Hybrid Sequencing Reveals Novel Features in the Transcriptomic Organization of Equid Alphaherpesvirus 1

Dóra Tombácz[1#], Gábor Torma[1#], Gábor Gulyás[1#], Ádám Fülöp[1], Ákos Dörmő[1], István Prazsák[1], Zsolt Csabai[1], Máté Mizik[1], Ákos Hornyák[2], Zoltán Zádori[2], Balázs Kakuk[1], Zsolt Boldogkői[1*]

[1] Department of Medical Biology, Faculty of Medicine, University of Szeged, Szeged, Hungary

[2] Institute for Veterinary Medical Research, Centre for Agricultural Research, Budapest, Hungary

# These authors contributed equally

* Corresponding author: boldogkoi.zsolt@med.u-szeged.hu

DT: tombacz.dora@med.u-szeged.hu

GT: torma.gabor@med.u-szeged.hu

GG: gulyas.gabor@med.u-szeged.hu

ÁF: fulop.adam@med.u-szeged.hu

ÁD: dormo.akos@med.u-szeged.hu

IP: prazsak.istvan@med.u-szeged.hu

ZC: csabai.zsolt@med.u-szeged.hu

MM: mizik.mate.levente@med.u-szeged.hu

ÁH: akos1526@gmail.com

ZZ: zadori.zoltan@vmri.hu

BK: kakuk.balazs@med.u-szeged.hu

ZB: boldogkoi.zsolt@med.u-szeged.hu

Keywords: Equid alphaherpesvirus 1, EHV-1, transcriptome, replication origin, long-read sequencing, nanopore sequencing, direct RNA sequencing, Illumina sequencing

Summary

In this study, a structural profiling of equid alphaherpesvirus 1 (EHV-1) transcriptome was carried out using next-generation (Illumina) and third-generation (Oxford Nanopore Technologies) sequencing platforms. We annotated the canonical mRNA molecules and their isoforms, including transcript start and end site isoforms, and splice variants. Additionally, a number of putative 5′-truncated mRNAs containing shorter in-frame ORFs were detected. We also demonstrated that EHV-1 produces a high number of non-coding transcripts, including antisense and intergenic RNAs. One of the most remarkable features of the EHV-1 is the generation of abundant fusion transcripts some of which encoding chimeric polypeptides. We observed a higher number of splicing and transcriptional overlaps than in related viruses. Additionally, we found that many upstream genes of tandem gene clusters have their own transcript end sites (TESs) besides the co-terminal TESs, which is rare in other alphaherpesviruses. We show here that the replication origins (OriS and OriL)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2022 doi:10.20944/preprints202206.0219.v1

of the virus are co-localized with promoter sequences and overlap with specific RNA molecules. Furthermore, we discovered a novel non-coding RNA (designated as NOIR) that overlaps the 5′-ends of the longer transcript variants encoded by the two main transactivator genes ORF64 and 65 bracketing the OriL. These all suggest the existence of a central regulatory system which controls the genome-wide transcription and the replication through a mechanism based on the interference between the machineries carrying out the synthesis of DNA and RNA.

Introduction

Equid alphaherpesvirus 1 (EHV-1, formerly known as equine herpesvirus 1) belongs to the *Varicellovirus* genus of herpesviruses (O'Callaghan et al., 1994; Oladunni et al., 2019). EHV-1 is an important veterinary pathogen causing severe losses in equine industry throughout the world. The most common symptoms of EHV-1 infection are upper respiratory tract disease, abortion in pregnant mares, neonatal death, and fatal myeloencephalopathy (Carroll and Westbury, 1985; Allen and Bryans, 1986; Patel and Heldens, 2005). The virus has an approximately 150 kbp linear double-stranded DNA genome, with 56.7% GC content (Roizmann et al., 1992). The EHV-1genome is composed of two unique regions, the unique short (US) surrounded by a long inverted repeat region (IR), and the unique long (UL) flanked by a short IR (Roizmann et al., 1992). The viral genome contains 80 open reading frames (ORFs) encoding 76 protein-coding genes (four genes are located at the IR region (Telford et al., 1992). EHV-1 codes for 5 genes (ORF1, 2, 67, 71, and 75) of which no homologs can be found in other alphaherpesviruses with annotated genomes (Allen et al., 2004). Similar to other alphaherpesviruses, EHV-1 can productively infect the cells or enter latency in specific sensory nerve cells (Paillot et al., 2008). The virus can infect the susceptible cells through two different mechanisms, including endocytosis or fusion between the host cell membrane and the viral envelope (Frampton et al., 2007). Recognition of target cells by EHV-1 is a receptor-dependent process mediated by the viral glycoproteins gC, gB, and gD (Osterrieder, 1999).

The EHV-1 genes, classified as immediate-early (IE), early (E) and late (L), are expressed in a well-ordered cascade controlled by the viral transcription factors (ORF5, 12, 63, 64, and 65) (Caughman et al., 1985; Kim et al., 2003; Derbigny et al., 2002; Kim et al., 2006). The IE genes of herpesviruses can be transcribed in the absence of *de novo* viral protein synthesis. The ORF64 [homologous to the *ie180* gene of pseudorabies virus (PRV) and *icp4* gene of herpes simplex virus 1 (HSV-1)] is the only EHV-1 IE gene (Smith et al., 1992). Most of the viral E genes encode enzymes that are needed for the DNA replication. The L genes specify the structural proteins of the virion, including capsid and spike proteins. The late genes can be further subdivided into leaky late (L1) and true late (L2) based on whether their expression is dependent on the DNA replication (Gray et al., 1987).

Short-read sequencing (SRS) and long-read sequencing (LRS) platforms have proved to be exceptionally successful in the analysis of the structural aspects of the transcriptomes. The Illumina technique has a high base accuracy and coverage, but due to its short-read-based assembly it is inefficient for the identification of the transcription start sites (TSSs), transcription end sites (TESs), and splice sites (especially of the alternative TSSs, TESs and splice sites), the embedded transcripts, the multigenic RNA molecule and also the parallel transcription overlaps (Boldogkői et al., 2019A). LRS platforms developed by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) (Workman et al., 2019; Marx, 2021), however, are able to determine full-length cDNA or native RNA sequences at the price of a lower throughput and higher sequencing error rates (Laver et al., 2015; Rhoads and Au, 2015; Irimia et al., 2014; Tombácz et al., 2016). LRS is superior to the SRS in the detection of transcript isoforms and multigenic RNA molecules. Although ONT-based approach is hampered by a high error rate, this does not present an important obstacle in transcriptome research in the case of high data coverage and if well-annotated genomes are available.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2022 doi:10.20944/preprints202206.0219.v1

Direct RNA sequencing (dRNA-Seq) has become the golden standard in RNA sequencing (Braspenning et al., 2020) because it is thought to be able to circumvent the production of non-specific reads caused by reverse transcription, second strand synthesis, or PCR amplification. Furthermore, dRNA-Seq preserves read orientation, and allows the detection of RNA modifications (Balázs et al., 2019; Luo and Taylor, 1990; Cocquet et al., 2006). However, dRNA-Seq method is incapable for capturing complete full-length transcripts, since 15-30 bps sequences from the 5′-termini and in many cases also the poly(A)-tails were missing from the reads (Workman et al., 2018; Moldován et al., 2017). Another limitation of the dRNA-Seq technique is its relatively low throughput compared to cDNA sequencing. Additionally, we observed that dRNA-Seq produced certain transcripts, which were undetected by other techniques, and conversely, true transcripts detected by cDNA sequencing were unidentified using dRNA-Seq (Moldován et al., 2020). An integrated approach including SRS, LRS, and also the various library preparation techniques is able to circumvent the above problems, and to provide a highly efficient and reliable method in transcriptome research.

Besides SRS (Oláh et al., 2015), herpesvirus transcriptomes have been analyzed by various LRS techniques, including synthesis-based sequencing (from PacBio) [PRV: Tombácz et al., 2016; Moldován et al., 2018A; Epstein-Barr virus: O'Grady et al., 2016; human cytomegalovirus: Balázs et al., 2017A; HSV-1: Tombácz et al., 2019], nanopore sequencing (from ONT) [varicella-zoster virus (VZV): Prazsák et al., 2018; PRV: Torma et al., 2021) and LoopSeq single-molecule synthetic long-read sequencing (from Loop Genomics) on Illumina platform (Bovine alphaherpesvirus 1 (BoHV-1): Moldován et al., 2020].

EHV-1 transcriptome has already been sequenced using an SRS technique (Zarski et al., 2021) however, in this work, only the transcriptional activity of the genomic regions has been reported without the annotation of viral transcripts. Our objective in this present study was to provide a comprehensive transcriptome annotation of a field isolate EHV-1 using ONT MinION and Illumina MiSeq platforms. We applied amplified cDNA sequencing (Illumina), direct cDNA sequencing (dcDNA-Seq, ONT), as well as direct RNA sequencing (dRNA-Seq, ONT).

Results

Decoding the architecture of the EHV-1 transcriptome using a multi-technique approach

In this study we carried out RNA sequencing using a dual SRS-LRS (Illumina/ONT) approach for profiling the poly(A)+ fraction of the EHV-1 lytic transcriptome. We applied multiple library preparation techniques, including cDNA and native RNA-based methodologies (**Figure 1**). We also prepared Terminator enzyme-based libraries for both dcDNA-Seq and dRNA-Seq approaches. Mapped reads were analyzed for transcript annotation using the LoRTIA pipeline developed in our laboratory (Balázs et al., 2019). We mapped the reads for both the EHV-1 (NC_001491.2) and the host genome (GCF_000003625.3) using the Minimap2 alignment tool with default parameters. Read statistics is available in **Table 1** and in **Supplementary Table S1**.

The criterion for accepting sequence ends as true TSSs or TESs was the presence of at least three independent reads which contained the same 5′-, or 3′-ends, respectively. A splice site was accepted if canonical GT/AG and GC/AG splice junction sequences were present and if the same splice site occurred in at least four independent reads of which at least one was dRNA-Seq result. A sequencing read was accepted as a true transcript if it contained previously annotated TSS and TES. We considered the most abundant transcript generated from a viral gene as the canonical RNA, and as transcript isoforms of those TSS, TES or splice variants which are produced in a lower abundance (**Figure 2, Supplementary Figure S1, Supplementary Table S2**). This rule was not applied to the 5′-truncated transcripts (nested mRNAs). For the annotation of these transcripts, we introduced the following additional criterion: the proportion of such transcripts had to reach at least 5% of the host

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2022 doi:10.20944/preprints202206.0219.v1

mRNAs into which they are embedded. If multiple 5′-truncated transcripts with different in-frame ATGs were encoded within a host gene, they were considered as distinct putative mRNAs (and also genes). Transcripts with the same 5′-truncated ORFs but different TSSs are considered as transcript isoforms.

