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 combining CAGE-Seq and dcDNA-Seq, we identified and validated approximately 200 novel transcripts, refining the EHV-1 transcriptome annotation and uncovering 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 LRS technology based on ONT is highly suitable for direct 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, dRNA-Seq has limitations in accurately identifying transcription start sites (TSSs) due to 5' truncation caused by motor protein stalling during sequencing.

To overcome these limitations and refine the EHV-1 transcriptome annotation, we integrated cap analysis of gene expression sequencing (CAGE-Seq) with direct cDNA sequencing (dcDNA-Seq) in the current study. CAGE-Seq offers high-resolution mapping of TSSs, while dcDNA-Seq provides full-length transcripts without the 5' truncation issues of dRNA-Seq. This combined approach allowed us to validate transcripts that were identified but not included (due to the strict filtering criteria used) in our previous annotation and discover additional transcript isoforms with greater accuracy.

Furthermore, by utilizing dcDNA-Seq, we were able sequence 27 samples, biological triplicates collected at nine time points ranging from 1 to 48 hours post-infection (1, 2, 4, 6, 8, 12, 18, 24, and 48 hpi). This extensive temporal sampling allowed us to capture the full dynamics of viral gene expression throughout the complete infection cycle. By analyzing these time-resolved data, we clustered the genes based on their expression dynamics into de novo kinetic classes.

The primary novelty of our work lies in investigating the temporal dynamics of viral gene expression throughout the infection cycle and clustering the genes into de novo kinetic classes using direct cDNA sequencing (dcDNA-Seq). Additionally, by integrating cap analysis of gene expression sequencing (CAGE-Seq) data for reannotation purposes, we refined existing annotations, validated previously excluded transcripts, and uncovered previously unrecognized transcript isoforms with greater accuracy. This comprehensive analysis provides new insights into the complexity and regulation of the EHV-1 transcriptome, enhancing our understanding of viral gene regulation and potentially informing strategies for intervention.

## **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 4**). 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 4**).

**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 and Data Analysis**

*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] and other R-packages from the Bioconductor repository [33]. This was then used to count the 3'- and 5'-ends per nucleotide and the coverages. The GFF-compare script [34] was used to count the reference transcripts in the samples. But because this tool tends to assign shorter transcript isoforms, which are contained within another transcript, to the longer one, we ran this tool iteratively for each reference transcript separately and the results were merged together and for each query alignment the best hit was selected, i.e. that reference transcript, the ends of which show the smallest distances. For counting the reference isoforms, only the hits with “equal to reference” were kept, with a distance cutoff of 10 nt-s for both ends. R-packages such as rtracklayer [35] was used to export and import .gff3 files. The moanin R-package (<https://nellev.github.io/moanin/>) was used to cluster together genes with similar expression profiles.

*CAGE sequencing*

The STAR aligner (version 2.7.3 a) [36] 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 [37] 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.

## **Results**

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

In this study, we employed direct cDNA sequencing (dcDNA-Seq) to conduct a time-course transcriptomic analysis of EHV-1 and the reannotation of it using mainly the ONT MinION platform, amended with cap analysis of gene expression sequencing (CAGE-Seq) on Illumina MiSeq platform.

The lrRNA-Seq techniques are recognized to introduce technical noise due to library preparation, sequencing, and mapping errors. To overcome this, we previously implemented native RNA sequencing using the ONT platform [28]. These data were used for the validation of novel dcDNA-Seq results.

Although the dcDNA-Seq technique can precisely map TESs and splice sites of transcripts, it also produces a significant number of transcription reads with incomplete 5' ends. To circumvent this problem, we employed cap analysis of gene expression sequencing (CAGE-Seq) on Illumina MiSeq platform.

We mapped the reads to the EHV-1 genome (NC\_001491.2) and the alignments underwent adapter identification, TSS and TES finding through the LoRTIA pipeline, which was developed in our laboratory [18].

To confirm sequencing reads as genuine transcripts, we required a minimum of three independent reads sharing identical TSSs and TESs.

