COMPREHENSIVE CHARACTERIZATION OF VIRAL TRANSCRIPTOMES USING LONG-READ SEQUENCING

PhD Thesis

Norbert Moldován

Szeged

- 2020 -

COMPREHENSIVE CHARACTERIZATION OF VIRAL TRANSCRIPTOMES USING LONG-READ SEQUENCING

PhD Thesis

Norbert Moldován

Department of Medical Biology

Doctoral School of Interdisciplinary Medicine

Faculty of Medicine

University of Szeged

Supervisor: Prof Zsolt Boldogkői, PhD, DSc

Szeged

- 2020 -

1. List of publications:

1.1. Publications directly related to the subject of the thesis

I. Moldovan, N. et al., 2018. Third-generation Sequencing Reveals Extensive Polycistronism and Transcriptional Overlapping in a Baculovirus. *SCIENTIFIC REPORTS*, 8.
IF: 4.011

II. Boldogkői, Z. et al., 2019. Long-Read Sequencing – A Powerful Tool in Viral Transcriptome Research. *TRENDS IN MICROBIOLOGY*, 27(7), pp.578–592.
IF: 11.974

III. Tombácz, D. et al., 2019. Multiple Long-read Sequencing Survey of Herpes Simplex Virus Dynamic Transcriptome. *FRONTIERS IN GENETICS*, 10.
IF: 3.517

1.2. Other related publications

IV. Balázs, Z. et al., 2019. Template-switching artifacts resemble alternative polyadenylation. BMC GENOMICS, 20.
IF: 3.501

V. Boldogkői, Z., Balázs, Z., et al., 2019. Novel Classes of Replication-associated Transcripts Discovered in Viruses. RNA BIOLOGY, 16(2), pp.166–175.
IF: 5.477

VI. Boldogkői, Z. et al., 2018. Transcriptome-wide analysis of a baculovirus using nanopore sequencing. SCIENTIFIC DATA, 5.
IF: 5.929

VII. Moldován, N. et al., 2017. Multi-platform Analysis Reveals a Complex Transcriptome Architecture of a Circovirus. VIRUS RESEARCH, 237, pp.37–46.
IF: 2.736

VIII. Moldován, N. et al., 2018. Multi-platform Next-generation Sequencing Identifies Novel RNA Molecules and Transcript Isoforms of the Endogenous Retrovirus Isolated from Cultured Cells. FEMS MICROBIOLOGY LETTERS, 365(5).
IF: 1.994

IX. Moldován, N. et al., 2018. Multi-Platform Sequencing Approach Reveals a Novel Transcriptome Profile in Pseudorabies Virus. FRONTIERS IN MICROBIOLOGY, 8.
IF: 4.259

X. Prazsák, I. et al., 2018. Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus. BMC GENOMICS, 19.
IF: 3.501

XI. Tombácz, D., Balázs, Z., et al., 2017. Characterization of the Dynamic Transcriptome of a Herpesvirus with Long-read Single Molecule Real-Time Sequencing. SCIENTIFIC REPORTS, 7.
IF: 4.011

XII. Tombácz, D., Csabai, Z., et al., 2017. Long-Read Isoform Sequencing Reveals a Hidden Complexity of the Transcriptional Landscape of Herpes Simplex Virus Type 1. FRONTIERS IN MICROBIOLOGY, 8(JUN).
IF: 4.259

XIII. Tombácz, D., Prazsák, I., et al., 2018. Lytic Transcriptome Dataset of Varicella Zoster Virus Generated by Long-read Sequencing. FRONTIERS IN GENETICS, 9.
IF: 3.517

XIV. Tombácz, D., Donald, S., et al., 2018. Transcriptome-wide survey of pseudorabies virus using next- and third-generation sequencing platforms. SCIENTIFIC DATA, 5.
IF: 5.929

Cumulative IF: 64.615

2. Table of contents

3. Introduction

The phenotype of an organism is determined by the expression of its genes, which is modified by the environment. Crick outlined the central dogma of biology (Crick, 1958, 1970), explaining the flow of information from genes to proteins, the carrier being the RNA. The sum of the RNA molecules in a given time point, called the transcriptome characterizes the cell in which they were expressed. Subsets of these RNAs determine the cell's type and regulates its biological functions. Hence their central role, RNAs can be used to understand the functional elements of a genome and to interpret development and disease.

In the past decades, a huge variety of RNAs have been proposed, each having a separate function. They can be divided into four groups on the bases of their functions:

1. Messenger RNAs (mRNAs) are the carriers of the information involved in building the cell's proteins. These RNS are ubiquitous in all living organisms and viruses.
2. Transfer RNAs deliver amino acids, the building blocks of proteins to ribosomes, and bear the information necessary for the correct assembly of their "luggage" in the form of anticodons.
3. Regulatory RNAs form a numerous group, involved in every part of the cell's life, from DNA replication, through transcription and RNA turnover to the apoptosis of the cell. These RNAs form a diverse group, with many sizes and functions. RNAs shorter than 200 bp conventionally are termed small non-coding RNAs (snRNA) (Brosnan and Voinnet, 2009), while those longer than 200 bp are long non-coding RNAs (lncRNA) (Ransohoff, Wei and Khavari, 2018). These noncoding types have their own subtypes with particular functions.
4. Some viruses use RNA as their genetic material. The viral genome can be single-stranded or double-stranded. Given that it is single-stranded they can serve as the template of the protein synthesis (positive-strand RNA viruses), or their genome can be transcribed by RNA dependent RNA polymerase into mRNA (negative-strand RNA viruses) (Baltimore, 1971). A small group of

RNA viruses are ambisense, meaning they are negative-strand with a capability of translating some genes from their genome, without transcription (Nguyen and Haenni, 2003).

In addition to the 7-methylguanylate (m^7G) cap RNAs contain more than 60 other modified bases (Zhao, Roundtree and He, 2018), including other methylated bases, like m^6A, m^1A and m^5C, the isomerised pseudouridine (Ψ) or the oxidized forms of m^5C (hm^5C) (Roundtree *et al.*, 2017). Previous modification studies detected RNA methylation in simian virus-40 (Lavi and Shatkin, 1975), adenoviruses (Sommer *et al.*, 1976) and the human alphaherpesvirus 1 (HSV-1) (Moss *et al.*, 1977). Despite the wide range of known nucleotide modifications, their functions are poorly understood. Several recent studies reported on the role of 6-methyladenosine in the replication and virus-host interactions of the HIV-1 (Lichinchi *et al.*, 2016; Tirumuru *et al.*, 2016), and the RNA synthesis of *Flaviviridae (Gokhale et al., 2016).* Viral RNAs are methylated by the cell's methyltransferases present in both the cytoplasm and the nucleus (Gokhale *et al.*, 2016) as well as viral methyltransferases (Ho, Gong and Shuman, 2001; Tao *et al.*, 2013).

3.1. Sequencing the transcriptome

Our understanding of the genomic functions has been shaped by an endeavour of huge proportions, that started more than a decade ago, through the development of the RNA sequencing (RNA-seq) technology (Emrich *et al.*, 2007). RNA-seq has become a ubiquitous tool used for analysing the quantitative changes of gene expression between experimental groups (differential gene expression or DGE) (Young *et al.*, 2012) or during longitudinal sampling of microorganisms and tissues (Hubbard *et al.*, 2013). It allows the structural characterization of the transcriptome, uncovering alternative splicing events (Wang *et al.*, 2008) and length isoforms (Depledge *et al.*, 2019). It broadened our understanding of the regulation of gene expression by non-coding RNAs (Djebali *et al.*, 2012; Morris and Mattick, 2014). Additionally, it is used to detect the increasing number of RNA modifications (Schaefer, Kapoor and Jantsch, 2017; Liu *et al.*, 2019). To date, a wide range of methods and downstream analysis workflows has been developed to accomplish these divergent purposes, all using RNA-seq as their common technology.

The first instances of RNA sequencing took place during the early times of Sanger sequencing, also known as first-generation sequencing (Adams *et al.*, 1991, 1995). This platform suffered from several technical difficulties due to sequencing length constraints and low throughput. To enable higher throughput, more precise isoform detection and to allow quantitative analysis, tag-based techniques were developed, such as the serial analysis of gene expression (SAGE) (Velculescu *et al.*, 1995) or cap analysis gene expression (CAGE) (Shiraki *et al.*, 2003). These methods are still being used with some extent by the virology community, on modern sequencing platforms (Wyler *et al.*, 2017; Djavadian, Hayes and Johannsen, 2018).

Innovations in microfluidics and nanotechnology brought the era of the next-generation sequencing platforms. These are capable of sequencing millions of cDNA molecules at the same time (massively parallel sequencing), reducing the total time and cost of sequencing and enormously increasing the amount of output information. In contrast to first-generation platforms, the second-generation is sensitive to the expression levels of splice isoforms and can be used to find novel genes and non-coding RNAs (Wang, Gerstein and Snyder, 2009). Next-generation sequencing (NGS) platforms use a variety of techniques for the detection of incorporated nucleotides in a newly synthesized complementary polynucleotide chain, from the measurement of pH changes (Ion Torrent) to the detection of fluorescence (Roche 454 / Illumina). All require the reverse transcription of RNA into cDNA, and their sequencing length is very limited.

We are in the midst of the third revolution of sequencers, which are capable of producing long reads from individual molecules while maintaining high throughput. Pacific Biosciences developed a new platform based on a previously explored concept: the RS, RSII, Sequel and Sequel II systems use fluorescently labelled nucleotides and a DNA polymerase for sequencing. The synthesis of a novel chain of DNA from these labelled nucleotides is performed in a nanophotonic confinement ca. 20 zeptoliters in volume. The polymerase enzyme is immobilized on the bottom of this tiny chamber. The size of the confinement makes it possible to observe the incorporation of each nucleotide in the growing strand, resulting in the so-called “Real-time Sequencing” specific to this platform (McCarthy, 2010). The system doesn't require the amplification of starting material, it is capable to accept native DNA (Coupland *et al.*, 2012) and theoretically native RNA as input and is capable of detecting modified nucleotides, by measuring the changes in the incorporation times. However, this latter application is not well documented, there are only a few tools available for data analysis, and the results are hard to interpret. A new competitor in the sequencing industry is Oxford Nanopore Technologies. They based their approach to sequencing on totally new grounds by the use of protein nanopores embedded in a synthetic membrane. A ratcheting molecule is attached to the DNA or RNA to-be-sequenced during the library preparation, which unwinds and splits the double-stranded DNA [or in case of direct RNA sequencing the cDNA-RNA hybrid (Ayub *et al.*, 2013)] into separate strands, threads one strand into the nanopore, and controls the speed of the strand passing through the pore. The detection of nucleotides is based on the changes in the current, caused by the nucleotides in the nanopore, which affect the ionic flow through the pore (Meller, Nivon and Branton, 2001; Branton *et al.*, 2008). The specifics of each platform mentioned in this section can be found in **Table 1**.

Table 1. **Comparison of sequencing platforms.**

	Sequencing Method	Read Length (base)	Run Time	Gb/ Run	Advantage	Disadvantage
ABI 3730	Chain termination	500	3h	0,00 05	High accuracy	High running cost, low throughput
Roche 454 GS FLX+	Pyrosequencing	700	23h	0.7	Average read length	High homopolymer error rate
Illumina MiSeq	Sequencing by synthesis (detecting fluorescence)	2*300	2d	15	Low error rates, run cost	Short read length
Thermo Fisher Ion Proton	pH change detection	200	2h	100	Rapid runs	High homopolymer error rate
PacBio Sequel	Real-Time Sequencing (detecting fluorescence)	up to 15,000	20h	160	Low error rate in consensus, long read length	High capital cost

ONT MinION	Nanopore sequencing (detecting voltage changes)	up to 1M	Flexible	30	Long read length, low capital cost	Relatively high error rate

3.2. Preparing sequencing libraries for RNA-seq

The first step of all RNA-seq protocols is the extraction of the RNA from the cells infected by the virus. Depending on the goals of the experiment several extraction and selection methods has been developed (Kumar *et al.*, 2017). Extraction methods start with mechanical or chemical homogenization of the cells, followed by separation of proteins and nucleic acids usually by a centrifuge and silica membrane. Additional enzymatic treatments can be involved to minimize DNA contamination of the sample. This sample is coined as the *total RNA*, as it contains all the RNA fractions mentioned above, and it can be used for sequencing library preparation as it is or can be due for selection. RNA selection is recommended in most cases, as mRNAs and ncRNAs form only 1-5% of the total RNA pool, the other 95% being composed of ribosomal RNAs and transfer RNAs (of the host). Most of the selection methods are based on the retention of RNA molecules by hybridization with a single-stranded oligonucleotide anchored to a substrate. The oligonucleotides can be homopolymer T-s, for the retention of the polyadenylated fraction alias poly(A) selection, or sequence-specific oligos for filtering out ribosomal RNA-s, coined by the term: ribo-depletion.

Short-read sequencing technologies require the fragmentation of RNAs, while this process is not needed for platforms capable of sequencing long reads.