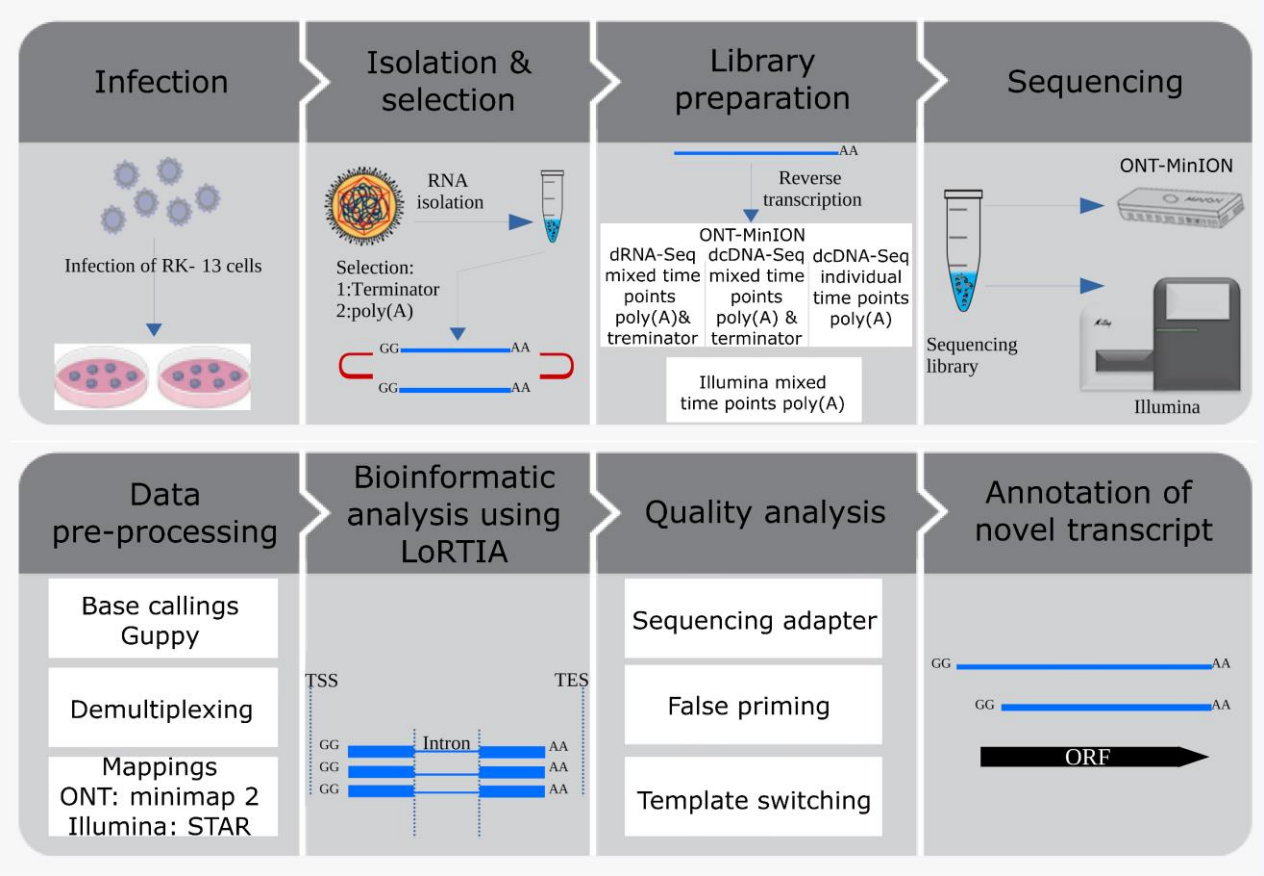

Figure 1. Workflow. *This figure shows the steps of our analysis starting from the infection of RK-13 cells with a field isolate of EHV-1 and ends with the annotation of transcripts.*

The LoRTIA program checks the quality of sequencing adapters and poly(A) sequences and filters out false TSSs, TESs, and splice sites generated by RNA degradation, RT, second strand synthesis, PCR amplification, or false priming in sequencing (Balázs et al., 2019). In order to have higher confidence for the validity of the annotated LoRTIA transcripts, additional stringent filtering criteria were used.

Promoter motifs, poly(A) signals, transcript start and end sites

We identified 84 TATA boxes with an average distance of 30.86 bps from the TSSs, 195 GC boxes with an average distance of 60.35 bps from the TSSs, and 43 CAAT boxes with an average distance of 112.41 bps from the TSSs. We found that the +1 position of sequences containing TATA box are enriched in G bases, while in TATA-less sequences not only the +1, but also the +2 position is GC-rich (**Figure 3**). The GC enrichment in HSV-1 VP5 promoter has already described (Huang et al., 1996). Out of the 99 annotated TESs, 52 have consensus polyadenylation signal (PAS). In accordance with the eukaryotic splice site consensus sequences, we identified A/C cleavage sites and U/G downstream elements at the transcripts containing PASs, while no such consensus sequences were detected in PAS-less transcripts. **Figure 4** illustrates the distribution of TSSs (**Panel A**) and TES (**Panel B**) along the EHV-1 genome. Here, we demonstrate that transcription starts and ends at multiple closely spaced points (typically within a ± 25 bp interval), which are termed as TSS (**Panel C**) or TES (**Panel D**) clusters. A canonical TSS or TES is considered to be the most abundant transcript end. This study detected a high level of a TSS polymorphism in each viral transcript. The long 5′-UTR isoforms of EHV-1 appear to be longer on average and are produced in higher

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2022 doi:10.20944/preprints202206.0219.v1

proportion than in other herpesviruses. In contrast to the poxviruses (Tombácz et al., 2021) and the baculoviruses (Moldován et al., 2018B), herpesvirus transcripts exhibit a low level of TES polymorphism. Our investigations confirmed this phenomenon also in EHV-1: except the case of premature termination of mRNAs and independent TESs of the upstream genes in tandem clusters (see below), the usage of alternative TSSs is also rare in EHV-1.

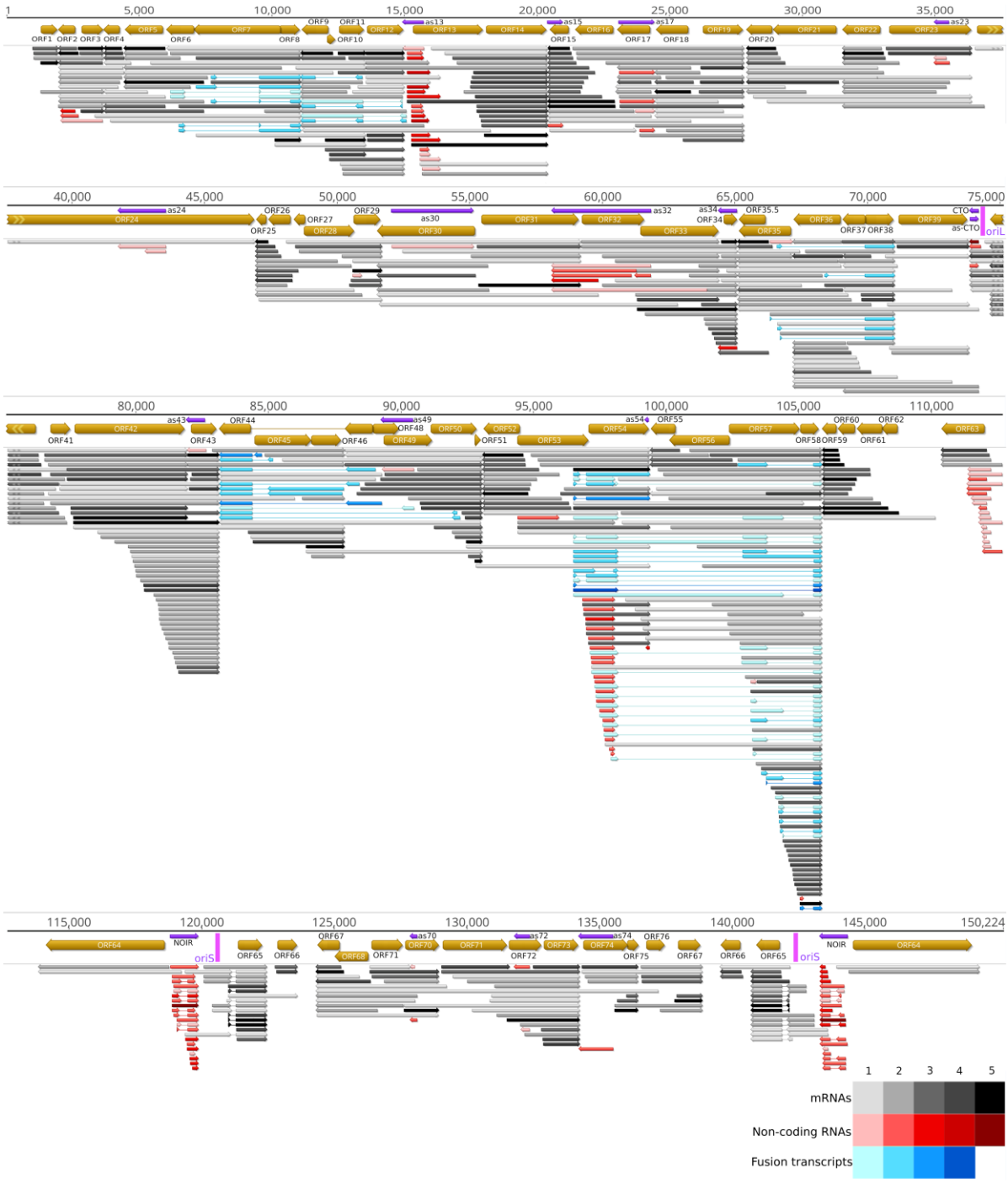

Figure 2. EHV-1 transcriptome. *For the sake of clarity, we removed the 5'-trunced monocistronic transcripts from this figure, but Supplementary Figure S1 contains them. The color code is defined in the figure. The shade of the colors corresponds to the abundance: 1: 1-9 reads, 2: 10-49 reads, 3: 50-199 reads, 4: 200-999 reads, 5: >1000 reads.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2022 doi:10.20944/preprints202206.0219.v1