Similarly, we recognized a new splice site as valid only if it appeared in at least three separate reads. We also set the detection by dRNA-Seq [28] as the criterion for the annotation of TESs and splice sites.

We termed the most abundant transcripts from a viral gene as the canonical mRNAs and its TSS and TES as the canonical TSS and TES, respectively.

In most instances, these transcripts were observed to be significantly more prevalent than other transcript isoforms, including those with variations in TSSs, TESs, and splice variants.

\*\*\*

In this study, we employed direct cDNA sequencing (dcDNA-Seq) to conduct a time-course transcriptomic analysis of EHV-1 and the reannotation of it using mainly the ONT MinION platform, amended with cap analysis of gene expression sequencing (CAGE-Seq) on Illumina MiSeq platform.

In our previous study [28], strict filtering criteria were applied to ensure the highest confidence in transcript identification, which resulted in the exclusion of some transcripts that were detected but not included in the final annotation. These criteria were necessary due to the limitations of dRNA-Seq, particularly regarding accurate TSS identification and potential sequencing errors. During this work, we integrated CAGE-Seq data with dcDNA-Seq reads to overcome these limitations. CAGE-Seq provided high-resolution mapping of TSSs, allowing us to validate the 5' ends of transcripts with greater accuracy. By aligning the 5' ends of dcDNA-Seq reads with CAGE-Seq TSS clusters, we were able to confirm the authenticity of previously excluded transcripts and include them in our refined annotation. In addition to those, we annotated novel transcripts as well. HOW MANY?

In addition, we carried out the temporal dynamics of viral gene expression throughout the infection cycle. To achieve that, we mapped the dcDNA reads to the EHV-1 genome (NC\_001491.2) and the alignments underwent adapter identification, TSS and TES finding through the LoRTIA pipeline, which was developed in our laboratory [18]. For the kinetic analysis we used the canonical transcripts, TSSs and TESs of each viral gene. We termed the most abundant transcripts (for a viral gene) as the canonical mRNAs and its TSS and TES as the canonical TSS and TES, respectively. We analysed the gene expression kinetics, clustered them into de-novo kinetic classes, based on their expression curves and compared these classes with the traditional kinetic classes.

This combined approach not only enhanced the accuracy of TSS identification but also allowed us to construct full-length transcripts by associating validated TSSs with TESs identified in our previous work. As a result, we identified and validated approximately 303 additional transcripts, providing a more comprehensive and detailed map of the EHV-1 transcriptome. Additionally, we

**The ORF64 is the only EHV-1 IE gene**

It has been previously established that ORF64 is the only IE gene of EHV-1 (Smith et al., 1992). 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 1**). In pseudorabies virus (PRV), the closest relative of EHV-1 with an annotated transcriptome, the homologous gene (*ie180*) is also the only IE gene [38]. 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 utilized CAGE-Seq on an Illumina MiSeq platform to detect the TSSs of EHV-1 transcripts with high resolution. We integrated the CAGE-Seq data 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 2). The results showed that the support for reference transcripts varied across different levels of significance. Specifically, 251 transcripts received the highest level of support (\*\*\*), indicating robust validation by CAGE-Seq and dcDNA-Seq integration. Medium support (\*\*) was found for 47 transcripts, while the lowest level of support (\*) was seen in 51 transcripts. WHAT THE ASTERIXES MEAN?

Subsequently, those dcDNA-Seq reads that could not be assigned to transcripts in the previous annotation, as their 5` ends differed with more than 10 nts, were used to refine and expand it. To achieve this, we aligned the 5` ends of reads containing the correct 5' adapter sequences (as determined by the LoRTIA pipeline) with TSS clusters identified from the CAGE-Seq data using the CAGEfightR package.