RNAs are fragile molecules, are easily degraded by ribonucleases. To circumvent this, and to allow the amplification of RNA sequences, most RNA-seq protocols rely on the reverse transcription (RT) of the RNA molecule into complementary DNA (cDNA). The RT can be followed by the enzymatic removal of RNAs and the synthesis of a the second strand of the cDNA. This can be used in workflows relying on non-amplified cDNA as starting material, like the Iso-seq protocol of PacBio or the direct cDNA protocol of ONT. Other workflows imply the amplification of the cDNAs using PCR. These are adequate for target enrichment, like in some of the 5’ cap-selection methods used for full-length cDNA sequencing and transcriptional start site profiling (Schmidt and Mueller, 1999), and in some cases, the amplification step is necessary for the incorporation of sequencing adapters. At the end of these steps, the library is ready for loading on the sequencing units.

3.3. Sequencing the RNA directly

While reverse transcription of RNAs into cDNAs and subsequent PCR is a common practice and for most platforms is a necessary step prior to sequencing, it conceals information regarding modified ribonucleotides and can lead to many RT and PCR-related artefacts.

The primer annealing step of PCR can form three kinds of duplexes. When the concentration of the primer is high, there is a high chance of duplexes formed by the primer and the single-stranded cDNA, however, in later cycles of the PCR the primer is used up, and homoduplexes and heteroduplexes are more abundant. Homoduplexes are formed between one cDNA strand and its matching reverse complementary strand, and these will result in no amplification, however, heteroduplexes, which are formed between a cDNA strand and another cDNA with a complementary sequence fragment can lead to the formation of chimeric sequences (Qiu *et al.*, 2001). Two other mechanisms have been described leading to the formation of chimaeras. Incompletely extended primers in the later phase of the PCR, when they can compete with the original primers hybridize with partially homologous sequences and can act as primers themselves (Pääbo, Irwin and Wilson, 1990), while during the extension phase the DNA polymerase or the reverse transcriptase can jump from one sequence to another homologous sequence. This phenomenon can occur between separate cDNA strands (strand switching), resulting in chimaeras, or on one cDNA template (template switching) resulting in missing stretches of cDNA, identified later as (false) introns (Odelberg *et al.*, 1995). Random events during RT and PCR can also lead to false RNA isoforms. Missanealing of the oligo(dT) primers can lead to the identification of false Transcriptional End Sites (TESs) (Balázs *et al.*, 2019), while incomplete RTs can interfere with the identification of transcriptional start sites (TSSs). Additionally, the misincorporation of nucleotides by reverse transcriptase or DNA polymerase can alter primer binding or downstream analysis of the sequences (Kwok *et al.*, 1990; Reiss *et al.*, 1990). RT and PCR biases not only affect the detection of RNA isoforms but their quantitative analysis too. Sequence-specific primers are affected by biases in primer binding energy if the target sequence is not known with base-precision. Even one mismatch to the target results in low efficiency in amplification, obscuring gene expression analysis (Polz and Cavanaugh, 1998). Due to the rehybridization bias, the amplification of abundant sequences is hindered, resulting in a 1 to 1 ratio after many PCR cycles (Mathieu-Daude, 1996). This leads to false relative abundance estimations.

The elimination of PCR by the capability of third-generation platforms of sequencing single molecules without amplification (coined direct cDNA sequencing) offers the solution for PCR-related artefacts, while the capability of the MinION platform to sequence the RNA directly removes the possibility of RT-related artefacts.

MinION sequencing requires the DNA or RNA to be free of secondary structures to prevent strains and stresses on the molecule, which can alter its constant speed through the nanopore. For this, the dRNA sequencing protocol employs an RT step at the beginning, however, the produced cDNA molecule will not be sequenced. The RT, which can be oligo(dT) or randomly primed, is followed by the ligation of the sequencing adapter to the 3' end of the RNA, and loading of the sample on the flow cell. RNAs are always sequenced in the 3' to 5' direction on the flow cell. The dRNA-seq technology currently falls behind the cDNA or DNA sequencing technologies of MinION in throughput, because the sequencing speed of RNA is six-times slower (Garalde *et al.*, 2018). The resulting reads are of a lower quality than those of the cDNA-seq of MinION, but this does not pose a problem for isoform detection if a well-annotated reference genome is known. Another, more serious issue with the technology is that it is unable to sequence the last ca. 30 nt region on the 5' end of the RNA. This is because the ratcheting molecule releases the RNA strand before it passes through the pore, resulting in a rapid slippage of the molecule through the pore, producing an obscured signal, which cannot be base called. In some cases, the poly(A) tail is also missing from the reads, caused by the muddling of the signal by the attached sequencing adapter, which is DNA.

The strength of dRNA-seq technology currently lies in that it lacks the library preparation steps known to produce false introns or chimaeras, and that it preserves the information regarding modified nucleotides.

3.4. Annotating the transcriptome

After sequencing, the data is in raw format and needs to be base called. This usually happens on the sequencer or in the cloud for the Illumina and PacBio platforms, while ONT platforms tend to have their base callers installed on their sequencers or on a user-provided PC. In any case, the resulting *fastq* files contain base called reads, the quality information of each nucleotide, and additional information about the read, if necessary. Several preprocessing steps can be implemented in order to clean the data of any unnecessary sequences, which can spoil the transcript assembly, annotation or further analysis. Run statistics and read quality is reported by every platform during the sequencing, however, it is a good practice to check the quality of a small batch of the reads before further processing, to avoid superfluous work. Preprocessing starts with demultiplexing if several samples were loaded for sequencing. This is followed by trimming of indexes, adapters and barcodes from the end of the reads. Trimming must be avoided when TSS and TES annotation is based on the presence of indexes. If the goal is to assemble the viral but not the host's transcriptome, the data can be decontaminated, by removing reads not mapping to the viral genome. Chimaeras can be removed, however, this will prevent the discovery of fusion transcripts.

For short-read sequencing (SRS) technologies preprocessing is succeeded by the assembly of the transcripts, which can be *de novo (Grabherr et al., 2011)* or based on their mapping to a viral genome and gene annotations (the "mapping first approach") (Trapnell *et al.*, 2010). This process involves the aggregation of short reads based on homologous overlaps, which is computationally demanding and can lead to false splice isoform discoveries when reads can't span an entire exon (**Figure 1.**). This is not an issue for long-read sequencing (LRS) technologies, where the entire length of an RNA is sequenced at once, and the assembly is usually not part of the transcript annotation process. The next step for both SRS and LRS platforms is the annotation of TSSs, TESs and splice sites, followed by the annotation of whole transcripts.

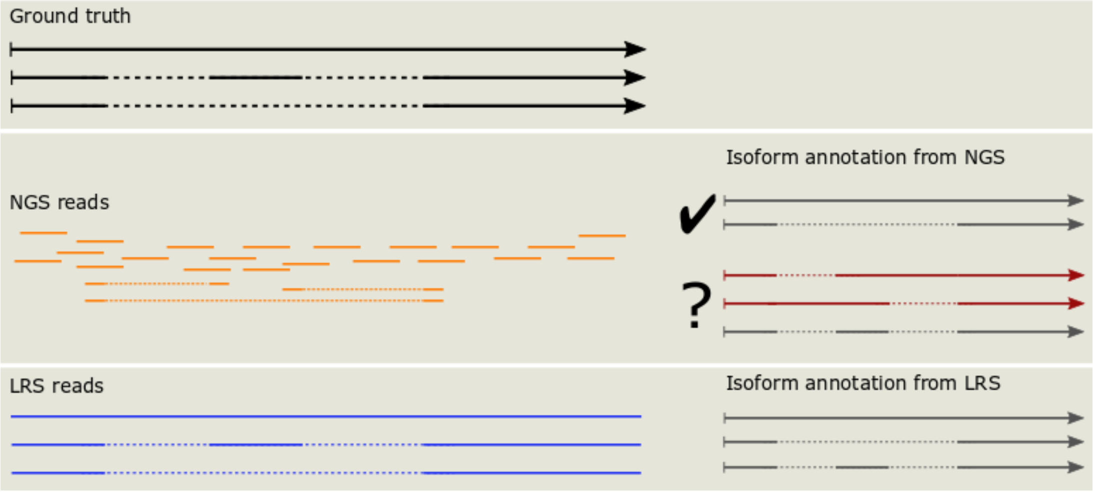


Figure 1. The read-length-exon-length problem of NGS and the ability of LRS to sequence full-length splice isoforms. Ground Truth represents the RNA which will be sequenced. Dashed lines are introns while solid lines are exons. NGS reads are outlined in orange while LRS reads in blue. Real isoforms detected are outlined in the right column by grey arrows, while artefacts resulting from the assembly process are in red.

3.5. Detecting RNA modifications

Several methods for the detection of non-canonical RNA nucleotides coupled with RNA-seq technologies have paved the way for transcriptome-wide modification mapping, although the precise detection of modified nucleotides remains a highly challenging task. The first trials applied antibody-based immunocapturing (MeRIP) (Delatte *et al.*, 2016; Amort *et al.*, 2017; Sinclair *et al.*, 2017) and UV cross-linking (miCLIP) (Hussain *et al.*, 2013), where regions surrounding the modified nucleotide were coupled and retained by modification-specific antibodies, while the unmodified fraction was degraded. After deep sequencing, the position of the modified nucleotides could be annotated based on the coverage of the reads. Variants of this technique are still widely used today.

Another approach includes the reverse transcriptase's tendency to arrest at modified nucleotides or to incorporate non-Watson-Crick compatible dNTPs into the nascent cDNA (Tserovski *et al.*, 2016). The results of RT signature-based methods have proven to be highly dependent on the reaction conditions (Hauenschild *et al.*, 2015), which can be improved using engineered enzymes (Aschenbrenner *et al.*, 2018).

Chemically induced alterations of modified bases were first used for DNA modification detection, and later adapted for RNA modifications too. One of the better-known methods is the Michael addition to cytidines leading to their deamination, alias bisulfite treatment and sequencing (Schaefer *et al.*, 2008). The method is based on the inertness of $5^{m}C$ to deamination, leading to preserved Cs at modified locations.

With the development of direct RNA sequencing, the explicit detection of modified nucleotides became possible. At the moment ONT manufactures the only platforms that are capable to sequence the RNA directly. In theory, modified nucleotides produce a distinct and characteristic electric signal when sliding through the nanopore; however, the calling of these bases is problematic, as signal recognition is done by machine learning algorithms, which need a ground truth. This requires the synthetic or *in vitro* production of sequences with specifically modified bases at known position used for the training of the algorithm (Stoiber *et al.*, 2016).

3.6. Examples of viral transcriptome sequencing

Our group sequenced and analysed several human (Balázs *et al.*, 2017; Tombácz *et al.*, 2017, 2019; Prazsák *et al.*, 2018) and non-human (Tombácz *et al.*, 2014, 2016, 2018; Oláh *et al.*, 2015; Moldován *et al.*, 2018; Csabai *et al.*, 2019) pathogenic and endogenous (Moldován, Balázs, *et al.*, 2017; Szűcs *et al.*, 2017) virus genomes and transcriptomes in the previous six years, using both NGS and LRS technologies. Here I present the broader biology of three of these viruses, with the intent of focusing on their transcript analysis in the remaining part of the thesis.

The HSV-1

Herpes simplex virus type 1 is one of the most studied viruses. It belongs to the *Alphaherpesvirinae* subfamily of the *Herpesviridae* family. Its relatives are widespread among vertebrate groups, from fish and amphibians (family *Alloherpesviridae*) (Hanson, Dishon and Kotler, 2011), through reptiles (Hoff and Hoff, 1984) and birds (Kaleta, 1990) to mammals (Pomeranz, Reynolds and Hengartner, 2005; El-Mayet *et al.*, 2019; Hussey, 2019). Cold sores are among the most common symptoms of HSV-1 infection, with recurring blisters mainly on the lips, caused by reactivation from viral latency. In immunocompromised patients, HSV-1 may cause acute encephalitis. Herpesviruses' ability to establish lifelong latency within its host organisms significantly contributed to their evolutionary success: according to WHO's estimates, more than 3.7 billion people under the age of 50 are infected with HSV-1 worldwide (Looker *et al.*, 2015).

The 152 kbp long linear, double-stranded DNA genome of the HSV-1 is enclosed into an icosahedral capsid, wrapped into a lipid envelope. The viral genome consists of a Unique Long and a Unique Short region, both flanked by inverted repeat region (IRL and IRS) (Macdonald *et al.*, 2012) (**Figure 2.**). During lytic infection the 72 genes are transcribed in three kinetic classes: immediate-early (IE), early (E) and late (L), by the host cell's RNA polymerase (Harkness, Kader and DeLuca, 2014). The transcription of IE genes is promoted by VP16, a protein carried by the capsid (Batterson and Roizman, 1983). This results in cascade-like gene expression, whereas the transcription of E genes is dependent on the presence of IE, while the transcription of L genes on the presence of E genes (Honess and Roizman, 1975). Genes of the three kinetic classes share common promoters. IE and E genes for the example have a TATA box upstream their TSS, which serve as a recognition site for VP16, and cellular transcription activators. While the expression of L genes is strongly impacted by the presence of an initiator sequence at their +1 site, besides their TATA promoter (Guzowski and Wagner, 1993). Previous studies of HSV-1 revealed 89 mRNAs (Rajcáni, Andrea and Ingeborg, 2004), 10 ncRNAs (Hu *et al.*, 2016) and 18 miRNAs (Du *et al.*, 2015), with one of our previous studies uncovering 142 additional transcripts (Tombácz *et al.*, 2017). The highly compact nature of the viral genome, with several overlapping transcript isoforms, present a challenge for NGS as shown previously. In this work, I will use our results on the HSV-1 transcriptome to demonstrate the suitability of LRS technologies for transcript isoform detection and annotation.