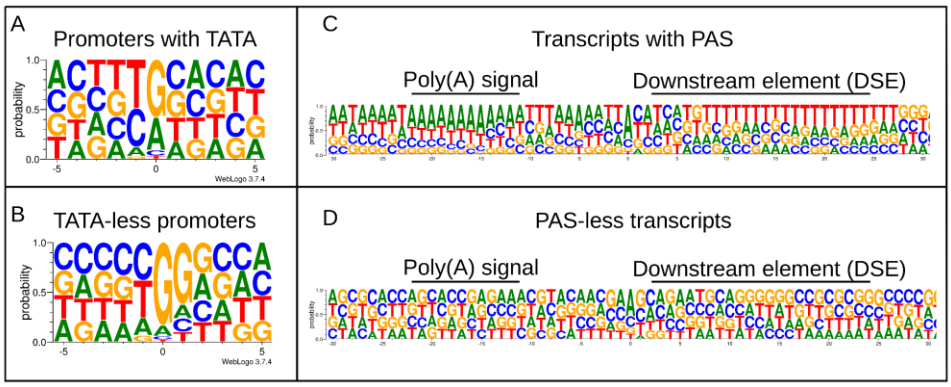

Figure 3. TATA boxes and poly(A) signals (A) *Genomic surrounding of TSSs with TATA box within a ± 5 bp interval. The first letter of TSSs (position 0) is enriched with G/A bases, while the − 1 position contains mainly C/T bases.* ***(B)*** *A) Genomic surrounding of TSSs without TATA box within a ± 5 bp interval. The 0 and +1 TSS positions are enriched with G letters* ***(C)*** *Sequence motifs of transcripts containing polyadenylation signals.* ***(D)*** *Sequence motifs of transcripts without polyadenylation signals.*

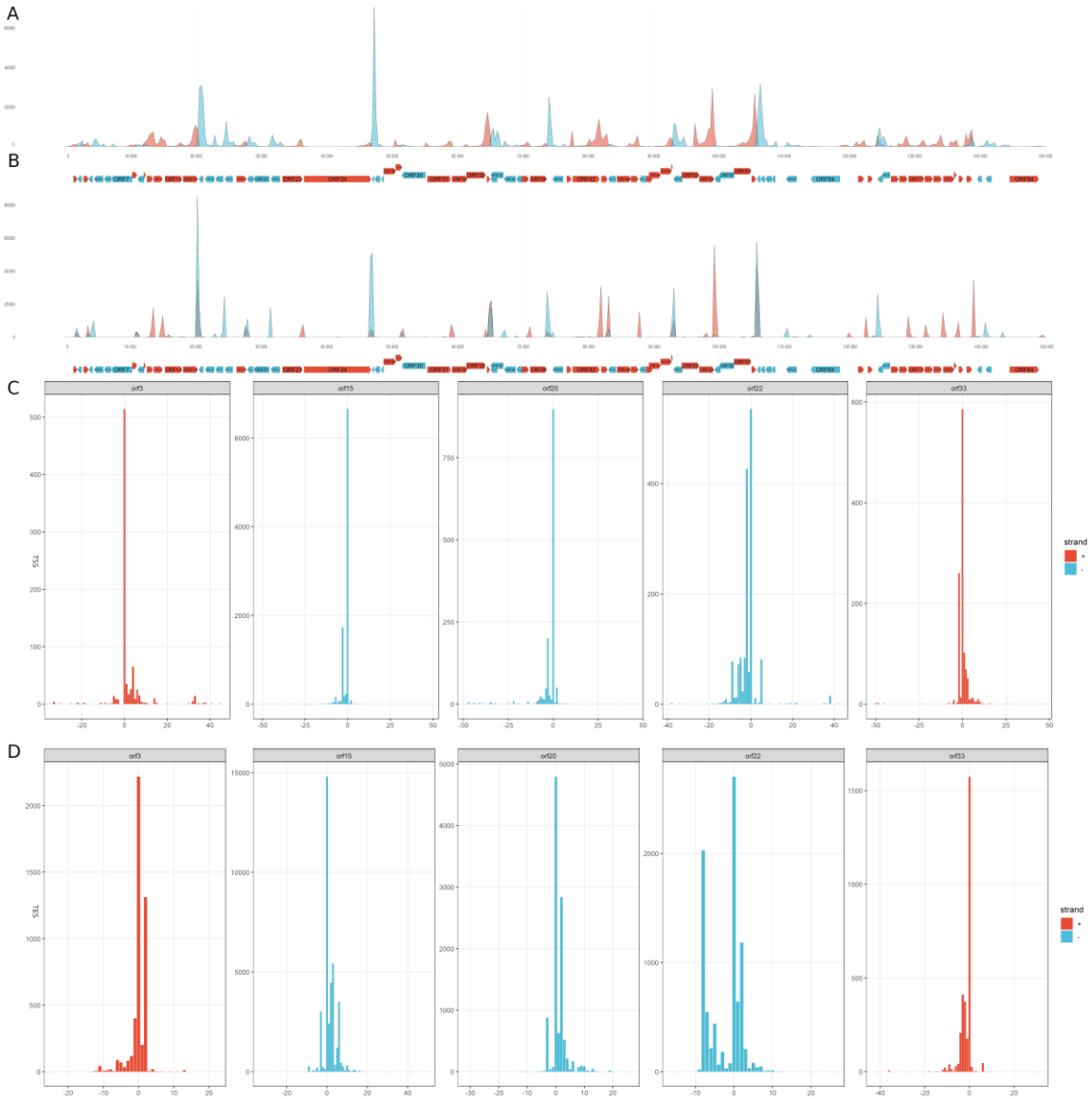

Figure 4. Transcription start and end sites (a) G*enome-wide localization of TSSs. The relative amount of the TSSs is calculated from a mixed time-point sample. (b) Genome-wide localization of TESs. The relative amount of the TSSs is calculated from a mixed time-point sample. (c) TSS clusters illustrated by 5 examples (d) TES clusters illustrated by 5 examples*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2022 doi:10.20944/preprints202206.0219.v1

Canonical mRNAs

In this part of the work we annotated the canonical transcripts by identifying their TSS and TES clusters and splice sites (**Supplementary Table S2**). The canonical transcripts were defined as the highest abundance RNA isoform specified by a given protein-coding or a non-coding gene. Transcripts containing 5′-truncated ORFs were not considered in this calculation because the short RNA molecules are overrepresented due to the preference of LRS toward the sequences falling into this size range. We were able to identify canonical transcripts for every EHV-1 gene.

Putative nested genes

Long-read sequencing allows the clear distinction between the larger host and the smaller embedded transcripts. In this part of the study we identified putative 5′-truncated mRNAs that lack certain upstream parts of the gene including the canonical ATG, but contain one or more downstream in-frame ATGs (**Supplementary Figure S1, Supplementary Table S2**). The shorter ORFs, if translated, would encode N-terminal truncated polypeptides. We term these ORFs as 'in-frame ORFs' (ifORFs), the carrying gene as putative nested gene and the transcripts as putative nested mRNAs. The nested mRNAs are also polymorphic regarding the length of their 5′-untranslated regions (UTRs), which are the coding part of the larger host genes. We use the name 'ifORF' if the truncated in-frame ORF was detected within monocistronic genes. However, for ifORFs located within the 5′-UTR of RNAs encoded by the downstream genes, it is difficult to decide whether these ifORF-containing transcripts are just long 5′-UTR isoforms of the downstream gene, or if they are translated. It is an important issue, since in the latter case the downstream gene would not be translated. Whether these 5′-truncated ORFs have true coding capacity and are all biological products remains to be determined.

Non-coding transcripts

The non-coding RNA (ncRNA) molecules include those transcripts which do not contain functional ORFs (**Supplementary Table S2**). Most of the detected EHV-1 non-coding transcripts are long ncRNAs (lncRNAs) which, by definition, are made of more than 200 nucleotides. We also identified short ncRNAs (sncRNAs) although our approach is not optimal for the detection of this transcript type, especially in the case of molecules comprised of less than 50 nucleotides, such as microRNAs. The non-coding transcripts have their own promoters and are located in either intragenic, or in intergenic positions, or overlap the mRNAs in antiparallel manner.

Ten canonical antisense RNAs (asRNA) and their isoforms (altogether 27 asRNAs) were detected in this work. While asRNAs are controlled by their own promoters, antisense segments can also be part of mRNAs as a result of convergent or divergent transcriptional overlaps between adjacent or even distal antiparallel genes. We identified as70 transcript overlapping the ORF70, which is a homolog of PRV US4-AS, but we did not detect the PRV AZURE transcript that runs opposite to the *us3* (homolog of ORF69) and *us4* (homolog of ORF70) genes (Torma et al., 2021). EHV-1 expresses a higher number of asRNAs than other alphaherpesviruses including PRV, its close relative. Most of the identified asRNAs have not been detected in other alphaherpesviruses. Due to the extensive transcriptional overlaps, practically, both DNA strands are transcriptionally active throughout the entire viral genome. Intriguingly, Coding Potential Calculator (CPC2; Kang et al., 2017) analysis gave the result that the small (average: 141.44 bp) ORFs of 9 asRNAs resemble to the coding sequences of the vertebrate organisms, therefore they might have coding potential (**Supplementary Table S3**).

We detected two very abundant groups of intergenic non-coding transcript, the NOIR and the CTO-S (the latter is discussed in the next section). Both transcripts have homologs in PRV (NOIR-1 and CTO-S, respectively), but not in other herpesviruses with annotated transcriptomes. Possibly, the NTO2-4 transcripts of VZV described by our research group (Prazsák et al., 2018) have a similar function as the NOIR transcripts, but in contrast to the NOIR, they are located within the canonical

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2022 doi:10.20944/preprints202206.0219.v1

ORF62 (*icp4* homolog). While in PRV, the NOIR-1 has a single splice variant, the EHV-1 homolog has two splice isoforms besides the unspliced RNA (**Figure 5A**). The function of this RNA gene is completely unknown. No homolog of the low-abundance PRV *noir-2* non-coding gene was detected in the EHV-1 genome.