The majority of these transcripts were already found during our previous transcriptome annotation, but due to the stringent filtering criteria applied there, they were excluded from the final published transcript list. Now, the integration of CAGE-Seq data allowed us to validate their TSSs with higher confidence. Additionally, we included novel transcripts that were identified and validated through this combined approach. As some TSS clusters identified from the CAGE-Seq data were broad (occasionally exceeding 200 base pairs), we refined the clusters using the dcDNA-Seq reads to identify distinct TSS peaks within them. We then constructed the final transcripts by combining dcDNA-Seq reads based on their shared TESs (from our previous annotations) and the refined TSS peak positions. Transcripts were included in the final annotation if they met stringent criteria: they had to be supported by at least five dcDNA-Seq reads in the time point samples, and their 5' ends had to align with validated TSSs from the CAGE-Seq data.

Through the integration of CAGE-Seq and dcDNA-Seq data, we identified and validated a total of 303 additional transcripts, comprising both previously excluded transcripts and completely novel ones. These newly annotated transcripts include novel TSS variants of known genes and additional transcript isoforms, enhancing the comprehensiveness of the EHV-1 transcriptome annotation. The transcripts were classified into various categories based on their structural and functional characteristics (**Table 1**). Notably, we identified 137 putative mRNAs, which contain 3'-coterminal but 5'-truncated (in-frame) variants of the canonical ORF, potentially encoding truncated proteins. We also identified 26 non-coding RNAs and 8 non-coding-short RNAs, 38 short monocistronic transcripts and 10 short multicistronic transcripts (which span multiple genes). In addition, we annotated 42 long monocistronic transcripts and 10 long multicistronic transcripts. In contrast to the putative mRNAs, the long and short variants encode the same ORF as the canonical transcript, but have either extended or truncated 5' untranslated regions (UTRs), compared to canonical transcripts.

—including the CTO-L and CTO-L-2 transcripts—

These findings significantly expand the EHV-1 transcriptome annotation, providing a more detailed and nuanced understanding of the viral gene expression landscape. Details on the CAGE-Seq results, combined with the updated list of transcripts and their counts in each sample, can be found in **Supplementary Table 2**. This table indicates which transcripts are novel additions compared to our previous study and their levels of support based on our validation criteria.

**Kinetic characterization of TSSs, TESs and splice sites of EHV-1**

Here, we explored the dynamic features of TSSs (**Figure 1** and **Supplementary Figure 1)**, TESs (**Figure 2** and **Supplementary Figure 2**) and splice sites (**Figure 3**) of EHV-1 transcripts over the course of infection.

***Time Points Corresponding to IE, E, and L Genes***

EHV-1 genes are categorized into immediate-early (IE), early (E), and late (L) genes based on their temporal expression patterns during infection. Immediate-early genes are expressed immediately upon infection without the need for de novo protein synthesis, early genes are expressed before viral DNA replication, and late genes are expressed after the onset of viral DNA replication.

In our experiments, we collected samples at multiple time points post-infection (hpi): 1, 2, 4, 6, 8, 12, 18, 24, and 48 hpi. For the purpose of this study, we consider the following time points: immediate-early (IE) genes: expressed at 1–2 hpi; early (E) genes: expressed predominantly at 2–6 hpi; and late (L) genes: expressed predominantly after 6 hpi, reaching maximal expression at 8–12 hpi and beyond. This is in line with previous categorizations of alphaherpesviral kinetic classes.

***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 1**). 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.

However, several TSS dynamics did not align with their expected kinetic class, suggesting complex transcriptional behavior. For instance, ORF38, traditionally classified as a late gene, exhibited TSS activity peaking at 6 hpi, a pattern more characteristic of early genes. Similarly, ORF45, typically a late gene, showed a peak at 12 hpi but also had significant TSS activity at 48 hpi, deviating from the expected late gene profile. ORF45 encodes a glycoprotein involved in cell-to-cell spread, indicating a potential role in both initial and later stages of infection. ORF54, classified as an early gene, displayed a late peak at 24 hpi, which is unusual for early genes that generally peak within the first few hours of infection. This suggests that ORF54, may require prolonged activity to counter host defenses. ORF67, categorized as a late gene, exhibited an early peak at 2 hpi, followed by fluctuations that are not typical for its kinetic class. Additionally, ORF70, expected to show late kinetics, peaked at 1 hpi and again at 12 hpi, displaying a bimodal pattern not consistent with late gene expression. ORF70 is a capsid protein, and its early expression could be linked to initial assembly processes or other regulatory functions.

