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

Ph.D. Thesis

Islam Almsarrhad M.Sc.

University of Szeged
Faculty of Medicine
Department of Medical Biology
Doctoral School of Multidisciplinary Medicine

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

Szeged
2023

1. INTRODUCTION

1.1 Regulation of herpesvirus transcription

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

1.2 DNA replication

DNA replication varies significantly among life's domains, but they also have many common features. Prokaryotic genomes have a single site where DNA synthesis begins (called replication origin; Ori) that is determined by consensus sequences, while eukaryotic genomes usually have tens of thousands of Oris that are defined by their chromatin structure. Viruses have one or a few Oris, which are specified by a combination of structural properties and sequence specificity (typically AT-rich regions) of the particular DNA segment. The replication of eukaryotic genomes starts with a binding of the origin recognition complex (ORC) to the Ori. The function of ORC is to serve as a platform for the assembly of the replisome, which consists of a wide range of proteins such as DNA helicase, DNA polymerase (DNP), topoisomerase, primase, DNA gyrase, single-stranded DNA-binding protein (ssDBP), RNase H, DNA ligase, and telomerase enzymes. Several proteins that are essential for DNA replication are encoded by herpesviruses. For instance, an origin-binding protein (OBP) (*ul9*), an ssDBP (*ul29*), two DNPs (*ul30* and *ul42*), and three helicase/primase enzymes (*ul5*, *ul8*, and *ul52*) are coded by HSV-1. Some viral factors that play roles in nucleotide metabolism [ribonucleotide reductase (*ul39/40*), thymidine kinase (*ul23*), uracil glycosylase (*ul12*), deoxyuridine triphosphatase (*ul50*), alkaline nuclease (*ul12*)] allowing herpesvirus replication in non-dividing cells.

1.3 Overlapping viral transcripts

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

1.4 Non-coding RNAs regulating DNA replication

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

1.5 Replication origin-associated herpesvirus transcripts

Replication-origin-associated RNAs (raRNAs) have been identified in all three subfamilies of herpesviruses. These transcripts have been previously characterized in βHVs and γHVs but were mostly overlooked in αHVs until recently. One example of a non-coding raRNA in HCMV is RNA4.9, which originates from the OriLyt and has multiple functions in regulating viral DNA replication and gene expression. RNA4.9 can form DNA: RNA hybrids and modulate the level of ssDBP encoded by *ul57*. RNA4.9 may also have other roles in cis and trans, such as repressing the major IE promoter during latency. This discovery indicated that HCMV possesses a distinct method of managing replication, for which there's no proven evidence. Another vital non-coding raRNA crucial for HCMV DNA replication is the SRT (smallest replicator transcript), which is also located at the OriLyt. Two more raRNAs (vRNA-1 and vRNA-2) overlapping the OriLyt have been described in HCMV. Rennekamp and Lieberman reported that a ncRNA designated *BHLF1* forms an RNA: DNA hybrid at the OriLyt region of EBV. Additionally, a bidirectional promoter and a highly structured RNA were identified within this region of EBV. The role of this subsequent transcript is to aid the viral EBNA1 and HMGA1a proteins in attracting ORC. Two co-terminal lncRNAs with the same end near the OriS of HSV-1 were also described earlier. Long-read sequencing (LRS) techniques have enhanced transcript research and aided in the identification of new viral transcripts and their TIs, such as splice, TSS, and TES variants. These studies have detected several lncRNAs close to the OriS and OriL regions of αHVs. However, their precise function remains unknown, so labeling them as 'replication RNAs' would suggest a specific involvement in DNA replication. Moreover, the genes around the Oris produce TIs with long 5′-untranslated regions (5′ UTRs – TSS isoforms), or 3′ UTRs (TES isoforms) that overlap with the replication origin.

RESULTS

1.6 Multiplatform sequencing for characterization of viral transcripts

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

In the first part of the study, the precise locations of canonical viral RNA transcripts, including Transcription Initiation sites (TIs), Transcription Start Sites (TSSs), Transcription End Sites (TESs), and splice variants, were annotated. Previous annotations were validated or adjusted by combining and reevaluating sequencing datasets. Cis-regulatory elements for various viral RNA transcripts were identified, consistently featuring promoter elements (TATA boxes) within Ori sequences. The study aimed to provide a comprehensive understanding of the complexity of Transcript Origins (TOs) in the scrutinized genomic areas. In cases where viruses lacked dRNA-Seq and/or CAGE data, more stringent standards were applied for transcript annotation, resulting in reduced transcript diversity. The study also determined the relative abundances of transcripts in viruses with sufficient data and utilized multi-time-point real-time RT-PCR analysis to monitor the expression kinetics of the three most crucial long non-coding RNAs (lncRNAs) in PRV. To confirm the lncRNAs and longer mRNA variants of PRV, BoHV-1, EHV-1, HSV-1, and KSHV, we used RT-PCR. We also used native RNA sequencing to check the cDNA sequencing results.