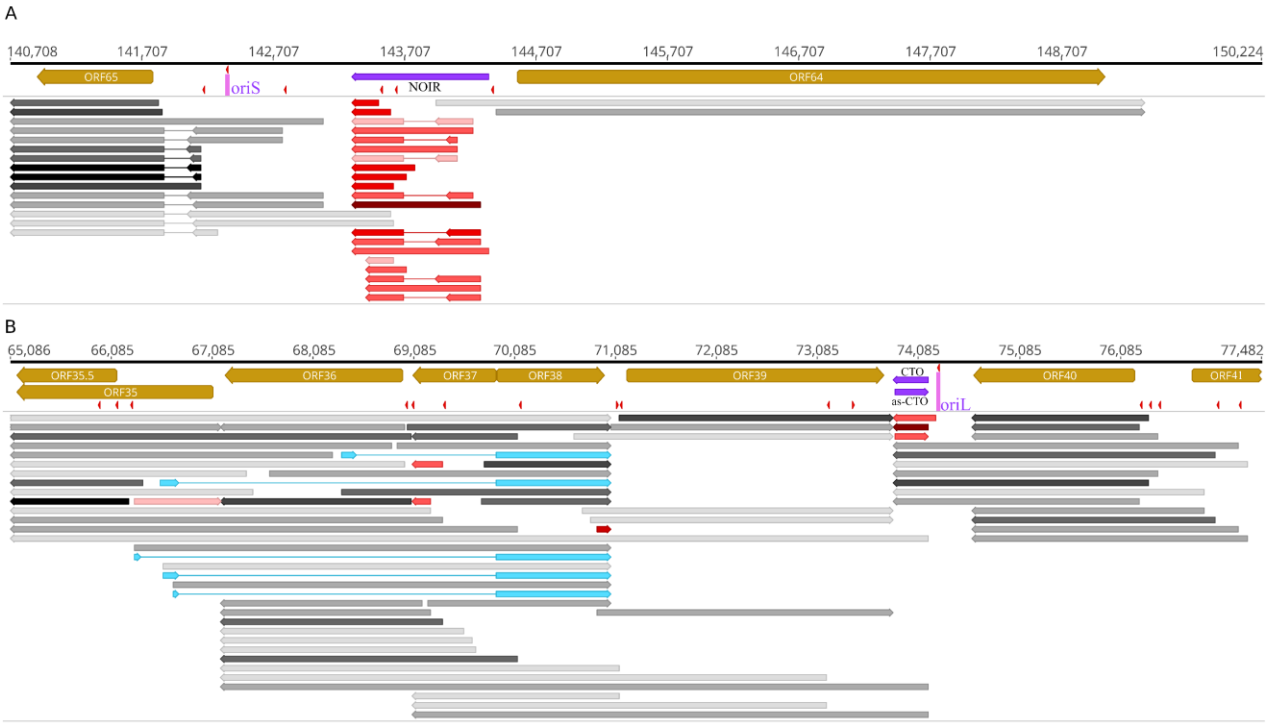

Figure 5. Transcription near the replication origins

A. OriS: ORF64-65 region
B. OriL: ORF35-41 region

The color code is defined in the figure. The shade of the colors corresponds to the abundance: 1: 1-9 reads, 2: 10-49 reads, 3: 50-199 reads, 4: 200-999 reads, 5: >1000 reads.

One type of intragenic ncRNAs are those ones which share their promoters with the mRNAs but lack the STOP codon due to the premature transcription termination. These transcripts are designated as 'non-coding start' (**ncs**). Similar to the mRNAs, the 'ncs' transcripts have the same alternative TSSs. One of the specialties of EHV-1 transcriptome is the presence of abundant 'ncs' transcripts, e.g., ORF13-ncs, ORF53-ncs, and ORF63-ncs. Additionally, intragenic ncRNAs without functional ORFs can be the result of 5′-truncation of mRNAs. This type of ncRNAs is termed as 'non-coding coterminal' (**nct**). Some of the monocistronic 5′-truncated transcripts discussed in the previous section may also be 'nct' transcripts. Since the frequency of false TSSs in short reads are higher than in longer reads (Moldován et al., 2020), we accepted an 'nct' transcript as true if its abundance reached the 5% of the canonical mRNA into which it is embedded. The 'non-coding start and stop' (**ncss**) transcripts are also intragenic but lack both the TSS and TES of the host mRNA. An example of this transcript type is the ORF54-ncss.

Replication origin-associated transcripts

Replication origin-associated RNAs (raRNAs) are mapped near the genomic location of the replication origins (Oris). These transcripts have been identified in every examined virus, including alphaherpesviruses (Boldogkői et al., 2019B). Many raRNAs overlap the Oris, while others are terminated in their close vicinity (Boldogkői et al., 2019B). In herpesviruses these transcripts can be non-coding or the longer TSS or TES variants of mRNAs are encoded by one of the neighboring genes or both (Prazsák et al., 2018). Similar to PRV, EHV-1 contains a single OriL at the unique long region and two OriS at the repeat region. Like in other alphaherpesviruses, the raRNAs

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2022 doi:10.20944/preprints202206.0219.v1

overlapping the OriS are the long 5′-UTR isoforms of ORF65 homologous to the *us1 (icp22)* gene of alphaherpesviruses (**Figure 5A**). EHV-1 encodes the very abundant PRV homolog, the CTO-S transcripts from an RNA gene located near the OriL (**Figure 5B**). The CTO-S transcript has very long TES isoforms being co-terminal with the ORF35, 36 and 37 genes. This transcript is a complex RNA (cxRNA) molecule because it contains genes (ORF38 and 39) with antiparallel orientation, and it is likely ncRNAs because its first ATG (of ORF 37) is too far from its TSS (**Figure 5B**). A transcript, antiparallel to CTO-S, was also detected. Furthermore, the *ul21* homolog of EHV-1 (ORF40) codes for a TES isoform (also termed as CTO-L), which is co-terminal with the canonical CTO-S. However, we could not detect the PRV homolog CTO-M for which the reason may be the relatively low data coverage at this region. Intriguingly, the TATA box of the longer CTO-S isoform is co-localized with the OriL. Likewise, we detected a TATA box within the OriS and identified the transcript which is likely to be controlled by this promoter element. It is possible that the NOIR transcripts have also a direct or indirect role in the regulation of replication (see Discussion for explanation).

Multigenic transcripts

Multigenic transcripts contain two or more genes on an RNA molecule.

Polycistronic transcripts

A characteristic feature of the organization of herpesvirus genomes is that tandem genes are transcribed as polycistronic transcripts in the following pattern: ‘abcd’, ‘bcd’, ‘cd’, ‘d’, where ‘a’ is the most upstream and ‘d’ is the most downstream gene. However, in contrast to the prokaryotic polycistronic RNAs, in herpesviruses only the most upstream gene is translated. If an ifORF is located in the most upstream gene of a multigenic transcript, functional analysis is needed to determine whether this transcript is only a long TSS variant, or the 5′-truncated ORF is translated. One of the major findings of this study is that in EHV-1 many upstream genes of the tandem gene clusters have their own TESs besides the co-terminal TESs.

Complex transcripts

CxRNA molecules contain multiple genes in which at least one gene is oriented in antiparallel direction. Those transcripts in which the most upstream gene stands in antiparallel orientation are probably non-coding because of the long distance between the TSSs and the canonical ATGs. Despite this possibility we labeled them as coding transcripts with very long 5′-UTR in **Figure 2**. Another distinctive feature of EHV-1 is the common occurrence of relatively abundant very long cxRNA molecules. Altogether, we identified 81 cxRNAs, which is obviously an underestimation of the real number.

Splicing, splice isoforms and fusion transcripts

We used very stringent criteria for the identification of splice sites: besides their incidence in at least three distinct samples prepared by different techniques, we also requested the presence of splice consensus sites and the detection by dRNA-Seq. Alphaherpesviruses produce much less spliced transcripts and a lower variety of splice isoforms than other herpesviruses. However, in EHV-1, we detected complex splicing patterns even in those transcripts which are unspliced in the related viruses (**Figure 6**). The most intriguing spliced transcripts are the fusion RNAs (fRNAs). One type of fRNAs utilizes the genomic segments from as 5′-UTR of adjacent or more distal genes standing in an opposite direction (e.g., ORF8 in **Figure 6A**). The pre-mRNAs of these transcripts are complex transcripts. ORF8 (*ul51*) utilizes some parts of ORF6 and ORF7 as 5′-UTR in various combinations. Intriguingly, a truncated coding sequence of ORF8 is also produced. A similar complex splicing pattern is also observable in ORF9 (*ul50*). The pre-mRNAs of the fusion transcripts expressed from the ORF35-38 region (**Figure 6B**) are cxRNAs because the first three genes stand in an antiparallel orientation relative to the ORF38. These transcripts contain the entire coding region of ORF38 and various 5′-UTR segments from the ORF35 and ORF36 genes. The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2022 doi:10.20944/preprints202206.0219.v1

ORF44 (*ul15*) gene is encoded in a special way in alphaherpesviruses: the continuity of its ORF is disrupted by two other genes (ORF45/*ul17* and ORF46/*ul16*), which are spliced out from the mature ORF44 transcripts. The downstream exon is also expressed independently (**Figure 6C**). The EHV-1 ORF44 is encoded an even more intricate manner: many of these transcripts contain a much longer intron which encompass the entire ORF48 and 49 genes and a large part of ORF50 gene. The ORF53-58 (*ul9-4*) region exhibits the highest complexity (**Figure 6D**). In this region genes produce fusion proteins in various combinations. This genomic segment can also be used for exemplifying the 5′-truncated ORFs, the intragenic ncRNAs, the complex transcripts, and also the independent termination of the upstream members of tandem gene clusters. The fusion transcripts and the chimeric proteins are illustrated in **Figure 6E**. Many fusion results in-frame chimeric protein molecules, but in some cases the second or third exon is not at the same reading frame as the upstream exon. However, in this latter case, a close stop codon is located in the new reading frame.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2022 doi:10.20944/preprints202206.0219.v1