These observations held true even when normalizing against the host read counts.

***TES expression kinetics***

TES dynamics are illustrated **in Figure 2** and **Supplementary Figure 2**. Similar to their TSS, early genes such as ORF20, ORF21, ORF30, ORF31, and ORF63, as well as late genes including ORF11, ORF14, ORF22, and ORF73, exhibited consistent peak activities in both TES and TSS dynamics. Early genes peaked within the first few hours post-infection, while late genes reached their maxima between 8 to 12 hpi.

Genes such as ORF32, ORF51, and ORF19 showed consistent anomalous dynamics in both TSS and TES. Specifically, ORF32 and ORF51 peaked earlier than expected for their late classification, and ORF19 peaked much later than typical for its early classification.

We observed several genes where the TSS dynamics differed from the TES dynamics. For example, ORF38, traditionally a late gene, exhibited an early TES peak at 6 hpi, differing from its TSS pattern. ORF50, also expected to follow late kinetics, showed peak TES activity at 6 hpi, which contrasts with its TSS dynamics. ORF45, despite showing anomalous TSS behavior with peaks at 12 and 48 hpi, did not exhibit similar TES anomalies. ORF54, which had a late TSS peak at 24 hpi, displayed consistent TES dynamics with its classification, suggesting that its regulation is tightly regulated. ORF67, showing an early TSS peak at 2 hpi, had TES dynamics consistent with late gene behavior. ORF70, expected to show late kinetics but had bimodal TSS peaks, did not exhibit TES anomalies.

~~The variance between TSS and TES kinetics in some genes can be attributed to the multicistronic nature of herpesvirus transcripts, wherein most RNA molecule contains multiple genes, leading to complex regulatory mechanisms. Additionally, a single TSS can be associated with several TESs, and one TES can serve as the endpoint for multiple TSSs. While we used canonical TSSs and TESs in this part of the study, the differences might also be due to non-canonical ones. Therefore, in the next phase of the work, we conducted an isoform analysis using LRS to differentiate and quantify the isoforms from each gene and their dynamics.~~

***Analysis of the Link Between TSS and TES Sites***

To investigate the 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.

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 TESs, as well as multicistronic and overlapping transcripts. This complexity means that a single TSS can be associated with multiple TESs and vice versa.

For example, in the case of ORF50, the early TES peak at 6 hpi could be contributed to the coterminal TES of ORF50 and ORF51

For example, in the case of ORF38, the early TES peak at 6 hpi suggests that transcripts terminating at this TES are produced earlier than the main TSS activity would indicate. This may be due to transcripts initiating upstream of ORF38 but terminating at its TES, thereby contributing to the early TES activity. Similarly, for ORF50, the peak in TES activity at 6 hpi indicates early termination events that are not directly linked to its main TSS activity.

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.

***Dynamics of Spliced Transcript expression***

The splice sites of EHV-1 transcripts were previously identified in our laboratory using native RNA sequencing. We found splice sites in the following genes: ORF8, ORF9, ORF38, ORF47-44, ORF53, ORF54, ORF58, ORF65, and the NOIR family of non-coding transcripts. **Figure 3** and **Supplementary Figure 3** shows these genes. We found that for several genes there was a difference in the ratio of spliced vs. non-spliced transcripts during the course of the infection. In the case of ORF9 and ORF38, this ratio elevated rather steadily from 0 at 1-4 hpi (all transcripts are non-spliced in the gene) to 30% spliced at 18-24 hpi. The ratio in ORF65 was quite the opposite, it dropped from about 90% spliced to 25%, with a mid-time peak at 6 hpi. The spliced NOIR transcripts first decreased and then, after 12 hpi, started to increase and eventually reached the initial 40% at the end of the experiment (48 hpi).