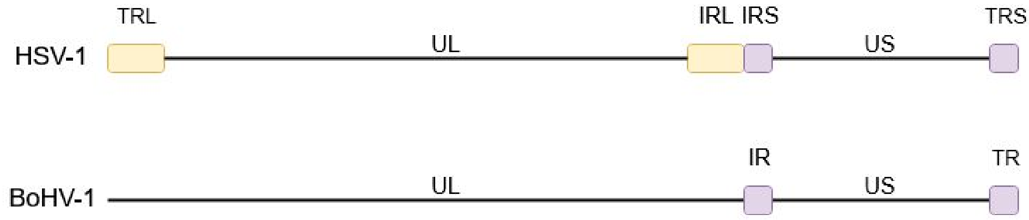


Figure 2. The genomic organization of HSV-1 and BoHV-1. Black lines represent the two main genomic regions, the Unique Long (UL) and Unique Short (US) region; rectangles represent the repeat regions: the Terminal Repeat of the UL (TRL), and of the US (TRS) and the Internal Repeat of the UL (IRL) and the US (TRS).

The BoHV-1

A non-human pathogenic relative of the HSV-1 is the *Bovine alphaherpesvirus type 1* (BoHV-1), affecting feedlot and dairy herds all over the world. The virus causes respiratory and fertility-related symptoms in cattle and is the main causative agent of infectious bovine rhinotracheitis (IBR) (Muylkens *et al.*, 2007). Similarly to other herpesviruses, its primary replication occurs in the mucosal surfaces of the host, after which the virions ascend towards the central nervous system, preferentially through the trigeminal nerve, the latent infection usually developing in the trigeminal ganglion (Gerdts *et al.*, 2000). The BoHV-1's 135 kbp long double-stranded DNA genome has a high, 72% GC content, and is comprised of two unique sequences, the UL and the US, the latter being bracketed by inverted repeat regions (the internal INR and the terminal TRL repeats) (d'Offay, Fulton and Eberle, 2013) (**Figure 2.**). The 73 genes of the virus are mostly homologous to other alphaherpesviruses' genes and their nomenclature follows that of the HSV-1's. Two peculiar exceptions from this are the circ gene, which is common in *Varicelloviruses*, and the ul0.5, which is specific to BoHV-1 (Delhon *et al.*, 2003).

The BoHV-1 infection starts with the viruses attachment to the cell surface structures through two of its glycoproteins gB and gC (Li *et al.*, 1995, 1996). This is followed by the stable binding of the virion to gD to a wide range of cellular receptors (Campadelli-Fiume *et al.*, 2000). After attachment, the viral envelope fuses with the cell membrane, and the viral particles are carried by dynein motor complexes towards the nuclear pores (Döhner *et al.*, 2002). There are several important proteins shed during the centripetal cytosolic transport, like the tegument protein encoded by ul41 (vhs: virion host shutoff), which plays a major role in early protein synthesis of BoHV-1, by shifting the host's protein synthesis towards the viral proteins (Hinkley *et al.*, 2000). The VP16 tegument protein (or trans-inducing factor α) (gene ul48) is responsible for the initiation of the expression of the immediate-early (IE) genes. The viral DNA is ought to circularize in the nucleus before the start of transcription (Fraefel *et al.*, 1993). The kinetics of the BoHV-1 gene expression are similar to other alphaherpesviruses', giving rise to immediate-early (IE), early (E), and late (L) genes. IE genes regulate the viral gene expression, E genes initiate and direct replication, while L genes are involved in the virion morphogenesis. During the IE phase BICP0, BICP4, BICP22 and the CIRC transcripts are expressed. The products of the first three serve as activators for all viral promoters, (Saydam *et al.*, 2004, 2006) while the function of CIRC, which translates into a small, myristylated protein is unclear (Fraefel, Ackermann and Schwyzer, 1994). Recently, Pokhriyal et al. (Pokhriyal *et al.*, 2018) found that three other genes (ul21, ul33 and ul34) are also expressed during the IE phase of the infection in cycloheximide-treated cells. These genes have a TATA box but lack the OCT-1 binding site characteristic of conventional IE genes. As the proteins of these genes conventionally are late products (UL21 is a tegument protein; UL3 directs DNA packaging, while UL34 the nuclear egress), their function during the early phases of the viral life cycle is unknown.

In my thesis, I intend to show the use of single-molecule LRS technologies in viral RNA quantification and the characterization of viral gene expression. Additionally, I want to demonstrate the suitability of these techniques in isoform quantification.

The AcMNPV

The *Autographa californica multiple nucleopolyhedrovirus* (AcMNPV) is an insect virus from the *Baculoviridae* family (Blissard and Rohrmann, 1990). It is widely used in recombinant protein expression systems as a gene delivery vector and as a biopesticide (Hu, 2005). Insects get infected through their digestive systems, by consuming vegetation contaminated by viral occlusion bodies, which are resistant protein structures containing the virus. Following ingestion, the virus attaches and enters the endothelial cells of the host, from where it infects the insect's whole system through budding (Rohrmann, 2019). Its 133 kbp double-stranded circular DNA genome harbours 156 tightly-spaced open reading frames. As our group and others have demonstrated, the closeness of the ORFs causes overlapping among many of the AcMNPV's transcripts (Chen *et al.*, 2013; Moldován *et al.*, 2018). This also encumbers the study of transcript isoforms using SRS technologies. Viral gene expression is grouped into three distinct phases: early (E), late (L) and very late (VL). The promoter of early genes is homologous to the canonical TATA box and is recognised by the host's transcription machinery, with a frequent transcriptional initiation site being a CAGT (the +1 nt is underlined) sequence, which is common among insects (Kogan, Chen and Blissard, 1995). L and VL genes are transcribed by the virus' own RNAP and start at a very conserved TAAG (the +1 nt is underlined) motif also known as the late initiation sequence (LIS) (Garrity, Chang and Blissard, 1997).

Using the dRNA-seq data obtained from our study of the AcMNPV (Boldogkői *et al.*, 2018) I will show the potential usage of LRS to detect RNA nucleotide modifications on native RNA molecules.

4. Aims

In this work, I will demonstrate the potential of long-read sequencing technologies in the analysis of viral transcriptomes, using three viruses studied by our team. Taking as an example the transcriptome of the HSV-1 I show the capability of third-generation sequencing technologies to explore the isoform diversity of viral RNAs and to compare the transcriptome structurally, the BoHV-1 will be used to indicate their suitability in the analysis of viral gene expression, while for the AcMNPV I will focus on their ability to detect RNA modifications. Although these will be my points of convergents, for the sake of demonstration I will use some parts of the study on AcMNPV in the chapter exploring isoform detection and annotation.

5. Materials and methods

5.1. Cells and viruses

Every step of this section was carried out according to the relevant guidelines and regulations for virus propagation and decontamination.

HSV-1

The strain KOS of HSV-1 was propagated on the Vero cell line, an immortalized kidney epithelial cell line isolated from the African green monkey (*Chlorocebus sabaeus*). Cells were cultivated in Dulbecco's modified Eagle medium supplemented with 10% fetal bovine serum (Gibco Invitrogen) and 100 μl penicillin-streptomycin 10K/10K mixture (Lonza)/ml in a 5% CO_2 atmosphere at 37°C. The viral stocks were prepared by infecting rapidly-growing semi-confluent Vero cells at a multiplicity of infection (MOI) of 1 plaque-forming unit (pfu)/cell, followed by incubation until a complete cytopathic effect was observed. To harvest the virions the infected cells were frozen and thawed three times, and then centrifuged at 10,000 ×g for 15 min. For the sequencing studies, Vero cells were infected with MOI = 1 and incubated for 1 h. This was followed by removal of the virus suspension and phosphate-buffered saline (PBS) washing step. Next, the cells were supplied with a fresh culture medium and were then incubated for 1, 2, 4, 6, 8, 10, 12, or 24 h in the media described above.

BoHV-1

Madin Darby Bovine Kidney (MDBK) cells were incubated at 37°C in a humidified incubator with 5% CO_2, and cultured in Dulbecco's modified Eagle medium (DMEM) supplemented with 5% (v/v) fetal bovine serum, 100 U/mL penicillin, and 100 μg/mL streptomycin. Confluent cultured cells were infected in three replicates with the Cooper isolate (GenBank Accession # JX898220.1) of Bovine Herpesvirus 1.1 (BoHV-1) at MOI = 5 IU. After one hour of incubation at 4°C, cells were washed with PBS and 15 ml DMEM to each culture. Cells were incubated for further time points at 37°C, 5% CO_2. At each p.i. time point (1, 2, 4, 6, 8, 12 h) the infected cultures were harvested and stored at -80°C until use.

AcMNPV

The Sf9 epithelial cell line (derived from the parental *Spodoptera frugiperda* cell line) was used for the propagation of a LacZ expressing recombinant *Autographa californica multiple nucleopolyhedrovirus* (AcMNPV) in Sf-900 II SFM insect cell culture medium (Thermo Fisher Scientific). Cultivation of the infected cells was carried out in Corning Spinner Flasks (Sigma Aldrich/Merck) at 26 °C, the speed of the shaker was set to 70 rpm. The LacZ gene is inserted to the promoter region of the polh gene (βgal-AcMNPV). The virus stock was obtained from SOLVO Biotechnology Inc. (Szeged, Hungary). Cells were infected with a multiplicity of infection of 2 plaque-forming unit/cell. After infection, cells were incubated for 0, 1, 2, 4, 6, 16, 24, 48 or 72 h, and after each interval 5 ml of sample was pipetted out of the flask, centrifuged at 2,000 rpm at 4°C, followed by washing with PBS and centrifuged again by setting the same parameters. Samples were stored at −80°C until use.

5.2. RNA isolation and library preparation

RNA isolation was carried out with the NucleoSpin® RNA kit (Macherey-Nagel) according to the manufacturer's guidance for all three viruses. In short, infected cells were collected by centrifugation and the cell membrane was disrupted by the addition of lysis buffer (derived from the kit). Genomic DNA was digested by treatment with RNase-free rDNase solution (supplied with the kit). Samples were eluted in a total volume of 50 μl nuclease-free water. To eliminate residual DNA contamination samples were treated with TURBO DNA-free Kit (Thermo Fisher Scientific). The RNA concentration was measured using the Qubit Fluorometer (v.2.0 for HSV-1 and AcMNPV and v4.0 for BoHV-1), using the Qubit RNA BR Assay Kit (Thermo Fisher Scientific).

Three RNA batches were then prepared for HSV-1 and AcMNPV and two for BoHV-1, as follows: total RNA samples of the HSV-1 and BoHV-1 were pooled per virus and were subjected to rRNA depletion for the random primed sequencing; the poly(A)$^+$ RNA fraction of the samples of HSV-1, BoHV-1 and AcMNPV were selected for polyA-sequencing, while a third total RNA sample was put aside from the HSV-1 and the AcMNPV and was pooled per virus for the 5' cap selection protocol.

Sequencing libraries for the PacBio RSII and Sequel platforms

The Clontech SMARTer PCR cDNA Synthesis Kit was used for cDNA preparation according to the PacBio Isoform Sequencing (Iso-Seq) protocol. For the analysis of relatively short viral RNAs, the 'No-size selection' method was used and samples were run on the RSII and Sequel platforms. The reverse transcription (RT) reactions were primed by using the oligo(dT) from the SMARTer Kit. However, we also used random primers for a non-size selected sample to detect non-polyadenylated RNAs. The cDNAs were amplified by the KAPA HiFi Enzyme (KAPA Biosystems), according to PacBio's recommendations. The SMRTbell libraries were generated using PacBio Template Prep Kit 1.0. For binding the DNA polymerase and annealing the sequencing primers, the DNA/Polymerase Binding Kit P6-C4 and v2 primers, as well as the Sequel Sequencing Kit and v3 primers were used for the RSII and Sequel sequencing respectively. The polymerase-template complexes were bound to MagBeads with the PacBio MagBead Binding Kit.

Poly(A) selected cDNA sequencing libraries for the ONT MinION platform

The 1D Strand switching cDNA by ligation protocol from ONT was applied for sequencing HSV-1 cDNAs on the MinION platform. The ONT Ligation Sequencing Kit 1D (for HSV-1 and AcMNPV: SQK-LSK108; for BoHV-1: SQK-LSK109) was used for the library preparation with the recommended oligo(dT) primers, or custom-made random hexamer oligonucleotides, as well as the SuperScipt IV enzyme for the RTs. The cDNA samples were subjected to PCR reactions with KAPA HiFi DNA Polymerase (Kapa Biosystems) and Ligation Sequencing Kit Primer Mix (part of the 1D Kit).

The NEBNext End repair/dA-tailing Module (New England Biolabs) was used for end-repair, whereas the NEB Blunt/TA Ligase Master Mix (New England Biolabs) was utilized for the adapter ligation. The enzymatic steps (e.g.: RT, PCR, and ligation) were carried out in a Veriti Cycler (Applied Biosystems) according to the 1D protocol. The Agencourt AMPureXP system (Beckman Coulter) was used for the purification of samples after each enzymatic reaction. The quantity of the libraries was checked using the Qubit Fluorometer (v.2.0 for HSV-1 and AcMNPV and v4.0 for BoHV-1) and the Qubit (ds)DNAHS Assay Kit (Life Technologies).