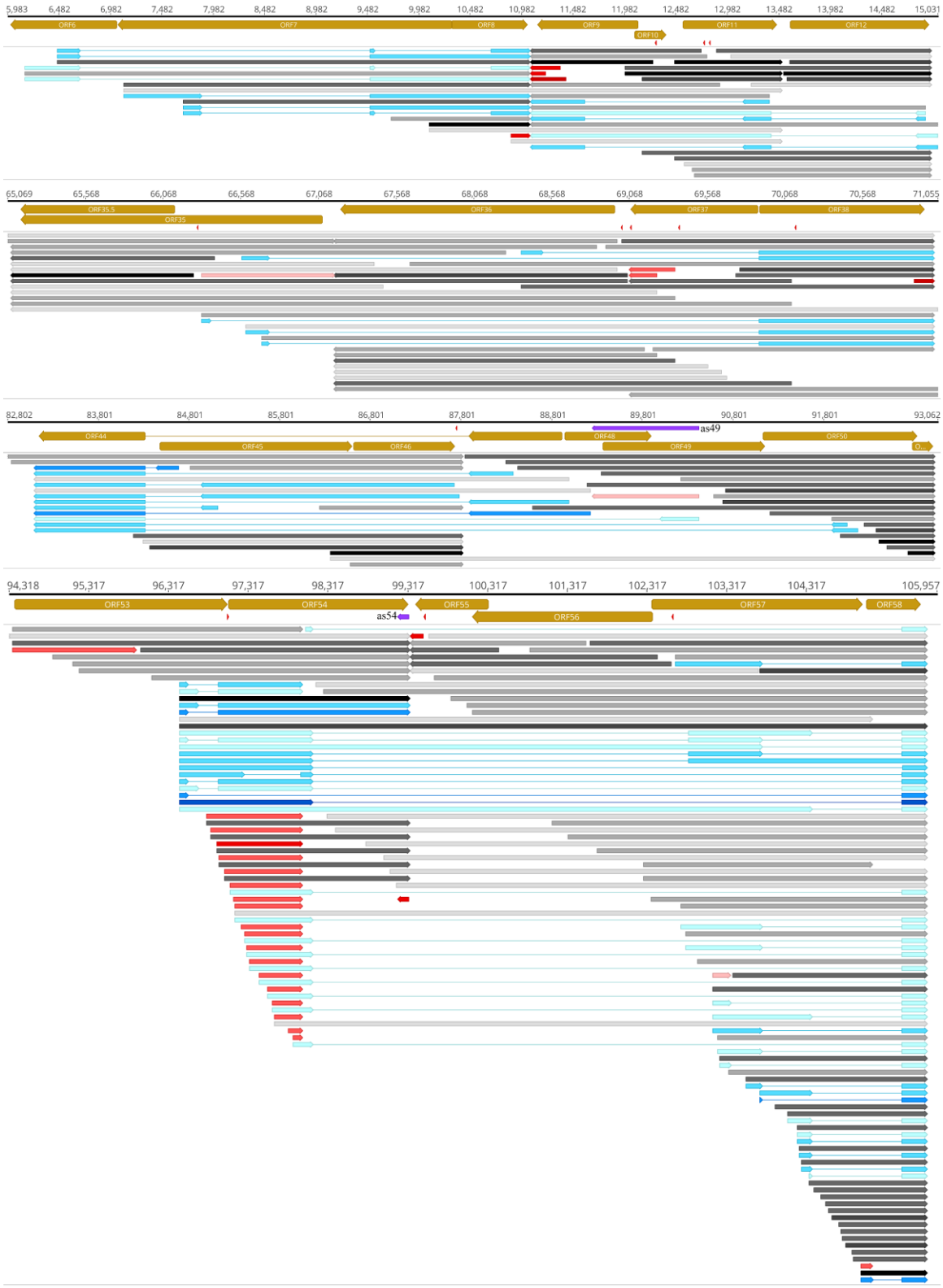

Figure 6. Splicing and fusion transcripts

A. ORF6-12
B. ORF35-38
C. ORF44-50
D. ORF53-58
E. Fusion transcripts and chimeric proteins

The color code is defined in the figure. The shade of the colors corresponds to the abundance: 1: 1-9 reads, 2: 10-49 reads, 3: 50-199 reads, 4: 200-999 reads, 5: >1000 reads.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2022 doi:10.20944/preprints202206.0219.v1

Transcriptional overlaps

The orientation of gene pairs can be parallel (→→), convergent (→←), or divergent (←→). The transcriptional overlaps of parallel and convergent partner genes are produced by transcriptional readthrough, whereas divergent overlaps are generated by the overlap of 5′-UTRs of divergently oriented gene products. **Supplementary Figure S3** shows that practically every divergent gene pair produces transcripts with extensive head-to-head overlaps. Another characteristic feature of the EHV-1 transcriptome is the occurrence of very long transcriptional overlaps, encompassing several genes. The canonical transcripts encoded by convergent gene pairs do not overlap each other, but they occasionally produce transcriptional read-throughs ('soft' overlap). The ORF 29/30 (*ul31-30*) gene pair is an exception to this rule because their canonical transcripts overlap each other ('hard' overlap), which is similar to other alphaherpesviruses.

Discussion

The past decade has seen a rapid progress of sequencing technologies. Third-generation LRS approaches led to a paradigm shift in genome and transcriptome research, especially in small-genome organisms (Tombácz et al., 2014; Depledge and Breuer, 2021). The transcriptomic architecture of viruses proved to be much more complex than previously expected (Boldogkői et al., 2019A). A large variety of overlapping transcripts, including long 5′-UTR isoforms, polycistronic and complex transcripts, 5′-and 3′-truncated mRNAs with in-frame ORFs, and read-through transcripts have been discovered (Balázs et al., 2017B; Prazsák et al., 2018; Tombácz et al., 2020; Torma et al., 2021).

It has recently been shown that nested genes embedded into a larger canonical gene represent a more common phenomenon in viruses than thought earlier (Moldován et al., 2018A; Tombácz et al., 2018, Torma et al., 2021). The SRS technique is inefficient in identifying the nested mRNAs, that is why they had gone undetected before. Ribosome profiling and other genome-wide translational analyses are needed to perform for confirming the translation of this transcript type.

Although an increasing number of lncRNAs have already been discovered, they remain poorly characterized mechanistically. These RNAs were previously regarded as 'transcriptional noise' owing to lack of protein-coding capacity, emerging evidence indicates that they play a wide variety of functions acting through distinct pathways (Statello et al., 2021). In this work, we identified a variety of non-coding RNAs, including intra- and intergenic transcripts and antisense RNAs. The asRNAs are encoded by the complementary DNA strands of protein-coding genes and are controlled by their own promoters. However, mono- and polycistronic mRNAs and complex transcripts can also have antisense parts that are generated by either transcriptional read-through events between convergent gene pairs, or by head-to-head overlaps between divergent neighboring or distal genes. We identified 27 asRNA clustered in 10 groups, which is more than described in the related viruses.

It has been shown that 72% of mammalian replication origin-associated transcripts are controlled by active promoters specifying the expression of protein coding or non-coding genes (Dellino et al., 2013). Similar raRNA molecules have also been detected in viruses (Boldogkői et al., 2019B). It has been shown in human BK polyomavirus that an raRNA specimen significantly inhibits the replication of the virus through interfering with the RNA primer synthesis by binding simultaneously to both sense and antisense DNA strands within the Ori region (Tikhanovich et al., 2011). These Ori-associated transcripts undergo rapid evolution even within alphaherpesviruses. While the location of OriS is conserved in alphaherpesviruses (between the *icp4* and *us1* genes), OriL is either missing (in VZV and BoHV-1) or are mapped to different genomic locations: between *ul29* and *ul30* genes in HSV-1, and between *ul21* and *ul22* genes in PRV and EHV-1. The raRNAs of OriS include transcripts that are the long TSS isoforms of US1 transcripts, which overlap the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2022 doi:10.20944/preprints202206.0219.v1

origin of replication, or are initiated closely to it (in all alphaherpesviruses). The raRNAs of OriL include the long TES isoform of UL21 transcripts, and the CTO-S ncRNAs mapped downstream of the *ul21* gene (in PRV and EHV-1). We observed that promoter elements of viral transactivator genes are co-localized with both OriS and OriL (such as in mammalian cells: Dellino et al., 2013), which suggests a co-regulation of the initiation of transcription and replication. The ori-overlapping RNAs may regulate the later phases of replication. The precise function of raRNAs of alphaherpesviruses, however, remains to be ascertained.

The EHV-1 NOIR represents an intriguing group of ncRNAs. Its homolog (NOIR-1) has been described in PRV (Tombácz et al., 2016), but no convergent partner (NOIR-2 in PRV) was detected in EHV-1. The canonical form of these EHV-1 transcripts overlaps both the long 5′-UTR isoforms of ORF64 (*rs1/icp4* of HSV-1) and ORF65 (*us1/icp22* of HSV-1). We can speculate that the expression and/or the transcripts themselves might affect the activity of these transcription factor genes, which, if so, it would have a role in the regulation of genome-wide gene expression and also in the DNA replication. In other words, the *icp4-us1* genomic region of alphaherpesviruses might be the center for the viral regulatory mechanisms: the viral transcription factors and the ncRNAs at this locus might coordinate the onset and progression of both the replication and the global gene expression by physical interactions of their apparatuses, including collision, as well as competition for the promoters and Oris. The raRNAs are very likely not only by-products of an interference-based mechanism, but they also have function through e.g., forming a DNA-RNA hybrid at the Ori region (Tai-Schmiedel et al., 2020).

Our results show that splicing events in EHV-1 are more frequent than in the related alphaherpesviruses. We detected relatively abundant fusion transcripts, some of which encode chimeric proteins in various exon combinations. Other fusion transcripts are the results of the combination of 5′-UTR sequences of one or more upstream genes with complete or 5′-truncated form of ORF of one or more downstream genes. The 5′-UTR sequences of the fusion transcripts in many cases are derived from the antiparallel strand of upstream genes. Low-abundance fusion genes in alphaherpesviruses have also been described by others (Braspenning et al., 2021).

Polycistronism is widespread in bacteria and in viruses, but it is extremely rare in eukaryotic organisms. In prokaryotes and bacteriophages, the Shine-Dalgarno sequence (ribosomal binding site in the mRNAs) allows the translation of every gene on the polycistronic RNA molecules. Many small-genome eukaryotic viruses have evolved a variety of mechanisms to tackle this problem, which include the use of internal ribosome entry sites, ribosomal frameshifting, or leaky ribosomal scanning (Stacey et al., 2000). Co-oriented herpesvirus genes tend to be organized into gene clusters expressing transcripts with common downstream sequences and distinct 5′-exons in the following pattern: ‘abcd’, ‘bcd’, ‘cd’ and ‘d’, where ‘a’ is the most upstream and ‘d’ is the most downstream gene. The function of multigenic transcripts in large DNA viruses has not yet been ascertained because with a few exceptions [uORFs are translated addition to the canonical ORFs (Vilela et al., 2003; Kronstad et al., 2013)], no translation has been described from the downstream genes.

Polycistronic EHV-1 RNAs encoded by tandem genes represent parallel overlaps. In other alphaherpesviruses, the majority of upstream genes of tandem gene clusters do not produce monocistronic transcripts, or if so, these RNA molecules are expressed in very low abundance. We found a more extensive use of alternative TESs by the upstream genes of EHV-1. Similar to PRV and HSV-1, we detected a ‘hard’ overlap between the ORF 29/30 (*ul31/30*), but not between ORF54/55 [such as in PRV: ul8/7 (Tombácz et al., 2016)] or ORF58/60 [such as in HSV-1: ul4/3 (Tombácz et al., 2017)]. We did not observe an increased extent of convergent overlaps and readthroughs, however, the divergent overlaps were found to be more extensive in EHV-1 than in other alphaherpesviruses. This genomic organization suggests the existence of genome-wide transcriptional interference, caused by the collision and/or competition of the transcription apparatuses, which represent a novel level of genetic regulation (Boldogkői et al., 2012). The role

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2022 doi:10.20944/preprints202206.0219.v1

of genomic context in the regulation of gene expression has recently been described in yeast (Gilet et al., 2020; Brooks et al., 2022).