**Kinetic Profiling of EHV-1 Transcripts**

The kinetic profiles of canonical viral transcripts were analysed using two distinct normalization methods. For one approach, the measurement of a specific EHV-1 transcript at a given time point was assessed relative to the total viral read counts at that same time point (**Figure 4**). In the other approach, normalization was conducted against the total host read counts (**Figure 5**). Additionally, we carried out VST (variance-stabilizing transformation) to normalize the expression levels of Canonical EHV-1 Transcripts Over the Infection (**Supplementary Figure 4**).The total coverage of the viral genome, as calculated from reads whose orientation could be determined, is illustrated in **Supplementary Figure 5**. The total host and viral read count for each sample is presented in **Supplementary Table 3**. We carried out a clustering approach to group together genes with similar expression curves, based on the viral read count-normalized TSS-TES dynamics (**Supplementary Figure 6**). These observations underscore the complexity and potential variability in the transcriptional regulation of EHV-1, suggesting that some genes may have multifaceted roles or be subject to regulatory mechanisms that deviate from the conventional kinetic classifications. The tightly regulated and staggered transcriptional program of EHV-1 orchestrates the sequential expression of its genome to optimize replication and assembly.

**Dynamics of transcriptional overlaps and transcript isoforms**

In this part of the study, we examined the kinetics of transcriptional overlaps between adjacent and distal genes and the transcript isoforms of several genes. The transcripts identified by the LoRTIA program are illustrated in **Figure 6**. We calculated the ratio of each transcript isoform for each gene in the samples separately. The ORF11 showed an increased expression for its 5' truncated isoform (ORF11.5), which then decreased, but after the mid-infection, increased again (**Figure 7**,top left panel), while ORF40 showed a rather steady increase in its alternatively terminated (AT) isoforms (**Figure 7,** bottom left panel). The ORF13 exhibited higher expression levels of its ORF13-14 long variant early on, while the canonical ORF13-14 transcript predominated from 18 hpi (**Figure 7**, top right panel). The ORF14 exhibited the canonical isoform expression between 1-4 hpi, followed by the long variant from 6-18 hpi, after which the canonical isoform became the most abundant again (**Figure 7**,middle left panel). The ORF19 showed the canonical form up to 18 hpi, the 18-19 complex at 24 hpi, and a long variant of The ORF19 at 48 hpi (**Figure 7,** middle right panel). For ORF54, the short version was most prominent between 2-6 hpi, with the canonical form becoming predominant from 8 hpi onwards (**Figure 7,** bottom right panel).

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

The replication origin-associated RNAs (raRNAs) are situated close to the replication origins (Oris) within herpesviruses. Compared to our previous publication [29], CAGE-Seq analysis enabled us to update both CTO-S13, located at Ori-L (**Supplementary Figure 7a**) and identified exclusively in EHV-1 and PRV, as well as NOIR [11], located at OriS (**Supplementary Figure 7b**) 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 [39]. 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 [21,40–42]. Recent studies have demonstrated that the occurrence of nested genes within a larger canonical gene is more prevalent in viruses than previously believed [23,42,43]. 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 [44]. 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 [45]. 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 [46]. 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 [47,48].

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 important problem with lrRNA-Seq techniques is that they utilize bioinformatics methods which are geared towards identifying the most common transcript isoforms, both main and alternative. These methods typically involve clustering of reads, focusing on transcripts with higher expression levels, and use corrections based on existing reference annotations to identify transcripts. However, this approach tends to overlook transcripts that are present in low abundance, which may be dismissed as biological noise despite their potential functional significance. A significant challenge in this area is the lack of software tools that can accurately distinguish between actual RNA molecules and technical artifacts. Consequently, our study focused on annotating the ends of main transcripts but also included information about the TSSs and TESs that are less abundant.

In this work, we carried out the kinetic characterization of the TSSs, TESs, splice sites and the transcripts themselves including both mRNAs and ncRNAs (**Figures 1-5**). In certain genes, we observed differences in TSS and TES kinetics, likely due to the multicistronic nature of herpesvirus transcripts, where each RNA molecule typically contains multiple genes, leading to complex regulatory patterns, with a single TSS connected to several TESs and a single TES associated with multiple TSSs. Additionally, we demonstrated a temporal shift in alternative splicing patterns. The functional significance of these findings remains to be determined.