Cap selection followed by cDNA sequencing on the ONT MinION platform

The TeloPrime Full-Length cDNA Amplification Kit (Lexogen) was used for generating cDNAs from 5’ capped polyA$^{(+)}$ RNAs. RT reactions were carried out with oligo(dT) primers (from the kit) (for HSV-1 and AcMNPV) or random hexamers (custom made) (for HSV1 only) using the enzyme from the kit. A specific adapter (capturing the 5’ cap structure) was ligated to cDNAs (25°C, overnight), then the samples were amplified by PCR using the Enzyme Mix and the Second-Strand Mix from the TeloPrime Kit. The reactions were performed in a Veriti Cycler and the samples were purified on silica membranes (TeloPrime Kit) after the enzymatic reactions. The Qubit 2.0 and the Qubit dsDNA HS quantitation assays (Life

Technologies) were used for measuring the concentration of the samples. A quantitative PCR reaction was carried out for checking the specificity of the samples using the Rotor-Gene Q cycler (Qiagen) and the ABsolute qPCR SYBR Green Mix from Thermo Fisher Scientific. A gene-specific primer pair (HSV-1 *us9* gene, custom made) was used for the test amplification. The PCR products were used for ONT library preparation and sequencing. The end-repair and adapter ligation steps were carried out as described in the previous section.

Direct cDNA library for the ONT MinION platform

ONT's Direct cDNA (dcDNA) sequencing protocol together with the Direct cDNA Sequencing Kit (SQK-DCS109) was used to generate non-amplified cDNA libraries for the MinION. Briefly: the Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific) together with the SSP and VN primers (supplied in the kit) were used for the first-strand synthesis. The degradation of the RNA was performed with RNase Cocktail Enzyme Mix (Thermo Fisher Scientific), followed by the second strand synthesis using 2x LongAmp Taq Master Mix (New England Biolabs). The NEBNext End repair/dA-tailing Module (New England Biolabs) was used for the end-repair, whereas the NEB Blunt/TA Ligase Master Mix (New England Biolabs) was utilized for the adapter ligation. The enzymatic steps (e.g.: RT, RNA degradation, and ligation) were carried out in a Veriti Cycler (Applied Biosystems) according to the protocol. The RNA-DNA hybrid and the cDNA was purified between each step by using Agencourt AMPure XP magnetic beads (Beckman Coulter), treated with RNaseOUT Recombinant Ribonuclease Inhibitor (Thermo Fisher Scientific). The quantity of the libraries was checked using the Qubit Fluorometer (v4.0) and the Qubit (ds)DNA HS Assay Kit (Life Technologies).

Direct RNA library for the ONT MinION platform

The ONT's Direct RNA sequencing protocol and the ONT Direct RNA Sequencing Kit (for HSV-1 and AcMNPV: SQK-RNA001; for BoHV-1: SQK-RNA002) were used to prepare the native RNA sequencing library. The first-strand cDNA was synthesized by SuperScript IV Reverse Transcriptase (Thermo Fisher Scientific) using an RT adapter with T_{10} nts and the mRNA mix. Adapters (supplied by the kit) were ligated using T4 DNA ligase (New England Biolabs). The RNA-DNA hybrid was purified between each step by using Agencourt AMPure XP magnetic beads (Beckman Coulter) and treated with RNaseOUT Recombinant Ribonuclease Inhibitor (Thermo Fisher Scientific). Sample concentration was determined using a Qubit Fluorometer (v.2.0 for HSV-1 and AcMNPV and v4.0 for BoHV-1) and the Qubit DNA HS Assay Kit (Thermo Fisher Scientific).

Barcoding of the ONT MinION libraries

Samples intended for expression analysis from HSV-1 and AcMNPV and every sample from the BoHV-1 were barcoded prior to the end repair step. The 1D PCR barcoding (96) genomic DNA protocol was followed for HSV-1 and AcMNPV cDNA time separated samples and the BoHV-1 cDNA pooled libraries using ONT's PCR Barcoding (96) kit, while the Direct cDNA Native Barcoding Protocol for the BoHV-1 dcDNA libraries, using ONT's Native Barcoding (12) Kit.

The 1D PCR barcoding (96) genomic DNA protocol in short: 10 µl of samples were mixed with 6.5 µl of Barcode Adapter (from bc25 to bc33; supplied in the kit) and 17 µl Blunt/TA Ligase Master Mix (New England Biolabs). After 10 min incubation at room temperature, the samples were purified by using AMPure XP magnetic beads (Beckman Coulter). PCR amplification was carried out using KAPA HiFi DNA Polymerase (Kapa Biosystems) and 1 µl from one of the PCR Barcodes (supplied in the kit), as recommended by the 1D PCR barcoding protocol.

The Direct cDNA Native Barcoding Protocol in short: 22.5 µl of end-prepped cDNA was mixed with 2.5 µl of Barcode Adapter (from bc1 to bc12; supplied in the kit) and 25 µl of Blunt/TA Ligase Master Mix(New England Biolabs). The mixture was left to incubate at room temperature for 10 min, after which it was purified using AMPure XP magnetic beads (Beckman Coulter).

The experimental layout is outlined in **Figure 4**.

5.3. Sequencing

For sequencing on the PacBio platforms, libraries were loaded onto the RSII SMRT Cell 8Pac v3 or Sequel SMRT Cell 1M. The movie length was 240 or 360 min *per* SMRT Cell for the RSII, while 600-min for the Sequel run.

For the ONT libraries, the samples were loaded on multiple R9.4 SpotON Flow Cells. To avoid cross-talk between the barcoded samples, libraries from the mock-infected, 1 and 2 h p.i. time points were sequenced on separate flow cells from all the other samples.

5.4. Read processing and analysis

The PacBio reads were base called and subreads were assembled using the pbccs tool (https://github.com/PacificBiosciences/ccs). The MinION reads were base called using either Albacore v.2.3.4 or Guppy v.3.4.5. Reads of HSV-1 were aligned to the genome from the NCBI Nucleotide database with the accession number of X14112.1, reads of BoHV-1 to AJ004801.1 while reads of AcMNPV to NC_001623.1 using Minimap2 (Li, 2018) with the `-ax splice -Y -C5` parameters. For every virus, the reference genome was duplicated before alignment to allow mapping of reads crossing the genomic junction. For the annotation and quantitation of TSSs, TESs, introns and transcript isoforms we used the LoRTIA software suite v.0.9 (https://github.com/zsolt-balazs/LoRTIA) developed in our lab. For the PacBio reads LoRTIA was ran with the parameters `-5 AGAGTACATGGG --five_score 16 --check_in_soft 15 -3 AAAAAAAAAAAAAAA --three_score 18 -s poisson -f True`, for the amplified MinION reads with `-5 TGCCATTAGGCCGGG --five_score 14 --check_in_soft 15 -3 GAAGATAGAGCGACA --three_score 14 -s poisson -f True`, while for the dcDNA and dRNA with `-5 TGCCATTAGGCCGGG --five_score 14 --check_in_soft 15 -3 AAAAAAAAAAAAAAA --three_score 14 -s poisson -f True`.

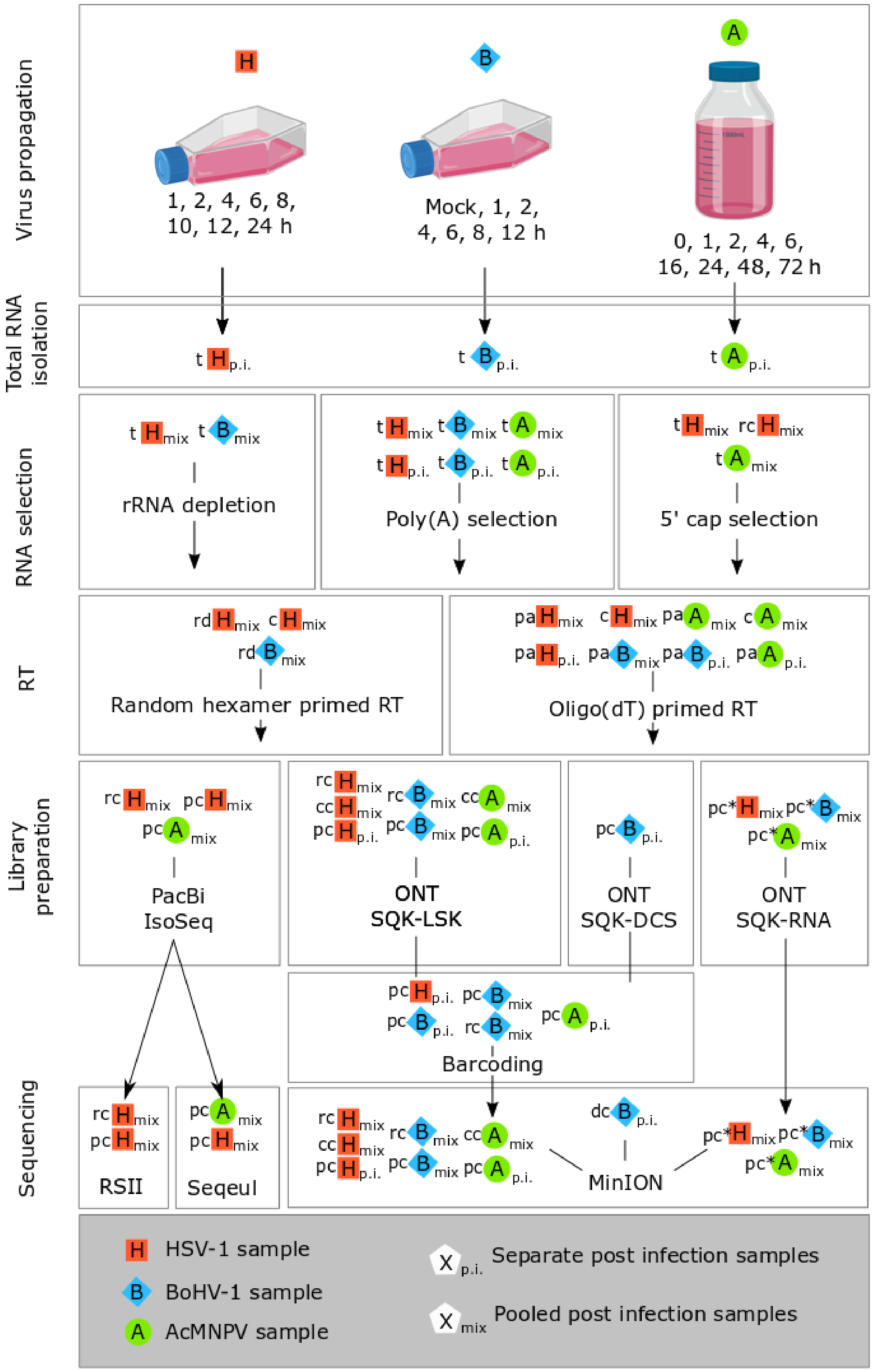


Figure 4. Experimental layout. Three cell lines were infected with either HSV-1, BoHV-1 or AcMNPV. Total RNA was isolated from the cells at different post-infection time points and was either pooled per virus or left "as is". Three RNA selection techniques were applied on the resulting samples followed by reverse transcription either by random hexamer primers or oligo(dT) primers. Sequencing

libraries were prepared for two PacBio platforms and one ONT platform. The following abbreviations apply for the samples: t: total RNA; rc: random primed cDNA; rd: rRNA depleted RNA; c: 5' cap selected RNA; pa: poly(A) selected sample; pc: oligo(dT) primed cDNA; pc*: oligo(dT) primed first strand + RNA; cc: 5' cap selected cDNA.

TSSs and TESs were considered if they were present in at least three samples, while an intron was considered only if it was detected by both dRNA and cDNA sequencing. We applied these criteria to exclude most of the false isoforms produced by RNA degradation or during RT and PCR.

The resulting annotations were visualized using IGV (Thorvaldsdóttir, Robinson and Mesirov, 2013), typed and named, according to the position of their TSS and TES relative to the most abundant or previously annotated isoform, and the presence or absence of introns. Cis-regulatory sequences were detected using our in-house script and were visualized using WebLogo v.3.0 (Crooks *et al.*, 2004)

The quantitative analysis of BoHV-1 genes and transcript isoforms started with read count-based filtering. Read counts of the isoforms that were present in at least 10 copies were normalization using a modified version of the Median Ration Normalization of the DESeq2 software suite (Wu *et al.*, 2019), and their average over the biological replicates was used for further analysis.

RNA base modification detection of the AcMNPV dRNA dataset was performed using Tombo v1.5 (Stoiber *et al.*, 2016).

6. Results

6.1. Sequencing and mapping statistics

The sequencing of the HSV-1 transcriptome yielded a total of 16,163,711 reads, 1,624,881 mapping to the viral genome. The sequencing of the BoHV-1 transcriptome yielded 26,663,352 total reads, 3,387,951 mapping to the viral genome. While the sequencing of AcMNPV yielded a total of 13,165,215 reads, 1,034,425 mapping to the viral genome. Detailed read statistics of each library can be found in **Table 2**. The highest average mapped read lengths were produced by the PacBio platform, followed by the MinION direct cDNA (dcDNA) and dRNA sequencing techniques (**Figure 5. Panel A.**). This can be due to a strong size-selection bias for longer reads during the library preparation and sequencing for both the RSII and Sequel machines, while the

MinION can sequence molecules in a wider length-range. The PacBio platform currently has higher accuracy compared to the MinION; however, this is only true for the circular consensus reads (or Reads of Insert; ROI) produced during base calling (**Figure 5. Panel B.**). The so-called subreads, from which the ROIs are assembled have a similar quality to the MinION reads (not shown in the figure). The MinION platform suffers from the inability to sequence long homopolymer runs, caused by the uniform signals in these regions.