Limitations of the study

One of the limitations of this study is that although the sequencing reads cover the entire EHV-1 genome, at certain loci the coverage is in insufficient for the precise annotation of the given genomic segment. However, it is not a critical problem in transcript identification. Another limitation of our approach is that the applied method is not optimal for the identification of sncRNAs (especially of microRNAs), and also of very long lncRNAs, therefore it does not provide a complete atlas of EHV-1 transcriptome. Additionally, some of the low-abundance transcripts may have gone undetected due to the given level of read coverage. Finally, LRS is biased toward 200-600 bp transcription reads, which therefore produces relatively large read coverage at this size range. Although, LoRTIA software suit filters out false transcripts, we cannot exclude that at this size range some of the identified TSSs and transcripts are non-biological but represent mere technical artifacts. The category of nested genes exists since many such RNA molecules encoded by them have already been detected. Each novel putative embedded mRNA has to be individually analyzed.

Materials and Methods

Cells and viruses

In this study we used a field isolate equid alphaherpesvirus 1 (EHV-1) strain MdBio (EHV-1-MdBio), which was isolated from the organs of an aborted colt fetus in the 1980's at Marócpuszta (Hungary). **Supplementary Figure S2** presents the phylogenetic tree of EHV-1 strains.

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 (Gibco) and 80 µg of gentamycin per ml (Gibco) at 37 °C in the presence of 5% CO_2. 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 the 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 various length of time. After the incubation, the culture medium was removed, and the infected cells were frozen at -80 °C until further use.

RNA isolation

RNA was extracted from the cells following the spin column-based protocol of the NucleoSpin RNA kit (Macherey-Nagel). First, the cells were incubated in a lysis buffer containing chaotropic ions, which inactivate RNases. DNA and RNA molecules were then bound to the silica membrane. DNase I was added to all samples to eliminate residual genomic (g)DNA contaminations. Total RNAs were eluted in nuclease-free water. The potential remaining gDNA contamination was removed with the use of TURBO DNA-free™ Kit (Invitrogen). The concentration of the samples was measured by using Qubit 4.0 fluorometer (Invitrogen) and Qubit Broad Range RNA Assay Kit (Invitrogen), while the Agilent TapeStation 4150 was used for the quality control. RIN scores ≥ 9.2 were used for cDNA production for further processes.

Purification of polyadenylated RNA

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2022 doi:10.20944/preprints202206.0219.v1

The Oligotex mRNA Mini Kit (Qiagen) was used to isolate the polyadenylated [poly(A)+] RNA fraction from the total RNA samples. In brief, RNase-free water was added to the samples to set their volume to 250 µL. Fifteen µL Oligotex suspension and 250 µL OBB buffer (both from the Qiagen kit) were added to the samples which then were heated to 70°C and incubated for 3 min followed by cooling down to 25 °C for 10 min. Next, the mixtures were centrifuged at 14,000×g for 2 min, then the supernatants were discarded. OW2 wash buffer (400 µL, from the Oligotex kit) was added to the samples, then the solution was loaded onto the spin columns from the Qiagen kit. Samples were span down at 14,000×g for 1 min. This washing step was repeated once, and finally, the polyadenylated RNA fraction was eluted from the membrane by adding 50 µl hot elution buffer (EB, Qiagen kit). RNA was eluted in 60 µl EB, then a second elution step was also carried out in order to maximize the yield.

rRNA removal

As an alternative to poly(A) purification, Ribo-Zero Magnetic Kit H/M/R (Epicentre/Illumina) was applied to enrich the mRNAs in the samples. In contrast to poly(A) purification, rRNA depletion eliminates the ribosomal rRNAs from the sample without the removal of the potential non-polyadenylated long non-coding RNAs and mRNAs. Five µg of a mixture of total RNA was used as starting material and was mixed with Ribo-Zero Reaction Buffer and Ribo-Zero rRNA Removal Solution. The sample was incubated at 68°C for 10 min, then at room temperature for 5 min, and finally it was added to the washed Magnetic Bead (225 µl, from the kit). After a short vortexing and incubation at room temperature for 5 min, the sample was heated to 50°C for 5 min. Mixture was placed on a magnetic stand, and the supernatant containing the rRNA-depleted RNA was collected and purified by following the AMPure XP Bead washing method.

Treatment with Terminator enzyme

A mixture of poly(A)+ RNA samples was treated with Terminator™ 5′-Phosphate-Dependent Exonuclease (Lucigen). This enzyme digests RNAs with 5′-monophosphate ends but not RNAs with 5′-triphosphate, 5′-cap, or 5′-hydroxyl groups at the 5′-end, therefore it helps in the enrichment of the 5′-ends of mRNAs. In short, the process was carried out with the addition of Terminator 10X Reaction Buffer A, RiboGuard RNase Inhibitor and Terminator Exonuclease (1 Unit) to the RNA mixture. Next, the sample was incubated at 30°C for 60 min, then the enzymatic reaction was terminated by the addition of 1 µL of 100 mM EDTA (pH 8.0). Finally, magnetic bead-based purification of the mRNA samples was carried out by using the Agencourt RNAClean XP beads (Beckman Coulter).

Illumina MiSeq short-read sequencing

The entire EHV-1 transcriptome was sequenced using a short-read sequencing approach. For this, the libraries were prepared from a mixture of poly(A)+ and rRNA-depleted samples using the NEXTflex® Rapid Directional qRNA-Seq Kit (PerkinElmer). The RNA fragmentation was carried enzymatically, using NEXTflex® RNA Fragmentation Buffer at 95°C for 10 min. This step was followed by the synthesis of the first cDNA strand with the addition of NEXTflex® First Strand Synthesis Primer. The mixture was incubated at 65°C for 5 min and then subsequently placed on ice. Reverse transcription was performed using NEXTflex® Directional First Strand Synthesis Buffer and Rapid Reverse Transcriptase enzyme and the following conditions: incubation at 25°C for 10 min, then at 50°C for 50 min, and termination at 72°C for 15 min. The synthesis of the second cDNA strand was performed at 16°C for 60 min via the addition of NEXTflex® Directional Second Strand Synthesis Mix (with dUTPs). The obtained double-stranded cDNAs were polyadenylated using the NEXTflex® Adenylation Mix. The adenylation reaction was performed at 37°C for 30 min, then it was terminated by incubating the samples at 70°C for 5 min. Molecular Index Adapters (part of the Kit) were ligated to the sample using the NEXTflex® Ligation Mix, which was performed at 30°C (10 min). Next, the cDNAs were amplified by PCR: for this, first, the NEXTflex® Uracil DNA Glycosylase was added to the samples, and the mixtures were incubated

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2022 doi:10.20944/preprints202206.0219.v1

at 37°C for 30 min, followed by heating at 98°C for 2 min, and were subsequently placed on ice. The following components were added to the mixtures: PCR Master Mix, qRNA-Seq Universal forward primer, and qRNA-Seq Barcoded Primer (sequence: AACGCCAT; all from the kit). The applied PCR protocol is summarized in **Supplementary Table S4.** AMPure XP Bead (Beckman Coulter) was used after each enzymatic step. Resuspension buffer (NEXTflex® Kit) was used for the final elution. Ten pM from the library mix was loaded to the reagent cassette and paired-end transcriptome sequencing was performed by MiSeq Reagent Kit v2 (300 cycles).

Qubit 4.0 fluorometer and the Qubit dsDNA HS Assay kit was used to quantify the library, while the quality of the sample was checked by using the Agilent TapeStation device and Agilent High Sensitivity D1000 ScreenTape. Average fragment size was 420 bp.

ONT MinION – dcDNA sequencing

Poly(A)+-enriched and poly(A)+-enriched plus Terminator-handled samples were used to generate libraries for direct cDNA sequencing on ONT MinION device. For this, the ONT Direct cDNA Sequencing Kit (SQK-DCS109) was used following the kit's manual. In brief, the RNA samples were mixed with the VN primer (VNP; from the ONT kit) and 10 mM dNTPs were mixed and incubated at 65°C for 5 min. Then, 5x RT Buffer, RNaseOUT (Thermo Fisher Scientific), and Strand-Switching Primer (SSP; from the ONT Kit) were added and the mixtures were warmed to 42°C and kept for 2 min. Generation of the first cDNA strand was carried out by the addition of Maxima H Minus Reverse Transcriptase enzyme (Thermo Fisher Scientific) to the samples. RT and strand-switching reactions were carried out at 42°C for 90 min, then the reactions were stopped by the heat inactivation of the enzyme at 85°C for 5 min. RNase Cocktail Enzyme Mix (Thermo Fisher Scientific) was used for the removal of RNA from the RNA-cDNA hybrids. This process was carried out at 37°C for 10 min.

For the second strand synthesis, the LongAmp Taq Master Mix [New England Biolabs (NEB)] and PR2 Primer (PR2P) were used. Details of PCR reactions are shown in **Supplementary Table S4**. End-repair and dA-tailing of the fragmented DNAs was performed employing NEBNext End repair /dA-tailing Module (NEB) at 20°C for 5 minutes and 65°C for 5 minutes. This step was followed by ligation of sequencing adapter at room temperature for 10 minutes, using the NEB Blunt /TA Ligase Master Mix (NEB). ONT dcDNA libraries were labeled using barcodes (**Supplementary Table S5**) from the ONT Native Barcoding (12) Kit as described by the manufacturer. The adapted and tethered cDNA libraries (200 fmol/flow cell) were purified and loaded onto the ONT R9.4.1 SpotON Flow Cells. Altogether five flow cells were used for dcDNA sequencing.

To avoid potential "barcode hopping," samples from the earlier time points were sequenced separately from the later time points.

AMPure XP Beads were applied after each additional enzymatic step. Samples were eluted in UltraPure™ nuclease-free water (Invitrogen), then their concentration was measured with Qubit 4.0 fluorometer and by using the Qubit dsDNA HS Assay kit (**Supplementary Table S6**).