1.6.A Alphaherpesviruses – OriS

In this part of the study, we discovered novel transcripts and TIs near the OriSs of α-HVs. The annotated transcripts, including lncRNAs like BoHV-1's OriS-RNA, HSV-1's OriS-RNA1, the *NOIR*-1 transcripts (in PRV, EHV-1, VZV, and SVV), and *NOIR*-2 transcript (in PRV). We found that in all examined αHVs, the very long 5′ TIs of transcription regulator genes (*us1* and *icp4*) of BoHV-1, EHV-1, HSV-1, and SVV overlap the OriS. We cannot rule out an opportunity that this is the case for other αHVs also overlapping the OriS, but they might have been missed due to their long size and low abundance. In the HSV-1, we detected a very complex splicing pattern of *US1* transcripts in αHVs. We also identified novel lncRNAs oriented antisense to the HSV-1 OriS-RNA1. The NOIR-1 family members exhibit a distinct arrangement in relation to the icp4 gene. Standard forms of these RNAs do not overlap with icp4, but longer NOIR-1 variants partially overlap with this crucial TR gene. In viruses like EHV-1, VZV, SVV, and likely PRV, an extended Transcription Initiation (TI) site from US1 originates from the noir-

1 gene promoter. In SVV, an elongated Transcription Start Site (TSS) variant of NOIR-1 overlaps with the regular ICP4 transcript, and both versions overlap with the OriS region. NOIR-1 is moderately expressed, while NOIR-2 has very low expression levels. In the VZV genomic area, five long non-coding RNAs (lncRNAs) labeled as NOIR-1A, -1B, -1C, -1D, and -1E were identified with distinct Transcription Start Sites (TSSs) and Transcription End Sites (TESs). TIs with TSSs align closely with TATA boxes within the OriSs in all six αHVs, suggesting the functionality of these promoter elements.

1.6.B Alphaherpesviruses – OriL

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

1.6.C Betaherpesviruses

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

1.6.D Gammaherpesviruses

The long 5′ UTR isoform of EBV *BCRF1* gene overlaps the OriP, Similarly, the long 5′ UTR variants of the *BHRF1* gene overlap OriLyt-L. One of these transcripts is also a spliced form of this gene. The promoter of the *BHLF1* gene is located within the Orilyt-L. We also report novel isoforms of lncRNAs that either have introns that overlap the OriLyt-R or start from the replication origin. We found that many ncRNAs of different lengths associated with OriLyt-L can be made from the same TSS besides the 1.4-kb ncRNA when KSHV is reactivated. The OriLyt-L is surrounded by short genes that code for proteins such as *K4.2, K4.1,* and *K4* on the left and *K5, K6,* and *K7* on the right side. Previous studies showed that *K4, K4.1,* and *K4.2* are expressed as mono-, bi-, and tri-cistronic mRNAs, but our analysis reveals a more complex expression pattern that includes unspliced RNAs of different lengths and spliced RNA variants. We also found that *K5/K6* genes can be expressed not only separately but also through splicing, which produces mRNAs with a first exon of varying length. Importantly, our results agree with previous genomics studies but also add to the number of different viral RNA transcripts that can be produced from the OriLyt-L locus, which can potentially increase the coding potential of viral mRNAs. KSHV latency locus is located between *K12* and *LANA* (*ORF73*), which codes for 4 protein-coding latent genes (*K12, K13, ORF72, ORF73*) and 12 pre-miRNAs. Here, we detected several lncRNAs that are antisense to the miRNA-coding genomic regions.