Table 2. Detailed read statistics of the sequencing libraries. Rows in red: PacBio data, rows in blue: MinION data. The following abbreviations were used for library types: o(dT): RT primed with oligo(dT) primers; rand: RT primed with random hexamer primers; cap: 5' cap selection performed during RT; dRNA: direct RNA sequencing; cDNA: sequencing of amplified cDNAs; dcDNA: sequencing of non-amplified cDNAs.

		Read count		Read length (nt) (Mean±SD)		Mismatch %	Insertion %	Deletion %
		Raw	Mapped	Raw	Mapped			
HSV-1	**RSII o(dT) cDNA**	232428	38476	1233 ± 623	1380 ± 517	1	0.4	1
	RSII rand cDNA	3921	496	920 ± 493	932 ± 460	1	0.2	1
	Sequel o(dT) cDNA	263006	80061	1970 ± 910	1931 ± 764	1	1	0.4
	dRNA	614667	24671	448 ± 561	928 ± 601	6	3	11
	Cap o(dT) cDNA	797571	41646	894 ± 729	762 ± 380	7	4	10
	Cap rand cDNA	3085128	8007	606 ± 700	418 ± 157	6	5	7
	Rand cDNA	1566744	39322	717 ± 555	831 ± 485	5	4	8
	o(dT) cDNA	2737762	65843	1028 ± 1103	1216 ± 715	7	5	11
BoHV-1	**dRNA**	916140	516294	970 ± 631	981 ± 702	4	3	5
	o(dT) cDNA	4717241	688617	1142 ± 817	677 ± 423	4	4	5
	Rand cDNA	5650274	352903	1053 ± 671	662 ± 398	4	4	4
	dcDNA	15379697	1830137	1439 ± 1160	1165 ± 870	4	6	5
AcMNPV	**Sequel pA cDNA**	47880	25371	2799 ± 1550	1519 ± 746	0.2	2	0.4
	dRNA	2139	2133	666 ± 412	665 ± 402	2	2	5
	Cap o(dT) cDNA	6862026	680006	726 ± 491	565 ± 282	3	3	4
	o(dT) cDNA	6253170	326915	841 ± 569	585 ± 434	4	5	5

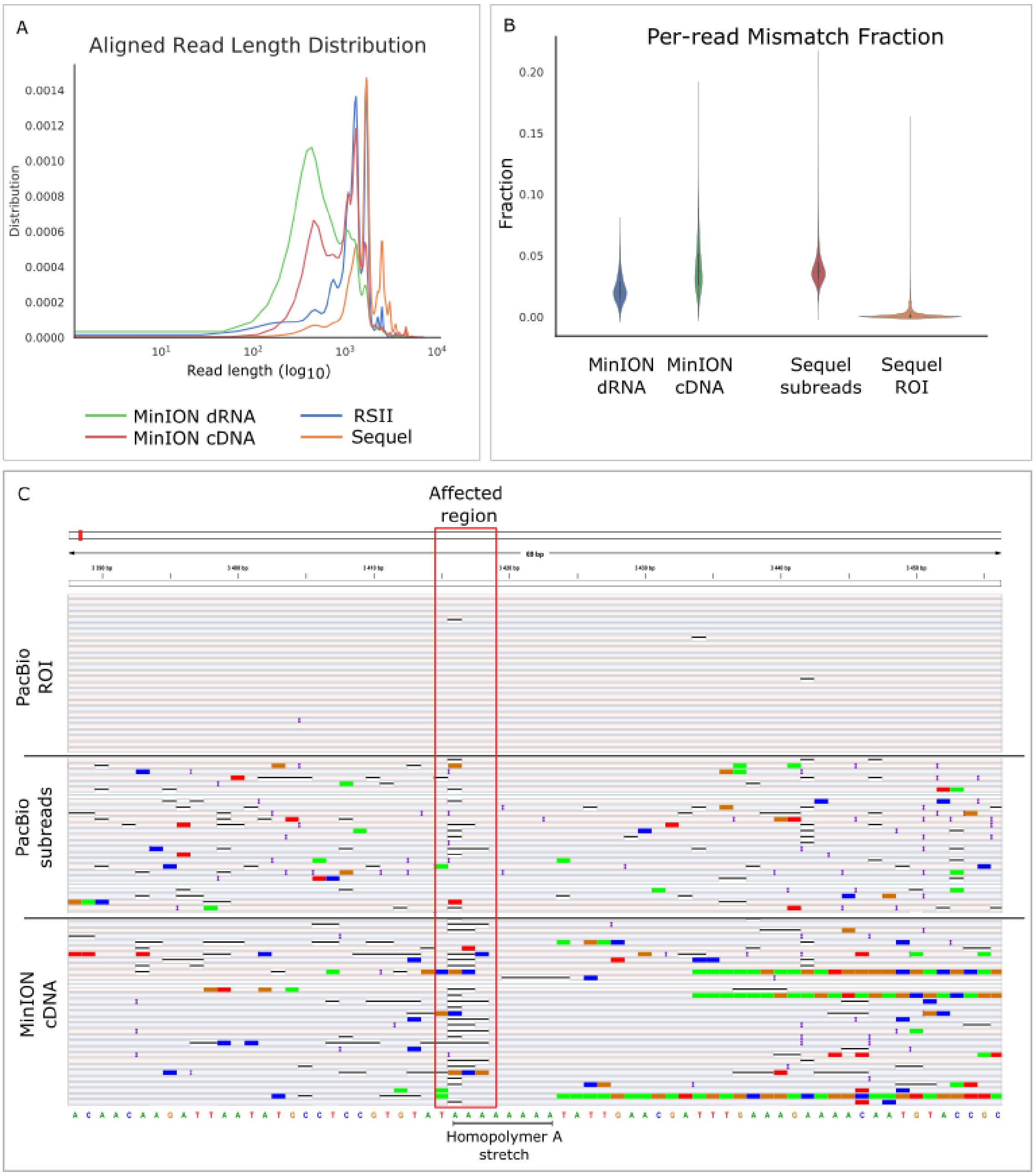


Figure 5. Read characteristics on PacBio and ONT platforms. A. The aligned read length distributions of ONT and PacBio platforms of HSV-1 reads. The RSII and Sequel platforms cover a narrower read length range than the MinION. B. Per-read mismatch fractions of ONT and PacBio platforms of AcMNPV reads. The MinION reads are noisier than Sequel consensus reads (ROIs), however individual Sequel subreads have the same amount of mismatches than MinION reads. C. The effect of homopolymer runs on reads from PacBio and ONT platforms. A homopolymer A stretch of the AcMNPV genome causing deletion in the MinION reads and Sequel subreads. However, this deletion is corrected by PacBio base caller during consensus generation and thus is not present in PacBio ROIs. Reads were visualised using IGV (Thorvaldsdóttir, Robinson and Mesirov, 2013).

These shortcomings are easily circumvented when mapping to a good quality genomic sequence, and thus are not restraining the use of MinION sequencing data in transcript analysis.

6.2. Transcriptional start and end sites

Eukaryotic transcription initiation is controlled by trans- and cis-acting elements, the later located in the close proximity of the TSS, however the sequence of the transcriptional initiation region (INR) is ambiguous, in most of the cases with only one conserved adenine in the +1 position (Javahery *et al.*, 1994; Fukue *et al.*, 2004). Nevertheless, the AcMNPV uses a much more conserved INR for its late transcription accomplished by the viruses own RNAP. The late initiator sequence (LIS) consists of a TAAG motif, with the first adenine (underlined) being the first nucleotide of late transcripts (Garrity, Chang and Blissard, 1997; Chen *et al.*, 2013). Using this knowledge, we could test the accuracy of our TSS discovery pipeline. We found that RNAs of the AcMNPV in samples taken after 6h p.i. have an increasing tendency to start at a TAAG motif **(Figure 6.)**.

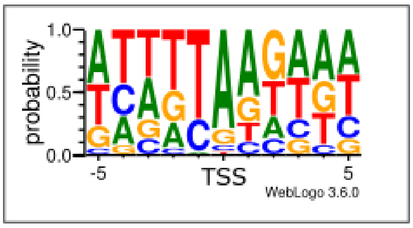


Figure 6. The sequence of the late transcriptional initiation site of the AcMNPV.

We also observed that a fraction of the reads is missing the sequencing adapter upstream from its 5’ end (in the soft clipped region). We hypothesized that these 5’ ends originate from mispriming, and thus represent artefacts produced during library preparation, and need to be discarded during TSS detection. To prove their artefactual nature we calculated the frequency of reads starting in a given position in the ±10 nt vicinity of the 5’ ends of the reads without the adapter and with the adapter in their 5’ soft clipped region. We found that the 5’ ends of the reads missing the sequencing adapter are scattered around the most abundant start position, while those with an adapter present usually start at the most abundant 5’ end position (**Figure 7 Panel A. and B.**).

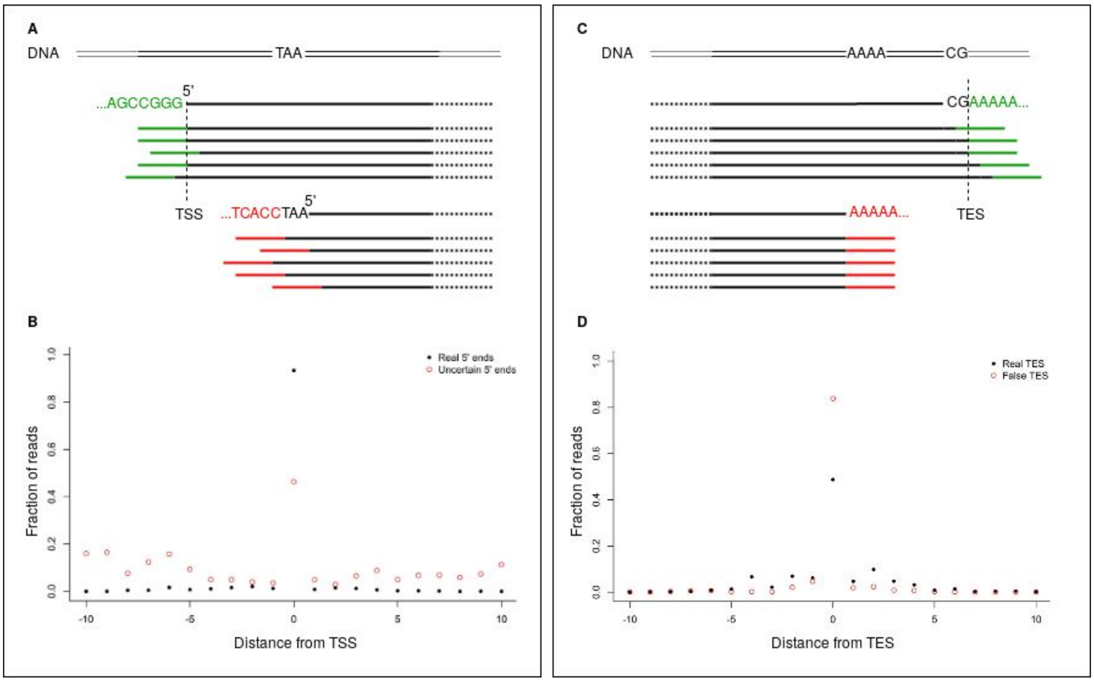


Figure 7. The source of false TSS and TES and their detection in the AcMNPV transcriptome. A. False TSSs form during library preparation by false priming or template switching. The variety of potential sequences causing false priming and the randomness of template switching causes high variance in the first nucleotide positions of questionable 5' ends. C. The formation of false TES caused by false priming of more than three adenines on the template. D. The false 3' end position is less variable than the natural variation of RNA cleavage, because of the sequence uniqueness giving rise to false priming.

The presence of promoters, conserved initiator regions and the variations of read start positions altogether suggest that our pipeline could distinguish TSSs from artefacts with good accuracy.

The termination of transcripts is regulated by more conserved cis-elements than their initiation. A polyadenylation signal (PAS) composed mainly of A-s and T-s can be found upstream in the close proximity of the cleavage site. The cleavage site itself has a tendency to be an A (usually preceded by a C), while downstream of the cleavage site a GU-rich region signals the end of the transcription. A-rich regions on RNAs can give rise to mispriming of the anchored oligo(dT) primer used during the RT, resulting in false TESs. We hypothesize that in cases where mispriming occurs the degenerate two nucleotide anchors together with the homopolymer T-s, cause a very strict false 3' end formation, compared to the natural variation in 3' formation of the cleavage and polyadenylation complex. We analysed the position of 3' ends of AcMNPV reads in the ±10 nt vicinity of the most abundant 3' ends and found that reads with a 3' end with more than 3 As tend to terminate in a single position more often than reads lacking a homopolymer A on its end (**Figure 7. Panel C. and D.**).

