells in a buffer containing chaotropic ions to deactivate RNases. This step facilitated the binding of DNA and RNA molecules to the silica membrane. To eliminate any residual genomic DNA (gDNA), all samples were treated with DNase I. The total RNA was then eluted in nuclease-free water. Further purification to remove any remaining gDNA was achieved using the TURBO DNA-free™ Kit from Invitrogen. The RNA concentration in the samples was determined using the Qubit 4.0 fluorometer and the Qubit Broad Range RNA Assay Kit, also from Invitrogen (**Supplementary Table S4**). Quality control was performed with the Agilent TapeStation 4150, and only samples with RIN scores equal to or greater than 9.2 were used for cDNA synthesis and subsequent experiments.

**Purification of polyadenylated RNA**

The poly(A)+ RNA fraction was extracted from the total RNA using the Oligotex mRNA Mini Kit by Qiagen. Initially, the volume of each sample was adjusted to 250 µL with RNase-free water. Then, 15 µL of Oligotex suspension and 250 µL of OBB buffer, both from the Qiagen kit, were added to the samples. The mixture was heated to 70°C for 3 minutes and subsequently cooled to 25°C for 10 minutes. After centrifuging at 14,000×g for 2 minutes, the supernatants were discarded. The samples were then washed with 400 µL of OW2 wash buffer from the kit and transferred to spin columns provided in the kit, followed by centrifugation at 14,000×g for 1 minute. This washing step was repeated. Finally, the polyadenylated RNA was eluted from the membrane using 50 µl of pre-heated elution buffer from the Qiagen kit, collected in 60 µl elution buffer, with a second elution step performed to maximize the yield (**Supplementary Table S4**).

**RNA Quantification**

For measuring total RNA, we used the Qubit RNA BR Assay Kit from Invitrogen (Carlsbad, CA, United States). To quantify the poly(A)+ fraction, the Qubit RNA HS Assay Kit, also from Invitrogen (Carlsbad, CA, United States), was employed. The final concentrations of these RNA samples were determined using the Qubit® 4 fluorometer.

**cDNA Quantification and Quality Assessment**

The concentrations of cDNA samples and sequencing-ready libraries were ascertained using the Qubit dsDNA HS Assay Kit from Invitrogen (Carlsbad, CA, United States). The quality of RNA, crucial for sequencing, was evaluated using the Agilent 2100 Bioanalyzer for PacBio sequencing, and the Agilent 4150 TapeStation System for MinION sequencing. Samples with RIN scores of 9.6 or higher were selected for cDNA synthesis.

**Cap Analysis of Gene Expression sequencing**

The Cap analysis of gene expression sequencing (CAGE-Seq) protocol has been previously described [29]. Briefly, using the CAGE™ Preparation Kit (DNAFORM, Japan), we performed CAGE-Seq on viral genomic regions employing three biological replicates. Initially, 5 µg of total RNA and the kit's RT primer were mixed and heated at 65 °C. SuperScript III Reverse Transcriptase (Invitrogen) and a trehalose/sorbitol mixture (from the kit) were used for first-strand cDNA synthesis, followed by oxidation of the Cap's diol groups and biotinylation. RNase I (from the kit) digested single-strand RNA. Biotinylated samples were then bound to Streptavidin beads, washed, and cDNAs were released and purified. RNase mixture treated the samples to digest any residual RNA. Streptavidin beads, coated with tRNA, were prepared for linker ligation. After reducing the sample volumes using the miVac DUO Centrifugal Concentrator (Genevac), 5′ and 3′ linkers were ligated, followed by Shrimp Alkaline Phosphatase (SAP) and USER enzyme treatments. The second cDNA strand was synthesized, treated with Exonuclease I, and samples were dried and resuspended in nuclease-free water. Single-stranded cDNA concentrations were measured using Qubit 2.0 and the Qubit ssDNA HS Assay Kit. Purification steps employed RNAClean XP and AmpureXP Beads at various stages. Pooled libraries with different barcodes were sequenced on a MiSeq instrument using v3 (150 cycles) and v2 (300 cycles) chemistries (Illumina). The final concentration and quality of the libraries were assessed using Qubit 4.0 with a 1X dsDNA High Sensitivity (HS) Assay and TapeStation, respectively.

**Library Construction and cDNA Sequencing Using ONT MinION**

Libraries for direct cDNA sequencing on the ONT MinION device were constructed using poly(A)+-enriched samples. We followed the protocol of the ONT Direct cDNA Sequencing Kit (SQK-DCS109), as outlined in the kit's manual. Initially, RNA samples were mixed with VN primer (VNP; from the ONT kit) and 10 mM dNTPs, and heated at 65°C for 5 minutes. This was followed by the addition of 5x RT Buffer, RNaseOUT (from Thermo Fisher Scientific), and Strand-Switching Primer (SSP; from the ONT Kit), and a subsequent 2-minute heating at 42°C. The first cDNA strand synthesis involved the Maxima H Minus Reverse Transcriptase enzyme (from Thermo Fisher Scientific), with the reaction occurring at 42°C for 90 minutes, and enzyme inactivation at 85°C for 5 minutes. RNA strands from RNA-cDNA hybrids were removed using the RNase Cocktail Enzyme Mix (from Thermo Fisher Scientific) at 37°C for 10 minutes. The second cDNA strand was synthesized using LongAmp Taq Master Mix [from New England Biolabs (NEB)] and PR2 Primer (PR2P), with PCR reaction specifics is described in [28]. DNA fragments were then processed for end-repair and dA-tailing using the NEBNext End repair/dA-tailing Module (NEB) at 20°C for 5 minutes, followed by 65°C for 5 minutes. This step was followed by adapter ligation using the NEB Blunt/TA Ligase Master Mix (NEB) at room temperature for 10 minutes. The ONT dcDNA libraries were barcoded as outlined in [28], and as per the ONT Native Barcoding (12) Kit instructions. The prepared cDNA libraries (200 fmol/flow cell) were purified and loaded onto ONT R9.4.1 SpotON Flow Cells, using a total of five flow cells for sequencing. To prevent "barcode hopping," samples from earlier and later time points were sequenced separately. After each enzymatic step, AMPure XP Beads were used for purification. The samples were then eluted in UltraPure™ nuclease-free water (from Invitrogen), and their concentration was measured using the Qubit 4.0 fluorometer and Qubit dsDNA HS Assay kit.

**Pre-Processing**

*dcDNA sequencing*

The raw current signals obtained from ONT-minION sequencing were basecalled to nucleotides with the Dorado-0.7.2 basecaller (<https://github.com/nanoporetech/dorado/>) using a quality threshold of 7. The resulting reads were aligned to the reference genome (accession number: NC\_001491.2) using the minimap2 [30] program. During the alignment with minimap2, the following settings were applied: -ax splice -Y -C5 -cs. To identify TSS, TES, and intron positions, we used the LoRTIA toolkit (<https://github.com/zsolt-balazs/LoRTIA>). For evaluating direct cDNA sequencing (dcDNA-Seq), the following settings were applied in the LoRTIA package: −5 TGCCATTAGGCCGGG –five\_score 16 –check\_in\_soft 15 –3 AAAAAAAAAAAAAAA –three\_score 16 –s Poisson –f true. A transcript was accepted when its 5' and 3' adapters were accurate, and in the case of 3' ends, false priming and template switching during intron identification were excluded. For introns, we accepted those annotated in dRNA sequencing for direct cDNA samples. For further analysis, we used an in-house developed R pipeline. Briefly, the “*stranded\_only.bam*” files from the LoRTIA output were imported into the R environment using Rsmatools [31] . A database was then built from it, containing the count of unique mapping positions and the information from the bam-files regarding LoRTIA’s adapter searching using data.table [32] R-package [33]. This was then used to count the 3'- and 5'-ends per nucleotide and the coverages.

## **Data Analysis**

*CAGE-Seq Analysis*

The STAR aligner (version 2.7.3 a) [34] was used to map the reads to the EHV-1 reference genome (NC\_001491.2), utilizing --genomeSAindexNbases 8 and default parameters. “Bam” files obtained from CAGE-seq were converted to BigWig format to detect 5′ end coverage. The CAGEfightR [35] package was used to determine TSS positions. The TSS clusters within a 10 nucleotides window were termed identical. Clusters with a “minimum pooled value” (--pooledcutoff=1) of 1 and below were excluded from the further analysis. We assigned TSS cluster from the CAGEFighteR output confidence by empirically combining support thresholds with quartile-based score bins: clusters with support ≤ 2 and scores in the bottom quartile (Q1) were classified as low confidence, while clusters with support of 3–5 and scores in Q2 or Q3 received moderate confidence. Clusters surpassing a support of 5 or having top-quartile (Q4) scores were categorized as high confidence, and those that partially met only one of these high-confidence criteria were also placed in the moderate group. This classification is heuristic rather than a formal statistical test; it relies on observed distribution patterns to label clusters as low, moderate, or high confidence.

*Reference Transcript Counting*

The GFF-compare tool [36] was used to count the reference transcripts (from [28]) in the samples. However, since this tool tends to assign shorter transcript isoforms – those contained within another transcript - to the longer one, we ran this tool iteratively for each reference transcript separately. The results were merged, and the best hit for each query alignment was selected based on the smallest distance between transcript ends. For counting reference isoforms, only the hits with “equal to reference” (“=”) were retained, with a distance cut-off of 10 nt-s for both ends. The R-packages rtracklayer [37] was used to export and import .gff3 files.