ONT MinION– dRNA sequencing

To avoid potential biases associated with reverse transcription and PCR, the direct RNA sequencing approach was also used for library preparation with the aim to detect and validate novel splice variants and 3′-UTR isoforms. Two RNA mixtures (1: Poly(A)+; 2: Poly(A)+ and Terminator-handled RNA) were mixed with RT Adapter (oligo dT-containing T10 adapter), RNA CS (both from the ONT kit; the latter was applied for monitoring the sequencing quality), NEBNext Quick Ligation Reaction Buffer, and T4 DNA ligase (both from NEB), and they were incubated at room temperature for 10 min. RT reactions were performed using dNTPs (NEB), 5 × first-strand buffer, DTT (both from Invitrogen) and UltraPure™ DNase/RNase-Free water (Invitrogen), and then the sample was mixed with SuperScript III enzyme (Thermo Fisher Scientific). RTs were carried out at 50°C for 50 min, and subsequently terminated at 70°C for 10 min. The samples were mixed with

RNA adapter (RMX; ONT kit), NEBNext Quick Ligation Reaction Buffer, T4 DNA ligase, and nuclease-free water and the ligation was carried out at room temperature for 10 min.

Agencourt RNAClean XP Beads were used to clean the RNA-cDNA hybrids, while the AMPure XP Beads were applied after each additional enzymatic step. Samples were eluted in UltraPure™ nuclease-free water (Invitrogen) between the reactions, while the ONT's elution buffer was applied for final elution. Following Qubit measurement (**Supplementary Table S7** shows the nucleic acid concentrations of libraries: RNA in dRNA-Seq and cDNA in cDNA-Seq), 100 fmol sample from the libraries was loaded onto two Flow Cells.

Pre-Processing and Data Analysis

Raw data were basecalled using guppy v3.4.5. Nucleotides were mapped to the reference genome (accession number: NC_001491.2) using the minimap2 software with the following option: -Y -C5–cs. SAMtools program was used for the conversion of SAM files to BAM files (https://academic.oup.com/bioinformatics/article/25/16/2078/204688). SeqTools, an in-house software was used for the identification of promoter elements and for the assembly of basic statistics (https://github.com/moldovannorbert/seqtools). The LoRTIA pipeline, developed in our laboratory, was used for the annotation of transcripts (https://github.com/zsolt-balazs/LoRTIA, v.0.9.9). The LoRTIA software was used for the identification of TSSs and TESs by checking the sequencing adapters and homopolymer As. Spurious TSSs and TESs – generated by RNA degradation, template switching or false priming – were filtered out. The significance of all qualified features was tested against the Poisson or the negative binomial distributions and the p-value is corrected using the Bonferroni method. The accepted TSSs, TESs and introns were assembled to transcripts using the Transcript_Annotator submodule of LoRTIA software. The very long, low abundance transcripts were manually identified. The Illumina reads were processed using the TrimGalore software using the following options: –paired –length 20 –quality 30 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), which was followed by mapping them the reference genome using the STAR 2.7.10a software. BAM file were visualized using the Geneious Prime 2022.0.2 (https://www.geneious.com) and IGV (Robinson et al., 2011) software. Variations in base compositions at certain positions compared to the reference genome were detected using the bam-readcount program (Khanna et al., (2022). Phylogenetic tree was assembled using the Genious software. EHV-1 sequences were aligned using the MAFT v7.450 software (Katoh and Standley, 2013). Phylogenetic tree was generated using the PhyML Geneious plugin (Guindon et al., 2010).

Author Contributions: Conceptualization, D.T., Z.Z. and Z.B.; Methodology, D.T., G.T., G.G. B.K. and Z.B.; Validation, D.T., Á.F., Á.D., Z.C. and Z.B.; Formal analysis, G.T., G.G., D.T., B.K., Á.F. and Z.B.; Investigation, D.T., Z.C., I.P., Á.D., M.M., Á.H., Z.Z.; Writing the original draft, D.T., G.T. and Z.B.; Writing, review and editing, Z.B.; Visualization, G.T., G.G. and D.T; supervision, Z.Z., F.M. and Z.B.; Funding acquisition, D.T., and Z.B. Supervision, Z.B. 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 128247 to ZB and FK 128252 to DT. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication. GG was supported by the UNKP-21-3-SZTE-51 New National Excellence Program of the Ministry of Human Capacities.

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

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

References

Allen, G.P., and Bryans, J.T. (1986). Molecular epizootiology, pathogenesis, and prophylaxis of equine herpesvirus-1 infections. Prog. Vet. Microbiol. Immunol. 2:78–144.

Allen, G., Kydd, J., Slater, J., and Smith, K. (2004). Equid herpesvirus-1 (EHV-1) and -4 (EHV-4) infections. In Infectious diseases of livestock, J.A.W. Coetzer and R.C. Tustin, eds. (Oxford Press: Cape Town), pp. 829–859.

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

Balázs, Z., Tombácz, D., Szűcs, A., Csabai, Z., Megyeri, K., Petrov, A.N., Snyder, M., and Boldogkői, Z. (2017B). Long-Read Sequencing of Human Cytomegalovirus Transcriptome Reveals RNA Isoforms Carrying Distinct Coding Potentials. Sci. Rep. 7, 15989. 10.1038/s41598-017-16262-z.

Balázs, Z., Tombácz, D., Csabai, Z., Moldován, N., Snyder, M., and Boldogkői, Z. (2019). Template switching artefacts resemble alternative polyadenylation. BMC Genomics 20, 824. 10.1186/s12864-019-6199-7.

Boldogkői, Z. (2012). Transcriptional interference networks coordinate the expression of functionally related genes clustered in the same genomic loci. Front. Genet. 3, 122. 10.3389/fgene.2012.0012.

Boldogkői, Z., Moldován, N., Balázs, Z., Snyder, M., and Tombácz, D. (2019A). Long-read sequencing – a powerful tool in viral transcriptome research. Trends Microbiol. 27, 578-592. 10.1016/j.tim.2019.01.010.

Boldogkői, Z., Balázs, Z., Moldován, N., Prazsák, I., and Tombácz, D. (2019B). Novel classes of replication-associated transcripts discovered in viruses. RNA Biol. 16, 166–175. 10.1080/15476286.2018.1564468.

Braspenning, S.E., Sadaoka, T., Breuer, J., Verjans, G.M.G.M., Ouwendijk, W.J.D., and Depledge, D.P. (2020). Decoding the Architecture of the Varicella-Zoster Virus Transcriptome. MBio 11, e01568-20. 10.1128/mBio.01568-20

Braspenning, S.E., Verjans, G.M.G.M., Mehraban, T., Messaoudi, I., Depledge, D.P., and Ouwendijk, W.J.D. (2021). The architecture of the simian varicella virus transcriptome. PLoS Pathog. 17, e1010084. 10.1371/journal.ppat.1010084.

Brooks, A.N., Hughes, A.L., Clauder-Münster, S., Mitchell, L.A., Boeke, J.D., Steinmetz, L.M. (2022). Transcriptional neighborhoods regulate transcript isoform lengths and expression levels. Science 375, 1000-1005. 10.1126/science. abg0162.

Carroll, C.L., and Westbury, H.A. (1985). Isolation of equine herpesvirus 1 from the brain of a horse affected with paresis. Aust. Vet. J. 62, 345–346. 10.1111/j.1751-0813. 1985.tb07660.x

Caughman, G. B., Staczek, J., and O'Callaghan, D. J. (1985). Equine herpesvirus type 1 infected cell polypeptides: evidence for immediate early/early/late regulation of viral gene expression. Virology 145, 49–61. 10.1016/0042-6822(85)90200-4.

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

Dellino, G.I., Cittaro, D., Piccioni, R., Luzi, L., Banfi, S., Segalla, S., Cesaroni, M., Mendoza-Maldonado, R., Giacca, M., and Pelicci, P.G. (2013). Genome-wide mapping of human DNA-replication origins: Levels of transcription at ORC1 sites regulate origin selection and replication timing. Genome Res. 23, 1–11. 10.1101/gr.142331.112.

Depledge, D.P., and Breuer, J. (2021). Varicella-Zoster Virus-Genetics, Molecular Evolution and Recombination. Curr. Top. Microbiol. Immunol. 10.1007/82_2021_238.

Derbigny, W.A., Kim, S.K. Jang, H.K., and O'Callaghan, D.J. (2002). EHV-1 EICP22 protein sequences that mediate its physical interaction with the immediate-early protein are not sufficient to enhance the trans-activation activity of the IE protein. 2685.. Virus Res. 84, 1-15. doi: 10.1016/s0168-1702(01)00377-x.

Frampton, A. R., Stolz, D. B., Uchida, H., Goins, W. F., Cohen, J. B., and Glorioso, J. C. (2007). Equine herpesvirus 1 enters cells by two different pathways, and infection requires the activation of the cellular kinase ROCK1. J. Virol. 81, 10879–10889. 10.1128/JVI.00504-07.

Gilet, J., Conte, R., Torchet, C., Benard, L., Lafontaine, I. (2020). Additional Layer of Regulation via Convergent Gene Orientation in Yeasts. Mol. Biol. Evol. 37, 365–378. 10.1093/molbev/msz221.

Gray, W.L., Baumann, R.P., Robertson, A.T., Caughman, G.B., O'Callaghan, D.J., and Staczek, J. (1987a). Regulation of equine herpesvirus type 1 gene expression: characterization of immediate early, early, and late transcription. Virology 158, 79–87. 10.1016/0042-6822(87)90240-6.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 2010;**59**(3):307–321. doi: 10.1093/sysbio/syq010.

Huang, C.J., Petroski, M.D., Pande, N.T., Rice, M.K., and Wagner, E.K. (1996). The herpes simplex virus type 1 VP5 promoter contains a cis-acting element near the cap site which interacts with a cellular protein. J. Virol. 70, 1898-904. 10.1128/JVI.70.3.1898-1904.1996.

Irimia, M., Weatheritt, R.J., Ellis, J.D., Parikshak, N.N., Gonatopoulos-Pournatzis, T., Babor, M., Quesnel-Vallières, M., Tapial, J., Raj, B., O'Hanlon, D., et al. (2014). A highly conserved program of neuronal microexons is misregulated in autistic brains. Cell 159, 1511–1523. 10.1016/j.cell.2014.11.035.

Kang, YL, Yang, DC, Kong L, Hou M, Meng, YQ, Wei L, Gao G. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Research*, Volume 45, Issue W1, 3 July 2017, Pages W12–W16, https://doi.org/10.1093/nar/gkx428

Khanna, A., Larson, D.E., Srivatsan, S.N., Mosior, M., Abbott, T.E., Kiwala, S., Ley, T.J., Duncavage, E.J., Walter, M.J., Walker, J.R., et al. (2022). Bam-readcount - rapid generation of basepair-resolution sequence metrics. J. Open Source Softw. 7, 3722. 10.21105/joss.03722.

Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30, 772–780. 10.1093/molbev/mst010.

Kim, S.K., Jang, H.K., Albrecht, R.A., Derbigny, W.A., Zhang, Y., and O'Callaghan, D.J. (2003). 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. 77, 2675– 2685. 10.1128/jvi.77.4.2675-2685.2003.