1.7 Non-coding RNAs mapping near the transcription regulator genes

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

1.8 Transcriptional overlaps of replication genes

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

1.9 Transcript validation using qRT-PCR

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

1.10 Epstein-Barr virus (EBV):

1.10.A Multiplatform profiling of the EBV transcriptome

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

1.10.B Novel monocistronic mRNAs with canonical ORFs

In this part of our work, we present the identification of 15 new monocistronic transcripts. Among these, we found unspliced versions of *BNRT10*, *BHLF1*, *BORF2*, and *BGLT18* transcripts, which were previously only known in their spliced forms. Additionally, we discovered ten monocistronic transcripts that contain complete open reading frames (ORFs), whereas previous descriptions only mentioned shorter isoforms with incomplete ORFs lacking an in-frame AUG codon. The genomic region of *BFRF3* has not been annotated with these transcripts yet, despite studies showing its transcriptional activity. Among our findings, we identified *BFRT3*, which fully overlaps with the *BFRF3* ORF and has a novel terminus.

1.10.C Splice junctions and introns

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

1.10.D mRNAs with altered coding potential

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

1.10.E Non-coding transcripts

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

2. DISCUSSION

With LRS technologies, we can now identify and accurately annotate transcripts and RNA isoforms, such as those with length and splice variants. These technologies include PacBio’s Single Molecule, Real-Time (SMRT) sequencing, ONT’s nanopore sequencing, and Loop Genomics’ LoopSeq synthetic LRS (which uses Illumina platform). They have been used alone or together with each other or with SRS to mark viral transcripts in herpesviruses from all three subfamilies (HSV-1; VZV; PRV; BoHV-1; HCMV; and EBV). We and other researchers have found out that different viruses have a hidden and complex network of genes that overlap with each other when they are transcribed into RNA molecules. It has been shown that the RNA molecules encoded by closely spaced genes overlap each other in a parallel, divergent, or convergent manner. This phenomenon implies an interaction between the transcription machinery at the TOs throughout the entire viral genome. We and others have previously demonstrated that in several viruses, the replication origins overlap with specific lncRNAs and with long 5′ or 3′ UTR isoforms of mRNA. We note that many of these long versions of mRNAs, 5′ UTR isoforms, may not code for proteins, because their start codons are far away from their TSSs. The only exceptions may be those transcripts whose large parts of the 5′ UTR are spliced out. Functional analyses discovered how some replication RNAs regulate DNA synthesis by forming RNA: DNA hybrids in numerous viruses.

In summary, even among closely related herpesvirus species, distinct strategies have evolved to create transcriptional overlaps (TOs) at the Oris and TR genes. This highlights the importance of TOs in controlling DNA replication and overall transcription. While the primary TR genes seem to regulate each other and the initiation of DNA replication at the OriS region of αHVs, in Simplexviruses, the principal replication genes at the OriL region may instead govern each other and DNA replication through mechanisms involving TOs. These potential mechanisms offer multiple layers of regulation beyond the traditional interaction of transcription factors with promoters. *ICP4* has been demonstrated to enhance the expression of *icp0* genes by binding to their promoter. Conversely, *ICP22* (the product of *us1*) suppresses the expression of both *icp4* and *icp0.* Additionally, *ICP0* transforms *ICP4* from a repressor into an activator of mRNA synthesis in HSV-1. Mutating the *us1* gene has a differential effect on the transcription kinetics of E and L genes. We believe that the significance of our findings extends far beyond our specific study and offers a more universal insight into how the control of herpes viral replication and transcription has developed over time in tandem.

Grasping these intricate relationships among different genes and regulatory components can offer valuable insights into the overall mechanisms that govern herpesvirus replication and gene activity. Delving deeper into these potential interactions and their functional importance could pave the way for novel therapeutic strategies to tackle herpesvirus infections.

3. Conclusions

RNA sequencing has been a pioneering technique in understanding viruses, significantly progressing the acquaintance of viral biology. After characterization of complete genetic sequence of viruses, transcriptomics offers productive visions into their genome structure, organization, and diversity, allowing the identification of viral genes responsible for important functions like replication, transcription, and protein synthesis. By pairwise analysis of viral transcriptomes, RNA sequencing exposes the evolutionary relationship between viruses. Allowing over times genetic-changes track down possible. The use of long-read sequencing (LRS) techniques has revolutionized transcriptomic research by revealing unexpected transcriptomic complexity in various organisms, including viruses. In this study, we used both newly generated and previously published LRS and short-read sequencing datasets to discover additional Ori-proximal transcripts in nine herpesviruses belonging to all of the three subfamilies (alpha, beta and gamma). We identified novel long non-coding RNAs (lncRNAs), as well as splice and length isoforms of mRNAs and lncRNAs. The analysis revealed an intricate network of transcriptional overlaps, suggesting the existence of a "super regulatory center" that controls both replication and global transcription through multilevel interactions between molecular components.