*TSS validation*

The clusters were then merged with the dcDNA dataset, containing transcript identity based on the alignment’s 5′ ends. The TSS clusters were then refined using the dcDNA read counts (with correct 5′ ends), as in some cases the clusters were as wide as 150 nts. For each TSS cluster a custom peak-analysis method was applied. In brief, the algorithm was provided with the per-position read counts (summed dcDNA 5′ ends) as weights, then instructed to group nearby positions into clusters if (**i**) their genomic distance was at most 5 bp and (**ii**) the cluster had not already reached its maximum size of 25 positions. Positions with zero coverage were excluded. This process ensured that any local “hotspots” of TSS usage were segregated into compact clusters for downstream analyses.

*Transcript Assembly, Classification and Validation*

Transcripts were reconstructed by pairing validated TSS peaks with corresponding TESs, applying stringent criteria: (i) hat is the transcripts were annotated to the novel TSS only if their 5′ ends were within ± 10 nts, and if their 3′ ends overlapped with a known TES (also within ± 10 nts). This method enabled us to combine the CAGE and dcDNA datasets to annotate the TSSs.

We constructed transcripts by pairing these refined TSS peaks with TESs: the transcripts were annotated to the novel TSSs if (i) its 5′ end to lied within ±10 nt of the refined cluster, and (ii) the corresponding 3′ end had to overlap a known TES (also ±10 nt). This enabled us to integrate the CAGE and dcDNA datasets to annotate TSSs. Moreover, each transcript required stringent criteria to be annotated:

* At least three dcDNA-Seq reads sharing the same TSS and TES coordinates,
* Correct 5′ adapter sequences,
* Alignment with CAGE-Seq–derived TSS clusters, previously validated TESs.

Newly assembled transcripts were integrated with our prior annotation [28], allowing us to reintroduce previously excluded transcripts that now met refined criteria and add novel transcripts not previously detected. Validated transcripts were classified based on their structural and functional features, including coding capacity (e.g., putative mRNAs, non-coding RNAs), orientation (monocistronic or multicistronic), and variations in untranslated regions (e.g., truncated or extended isoforms).

We performed the evaluation of our previous dRNA dataset with the NAGATA software [38] also in order to validate our novel TSSs and the introns identified by LoRTIA as well. We applied the following settings for the NAGATA: *-m 1 -tg 2*, otherwise defaults. This configuration was chosen in order to annotate the rare TSS, TES, and introns in the dRNA data that were confirmed in the dcDNA data that had a significantly higher read count.

*Filtering Truncated ORF-carrying Transcripts (Putative mRNAs)*

In order to further filter TSSs of truncated ORF carry transcripts (*Putative mRNAs*), the identified TSSs had to be present in the NAGATA output within a 25 nt wobble (to account for the potentially missing 5′ ends), and additionally they had to be validated by CAGE their signal strength (5′ end read count) compared to their parent gene’s canonical TSS. That is, for a 5′-truncated isoform to be included, its TSS had to reach at least 5% expression relative to its canonical transcript.

*De novo Clustering of Genes by Dynamic Expression Patterns*

To identify groups of TSSs, TESs, and genes with similar temporal expression patterns, we performed *de novo* clustering on normalized gene expression data. Hierarchical clustering was conducted using the pvclust R package [39], which provides hierarchical clustering and bootstrap resampling for cluster assessment. The complete linkage method and an uncentered correlation distance measure were applied. Cluster stability and significance were assessed using 1000 bootstrap iterations, examining approximately unbiased (AU) p-values provided by pvclust. After generating a dendrogram, we evaluated a range of potential cluster solutions, from 4 to 15 clusters, and assessed their quality based on AU values and within-cluster sum of squares (WSS). Partitioning the data into 12 clusters provided a meaningful balance between resolution and interpretability. The hierarchical tree was cut into 12 clusters, each representing a distinct temporal expression pattern over the time course. Closely clustered genes formed larger clusters, while genes with unique kinetic patterns grouped independently.

## **Results**

## **Dynamic EHV-1 transcriptome: general considerations**

In this study, we conducted a comprehensive time-course transcriptomic analysis of EHV-1, primarily using direct cDNA sequencing (dcDNA-Seq) on the ONT MinION platform and incorporating cap analysis of gene expression sequencing (CAGE-Seq) on the Illumina MiSeq platform. Our previous work [28] relied on native RNA sequencing (dRNA-Seq) and stringent filtering criteria, which ensured high confidence but led to the exclusion of some detected transcripts that did not meet the strict validation requirements. One key limitation of dRNA-Seq was its reduced accuracy in identifying transcription start sites (TSSs), partly due to incomplete 5' ends in the sequencing reads.

To overcome these limitations, we integrated CAGE-Seq data with dcDNA-Seq reads. CAGE-Seq provides high-resolution TSS mapping, enabling more accurate identification and validation of 5' ends. By aligning dcDNA-Seq reads that carried correct 5' adapters and intact poly(A) tails with CAGE-derived TSS clusters, we confirmed the authenticity of previously excluded transcripts and added 141 novel ones. This integrative approach allowed us to refine our transcriptome annotation, detect additional transcript isoforms, and extend the scope of our EHV-1 transcript catalog.

For the kinetic analyses, we focused on canonical transcripts—defined as the most abundant isoforms for each viral gene—and their canonical TSSs and TESs. Using these canonical references, we examined the temporal dynamics of gene expression across multiple time points, clustered genes into *de novo* kinetic classes based on their expression curves, and compared these classes with the traditional IE/E/L framework.

Finally, beyond annotating full-length transcripts and refining TSS/TES definitions, we explored isoform switching and transcriptional overlaps, investigating how the virus orchestrates complex regulatory patterns over the infection cycle. In total, this integrated strategy identified and validated 277 additional transcripts and, by examining isoform switching, splicing dynamics, and transcriptional overlaps, significantly enhanced our understanding of the EHV-1 transcriptome and its intricate temporal regulation.

**The ORF64 is the only EHV-1 immediate-early** **gene**

It has been previously established that ORF64 is the only IE gene of EHV-1 [40]. However, our earlier study [28] and the current research have identified novel EHV-1 transcripts with previously unknown kinetic properties. To investigate whether the expression of these transcripts requires newly synthesized viral proteins, we treated RK-13 cells (ECACC: 00021715) with cycloheximide (CHX), a protein synthesis inhibitor, prior to infection with EHV-1. We administered CHX at concentrations of 20 and 100 μg/ml and collected samples at 6 and 8 hours post-infection (hpi). Subsequently, we performed long-read dcDNA sequencing. Our findings reinforced that ORF64 is the sole IE gene in EHV-1 (**Supplementary Table S1**). In pseudorabies virus (PRV), the closest relative of EHV-1 with an annotated transcriptome, the homologous gene (*ie180*) is also the only IE gene [39]. In contrast, other annotated alphaherpesviruses, including those in the Simplexvirus and Varicellovirus genera, have multiple IE genes among their members.

**Reannotation of EHV-1 transcriptome**

In this part of our study, we used CAGE-Seq on an Illumina MiSeq platform to detect the TSSs of EHV-1 transcripts with high resolution. The CAGE-Seq data were integrated with dcDNA-Seq reads to validate, refine, and expand our previous transcriptome annotation [28].

Firstly, we validated transcripts previously annotated in our laboratory using dRNA-Seq [28] against the newly acquired CAGE-Seq data (**Supplementary Table S2**). Among the examined transcripts, 251 received the highest level of support (\*\*\*), indicating robust validation, while 47 showed medium support (\*\*) and 51 had the lowest level (\*).

Next, we analyzed dcDNA-Seq reads that could not be assigned to previously annotated transcripts due to TSS differences. By aligning the 5' ends of these reads - validated by correct 5' adapter sequences (LoRTIA pipeline) - to TSS clusters identified by CAGE-Seq (CAGEfightR), we refined TSS annotations and identified distinct TSS peaks within broad clusters (>200 bp). Final transcripts were constructed by combining these refined TSSs with TESs from our previous annotations. Transcripts were included if they had at least five dcDNA-Seq reads with 5' ends aligning to validated TSSs from CAGE-Seq. This integrated approach enabled us to recover and validate 136 transcripts previously excluded due to stringent filtering, as well as identify novel transcripts. In total, we validated and included 169 additional transcripts, encompassing novel TSS variants and isoforms that enhance the comprehensiveness of the EHV-1 transcriptome. These transcripts were categorized as follows: 11 putative mRNAs (3'-coterminal but 5'-truncated variants), 26 non-coding RNAs (ncRNAs), 42 long monocistronic transcripts, 38 short monocistronic transcripts, 8 long multicistronic transcripts, and 10 short multicistronic transcripts. Long and short variants share the same ORF as the canonical transcript but differ in their 5' UTR lengths. *Putative mRNAs* may encode N-terminally truncated proteins; however, because nanopore-based methods often have less robust 5′ support and cytoplasmic recapping [40] can artificially generate truncated 5′ ends, we employed an additional filtering criterion. Specifically, for a 5′-truncated isoform to be included, its TSS had to be corroborated by dRNA-Seq data and achieve at least 5% expression (as determined by CAGE-Seq) relative to its canonical transcript.

Details on the previously annotated and novel transcripts, along with their levels of support based on our validation criteria in the CAGE-Seq and dRNA-Seq data and their counts across the dcDNA samples, are provided in **Supplementary Table S2**.

**Kinetic characterization of TSSs, TESs and canonical transcripts**

We next explored the dynamics of TSSs, TESs, and canonical transcripts over the infection cycle and compared these findings with the traditional IE, E, and L kinetic classes. The time points associated with IE, E, and L genes are as follows: IE genes are highly expressed at 1–2 hpi, E genes predominate between 2–6 hpi, and L genes are expressed after 6 hpi, with peak expression observed at 8–12 hpi and beyond. These time frames are consistent with established alphaherpesvirus kinetics.