LRS allows the detection of full-length RNA molecules. Besides the previously known co-oriented (parallel) transcriptional overlaps, we have shown extensive, genome-wide convergent and divergent overlaps (**Figure 6**).

The implications of our findings are significant for understanding the regulation of viral gene expression. The dynamic changes in transcript isoforms and overlaps indicate a complex regulatory network that allows EHV-1 to fine-tune gene expression throughout its replication cycle. The existence of various isoforms per gene implies that the virus is capable of adapting to distinct infection stages by generating diverse transcripts or, at times, protein variants, thus potentially optimizing its genomic utility and reacting more efficiently to the host's environment. 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.

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

**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. 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

39. 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

40. 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

41. 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

42. 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

43. 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

44. 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

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

46. 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

47. 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

48. 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

**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 2a and 2b. 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. Dynamics of Transcript Splice Group Ratios in Spliced EHV-1 Genes Over the Course of Infection**

This figure presents the splicing dynamics within EHV-1 for the genes (a) ORF9, (b) ORF38, (c) ORF65, and (d) noir, wherein spliced transcripts were identified. The transcripts shown on the right side were grouped into spliced and non-spliced groups, and the proportions of these group sums are visualized on the left side. The right side of each panel presents the transcript annotations, both spliced and non-spliced, along with their parent genes and the genomic location displayed below them. In the annotations, light red indicates positive-strand genes, light blue indicates negative-strand genes, dark red indicates positive-strand transcripts, and dark blue indicates negative-strand transcripts. On the left side of each plot, the temporal trends of these transcript groups are depicted, with averages and standard deviations (SD) calculated for each time point post-infection, based on read count data from the dcDNA-Seq. Blue dots and lines represent the sum of spliced and alternatively spliced transcripts, while orange dots and lines represent the summed proportion of non-spliced transcripts for each of the analyzed genes. Each data point is linked by lines to demonstrate the progression over time. 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.

**Figure 4. Kinetic Profiling of Canonical EHV-1 Transcripts Using Total Viral Read Counts for Normalization**

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 5. Temporal Analysis of EHV-1 Canonical Transcripts Normalized by Total Host Reads**

This figure presents the kinetic analysis of canonical EHV-1 transcripts, normalized by total host read counts per sample. The analysis included reads that align accurately with the canonical TSS at the 5' ends and TES at the 3' ends (allowing a deviation of +/- 10 nt for alignment discrepancies) of the gene. The counts of canonical transcripts for each gene were aggregated and then divided by the total host read count for each sample. The mean ratios for the three replicates for each gene are represented as points, and standard deviations (SD) as lines. The x-axis represents time post-infection (hours). The panels are color-coded to indicate the kinetic transcription phases: blue for immediate early (IE), orange for early (E), green for late (L), and red for unknown phases. This facilitates a visual comparison of transcriptional dynamics across the infection cycle, normalized by host reads.

**Figure 6. Transcription overlap between the EHV-1 transcripts identified by the LoRTIA program.** This figure demonstrates that the closely spaced transcripts form substantial transcriptional overlaps.

**Figure 7. Kinetics of Transcript Isoforms for Selected EHV-1 Genes**

This figure illustrates the dynamics of different transcript isoforms for the selected EHV-1 genes (a) ORF11, (b) ORF13, (c) ORF14, (d) ORF19, (e) ORF40, and (f) ORF54. The transcript isoforms, shown on the right side of each panel, 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. On the left side of each plot, these ratios are depicted as a function of time (x-axis: hours post-infection), with averages and standard deviations (SD) shown on the y-axis. Each data point is linked by lines to demonstrate the progression over time. The right side of each panel also includes the corresponding gene and transcript annotations, along with their parent genes and genomic locations displayed below them.