Using LRS and the LoRTIA software suite, we were able to annotate the transcriptional start sites (TSSs) and transcriptional end sites (TESs) of the HSV-1 transcripts with base-precision. We detected a total of 1,677 putative TSSs and 162 putative TESs. To further minimize false-positives we only accepted those TSSs or TESs as real that were present in at least three independent samples for the longest detected isoform, and five independent samples for the shorter isoforms. This stringent filtering resulted in 537 TSSs and 77 TESs of the HSV-1, which were further analysed. We performed promoter analysis on the -150 to +1 nt upstream region of the TSSs. We found that 4% of the TSSs were preceded by a CAAT box at a mean distance of 113.428 nt (σ = 15.471) and 11% by a TATA box at a mean distance of 30.373 nt (σ = 2.058). Our analysis of the ±5 nt vicinity of the TSSs shows that the most abundant initiator sequences of the HSV-1 were homologous with the eukaryotic initiators (Lim, 2004; Xi *et al.*, 2007) (C$\underline{\text{A}}$G, TSS position underlined) and are expressed in both early and late time points of the infection. However, in the late time points the fraction of Gs in the -1,+1 and +2 positions increased dramatically, resulting in a C$\underline{\text{G}}$G sequence (**Figure 8. Panel A. and B.**), which is part of the VP5 promoter discovered earlier (Huang *et al.*, 1996). These late initiators were present in 66% of the TSSs; however, they made up only 2.8% of the reads starting positions, possibly meaning a much lower rate of successful transcription initiations or a higher degradation rate for these transcripts.

The analysis of the ±50 nt surrounding of the TESs showed that 77% had a canonical polyadenylation signal (PAS) upstream at an average distance of 21.779 nt (σ = 5.558) and were present in both early and late samples. These TESs also had a canonical C/A cleavage site and a GU-rich downstream element (DSE) commune in eukaryotes. In contrast, TESs expressed only in the late phase of the infection tend to have no canonical PAS, cleavage site or DSE (**Figure 8. Panel C. and D.**). Also, these late TES were in low abundance, present in only 0.1% of all reads.

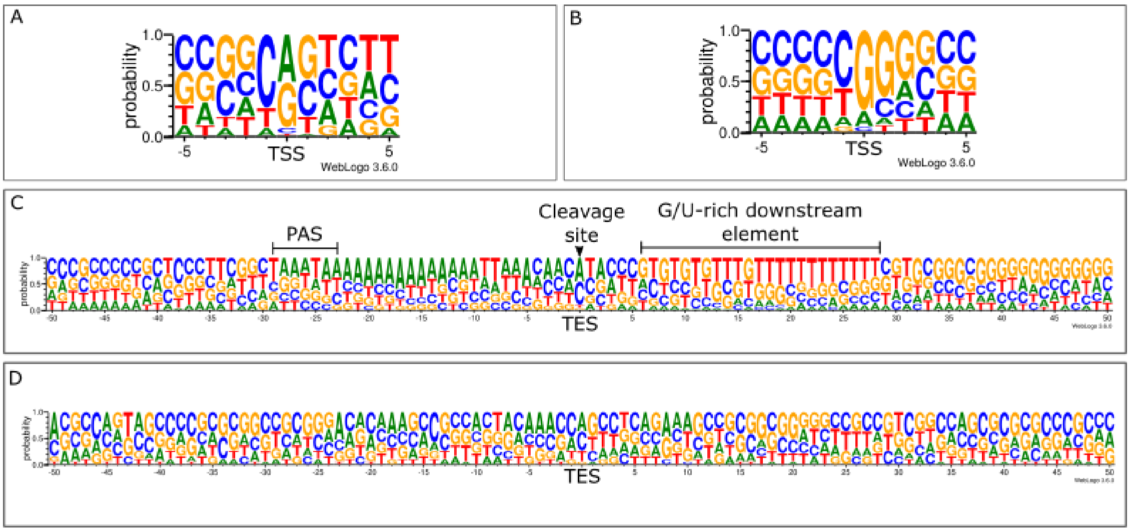


Figure 8. Initiators and polyadenylation signals of HSV-1. A. The initiator region of early and some of the late transcripts. B. The initiator region of late transcripts. C. The polyadenylation signal, cleavage site and downstream element of early and some of the late transcripts. D. No PAS or DSE could be detected for late transcripts.

6.3. Introns

As shown in the Introduction, gaps can form during sequencing library preparation, which can be falsely identified as introns during the analysis. To exclude these artefacts we applied the following criteria for gaps to qualify as introns:

1. introns needed to bare one of the most commune eukaryotic canonical splice junction sequences (GT/AG, GC/AG or AT/AC)
2. the splice junctions of the introns needed to represent at least one-thousands of the coverage
3. introns needed to be present in the dRNA-seq and in both MinION cDNA and PacBio cDNA samples

Following these criteria we identified 182 introns out of the 379 gaps detected by LoRTIA in the HSV-1 transcriptome, 155 carrying a canonical GT/AG, 25 a GC/AG and 2 an AT/AC splice junctions.

6.4. Annotating the viral transcriptome

The power of LRS lies in its capability to detect full-length transcripts, resulting in the discovery of a complex transcriptional landscape in many viruses (Balázs *et al.*, 2017; Moldován, Balázs, *et al.*, 2017; Moldován, Tombácz, *et al.*, 2017; Moldován *et al.*, 2018; Depledge *et al.*, 2019; Tombácz *et al.*, 2019).

We grouped novel transcript isoforms according to their TSS, TES and splice junction positions and their orientation.

Putative mRNAs

We found a total of 182 novel putative mRNAs, which are the 5' truncated isoforms of previously annotated transcripts. These transcripts all contain an in-frame ORF, which if translated, yield an N-terminally truncated isoform of the previously described protein.

Novel non-coding transcripts

Novel isoforms lacking an ORF were annotated as non-coding RNAs. We detected 18 ncRNAs 8 being novel. A previous study (Zhu *et al.*, 1999) suggested, that 0.7-kb LAT is not expressed in strain KOS of HSV-1, however, we could detect this transcript **(Figure 9. Panel A.)**. We detected antisense RNA activity in several genes (*rl1, rl2, ul1, ul2, ul4, ul5, ul10, ul14, ul15, ul19, ul23, ul29, ul31, ul32, ul36, ul37, ul39, ul42, ul43, ul44, ul49, ul50, ul53, ul54, us4, us5, us8*). These antisense transcripts were present in low abundance. According to our results, intergenic RNA expression is frequent in the HSV-1 transcriptome. We could identify three novel intergenic ncRNAs. The IGEN-1 coterminal with UL27-AT **(Figure 9. Panel B.)**, the IGEN-2 on the outer termini of the unique long region, and the RL2-LAT-UL1-2-3, a long isoform of LAT spanning multiple oppositely oriented genes. We found an intra-intronic ncRNA, the RL2I, which was expressed in low abundance (**Figure 9. Panel C.**).

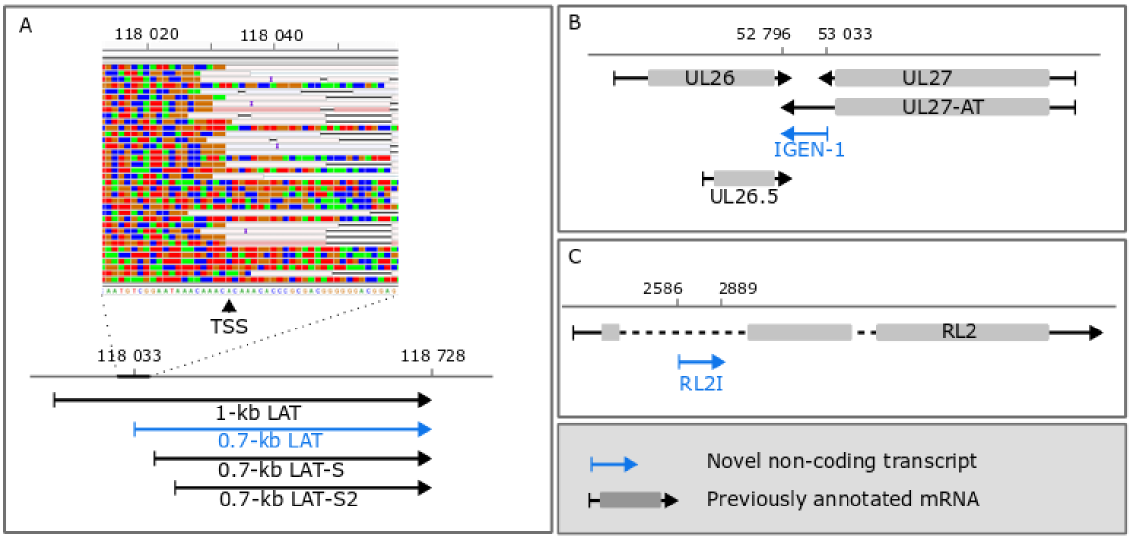


Figure 9. Non-coding RNAs of the HSV-1 transcriptome. A. The TSS of the 0.7-kb LAT visualised using IGV and the organization of the LAT region. B. The position of the intergenic IGEN-1 ncRNA. C. The position of the novel intra-intronic RL2I transcript.

TSS and TES isoforms

Genes with alternative promoters or PAS can produce longer or shorter UTR isoforms. We detected 53 genes producing TSS isoforms, 51 of which are protein-coding while 2 (the 0.7-kb lat and the rs1) non-coding. We detected that TSS isoforms are common amongst HSV-1 RNAs but expressed at low rates, however some genes like ul10 and ul19 express a high variety of 5’ UTR isoforms in relatively high abundance. The length of a 5’ UTR is affecting the number of upstream ORFs (uORFs) on mRNAs (**Figure 10**).

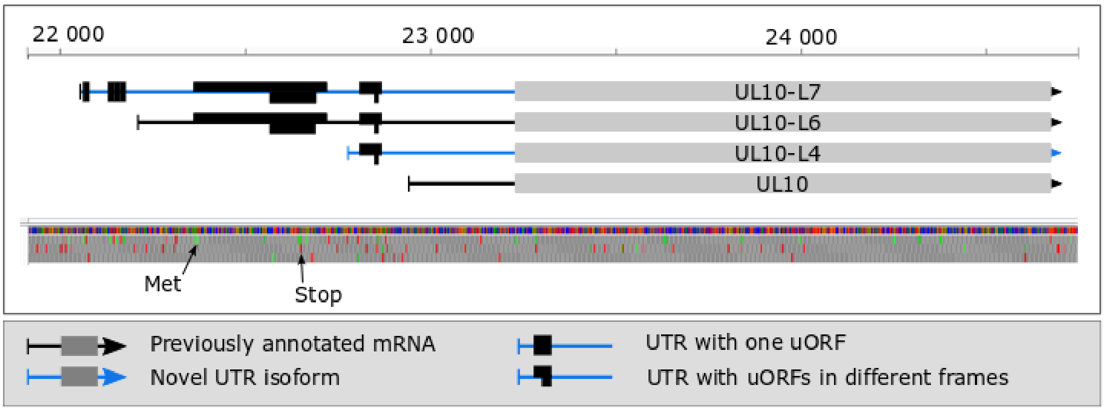


Figure 10. 5’ UTR isoforms of UL10. The number of upstream ORFs (uORFs) depends on the length of the 5’ UTR.

Using LRS we also detected transcripts with alternative termination. For example the TES of the reversely oriented UL27-AT-1 is 245 nt upstream from the TES of UL27. Some abundant genes, like the ul10 also express isoforms with longer 3’ UTRs.

Because the viral genome is tightly packed, these overhanging portions of the TSS and TES isoforms cause extensive overlaps, increasing the possibility of RNAP collisions. Additionally, long TSS isoforms of US1 and ORIS-RNA are responsible for overlapping the replicational origin, causing potential interaction between the transcriptional and replicational machinery.

Splice isoforms

Another advantage of LRS comes in detecting the diversity of intron combinations through their capability of sequencing full-length splice isoforms. NGS technologies struggle in this area, because of their inability to sequence longer than 300 nt exons **(Figure 1)**. Sequencing of cDNAs can be the source of false splice isoforms, because of the previously mentioned RT and PCR artefacts. The capability of the MinION platform to sequence the RNA directly eliminates the possibility of false priming or template switching, resulting in a more accurate isoform annotation.

We used the LoRTIA software suite to detect the splice donor and acceptor sites of the HSV-1 spliced transcripts. Following our selection criteria mentioned in section 6.3. we could verify the existence of 5 previously described splice sites in our annotated transcripts and we could detect 30 further sites, all novel. Among many, we discover two new splice variants of the UL34-35. One of the splicing events results in the deletion of the translational initiation site of ul34 ORF while the other causes a frameshift mutation, resulting in an unknown putative protein (**Figure 11. Panel A.**). Splicing of the polycistronic RL1-RL2 transcript also causes a frameshift mutation of the rl1 ORF (**Figure 11. Panel B.**). Additionally, we detected 8 splice isoforms of US1, all of the intron variants being present in the 5' UTR of the transcript.

Polycistronic and complex transcripts

LRS is also capable of detecting very long RNA molecules spanning multiple genes. These genes can all be oriented parallel with the transcripts, resulting in polycistronism, or one or more of the genes can stand in antisense direction compared to the RNA, resulting in so-called complex transcripts. We identified 201 multigenic transcripts in the HSV-1 transcriptome.