***TSS expression kinetics***

Our temporal expression analysis provided important insights into the regulation of EHV-1 genes over the course of infection (**Figure 1** and **Supplementary Figure S1**). For instance, early genes such as ORF20, ORF21, ORF30, ORF31, and ORF63 exhibited peak TSS activities as early as 2 hpi, followed by a gradual decline. Conversely, late genes, including ORF11, ORF14, ORF22, and ORF73, began to show substantial TSS activity starting from 4 hpi, reaching their maxima around 8 to 12 hpi. This pattern is consistent with the known function of these genes in either DNA synthesis (E genes) or in producing structural components necessary for virion assembly and egress (L genes). Detailed temporal profiling further elucidated this dynamic landscape by pinpointing specific TSS peak times for individual transcripts. For example, ORF32 showed an early peak at 2 hpi, ORF51 at 6 hpi, and ORF19 at 8 hpi, each followed by a characteristic decline. Additional examples include ORF18, which peaked at 8 hpi, ORF28, which showed a maximum at 6 hpi and then again at 8 hpi, and ORF50, which exhibited peak activity at 4 hpi.

To further clarify these patterns, we performed hierarchical clustering of TSS abundances throughout the infection. This approach grouped genes based on their expression trajectories rather than relying solely on predefined IE/E/L classifications. The clusters largely reinforced the TSS-based timing patterns: large clusters often predominantly contained L genes or E genes, reflecting the temporal shifts observed in the individual TSS analyses. For example, clusters dominated by late genes confirmed the existence of a robust late-expression phase, while clusters enriched in early genes validated an early wave of transcription closely following the IE stage. The kinetics of the TSSs grouped according to the *de novo* clustering are shown in **Supplementary Figure S2** and grouped according to traditional kinetic clusters in **Supplementary Figure S3.**

However, the clustering revealed that some genes did not fit neatly into the expected phases. As observed at the single-gene level, several traditionally late genes exhibited earlier-than-anticipated TSS peaks, while some early genes maintained or regained expression at later times. For example, ORF38, typically classified as late, displayed a TSS peak at 6 hpi - more characteristic of early kinetics - while ORF45, also considered late, peaked at 12 hpi and again at 48 hpi. Similarly, ORF54, an early gene, showed a peak at 24 hpi, well beyond the window typically associated with early functions. These "misaligned" genes were placed into mixed clusters containing both early and late markers, suggesting they may belong to transitional or intermediate regulatory states rather than strictly defined classes.

The clustering also highlighted small groups or outliers - single- or double-gene clusters - with unique timing patterns that do not align with the canonical IE/E/L framework. These outliers suggest that some genes may follow specialized regulatory circuits, contributing to the intricate temporal orchestration of viral gene expression.

Our analysis of TSS expression kinetics identified 12 distinct clusters, reflecting the temporal and functional profiles of viral gene expression. Cluster Cluster\_12 consists of ORF64, encoding the transcriptional regulator ICP4, underscoring its pivotal role in initiating viral transcription and, interestingly, ORF75 (US8A), traditionally considered late (likely due to detection in one replicate at 1 hpi).

Early-dominant TSS clusters, such as Cluster\_1 and Cluster\_3, contain genes like ORF20, ORF21, ORF30, ORF53, and ORF63, involved in nucleotide metabolism and genome replication, peaking early post-infection. Intermediate clusters featuring ORF19, ORF37, ORF55, and ORF76 bridge early and late phases, indicating overlapping or transitional expression profiles. Late-dominant clusters – most notably Cluster\_6 and Cluster\_7 – include genes (e.g., ORF22, ORF24, and ORF42 in Cluster\_6) and (e.g., ORF12, ORF13, and ORF48 in Cluster\_7) that encode proteins involved in virion assembly and packaging, peaking at 8–12 hpi. Overall, these patterns highlight a continuous, overlapping temporal landscape rather than strictly partitioned IE/E/L classes.

***TES expression kinetics***

Our examination of transcription end site (TES) dynamics (**Figure 2** and **Supplementary Figure S4**) reveals a complex and overlapping regulatory landscape, much like what we observed at the transcription start sites (TSSs). Many genes conform to their expected kinetic classes: early (E) genes such as ORF20, ORF21, ORF30, ORF31, and ORF63 exhibit TES peaks within the first few hours post-infection, while late (L) genes including ORF11, ORF14, ORF22, and ORF73 reach their maxima between 8 to 12 hours post-infection (hpi). This general pattern aligns with the known roles of early genes in DNA replication and late genes in virion assembly.

However, both the initial analysis of individual TES kinetics and the subsequent clustering based on TES usage reveal exceptions and overlapping dynamics that challenge the straightforward IE/E/L model. For example, ORF32 and ORF51, traditionally classified as late, exhibited earlier-than-expected TES peaks, while ORF19, categorized as early, showed a delayed TES maximum more characteristic of late genes. These anomalies suggest that the timing of transcript termination does not always correspond to the canonical temporal classes.

The clustering of TES expression profiles (**Supplementary Figures S5 and S6**) revealed distinct groups of genes with shared termination dynamics, providing insights into the complexity of viral transcription regulation. TESs often are groups of multiple co-terminal gene sets, reflecting the fact that many herpesvirus transcripts share common termination sites. As a result, individual TES clusters can encompass genes with various functions and temporal profiles, blurring the strict IE/E/L boundaries.

For example, clusters containing predominantly late-expressed structural and assembly genes (e.g., the set including ORF12, ORF13, ORF14 in Cluster\_6 or ORF17, ORF16, ORF15, ORF18 in Cluster\_8) underscore the coordinated late-phase production of capsid, tegument, and packaging proteins. Meanwhile, clusters enriched in early genes (e.g., ORF7, ORF30, and ORF63 in Cluster\_5) peak during the initial stages of infection, consistent with their roles in replication and regulation.

Mixed-phase clusters are particularly revealing. Sets like (ORF32, ORF33, ORF34) in Cluster\_7 or (ORF48, ORF49, ORF50, ORF51) in Cluster\_3 combine genes from both early and late classes, indicating that replication factors and structural proteins can share the same termination site. This co-termination creates overlapping kinetic patterns and emphasizes a more continuous temporal landscape rather than strictly segmented IE/E/L phases.

Smaller clusters also yield meaningful insights. Cluster\_11, which includes ORF64 (the sole IE gene ICP4), highlights its unique regulatory role at the TES level. Other clusters, such as those involving envelope glycoproteins and tegument proteins in Cluster\_10 or multi-gene E/L sets in Cluster\_1 and Cluster\_12, illustrate that transcripts with differing temporal classes can terminate together, further increasing transcriptional complexity.

This analysis underscores that termination events are not segregated by kinetic class. Instead, multiple genes—spanning different functional categories and expression timing—often converge at the same TES. This arrangement suggests that EHV-1 may leverage polygenic termination strategies to maintain flexibility and efficiency, ensuring that a variety of transcripts required at different infection stages can be readily produced from overlapping genomic segments.

***Linking TSS and TES sites***

To investigate the similarities and discrepancies between TSS and TES kinetics, we performed a detailed analysis of the linkage between TSS and TES sites using our long-read sequencing data. By matching TSSs to TESs on individual transcript molecules, we were able to assess whether the observed differences in kinetics were due to the presence of alternative transcript isoforms, multicistronic transcripts, or other factors. **Figure 3** shows each gene's abundance during the course of the infection, as assessed by the viral-read-normalized canonical transcript counts, according to their kinetic classes.

For genes where TSS and TES dynamics differed, our analysis revealed that the discrepancies could often be attributed to the complex transcriptional landscape of EHV-1. The virus produces a variety of transcript isoforms, including alternative TSSs and sometimes TESs, as well as multicistronic and overlapping transcripts **(Figure 4 and Supplementary Figure S7**). For example, both ORF38 and ORF50, expected to show late kinetics, exhibited early peaks, although their TSS peaked at 4 and 8 hpi, while their TES at 6 hpi. This misalignment is in the case of ORF38 the result of the elevated expression of other, mainly complex transcripts (overlapping ORF35-37 from the other strand) that terminate at the same TES, such as ORF37-ORF38-CX-Long-2 (**Figure 4A)** and possible transcriptional noise from other non-validated TSSs in this region. In the case of ORF50 (**Supplementary Figure S7B**), this could be attributed to a more complex differential transcript expression pattern consisting of mainly ORF50-ORF51-Canonic and ORF50-ORF51-PC-Long-2. Conversely, ORF67 (Late), which showed an early TSS peak, had TES dynamics more consistent with its late gene classification (**Supplementary Figure S7C**). And while its presence very early (hpi 1) could be attributed to transcriptional noise troublingthis very early time point, at 2 hpi it could not, but rather to the highly efficient early activation of its promoter. The discrepancy between its TSS and TES kinetics can be attributed to the large number (but individually low count) of potential mRNAs identified in this region, carrying truncated ORFs.

Our detailed mapping confirms that the discrepancies between TSS and TES kinetics are primarily due to the production of multiple transcript isoforms and the complex arrangement of transcription units in the EHV-1 genome. This highlights the importance of considering full-length transcript structures when interpreting gene expression dynamics.

Such discrepancies underscore the complexity of viral gene regulation at transcript endpoints and suggest that different levels of control - initiation and termination - may be modulated independently or influenced by overlapping transcriptional programs.

***Gene-level clustering of canonical transcripts***

Clustering canonical full-length transcripts (those matching both the defined TSSs and TESs) provides a cleaner, more canonical view of EHV-1’s transcriptional program. **Figure 5** shows the genes sorted into de novo kinetic clusters.