**Legends to Supplementary Tables**

**Table 1**. **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. The Putative mRNAs category includes transcripts that 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. The Non-coding RNAs consist of transcripts that do not contain any ORFs and are presumed to function as regulatory molecules within the viral genome. The Non-coding-short RNAs are shorter variants of non-coding RNAs with truncated transcript lengths, potentially representing distinct regulatory elements. Both Long monocistronic transcripts and Short monocistronic transcripts harbor the same ORF as the canonical transcript but differ in the length of their 5' untranslated regions (UTRs). The "long" monocistronic transcripts have extended 5' UTRs compared to the canonical transcripts, while the "short" monocistronic transcripts have truncated 5' UTRs. Similarly, the Long multicistronic transcripts and Short multicistronic transcripts span multiple genes and differ in their 5' UTR lengths. The "long" multicistronic transcripts have extended 5' UTRs, whereas the "short" multicistronic transcripts have truncated 5' UTRs compared to the canonical multicistronic transcripts. The Antisense-long transcripts are long transcripts that are antisense to known genes, potentially involved in regulatory functions such as gene silencing or modulation of gene expression. The number (N) indicates the count of transcripts in each category.

**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. 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 3.** **Temporal Pattern of Splice Sites of EHV-1 Genes During Infection**

This figure illustrates the splicing dynamics observed in the EHV-1 genes (a) ORF8, (b) ORF54, and (c) ORF58. The transcripts on the right side are categorized into spliced and non-spliced groups, with the proportions of these groups visualized on the left side. Each panel on the right provides the transcript annotations, highlighting both spliced and non-spliced variants, alongside their parent genes and corresponding genomic locations. The annotations are color-coded: light red represents positive-strand genes, light blue represents negative-strand genes, dark red corresponds to positive-strand transcripts, and dark blue to negative-strand transcripts. The left panels depict the temporal trends of these transcript groups, with points representing averages and lines indicating standard deviations (SD) for each time point post-infection. These calculations are based on read count data obtained from dcDNA-Seq. The sum of spliced and alternatively spliced transcripts is shown in blue, while non-spliced transcripts are displayed in orange. The data points are connected by lines to show progression over time. The analysis includes transcripts that precisely match, allowing deviations of +/- 2 nucleotides (nt) for splice junctions and +/- 10 nt for the start and end positions. Asterisks denote statistically significant transcripts, as determined by CAGE-Seq analysis.

**Supplementary Figure 4. Heatmap of VST-Normalized Expression of Canonical EHV-1 Transcripts Across Individual Samples Over Time**

This heatmap presents the VST (variance-stabilizing transformation) normalized expression levels of canonical EHV-1 transcripts across individual samples at various time points post-infection (1h, 2h, 4h, 6h, 8h, 12h, 18h, 24h, 48h). Only reads that overlap both the transcription start site (TSS) and transcription end site (TES) of each gene (based on canonical annotations) were included for analysis. The expression values are color-coded, with red indicating higher expression levels, white representing intermediate expression, and blue representing lower expression levels. Each column corresponds to an individual sample, grouped by hours post-infection (hpi), while the y-axis lists the EHV-1 open reading frames (ORFs). This heatmap provides a detailed visualization of the temporal expression dynamics of each ORF across individual biological replicates, facilitating the identification of temporal expression patterns during EHV-1 infection.

**Supplementary Figure 5. 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) *5,000 counts;* (b) *500 counts;* and (c) *50 limit*counts**.**

**Supplementary Figure 6. Dynamics of canonical transcripts, clustered into kinetic groups**

This figure illustrates the dynamics of each gene, calculated from reads that overlapped both the canonical TSS and TES of the gene. The counts were normalized to the total viral read count for each sample and then subjected to a clustering approach. Each panel represents a different cluster, displaying genes with similar dynamics. The x-axes indicate the sample time points, while the y-axes show the mean abundance values along with the standard deviation.

**Supplementary Figure 7. 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 1. 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 2. Verification of EHV-1 Transcripts Previously Annotated and Discovery of New TSSs through CAGE Sequencing**

**Supplementary Table 3. Read counts**

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