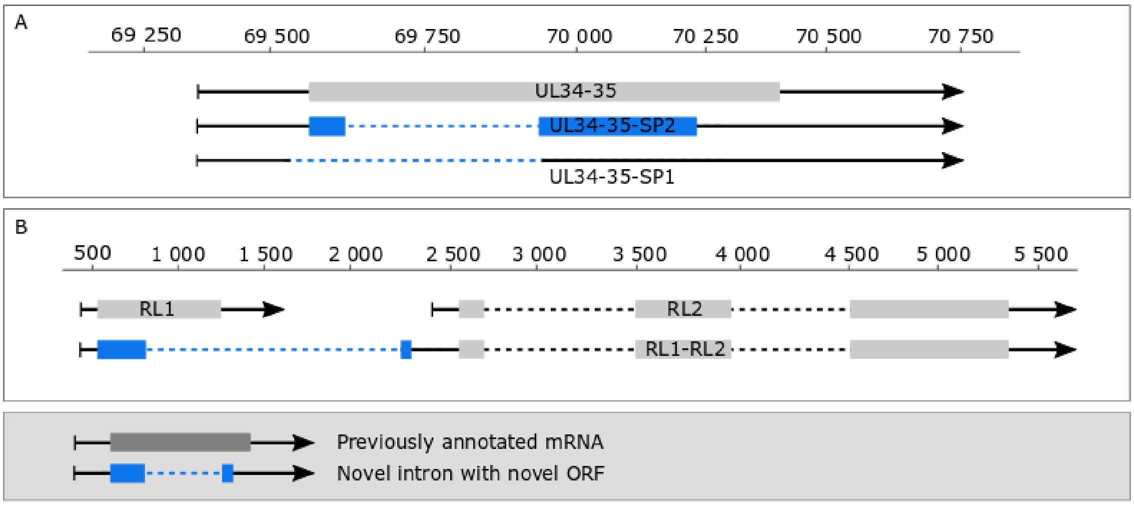


Figure 11. Novel splice isoforms. A. Two novel splice isoforms detected in UL34-35.

B. Novel splice isoform of RL1-RL2.

Amongst many, we detected a bicistronic transcript overlapping the RL1-RL2 and a long complex isoform spanning from the TSS of 0.7kb LAT to the TES of UL3.5.

Transcriptional overlaps

The increased number of multigenic transcripts, TSS and TES isoforms results in several novel overlaps in the HSV-1 transcriptome. These overlaps are the result of read-through events between parallelly or convergently oriented transcripts, or the long 5' UTRs of RNAs in divergent orientation. Practically all convergent genes produce overlaps caused by the RNAPII continuing its transcription for a short time following mRNA cleavage (Proudfoot, 2016). However, these residual RNA molecules are short-lived as exonucleases degrade them quickly. Another source of convergent overlaps is alternative termination and cleavage. We could observe transcripts produced by this phenomenon in low abundance in many HSV-1 genes.

6.5. Analysis of the viral gene expression

To show the suitability of LRS for quantitative analysis of gene expression we will focus on the transcriptome of a non-human pathogen, the BoHV-1. We sequenced the polyadenylated fraction of the viral RNAs from six post-infection time points during lytic infection in three biological replicates. This revealed that almost the entire BoHV-1 genome is transcriptionally active, including the genomic junction and the replicational origin.

TSS, TES and intron dynamics during lytic infection

The LoRTIA software suite annotated 823 TSS 135 TES and 25 introns with the filtering criteria described in *Materials and methods*. We found that both the number of TSSs and TESs increases two-fold during the first four hours of the infection, after which the increase is less steep. The abundance of TSSs with a TATA box shows a higher increase than those lacking a TATA box, the latter plateauing after 4h p.i. Also, TESs with a canonical PAS have a higher increase in abundance, and at 12 h p.i. their number exceeds TESs without a PAS three-fold (**Figure 12. Panel A. and B.**). The first TSSs and TESs were detected in the internal repeat (IR) region at 1h p.i. These features have canonical TATA boxes and PASs respectively. From 2h p.i. we could observe an increase in both TSS (**Figure 13. Panel A.**) and TES (**Figure 13. Panel B.**) activity in the US and several UL regions, with increased activity towards the later time points.

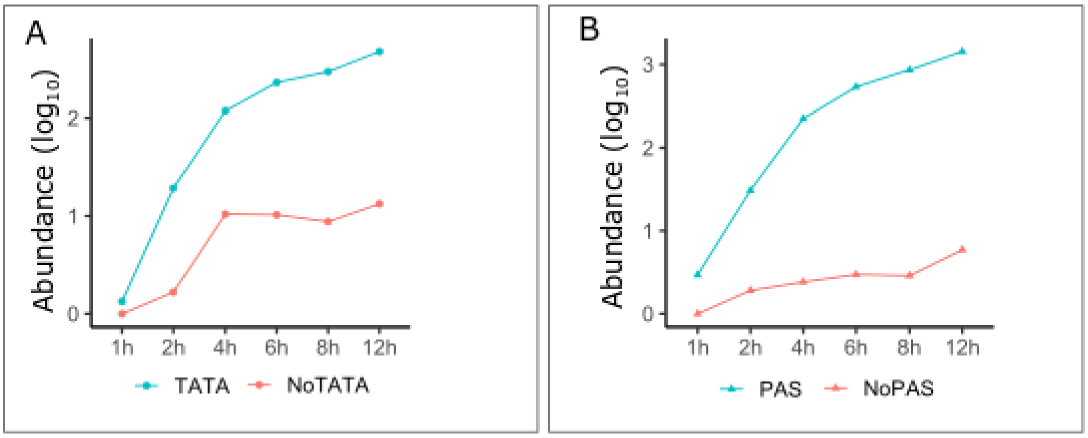


Figure 12. The change in viral TSS and TES abundance during the lytic infection of BoHV-1. A. TSSs with a TATA box are more abundant than those without a TATA box during the whole infection. B. TESs with a canonical PAS are more abundant than those without a PAS during the whole infection.

Five larger genomic stretches seem to be void of TSS, one being between 31,465 and 36,091 nt in ul36, the second between 105,527 and 108,569 nt in bicp4 the third between 108,844 and 111,379 nt in the middle of the IRS, the latter two being repeated in the TRS (**Figure 13. Panel A.**). Similarly, TESs are missing between 26,222 and 40,302 nt and from 58,768 to 67,255 nt, with several other smaller patches of the genome lacking a TES (**Figure 13. Panel B.**).

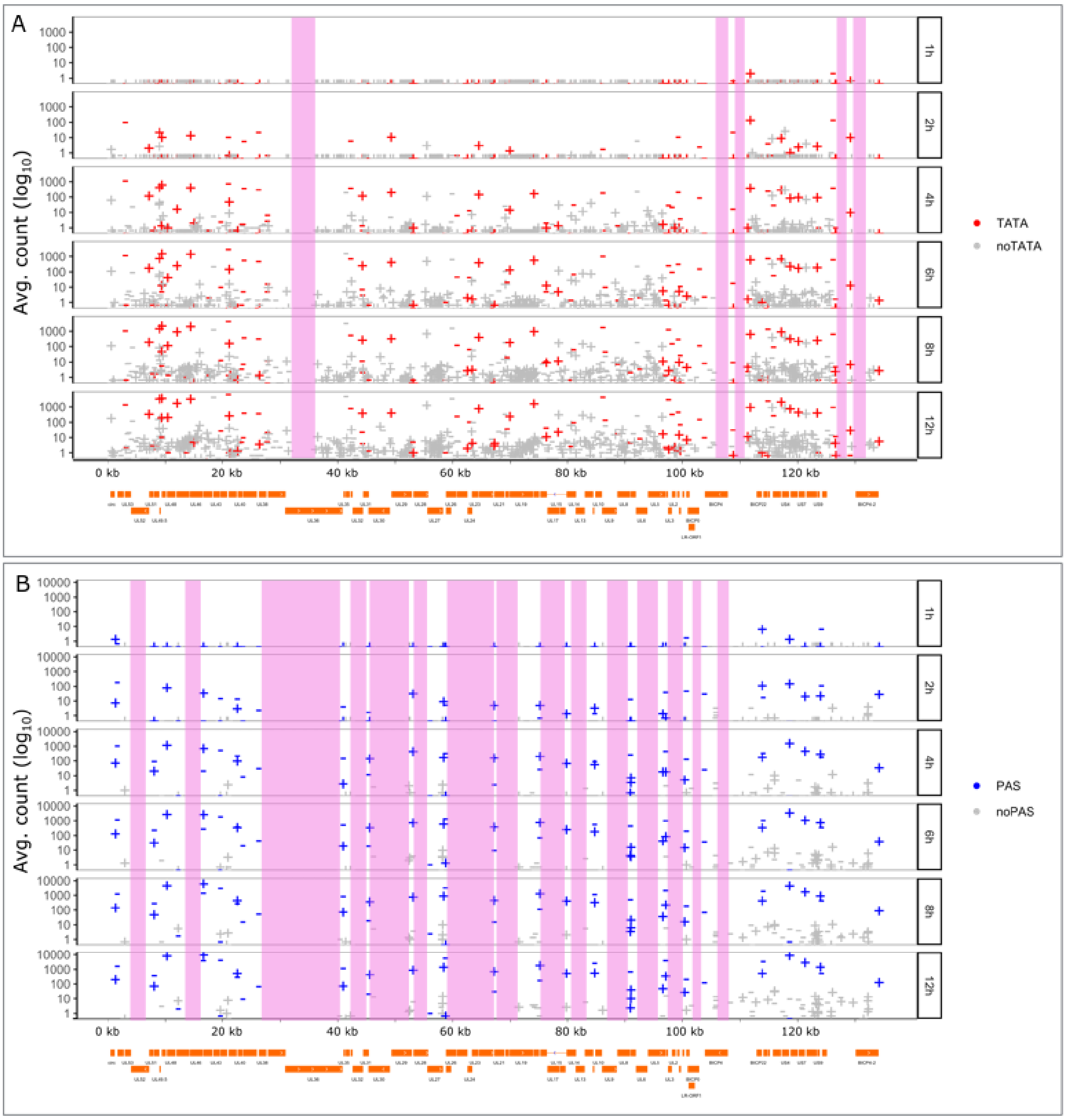


Figure 13. The location and the change of average counts of TSSs and TESs on the BoHV-1 genome during the lytic infection. Features in the sense orientation are marked by a "+" while those in the antisense orientation are marked by a "-". The orange rectangles represent CDS annotations. Purple areas represent genomic regions lacking TSSs or TESs. A. The first TSSs start to express just outside the US region, followed by a seemingly chaotic expression all along the genome in later time points. B The expression of TESs seems more orderly, with many genomic areas lacking a TES.

Gene expression analysis during lytic infection

To determine the expression profile of BoHV-1 genes we normalized the read counts derived by LoRTIA using a modified version of the Median Ration Normalization of the DESeq2 software suite (Wu *et al.*, 2019) for each transcript isoform, then selected the most abundant isoform as the representative transcript of each gene. Gene expression for the BoHV-1 follows the general pattern specific to alphaherpesviruses. We found that the viral genes create at least three temporal classes:

IE, E and L. Bicp0, bicp22 and circ are already present in the first hour of the infection **(Figure 14. Panel A and C)**. Circ is a myristylated tegument protein with unknown function (Fraefel, Ackermann and Schwyzer, 1994), its expression at the very beginning of the infection suggests a possible role in viral gene expression or DNA replication. It has to be noted, that the CIRC transcript we found to be the most abundant at every time-point is not spanning the genomic junction, and it is not the previously detected and described spliced isoform. We also detected this previously annotated isoform, although in very low abundance.

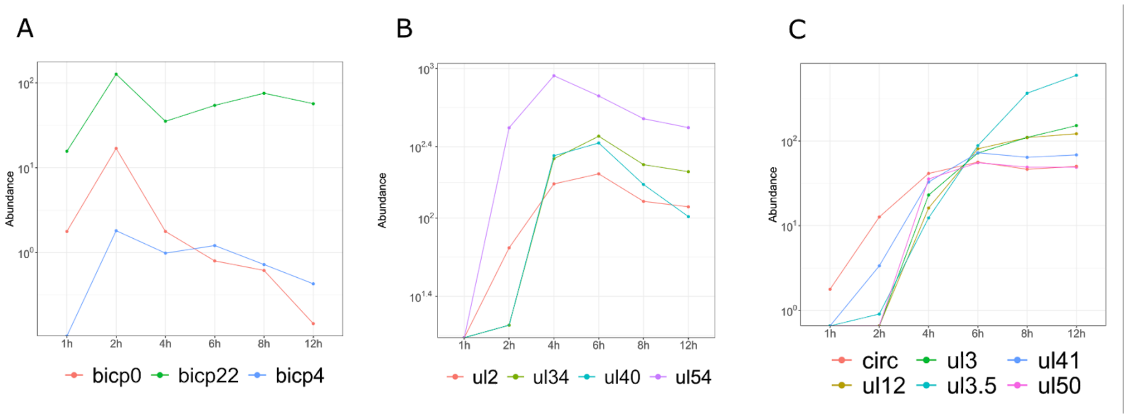


Figure 14. The expression characteristics of BoHV-1 genes during the lytic infection. A: Immediate early, B: Early genes, C Late genes. The abundances detected by LoRTIA were normalized using a modified version of the Median Ration Normalization.