Early and Mixed-Early Clusters: Cluster 1, a mix of early (E) genes (e.g., ORF20, ORF21, ORF31, ORF61), late (L) genes (e.g., ORF9, ORF38, ORF50), and genes with unknown timing, peaks around 2–4 hpi, indicating "leaky-late" activity against an early backdrop. Cluster\_3, also early-biased (e.g., ORF5, ORF7, ORF30, ORF53, ORF63), includes a few L genes (e.g., ORF10, ORF17), reinforcing the idea that certain late transcripts can be detected at low levels early on, blending replication and assembly factors in a transitional manner at around 2–4 hpi.

Robust Late-Dominant Clusters: Cluster 2 predominantly consists of L genes (e.g., ORF11, ORF14, ORF18, ORF26, ORF28, ORF29, ORF3, ORF39, ORF40, ORF68, ORF73, ORF76) with a few unknown-timing genes (e.g., ORF2, ORF75), aligning with a robust late-phase expression wave emerging after 6–8 hpi. Cluster\_5 is similarly composed of late structural and packaging components (e.g., ORF22, ORF23, ORF25, ORF33, ORF35.5, ORF36, ORF42, ORF43, ORF44, ORF46, ORF58, ORF62), steadily producing virion-related proteins during mid-to-late infection.

Cluster\_6 and Cluster\_7 are also late-dominated clusters but include notable exceptions. Cluster\_6, for instance, contains mainly late genes (ORF12, ORF13, ORF16, ORF35, ORF41, ORF48, ORF52, ORF54 (which is E), ORF57, ORF60, ORF71 (unknown), ORF72) involved in tegument formation and packaging. Its composition underscores the complexity within the late phase, where an early (ORF54) and an unknown gene (ORF71) appear alongside structural and assembly factors.

Smaller and Specialized Clusters: Cluster\_4 and Cluster\_12 each contain fewer genes and show mixed kinetics. Cluster\_4 includes L (e.g., ORF6, ORF67) and E (ORF65) genes, indicating subtle overlaps even in small sets. Cluster\_8, Cluster\_9, and Cluster\_10 further refine late gene subsets or highlight unique outliers.

Notably, Cluster\_10 hosts the sole immediate-early gene, ORF64 (ICP4). Although no canonical full-length transcripts were detected at 1 hpi—likely due to technical challenges capturing this long RNA intact so early—the clear isolation of ORF64 in its own cluster emphasizes its distinct temporal regulation (**Supplementary Figure S7D**).

Collectively, these clusters confirm that while the IE/E/L scheme provides a broad framework, actual gene expression patterns form a continuous and overlapping temporal gradient. **Supplementary Figure S8** shows the *de novo* cluster memberships of each gene, along with their kinetic classes. By integrating these analyses, we observe a consistent narrative: while many EHV-1 genes follow the classical IE, E, and L progression, a subset displays more complex or hybrid kinetics. Gene-level TSS analysis highlights individual timing anomalies, while clustering places these anomalies in a broader context, revealing that they form part of overlapping transcriptional waves rather than discrete phases. These findings suggest that EHV-1 gene regulation is multifaceted, with certain genes bridging temporal classes and potentially serving specialized regulatory or structural roles at unconventional times during infection.

**Dynamics of transcriptional overlaps, splicing and transcript isoforms**

***Dynamics of spliced transcript expression***

The splice sites of EHV-1 transcripts were previously identified in our laboratory using native RNA sequencing. We detected splice sites in the following genes: ORF8, ORF9, ORF38, ORF47-44, ORF53, ORF54, ORF58, ORF65, and within the NOIR family of non-coding transcripts. **Figure 4** and **Supplementary Figure S9** illustrate these genes. For several of them, we observed notable shifts in the ratio of spliced to non-spliced transcripts over the course of the infection.

In the gene ORF9, the combined ratio for spliced transcripts TR134 and TR172, sharing an intron and both carrying a 5`-truncated ORF, remained at 0% (mean = 0.0) from 1–8 hpi (**Figure 4C**). Their ratio began to rise at 12 hpi (mean = 1.03%, SD = 0.84%), increased significantly at 18 hpi (mean = 11.49%, SD = 5.65%), continued to increase at 24 hpi (mean = 12.85%, SD = 3.60%), and peaked at 48 hpi (mean = 17.10%, SD = 3.64%). Other non-spliced isoforms also elevated compared to the canonic ORF9 transcript, which dropped from 100% early at 2–4 hpi with 100% expression (mean = 100.0%), to only 41.69% (SD = 10.30%) by 48 hpi.

In the case of ORF38 (**Figure 4A**), we saw a very similar pattern, albeit the canonical transcript decreased in proportion even more, to 26.36% at 24 hpi and 27.70% at 48 hpi, reflecting a sharper decline compared to ORF9. This decrease in the canonical transcript was accompanied by a marked increase in the ratios of spliced transcripts TR886, TR888, TR889, and TR891, whose combined ratios rose significantly at 24 hpi and remained elevated at 48 hpi. These spliced transcripts share a common second exon (carrying the ORF) and an identical intron, but differ in their first exons, which define distinct 5'-UTRs. This suggests that the splicing process itself, rather than TSS variability, is tightly regulated and drives transcript diversity while contributing to the sharp decline of the canonical transcript.

The spliced transcripts TR148, TR150, TR152, TR154, and TR3 of ORF65 showed a common expression pattern. Their combined ratio peaked early at 2 hpi with 66.59%, remained high at 4 hpi (66.37%) and 8 hpi (59.76%), declined to 40.63% by 6 hpi, and decreased further through 12 hpi (59.05%), 18 hpi (15.97%), 24 hpi (17.22%), and 48 hpi (20.01%), reflecting their predominant expression during the early stages of infection (**Supplementary Figure S9A**).

These spliced transcripts share the same second exon, which carries the ORF, but differ in either their intron donor sites or the 5-prime ends of their first exons, indicating that the splicing process itself is controlled, contributing to transcript diversity and functional regulation during infection.

The canonical NOIR transcripts (NOIR-Canonic and NOIR-ORF65-PC-Canonic) showed higher expression at all time points, peaking at 12 hpi (46.08% vs. 3.79% for isoforms), while the splice isoforms contribute modestly throughout but increase relatively at later stages, reaching 14.98% at 24 hpi and 18.46% at 48 hpi compared to 35.02% and 31.15% for the canonical transcripts (**Figure 4D**). Similar to other genes, the spliced transcripts of NOIR share a common second exon and an identical intron acceptor site but differ in their intron donor sites and/or TSS, reflecting a controlled splicing process that regulates transcript diversity.

In ORF58, the canonical ORF58 transcript peaked at 8 hpi (75.68%) and declined steadily thereafter, ORF58.5 exhibited an early peak at 2 hpi (83.33%) before decreasing, and the spliced transcript TR1072 showed negligible expression early but increased at 24 hpi (12.61%) and 48 hpi (8.35%), highlighting distinct temporal expression patterns in this gene ().

***Dynamics of transcriptional overlaps and isoform switching in selected genes***

To gain a deeper understanding of the complexity of the transcriptional patterns in EHV-1, we analyzed the kinetics of transcriptional overlaps and the relative abundance of transcript isoforms of genes. By examining changes in isoform dominance over time, we uncovered dynamic patterns that highlight how the virus fine-tunes gene expression at various stages of infection. In **Figure 6**, we illustrate these transcripts and their shifting proportions, revealing that even canonical transcripts can be temporarily superseded by truncated or alternatively terminated isoforms, reflecting a finely controlled temporal program of viral gene expression. **Figure 7** and **Supplementary Figure S10** highlights the significant amount of convergent and divergent transcriptional overlaps in EHV-1, increasing as the infection proceeds.

The expression dynamics of ORF19 revealed a transition from early dominance of the canonical transcript, which peaked at 1 hpi (100%) and declined to 14.59% by 48 hpi, to increasing contributions from the combined complex isoforms (CX), overlapping completely the CDS of ORF18 on the other strand, which rose from negligible levels at 1–4 hpi to 41.44% at 12 hpi and peaked at 72.21% at 48 hpi (**Figure 6A**).

The canonical ORF40 transcript dominated expression at early stages, peaking at 6 hpi (97.78%) before declining sharply. Alternatively terminated (AT) isoforms contributed minimally during early stages but increased in abundance at later stages, with AT2 reaching a peak of 37.42% at 24 hpi, while other AT isoforms remained relatively low in expression throughout the time course (**Figure 6B**).

The canonical transcript of ORF13 (ORF13-ORF14-PC-Canonic) showed increasing expression from 2 hpi (8.33%) to a peak at 48 hpi (73.45%), while the long isoform (ORF13-ORF14-PC-Long) dominated early expression, peaking at 4 hpi (100%) and gradually declining to 26.22% at 48 hpi (**Figure 6C**).

In ORF14, transcript dynamics revealed distinct patterns and isoform switching among the canonical isoform, the long variants (L), and the truncated variant (ORF14.5, representing a 5′-truncated transcript). The canonical isoform dominated early expression, peaking at 63.89% at 2 hpi, and did not regain prominence later, showing a steady decline. The long variants (L) contributed modestly, with specific peaks such as L4 at 18 hpi (3.89%), while others like L1 and L3 remained low throughout, peaking below 1% by 48 hpi. The truncated variant ORF14.5 dominated the mid-phase, peaking at 78.54% at 6 hpi and remaining the most abundant transcript until 18 hpi, before declining to 24.44% at 48 hpi (**Supplementary Figure S9E**).