Kim, S.K., Ahn, B.C., Albrecht, R.A., and O'Callaghan, D.J. (2006). The Unique IR2 Protein of Equine Herpesvirus 1 Negatively Regulates Viral Gene Expression. J. Virol. 80, 5041–5049. 10.1128/JVI.80.10.5041-5049.2006.

Kronstad, L.M., Brulois, K.F., Jung, J.U., Glaunsinger, B. (2013). Dual short upstream open reading frames control translation of a herpesviral polycistronic mRNA. PLoS Pathog. 9, e1003156. 10.1371/journal.ppat.1003156.

Laver, T., Harrison, J., O'Neill, P.A., Moore, K., Farbos, A., Paszkiewicz, K., and Studholme, D.J. (2015). Assessing the performance of the oxford nanopore technologies MinION. Biomol. Detect. Quantif. 3, 1–8. 10.1016/j.bdq.2015.02.001.

Luo, G. X., and Taylor, J. (1990). Template switching by reverse transcriptase during DNA synthesis. J. Virol. 64, 4321–4328. 10.1128/JVI.64.9.4321-4328.1990.

Marx, V. (2021). Long road to long-read assembly. Nat. Methods 18, 125-129. 10.1038/s41592-021-01057-y.

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

Moldován, N., Tombácz, D., Szűcs, A., Csabai, Z., Snyder, M., and Boldogkői, Z. (2018A). Multi-platform Sequencing Approach Reveals a Novel Transcriptome Profile in Pseudorabies Virus. Front. Microbiol. 8, 2708. 10.3389/fmicb.2017.02708.

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

Moldován, N., Torma, G., Gulyás, G., Hornyák, Á., Zádori, Z., Jefferson, V.A., Csabai, Z., Boldogkői, M., Kalmár, T., Tombácz, D., et al. (2020). Time-course profiling of bovine herpesvirus type 1 transcriptome using multiplatform sequencing. Sci. Rep. 10, 20496. 10.1038/s41598-020-77520-1.

O'Callaghan, D.J., and Harty, R.N. (1994). Equine Herpesvirus. In Encyclopedia of Virology, R.G. Webster and A. Granoff A, eds. (Academic Press, Harcourt Brace & Company, San Diego, CA), pp. 423-429.

O'Grady, T., Wang, X., Höner Zu Bentrup, K., Baddoo, M., Concha, M., and Flemington, E.K. (2016). Global transcript structure resolution of high gene density genomes through multi-platform data integration. Nucleic Acids Res. 44, e145. 10.1093/nar/gkw629.

Oladunni, F.S., Horohov, D.W., and Chambers, T.M. (2019). EHV-1: A constant threat to the horse industry. Front. Microbiol. 10, 2668. 10.3389/fmicb.2019.02668.

Oláh, P., Tombácz, D., Csabai, Z., Póka, N., Prazsák, I. and Boldogkői, Z. (2015). Characterization of pseudorabies virus transcriptome by Illumina sequencing. BMC Microbiol. 15, 130. 10.1186/s12866-015-0470-0.

Osterrieder, N. (1999). Construction and characterization of an equine herpesvirus 1 glycoprotein C negative mutant. Virus Res. 59, 165–177. 10.1016/s0168-1702(98)00134-8.

Paillot, R., Case, R., Ross, J., Newton, R., and Nugent, J. (2008). Equine herpes virus-1: virus, immunity and vaccines. Open Vet. Sci. J. 2, 68–91. 10.2174/1874318808002010068.

Patel, J.R., and Heldens, J. (2005). Equine herpesviruses 1 (EHV-1) and 4 (EHV-4) – epidemiology, disease and immunoprophylaxis: a brief review. Vet. J. 170, 14–23. 10.1016/j.tvjl.2004.04.018.

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

Rhoads, A., and Au, K.F. (2015). PacBio sequencing and its applications. Genomics Proteomics Bioinformatics 13, 278–289. 10.1016/j.gpb.2015.08.002.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. 29, 24–26. 10.1038/nbt.1754.

Roizmann, B., Desrosiers, R., Fleckenstein, B., Lopez, C., Minson, A., and Studdert, M. (1992). The family Herpesviridae: an update. The Herpesvirus Study Group of the International Committee on Taxonomy of Viruses Arch. Virol. 123, 425–449. 10.1007/BF01317276.

Smith, R.H., Caughman, G.B., and O'Callaghan, J.D. (1992). Characterization of the Regulatory Functions of the Equine Herpesvirus 1 Immediate-Early Gene Product. J. Virol. 66, 936-945. 10.1128/JVI.66.2.936-945.1992.

Stacey, S.N., Jordan, D., Williamson, A.J.K., Brown, M., Coote, J.H., and Arrand, J.R. (2000). Leaky Scanning Is the Predominant Mechanism for Translation of Human Papillomavirus Type 16 E7 Oncoprotein from E6/E7 Bicistronic mRNA. J. Virol. 74, 7284–7297. 10.1128/jvi.74.16.7284-7297.2000.

Statello, L., Guo, CJ., Chen, LL. Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. Nat. Rev. Mol. Cell. Biol. 22, 96–118. 10.1038/s41580-020-00315-9

Tai-Schmiedel, J., Karniely, S., Lau, B., Ezra, A., Eliyahu, E., Nachshon, A., Kerr, K., Suárez, N., Schwartz, M., Davison, A.J., et al. (2020). Human cytomegalovirus long noncoding RNA4.9 regulates viral DNA replication. PLoS Pathog. 16, e1008390. 10.1371/journal. ppat.1008390.

Telford, E. A., Watson, M. S., McBride, K., and Davison, A. J. (1992). The DNA sequence of equine herpesvirus-1. Virology 189, 304–316. 10.1016/0042-6822(92)90706-u.

Tikhanovich, I., Liang, B., Seoighe, C., Folk, W.R., and Nasheuer, H.P. (2011). Inhibition of Human BK Polyomavirus Replication by Small Non-coding RNAs. J. Virol. 85, 6930–6940. 10.1128/jvi.00547-11.

Tombácz, D., Sharon, D., Oláh, P., Csabai, Z., Snyder, M. and Boldogkői, Z. (2014). Strain Kaplan of Pseudorabies Virus Genome Sequenced by PacBio Single-Molecule Real-Time Sequencing Technology. Genome Announc. 2, e00628-14. 10.1128/genomeA.00628-14.

Tombácz, D., Csabai, Z., Oláh, P., Balázs, Z., Likó, I., Zsigmond, L., Sharon, D., Snyder, M., and Boldogkői, Z. (2016). Full-length isoform sequencing reveals novel transcripts and substantial transcriptional overlaps in a herpesvirus. Plos One 11, e0162868. 10.1371/journal.pone.0162868.

Tombácz, D., Csabai, Z., Szűcs, A., Balázs, Z., Moldován, N., Sharon, D., Snyder, M., and Boldogkői, Z. (2017). Long-Read Isoform Sequencing Reveals a Hidden Complexity of the Transcriptional Landscape of Herpes Simplex Virus Type 1. Front. Microbiol. 8, 1079. 10.3389/fmicb.2017.01079.

Tombácz, D., Balázs, Z., Csabai, Z., Snyder, M., and Boldogkői, Z. (2018). Long-Read Sequencing Revealed an Extensive Transcript Complexity in Herpesviruses. Front. Genet. 9, 259. 10.3389/fgene.2018.00259.

Tombácz, D., Balázs, Z., Gulyás, G., Csabai, Z., Boldogkői, M., Snyder, M., and Boldogkői, Z. (2019). Multiple Long-read Sequencing Survey of Herpes Simplex Virus Lytic Transcriptome. Front. Genet. 10, 834. 10.3389/fgene.2019.00834.

Tombácz, D., Torma, G., Gulyás, G., Moldován, N., Snyder, M., and Boldogkői, Z. (2020). Meta-analytic Approach for Transcriptome Profiling of Herpes Simplex Virus Type 1. Sci. Data 7, 223 10.1038/s41597-020-0558-8.

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

Torma, G., Tombácz, D., Csabai, Z., Göbhardter, D., Deim, Z., Snyder, M., and Boldogkői, Z. (2021). An Integrated Sequencing Approach for Updating the Pseudorabies Virus Transcriptome. Pathogens 10, 242. 10.3390/pathogens10020242.

Vilela, C., and McCarthy, J.E.G. (2003). Regulation of fungal gene expression via short open reading frames in the mRNA 5′untranslated region. Mol. Microbiol. 49, 859–867. 10.1046/j.1365-2958.2003.03622. x.

Workman, R.E., Myrka, A.M., Wong, G.W., Tseng, E., Welch, K.C. Jr, and Timp, W. (2018). Single-molecule, full-length transcript sequencing provides insight into the extreme metabolism of the ruby-throated hummingbird Archilochus colubris. Gigascience **7**, 1-12. 10.1093/gigascience/giy009.

Workman, R.E., Tang, A.D., Tang, P.S., Jain, M., Tyson, J.R., Razaghi, R., Zuzarte, P.C., Gilpatrick, T., Payne, A., Quick, J., et al. (2019). Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat. Methods 16, 1297–1305. 10.1038/s41592-019-0617-2.

Zarski, L.M., Weber, P.S.D., Lee, Y., and Soboll Hussey, G. (2021). Transcriptomic Profiling of Equine and Viral Genes in Peripheral Blood Mononuclear Cells in Horses during Equine Herpesvirus 1 Infection. Pathogens 10, 43. 10.3390/pathogens10010043.

Supplementary Figures

Supplementary Figure S1. Genome-wide expression of EHV-1 transcripts

This figure shows the total transcripts of EHV-1 identified using the LoRTIA program suit. The color code is defined in the figure. The shade of the colors corresponds to the abundance: 1: 1-9 reads, 2: 10-49 reads, 3: 50-199 reads, 4: 200-999 reads, 5: >1000 reads.

Supplementary Figure S2. Transcriptional overlaps

Supplementary Figure S3. Phylogenetic tree of EHV-1 strains

Tables

Table 1. Statistics of EHV-1 reads

Supplementary Tables

Supplementary Table S1. The number of EHV-1 read counts and the average read-length obtained using different techniques.

Supplementary Table S2. List of EHV-1 transcripts

A. monocistronic transcripts; B. multicistronic RNA molecules; C. fusion transcripts; D. nested mRNAs; E. non-coding RNAs; F. manually annotated transcripts.

 doi:10.20944/preprints202206.0219.v1

Supplementary Table S3. Coding potential of ncRNAs

Supplementary Table S4. The applied PCR protocol

Supplementary Table S5. Barcodes

Supplementary Table S6. RNA concentrations

Supplementary Table S7. Nucleic acid concentrations of library samples