In the gene ORF11, the canonical ORF11 transcript peaked at 8 hpi (86.65%) before declining, while its truncated ORF-carrying ORF11.5 isoform showed an inverse pattern, peaking early at 4 hpi (74.72%) and decreasing steadily thereafter (**Supplementary Figure S9F**).

For ORF54, the short (S) isoform dominated early expression, peaking at 2 hpi (100%) and declining sharply after 6 hpi, while the canonical transcript emerged later, surpassing the short isoform at 8 hpi (48.85% vs. 22.47%) and dominating mid-to-late stages, peaking at 18 hpi (96.36%) and maintaining high levels through 48 hpi (93.17%). (**Supplementary Figure S9C**).

**Comparison of replication origin-associated transcripts of EHV-1 and PRV**

Replication origin-associated RNAs (raRNAs) are located near the replication origins (Oris) within herpesviruses. A novel ncRNA, named NOIR, has been discovered intersecting the 5′ ends of longer transcript isoforms associated with the major transactivator genes ORF64 and ORF65. This ncRNA is positioned near OriL and was previously identified in the published EHV transcriptome study [28]. Interestingly, a similar ncRNA was also found in PRV, where it resides between the IE180 and US1 (Unique Short 1) transactivator genes, similarly located adjacent to OriL. Compared to our previous publication [29], CAGE-Seq analysis enabled us to update both CTO-S13, located at Ori-L (**Supplementary Figure S11A**) and identified exclusively in EHV-1 and PRV, as well as NOIR [11], located at OriS (**Supplementary Figure S11B**) and found in Varicelloviruses. Since these are ncRNAs, their level of conservation is lower than that of the protein-coding regions.

## **Discussion**

## The last couple of years have witnessed significant advancements in sequencing technologies [41]. Full-length, lrRNA-Seq methods have revolutionized transcriptome research, particularly in organisms with small genomes. This has revealed that viral transcriptomic structures are far more complex than previously thought [11]. Discoveries include a wide array of overlapping transcripts, such as extended 5′-UTR isoforms, polygenic and complex transcripts (containing at least one gene on an opposite orientation), truncated mRNAs containing in-frame ORFs, and read-through transcripts [21,42–44]. Recent studies have demonstrated that the occurrence of nested genes within a larger canonical gene is more prevalent in viruses than previously believed [23,44,45]. Neither SRS nor LRS techniques alone are effective in detecting the products of these nested genes. This insufficiency stems from the tendency of lrRNA-Seq techniques to cause considerable 5' truncation in transcripts, potentially leading to the misidentification of TSSs [46]. To address this issue, we utilized CAGE sequencing via the Illumina MiSeq platform, a standard approach for identifying the 5' ends of capped RNA molecules. Although CAGE-Seq is generally reliable, it has the potential for detecting some fraction of degraded RNA. This possibility arises because, in mammalian cytoplasm degraded RNAs can be capped by special host enzymes [47]. Although these incomplete transcripts originate from biological processes, we attempted to minimize this form of 'noise' by establishing a high score threshold for recognizing them as true TSSs, produced through transcription.

Polycistronism is a common feature in bacteria and viruses, but it is rare in eukaryotes. In prokaryotes and bacteriophages, a ribosomal binding site on the mRNAs, known as the Shine-Dalgarno sequence, facilitates the translation of downstream genes in polycistronic RNA molecules. Many small-genome eukaryotic viruses have evolved various mechanisms, such as internal ribosome entry sites, ribosomal frameshifting, or leaky ribosomal scanning [48]. In herpesviruses, co-oriented genes often form clusters that produce transcripts with shared downstream sequences and unique 5′-exons, following a pattern like ‘abcd’, ‘bcd’, ‘cd’, and ‘d’, where ‘a’ is the most upstream gene and ‘d’ is the most downstream. The role of polygenic transcripts in large DNA viruses remains unclear, as translation from downstream genes has been rarely documented [49,50].

We have previously reported that EHV-1 exhibits more frequent splicing events compared to related alphaherpesviruses [28]. Transcripts of ORF44 (homolog of HSV ul15), ORF65 (homolog of HSV us1), and noir are spliced in other alphaherpesviruses as well. However, EHV-1 uniquely features splicing in different genomic regions, such as ORF6/12, ORF35/39, and ORF53/58. Additionally, the splicing observed in ORF44 extends to adjacent genomic areas, including ORF49/50, which is a distinctive characteristic of this virus.

An ongoing challenge in long-read RNA sequencing (lrRNA-Seq) studies is that most existing pipelines prioritize the most abundant isoforms and rely on reference annotations primarily tailored for eukaryotic transcriptomes, thus overlooking the complexities typical of viral genomes. Recent benchmarks, for instance, have focused on synthetic or mammalian data sets without evaluating the extensive overlap, alternative transcription start sites (TSSs), and co-terminal transcripts that frequently occur in viruses [51]. Tools such as StringTie2 and gff-compare, despite their utility in annotating standard eukaryotic transcripts, often collapse shorter isoforms into a single long isoform—an issue aggravated in viruses where numerous potential TSSs and overlapping ORFs create a high-density genomic landscape. NAGATA, a pipeline with reported applicability for viruses, has shown promise for native RNA (dRNA) sequencing data [38], but it explicitly discards 5′ soft-clipped reads, rendering it inapplicable to dcDNA-seq libraries that rely on 5′ adapter sequences to orient alignments. By contrast, our LoRTIA pipeline accommodates these 5′ adapter sequences, using them to determine strand information and effectively process dcDNA-seq data. Nevertheless, LoRTIA shares the fundamental challenge of distinguishing genuine 5′-truncated transcripts from artefacts introduced by incomplete reverse transcription or cytoplasmic mRNA recapping, the latter of which can produce novel 5′ ends that may be biologically meaningful [40] but risk being misidentified as legitimate viral TSSs if not carefully filtered.

In view of these constraints, our study focused on capturing canonical full-length viral transcripts while also monitoring alternative TSSs and transcription end sites (TESs) at lower abundance. By integrating multiple data sources (dcDNA-seq, dRNA-seq, and CAGE-seq) and using different tools—including LoRTIA for dcDNA libraries and NAGATA for dRNA data—we applied stringent criteria, particularly for 5′-truncated transcripts, to reduce false positives without overlooking potentially meaningful low-abundance isoforms. We also performed a peak analysis on TSS clusters identified from the CAGE data (CAGEfightR) using dcDNA read counts to refine 5′ boundaries. Together, these measures offer a robust, flexible strategy for annotating viral transcripts more accurately under current methodological and computational constraits.

In this work, we integrated dcDNA-Seq and CAGE-Seq data to create a comprehensive, time-resolved map of the EHV-1 transcriptome. By leveraging the strengths of each method - CAGE-Seq for high-resolution TSS identification and dcDNA-Seq for capturing full-length transcripts - we validated previously excluded transcripts and identified a substantial number of new isoforms. We refined TSS and TES annotations, discovered multiple spliced variants, and documented dynamic patterns of isoform switching, highlighting the EHV-1 transcriptome's complexity and fluidity beyond previous understanding. This integrated approach also enabled us to cluster genes into *de novo* kinetic classes, revealing overlapping temporal waves of expression that transcend the traditional IE/E/L framework.

The implications of our findings are significant for understanding the regulatory strategies of EHV-1. The presence of numerous transcript isoforms and intricate splicing dynamics suggests that the virus employs multiple layers of post-transcriptional control. Temporal shifts in isoform prevalence, along with alternative TSSs and TESs, indicate that EHV-1 fine-tunes gene expression throughout the infection cycle. The complexity arising from the vast number of transcript isoforms likely enables the virus to respond flexibly to host conditions, optimize resource utilization, and orchestrate the production of viral components for efficient replication and spread.

We believe that the significance of these isoforms arises not only from their coding capacity but also from their ability to physically inhibit the transcription of other genes through the process of transcriptional interference9. For instance, the early increase in the 5' truncated isoform of **ORF11** could indicate a regulatory mechanism to quickly produce a necessary protein without the full-length transcript, while the later increase may reflect a need for the complete protein function in later stages. On the other hand, the steady increase in **ORF40** AT isoforms (TES variants) suggests a role in interfering with the transcription initiation of its adjacent gene, the CTO, which is one of the most abundant non-coding transcripts.

Taken together, our results emphasize that EHV-1 gene regulation is governed by a sophisticated and multilayered transcriptional program. By delineating the full complement of viral transcripts and their temporal patterns - including splicing and isoform switching - we provide a solid foundation for future studies to unravel the molecular mechanisms underlying viral replication and pathogenesis.

## **Data availability:**

The sequencing datasets generated in this study are available at the European Nucleotide Archive under the accession: PRJEB52190 and PRJEB6233.

The R codes used to perform the analysis and generate the plots are available at: <https://github.com/Balays/EHV-1-dynamic>

## **References**

1. O’Callaghan DHR. Encyclopedia of Virology. Academic Press, Harcourt Brace & Company. 1994;5.

2. Carroll CL, Westbury HA. Isolation of equine herpesvirus 1 from the brain of a horse affected with paresis. Aust Vet J. 1985;62: 345–346. doi:10.1111/j.1751-0813.1985.tb07660.x

3. Webster RG, Granoff A. Equine Herpesvirus in Encyclopedia of Virology. 1994.

4. Roizmann B, Desrosiers RC, Fleckenstein B, Lopez C, Minson AC, Studdert MJ. The family Herpesviridae: an update. The Herpesvirus Study Group of the International Committee on Taxonomy of Viruses. Arch Virol. 1992;123: 425–449. doi:10.1007/BF01317276

5. Telford EA, Watson MS, McBride K, Davison AJ. The DNA sequence of equine herpesvirus-1. Virology. 1992;189: 304–316. doi:10.1016/0042-6822(92)90706-u

6. Allen, G., Kydd, J., Slater, J., and Smith, K. Equid herpesvirus 1 and equid herpesvirus 4 infections. Infect Dis Livest 2. 2004; 829–859.

7. Paillot R, Case R, Ross J, Newton R, Nugent J. Equine Herpes Virus-1: Virus, Immunity and Vaccines. The Open Veterinary Science Journal. 2008;2. Available: https://benthamopen.com/ABSTRACT/TOVSJ-2-68

8. Caughman GB, Staczek J, O’Callaghan DJ. Equine herpesvirus type 1 infected cell polypeptides: evidence for immediate early/early/late regulation of viral gene expression. Virology. 1985;145: 49–61. doi:10.1016/0042-6822(85)90200-4

9. Kim SK, Jang HK, Albrecht RA, Derbigny WA, Zhang Y, O’Callaghan DJ. Interaction of the equine herpesvirus 1 EICP0 protein with the immediate-early (IE) protein, TFIIB, and TBP may mediate the antagonism between the IE and EICP0 proteins. J Virol. 2003;77: 2675–2685. doi:10.1128/jvi.77.4.2675-2685.2003

10. Gray WL, Baumann RP, Robertson AT, Caughman GB, O’Callaghan DJ, Staczek J. Regulation of equine herpesvirus type 1 gene expression: characterization of immediate early, early, and late transcription. Virology. 1987;158: 79–87. doi:10.1016/0042-6822(87)90240-6

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

12. Laver T, Harrison J, O’Neill PA, Moore K, Farbos A, Paszkiewicz K, et al. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif. 2015;3: 1–8. doi:10.1016/j.bdq.2015.02.001

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

14. Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, et al. A highly conserved program of neuronal microexons is misregulated in autistic brains. Cell. 2014;159: 1511–1523. doi:10.1016/j.cell.2014.11.035

15. Tombácz D, Csabai Z, Oláh P, Balázs Z, Likó I, Zsigmond L, et al. Full-Length Isoform Sequencing Reveals Novel Transcripts and Substantial Transcriptional Overlaps in a Herpesvirus. PLOS ONE. 2016;11: e0162868. doi:10.1371/journal.pone.0162868

16. Moldován N, Tombácz D, Szűcs A, Csabai Z, Snyder M, Boldogkői Z. Multi-Platform Sequencing Approach Reveals a Novel Transcriptome Profile in Pseudorabies Virus. Frontiers in Microbiology. 2018;8. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2017.02708

17. Tombácz D, Moldován N, Balázs Z, Gulyás G, Csabai Z, Boldogkői M, et al. Multiple Long-Read Sequencing Survey of Herpes Simplex Virus Dynamic Transcriptome. Front Genet. 2019;10: 834. doi:10.3389/fgene.2019.00834

18. Balázs Z, Tombácz D, Csabai Z, Moldován N, Snyder M, Boldogkői Z. Template-switching artifacts resemble alternative polyadenylation. BMC Genomics. 2019;20: 824. doi:10.1186/s12864-019-6199-7

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

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

21. Balázs Z, Tombácz D, Szűcs A, Snyder M, Boldogkői Z. Long-read sequencing of the human cytomegalovirus transcriptome with the Pacific Biosciences RSII platform. Sci Data. 2017;4: 170194. doi:10.1038/sdata.2017.194

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

23. Moldován N, Torma G, Gulyás G, Hornyák Á, Zádori Z, Jefferson VA, et al. Time-course profiling of bovine alphaherpesvirus 1.1 transcriptome using multiplatform sequencing. Sci Rep. 2020;10: 20496. doi:10.1038/s41598-020-77520-1

24. Depledge DP, Srinivas KP, Sadaoka T, Bready D, Mori Y, Placantonakis DG, et al. Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. Nat Commun. 2019;10: 754. doi:10.1038/s41467-019-08734-9

25. Whisnant AW, Jürges CS, Hennig T, Wyler E, Prusty B, Rutkowski AJ, et al. Integrative functional genomics decodes herpes simplex virus 1. Nat Commun. 2020;11: 2038. doi:10.1038/s41467-020-15992-5

26. Braspenning SE, Sadaoka T, Breuer J, Verjans GMGM, Ouwendijk WJD, Depledge DP. Decoding the Architecture of the Varicella-Zoster Virus Transcriptome. mBio. 2020;11: e01568-20. doi:10.1128/mBio.01568-20

27. Rozman B, Nachshon A, Levi Samia R, Lavi M, Schwartz M, Stern-Ginossar N. Temporal dynamics of HCMV gene expression in lytic and latent infections. Cell Reports. 2022;39: 110653. doi:10.1016/j.celrep.2022.110653

28. Tombácz D, Torma G, Gulyás G, Fülöp Á, Dörmő Á, Prazsák I, et al. Hybrid sequencing discloses unique aspects of the transcriptomic architecture in equid alphaherpesvirus 1. Heliyon. 2023;9: e17716. doi:10.1016/j.heliyon.2023.e17716

29. Torma G, Tombácz D, Csabai Z, Almsarrhad IAA, Nagy GÁ, Kakuk B, et al. Identification of herpesvirus transcripts from genomic regions around the replication origins. Sci Rep. 2023;13: 16395. doi:10.1038/s41598-023-43344-y

30. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor. Bioinformatics. 2018;34: 3094–3100. doi:10.1093/bioinformatics/bty191

31. Morgan M, Pagès H, Obenchain V, H. N. Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. 2022.

32. Barrett T, Dowle M, Srinivasan A, Gorecki J, Chirico M, Hocking T, et al. data.table: Extension of “data.frame.” 2024. Available: https://cran.r-project.org/web/packages/data.table/index.html

33. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biology. 2004;5: R80. doi:10.1186/gb-2004-5-10-r80

34. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. F1000Res. 2020;9: ISCB Comm J-304. doi:10.12688/f1000research.23297.2

35. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics. 2009;25: 1841–1842. doi:10.1093/bioinformatics/btp328

36. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29: 15–21. doi:10.1093/bioinformatics/bts635

37. Thodberg M, Thieffry A, Vitting-Seerup K, Andersson R, Sandelin A. CAGEfightR: analysis of 5′-end data using R/Bioconductor. BMC Bioinformatics. 2019;20: 487. doi:10.1186/s12859-019-3029-5

38. Suzuki R, Terada Y, Shimodaira H (2019). pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling. R package version 2.2-0, https://CRAN.R-project.org/package=pvclust.

39. Tombácz D, Tóth JS, Petrovszki P, Boldogkői Z. Whole-genome analysis of pseudorabies virus gene expression by real-time quantitative RT-PCR assay. BMC Genomics. 2009;10: 491. doi:10.1186/1471-2164-10-491

40. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nat Rev Genet. 2016;17: 175–188. doi:10.1038/nrg.2015.16

41. Trotman JB, Schoenberg DR. A recap of RNA recapping. WIREs RNA. 2019; 10:e1504. https://doi.org/10.1002/wrna.1504

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

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

44. Torma G, Tombácz D, Csabai Z, Göbhardter D, Deim Z, Snyder M, et al. An Integrated Sequencing Approach for Updating the Pseudorabies Virus Transcriptome. Pathogens. 2021;10: 242. doi:10.3390/pathogens10020242

45. Tombácz D, Sharon D, Szűcs A, Moldován N, Snyder M, Boldogkői Z. Transcriptome-wide survey of pseudorabies virus using next- and third-generation sequencing platforms. Sci Data. 2018;5: 180119. doi:10.1038/sdata.2018.119

46. Calvo-Roitberg E, Daniels RF, Pai AA. Challenges in identifying mRNA transcript starts and ends from long-read sequencing data. bioRxiv. 2023; 2023.07.26.550536. doi:10.1101/2023.07.26.550536

47. Trotman JB, Schoenberg DR. A recap of RNA recapping. WIREs RNA. 2019;10: e1504. doi:10.1002/wrna.1504

48. Stacey SN, Jordan D, Williamson AJ, Brown M, Coote JH, Arrand JR. Leaky scanning is the predominant mechanism for translation of human papillomavirus type 16 E7 oncoprotein from E6/E7 bicistronic mRNA. J Virol. 2000;74: 7284–7297. doi:10.1128/jvi.74.16.7284-7297.2000

49. Vilela C, McCarthy JEG. Regulation of fungal gene expression via short open reading frames in the mRNA 5’untranslated region. Mol Microbiol. 2003;49: 859–867. doi:10.1046/j.1365-2958.2003.03622.x

50. Kronstad LM, Brulois KF, Jung JU, Glaunsinger BA. Dual Short Upstream Open Reading Frames Control Translation of a Herpesviral Polycistronic mRNA. PLOS Pathogens. 2013;9: e1003156. doi:10.1371/journal.ppat.1003156

51. Dong, X., Du, M.R.M., Gouil, Q. *et al.* Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures. *Nat Methods* **20**, 1810–1821 (2023). https://doi.org/10.1038/s41592-023-02026-3

**Author Contributions:** Conceptualization, D.T. and Z.B.; library preparation: ZC; CAGE-Seq, D.T.; dcRNA-Seq: D.T., Á.D.; Bioinformatic analysis, G.T., B.K., Á.F, GG. Writing the original draft, D.T., and Z.B.; Visualization, G.T., B.K. and D.T; Supervision, Z.B.; Writing the manuscript: Z.B., B.K., D.T. All authors have read and agreed to the published version of the manuscript.

**Funding**: This study was supported by the National Research, Development and Innovation Office grant: K 142674 to ZB and FK 142676 to DT. The APC charge was covered by the University of Szeged Open Access Fund: 7214. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

**Conflicts of Interest**: The authors declare no conflict of interest.

## **Legends to Figures**

**Figure 1. Kinetics of transcription start sites of EHV-1 detected by dcDNA-Seq and validated by CAGE-Seq.**

The time-course experiment utilizing dcDNA-Seq spanned 8 timepoints, ranging from 1 hour to 24 hours. TSSs were identified through LRS analysis and confirmed via CAGE-Seq. For each nucleotide, we counted the number of reads beginning at that position with their 5' ends. We included only those reads that had clear directionality, which was determined by the presence of 5' or 3' adapters. Data from all three replicates were combined.

(a) We then grouped the TSS signal strength values into 50-nt segments to illustrate the distribution of TSSs. The y-axis of the graph was automatically scaled to accommodate up to 500 read counts. An image with lower (5,000 read counts) and higher (50 read counts) resolution details can be found in Supplementary Figure 1a and 1b, respectively. In the representation, genes are indicated by arrows, and the distribution of TSSs is shown in different colors: red for the positive strand and blue for the negative strand. The bottom row of the image displays the CAGE-Seq counts.

(b)To illustrate the distribution of TSSs within each group, we normalized the signal strength by dividing it with the total signal strength observed for all samples in the same hour post-infection group (and the CAGE-Seq). This approach allows for a comparative analysis of TSS distribution relative to the group's overall viral read count.

**Figure 2. Kinetics of transcription end sites of EHV-1 detected by dcDNA-Seq and validated by dRNA-Seq.**

The time-course study covered 8 intervals, from 1 to 24 hours. The TESs were detected using oligo(dT) priming-based dcDNA sequencing, which subsequently were confirmed with dRNA-Seq. We counted the reads initiating from each nucleotide position with their 3' ends, focusing only on those with clear directional cues identified by the presence of 5' or 3' adapters. Data from all three dcDNA-Seq replicates were merged.

(a) The aggregated read counts were then summed into 50-nt blocks to illustrate the TES distributions. The graph's y-axis was set to automatically adjust, supporting up to 500 read counts. Images offering lower (5,000 read counts) and higher (50 read counts) resolution views are available in Supplementary Figures 4a and 4b. The diagrams mark genes with arrows and color-code the TSS distribution, using red for the positive strand and blue for the negative strand.

(b)To illustrate the distribution of TESs across each group, we adjusted the read counts at each position (prior to aggregating them into blocks) by the overall read count for all samples in the same time interval group. This approach enables a comparison of TES distribution against the total viral read count for the group.

**Figure 3. Kinetic profiling of canonical EHV-1 transcripts using total viral read counts for normalization according to kinetic classes**

This figure illustrates the kinetic profiling of canonical EHV-1 transcripts, utilizing the total viral read counts per sample for normalization. The analysis included only those reads that aligned with both the canonical TSS of genes at their 5' ends and the canonical TES of genes at their 3' ends (allowing a deviation of +/- 10 nt for both alignments). This method aggregated the counts of canonical transcripts for each gene in every sample. The mean values are represented as points, and standard deviations (SD) as lines, plotted on the y-axis as the ratio of transcript abundance for each gene. The x-axis represents time post-infection (hours). The panels are color-coded based on kinetic transcription phases: blue for immediate early (IE), orange for early (E), green for late (L), and red for unknown phases. This provides a visual distinction among different transcriptional dynamics throughout the infection.

**Figure 4. Dynamics of transcript isoform usage in splice-containing EHV-1 genes over the course of infection**

This figure presents the splicing dynamics within EHV-1 for the genes (a) ORF38, (b) ORF8, (c) ORF9, (d) NOIR, (e) ORF58 and (f) ORF53. The right side of each panel shows the transcript annotations, along with their parent genes and genomic locations displayed below them, with light red indicating positive-strand genes and light blue indicating negative-strand genes. The analysis focused on transcripts that matched exactly, allowing a deviation of +/- 2 nucleotides (nt) for splice junctions and +/- 10 nt for the start and end positions of transcripts. Asterisks indicate the CAGE-Seq significance level for each reference transcript. On the left side of each plot, the temporal trends of these transcript isoforms are depicted, with averages and standard deviations (SD) calculated for each time point post-infection, based on read count data from the dcDNA-Seq. Each data point is linked by lines to demonstrate the progression over time. The transcript isoforms, are color-coded according to their distinct isoforms, with these colors matching those used for the points and lines in the left panel. The isoform counts were normalized against the total number of isoform counts for each gene in each sample to calculate the ratio of each isoform. Isoforms on the right side are colored grey, if they not originate from the given gene and thus were not included in the isoform ratio calculation.

**Figure 5. Kinetic profiling of canonical EHV-1 transcripts using total viral read counts for normalization according to *de novo* kinetic clusters**

This figure illustrates the kinetic profiling of canonical EHV-1 transcripts, utilizing the total viral read counts per sample for normalization. The analysis included only those reads that aligned with both the canonical TSS of genes at their 5' ends and the canonical TES of genes at their 3' ends (allowing a deviation of +/- 10 nt for both alignments). This method aggregated the counts of canonical transcripts for each gene in every sample. The mean values are represented as points, and standard deviations (SD) as lines, plotted on the y-axis as the ratio of transcript abundance for each gene. The x-axis represents time post-infection (hours). Each cluster is colored according to its *de novo* kinetic cluster membership. The color-coding for the clustering is shown in the bottom right panel. This figure provides a visual distinction among different transcriptional dynamics, according to the gene’s relative abundance throughout the infection.

**Figure 6. Kinetics of transcript variants for isoform switching genes**

This figure illustrates the dynamics of different transcript isoforms for the selected EHV-1 genes (a) ORF11, (b) ORF19, (c) ORF13, (d) ORF40, (e) ORF14, and (f) ORF54.

The right side of each panel shows the transcript annotations, along with their parent genes and genomic locations displayed below them, with light red indicating positive-strand genes and light blue indicating negative-strand genes. The analysis focused on transcripts that matched exactly, allowing a deviation of +/- 2 nucleotides (nt) for splice junctions and +/- 10 nt for the start and end positions of transcripts. Asterisks indicate the CAGE-Seq significance level for each reference transcript. On the left side of each plot, the temporal trends of these transcript isoforms are depicted, with averages and standard deviations (SD) calculated for each time point post-infection, based on read count data from the dcDNA-Seq. Each data point is linked by lines to demonstrate the progression over time. The transcript isoforms, are color-coded according to their distinct isoforms, with these colors matching those used for the points and lines in the left panel. The isoform counts were normalized against the total number of isoform counts for each gene in each sample to calculate the ratio of each isoform. Isoforms on the right side are colored grey, if they not originate from the given gene and thus were not included in the isoform ratio calculation.

## **Legends to Supplementary Figures**

**Supplementary Figure 1. Kinetics of transcription start sites of EHV-1 detected by dcDNA-Seq and validated by CAGE-Seq.**

(a) *5,000 limit***.** Similar to Figure 1, this plot shows the 5' end distribution along the viral genome for each time-point group. The mean values for each time-point group were calculated and merged into 50-nt sized bins for visualization. The y-axis was limited to 5,000 counts.

(b) *50 limit***.** Similar to Figure 1, this plot shows the 5' end distribution along the viral genome in each time-point groups. The mean values for each time-point group was calculated and merged into 50-nt sized bins for the visualization. The y-axis was limited to 5,000 counts.

**Supplementary Figure 2. Kinetic profiling of canonical EHV-1 TSSs according to *de novo* kinetic clusters**

This figure illustrates the kinetic profiling of canonical EHV-1 TSSs, utilizing the total viral read counts per sample for normalization. The analysis included only those reads that aligned with the canonical TSS of genes at their 5' ends (allowing a deviation of +/- 10 nt). The mean values are represented as points, and standard deviations (SD) as lines, plotted on the y-axis as the ratio of TSS abundance for each gene. The x-axis represents time post-infection (hours). Each cluster is colored according to its *de novo* kinetic cluster membership. The color-coding for the clustering is shown in the bottom right panel. This figure provides a visual distinction among different transcriptional dynamics, according to the gene’s relative TSS abundance throughout the infection.

**Supplementary Figure 3. Kinetic profiling of canonical EHV-1 TSSs according to kinetic classes**

This figure illustrates the kinetic profiling of canonical EHV-1 TSSs, utilizing the total viral read counts per sample for normalization. The analysis included only those reads that aligned with the canonical TSS of genes at their 5' ends (allowing a deviation of +/- 10 nt). The mean values are represented as points, and standard deviations (SD) as lines, plotted on the y-axis as the ratio of TSS abundance for each gene. The x-axis represents time post-infection (hours). The panels are color-coded based on kinetic transcription phases: blue for immediate early (IE), orange for early (E), green for late (L), and red for unknown phases. This provides a visual distinction among different TSS dynamics throughout the infection.

**Supplementary Figure 4. Kinetics of transcription end sites of EHV-1 detected by dcDNA-Seq and validated by dRNA-Seq.**

(a) *5,000 limit.* Similar to Figure 2, this plot shows the 5' end distribution along the viral genome in each time-point groups. The mean values for each time-point group was calculated and merged into 50-nt sized bins for the visualization. The y-axis was limited to 50 counts.

(b) *50 limit.*Similar to Figure 2, this plot shows the 5' end distribution along the viral genome in each time-point groups. The mean values for each time-point group was calculated and merged into 50-nt sized bins for the visualization. The y-axis was limited to 50 counts.

**Supplementary Figure 5. Kinetic profiling of canonical EHV-1 TESs according to *de novo* kinetic clusters**

This figure illustrates the kinetic profiling of canonical EHV-1 TESs, utilizing the total viral read counts per sample for normalization. The analysis included only those reads that aligned with the canonical TSS of genes at their 5' ends (allowing a deviation of +/- 10 nt). The mean values are represented as points, and standard deviations (SD) as lines, plotted on the y-axis as the ratio of TES abundance for each gene. The x-axis represents time post-infection (hours). Each cluster is colored according to its *de novo* kinetic cluster membership. The color-coding for the clustering is shown in the bottom right panel. This figure provides a visual distinction among different transcriptional dynamics, according to the gene’s relative TES abundance throughout the infection.

**Supplementary Figure 6. Kinetic profiling of canonical EHV-1 TESs According to kinetic classes**

This figure illustrates the kinetic profiling of canonical EHV-1 TESs, utilizing the total viral read counts per sample for normalization. The analysis included only those reads that aligned with the canonical TES of genes at their 5' ends (allowing a deviation of +/- 10 nt). The mean values are represented as points, and standard deviations (SD) as lines, plotted on the y-axis as the ratio of TES abundance for each gene. The x-axis represents time post-infection (hours). The panels are color-coded based on kinetic transcription phases: blue for immediate early (IE), orange for early (E), green for late (L), and red for unknown phases. This provides a visual distinction among different TSS dynamics throughout the infection.

**Supplementary Figure 7. Kinetics of transcript isoforms for selected genes**

This figure illustrates the dynamics of different transcript isoforms for the selected EHV-1 genes (a) ORF23, (b) ORF51, (c) ORF67, (d) ORF64.

The right side of each panel shows the transcript annotations, along with their parent genes and genomic locations displayed below them, with light red indicating positive-strand genes and light blue indicating negative-strand genes. The analysis focused on transcripts that matched exactly, allowing a deviation of +/- 2 nucleotides (nt) for splice junctions and +/- 10 nt for the start and end positions of transcripts. Asterisks indicate the CAGE-Seq significance level for each reference transcript. On the left side of each plot, the temporal trends of these transcript isoforms are depicted, with averages and standard deviations (SD) calculated for each time point post-infection, based on read count data from the dcDNA-Seq. Each data point is linked by lines to demonstrate the progression over time. The transcript isoforms, are color-coded according to their distinct isoforms, with these colors matching those used for the points and lines in the left panel. The isoform counts were normalized against the total number of isoform counts for each gene in each sample to calculate the ratio of each isoform. Isoforms on the right side are colored grey, if they not originate from the given gene and thus were not included in the isoform ratio calculation.

**Supplementary Figure 8. Kinetics of transcript isoforms for selected genes**

This figure shows the *de novo* kinetic clustering of EHV-1 genes based on the viral read normalized canonic transcript counts (reads spanning from the canonic TSS tot the TES), compared to the traditional kinetic classification. The rows show the *de novo* cluster memberships, while the colors represent the kinetic classes.

**Supplementary Figure 9. Dynamics of transcript isoform usage in splice-containing EHV-1 genes over the course of infection**

This figure presents the splicing dynamics within EHV-1 for the genes (a) ORF65, (b) ORF44. The right side of each panel shows the transcript annotations, along with their parent genes and genomic locations displayed below them, with light red indicating positive-strand genes and light blue indicating negative-strand genes. The analysis focused on transcripts that matched exactly, allowing a deviation of +/- 2 nucleotides (nt) for splice junctions and +/- 10 nt for the start and end positions of transcripts. Asterisks indicate the CAGE-Seq significance level for each reference transcript. On the left side of each plot, the temporal trends of these transcript isoforms are depicted, with averages and standard deviations (SD) calculated for each time point post-infection, based on read count data from the dcDNA-Seq. Each data point is linked by lines to demonstrate the progression over time. The transcript isoforms, are color-coded according to their distinct isoforms, with these colors matching those used for the points and lines in the left panel. The isoform counts were normalized against the total number of isoform counts for each gene in each sample to calculate the ratio of each isoform. Isoforms on the right side are colored grey, if they not originate from the given gene and thus were not included in the isoform ratio calculation.

**Supplementary Figure S10. Dynamics of Total Coverage of Viral Reads.**

This figure shows the total coverage of the viral genome, calculated from reads, whose orientation were determined using either the 5' adapter, or the polyA-tail. The sub-plots show the data on different scales, limited to a read count of: (A) 5,000; (B) 500; (C) 50; and (D) using no predefined limit.

**Supplementary Figure S11. Comparison of replication origin-associated transcripts of three viruses**

This illustration compares the sequences of raRNAs [(a) CTO-S; (b) NOIR] from EHV-1 with those of two PRV strains (Kaplan and MdBio).

## **Legends to Supplementary Tables**

**Supplementary Table S1. Identification of immediate-early gene expression of EHV-1 using the inhibition of protein synthesis by CHX.**

In this experiment, we utilized two concentrations of CHX (20 and 100 mg/ml) and observed the effects at two different incubation durations (6 and 8 hours). Our findings unequivocally indicate that ORF64 is the sole immediate early gene of EHV-1. The numbers refer to the detected reads.

**Supplementary Table S2**. **Classification of newly annotated transcripts.**

This table lists the categories and counts of additional transcripts identified and validated through the integration of CAGE-Seq and dcDNA-Seq data.

Putative mRNAs: These transcripts harbor an open reading frame (ORF) and are 3'-coterminal but 5'-truncated (in-frame) variants of the canonical ORF of the given gene. "Putative" indicates that, while these transcripts contain an ORF, they may encode truncated proteins due to their shorter 5' ends compared to the canonical transcripts.

Non-coding RNAs: Transcripts that lack ORFs and are presumed to function as regulatory molecules within the viral genome.

Non-coding-short RNAs: Shorter variants of non-coding RNAs with truncated transcript lengths, potentially representing distinct regulatory elements.

Long monocistronic transcripts and Short monocistronic transcripts: These harbor the same ORF as the canonical transcript but differ in the length of their 5' untranslated regions (UTRs). "Long" monocistronic transcripts have extended 5' UTRs compared to the canonical transcripts, while "short" monocistronic transcripts have truncated 5' UTRs.

Long multicistronic transcripts and Short multicistronic transcripts: These span multiple genes and also differ in their 5' UTR lengths. "Long" multicistronic transcripts have extended 5' UTRs, whereas "short" multicistronic transcripts have truncated 5' UTRs compared to the canonical multicistronic transcripts.

Antisense-long transcripts: Long transcripts antisense to known genes, potentially involved in regulatory functions such as gene silencing or modulation of gene expression.

The number (N) represents the count of transcripts in each category.

**Supplementary Table 2. D of EHV-1 Transcripts previously annotated and discovery of new TSSs through CAGE Sequencing**

**A:** List of previously published transcripts, including dRNA data analyzed by CAGE-Seq and NAGATA. An additional Confirmed Putative column has been included, indicating the proportion of putative transcripts identified by CAGE-Seq relative to the canonical transcript. Only those putative transcripts reaching or exceeding 5% are marked in this column.

**B:** List of new transcripts identified by CAGE.

**C:** List of introns identified by LoRTIA in dRNA sequencing, validated using the NAGATA software.

**Supplementary Table 2. Verification of EHV-1 Transcripts and Discovery of Novel Transcripts Using CAGE Sequencing**

This table summarizes previously annotated EHV-1 transcripts, identifies new transcription start sites (TSSs) through CAGE sequencing, and validates transcripts through multiple sequencing approaches. Known transcription end sites (TESs) were linked with novel TSSs identified in this study to define new transcripts. Transcripts were further filtered and validated through dcDNA sequencing and dRNA-Seq, with a minimum relative abundance of 5% compared to the host gene in the CAGE dataset. Details of transcript features and splicing events are organized across three sheets:

Sheet A: Previously Published Transcripts

This sheet lists previously annotated transcripts, incorporating TSSs validated in this study using CAGE sequencing and dcDNA data.

Columns include:

Gene: Name of the associated gene.

Transcript Name: Identifier of the transcript.

Transcript Category: Type of transcript (e.g., monocistronic, multicistronic, antisense, or non-coding).

Sequencing Validation: Presence or absence of the transcript as identified by dRNA-Seq (NAGATA), CAGE-Seq, and dcDNA-Seq.

TSS Validation: Indicates whether the TSS was validated using CAGE data.

Confirmed Putative: Indicates novel putative transcripts meeting the 5% relative abundance threshold in CAGE data compared to the canonical transcript.

Start and Stop Coordinates: Genomic positions of the transcript.

Sheet B: Novel Transcripts

This sheet provides details on transcripts newly identified in this study through TSS-TES pairing:

Columns include:

Gene: Name of the host gene.

Transcript Category: Classification of novel transcripts (e.g., long, short, monocistronic, or multicistronic).

Sequencing Validation: Confirmation of these transcripts in dcDNA-Seq, dRNA-Seq, and CAGE-Seq.

Start and Stop Coordinates: Genomic positions of the novel transcripts.

Sheet C: Comparison of LoRTIA and NAGATA for Introns Annotation

This sheet lists introns annotated from dRNA sequencing data, comparing results obtained using LoRTIA and NAGATA.

Columns include:

Intron Donor and Acceptor: Genomic coordinates marking the start (donor) and end (acceptor) of splicing events.

Sequencing Validation: Presence or absence of annotated introns identified using both tools (LoRTIA and NAGATA) alongside dcDNA-Seq.

Comparison Notes: Highlights any discrepancies or agreements between the two methods in detecting specific splicing events.

**Supplementary Table 3. Read counts**

**Supplementary Table 4. Concentrations of total and poly(A)-selected RNAs**