Pokhriyal et al. reported (Pokhriyal *et al.*, 2018) the immediate early expression of three genes (ul21, ul33 and ul34) ought to be thought late. We could detect two of these (ul33 and ul34) to exhibit an early expression pattern, and in addition, found that 41 out of the 69 analysed genes are already expressing at 2h p.i., however, 66 only reach their maximum abundance following replication, suggesting that some early genes benefit from the onset of DNA replication. We detected genes involved in the replication of the viral genome to initiate their expression after the first hour, with detectable amounts of transcripts at 2 and 4h p.i. **(Figure 14. Panel B.)**. Twenty-three genes, including ul12, essential for the processing of the newly synthesised DNA and many structural components of the tegument and the capsid start their expression at 4h p.i. and continue to be expressed during the infection cycle **(Figure 14. Panel C.).** These represent the first wave of late genes and are encompassing the most abundant viral genes. The expression of only four of the viral genes (ul0.5, ul48, us2 and us8) seems to be dependent on DNA replication, with an expression start following 4h p.i. These genes could be considered a second wave of the late genes, and encode structural components of the virion and proteins playing role in the viral egress.

Quantitative analysis of the RNA isoforms

The structural complexity of the viral transcriptome was assessed using the LoRTIA software suite applying similar criteria as described in *Materials and methods* for the analysis of the HSV-1 transcriptome. Following transcript isoform annotation we evaluated their change in abundance. We observed heterogeneity in the expression of different isoform types. In the first hour of the infection, the transcriptome is composed of only monocistronic and spliced isoforms. However, starting from the second hour the isoform diversity increases dramatically, with alternatively terminating mRNAs being the only isoform type not appearing until 4h p.i.

Isoforms of many genes have a similar expression pattern **(Figure 15. panel A.)**, while in some cases different isoforms have a totally different expression **(Figure 15. panel B.)**. We detected a novel splice variant of UL40 the UL40-SP1, which results in a frameshift mutation and a putative protein with altered amino acid composition. The abundance of UL40-SP1 shows a constant and steady increase during the infection, while the non-spliced isoform, has a peak around 4-6h, after which it's abundance decreases. Intriguingly this decrease coincides with the increase in abundance of it's spliced isoform **(Figure 15. panel C.)**.

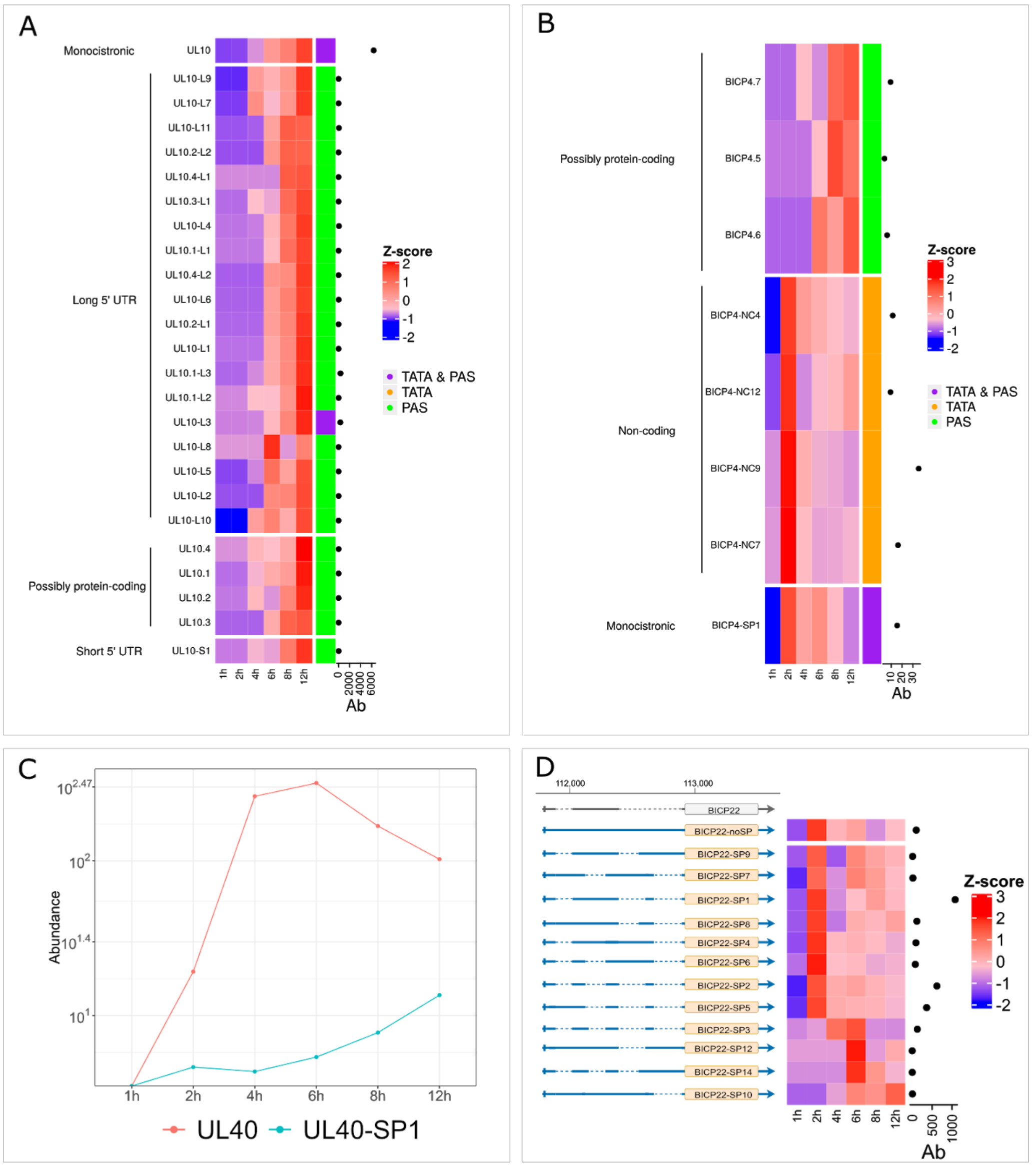


Figure 15. The expression characteristics of BoHV-1 transcript isoforms during the lytic infection. A. The isoforms of UL10 have a similar expression pattern; B. The most abundant splice isoform and the non-coding isoforms of BICP4 are expressed early, while the 5' truncated possibly protein coding isoforms are expressed in the later phase of the infection. C. The abundance of UL40 decreases as the count of it's splice isoform increases. D. The variance in expression of the differentially spliced BICP22 isoforms.

The BICP22 of BoHV-1 has several splice variants, all of the introns being present in the 5' UTR of the transcript. The most abundant isoform together with several others are expressed right after the start of the infection, while BICP22-SP3, BICP22-SP12, BICP22-SP14 and BICP22-SP10 have a peak in their expression at late time points (**Figure 15. Panel D.**). The presence of a canonical TATA box will increase the abundance of some UTR isoforms, for example, the UL10-L3 is 4 times more abundant than other length isoforms of UL10 lacking a canonical promoter.

6.6. RNA modifications

We performed 5^{m}C detection on the direct RNA sequencing data of AcMNPV using the Tombo software suite with the pre-trained model of the software. Tombo detected deviations from the normal signal in 30% of all C positions of the transcriptome. To reduce false-positive locations, we filtered every detection with a coverage less than 30 and with a modified fraction less than 30%, meaning that at least 9 reads out of 30 needed to show a deviation from the canonical signal produced by a non-modified cytosine in a specific position. We also removed every detection where the fraction of mismatches reached 30% because Tombo might confuse these positions, as mismatches deviate from the expected signal level. This resulted in 319 putative 5^{m}C sites in the AcMNPV transcriptome. No significant correlation was detected between the coverage and the fraction of methylated Cs (**Figure 16. Panel A.**).

We detected a potential methylation consensus sequence: UUACCG (the modified base underlined), which was present in low abundance in our filtered data but showed a good distribution of log-likelihood ratios (**Figure 16. Panel B.**) and the deviation from the canonical Cs in the signal was also distinguishable (**Figure 16. Panel C.**). In general, methylated cytosines were present in C and G-rich contexts. Twelve of the viral genes (ac-39k, ac-bro, ac-ctl, ac-odv-e25, ac-orf-58, ac-orf-73, ac-orf-74, ac-orf-75, ac-p40, ac-p6.9, ac-polhedryn and ac-vp39) had a coverage greater than the set threshold, we detected methylation in all of them. The AcMNPV genome is tightly packed, and the transcriptome is characterized by extensive polycistronism, thus most of the transcripts have only a short 5' or 3' UTR, which are usually populated by A/T-rich cis-acting elements, rendering most of the modified Cs in the ORFs. In the few cases where a longer UTR was present, we could detect several modified cytosines in these too.

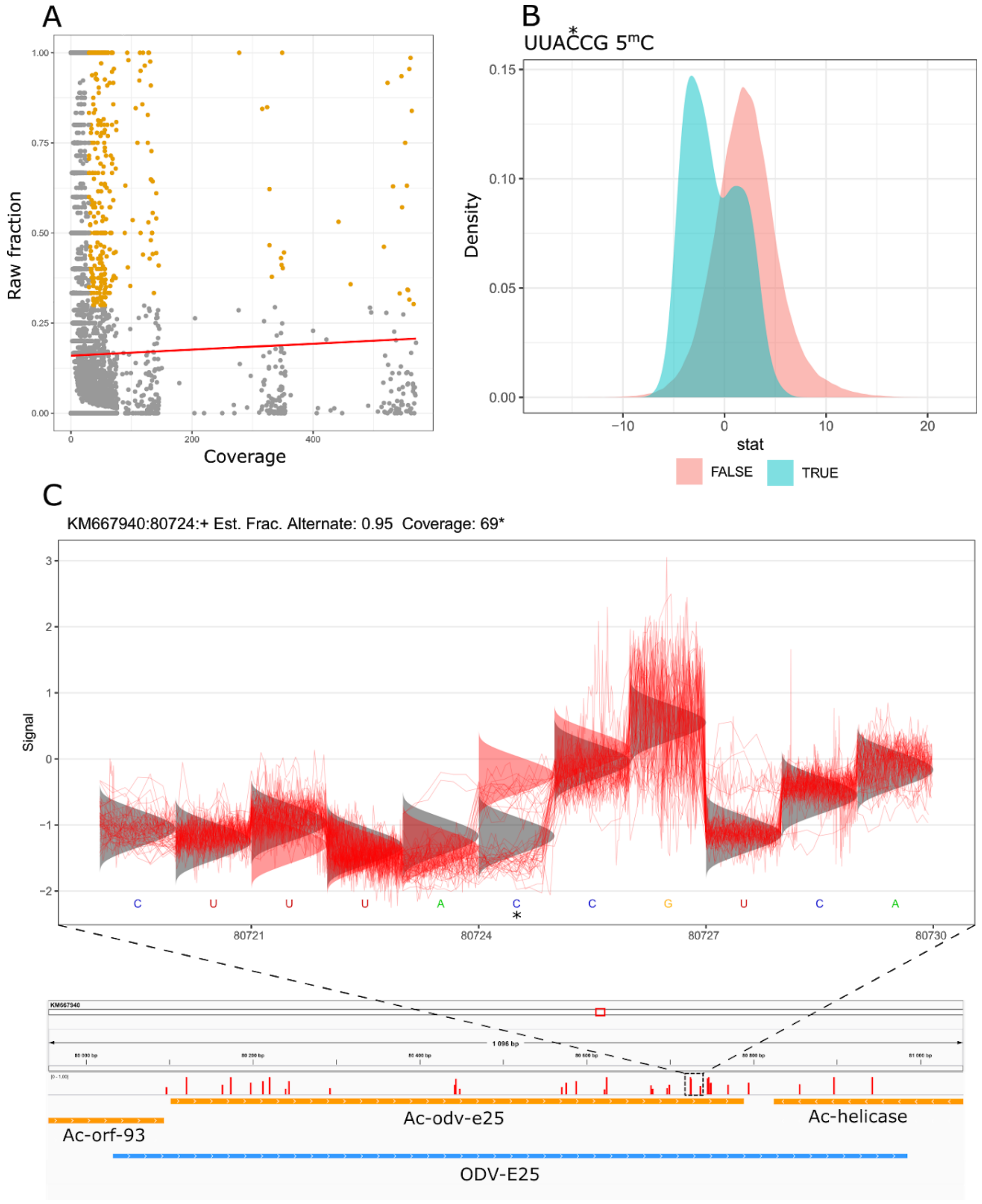


Figure 16. $5^{m}C$ detection on the transcriptome of AcMNPV. A. No significant correlation was detected between the coverage and the raw fraction of modified nucleotides. The yellow dots represent positions that were selected for further analysis. B. The plot of test statistics for the sequence UUACCG (the methylated position underlined) for motif-matching and non-motif-matching sites suggest a possible canonical site for the methyl-transferase. C. The signal levels surrounding a possibly methylated C (marked with an asterisk) on the ODV-E25 transcript. The red curves represent the electrical signal, the densities in grey represent the canonical signal level while the densities in red represent the alternative signal levels.

7. Discussion

In the last five years, LRS approaches uncovered that the viral transcriptome is structurally more complex than previously thought (Boldogkői *et al.*, 2019). Using the technologies provided by PacBio and ONT our group and others demonstrated, that this complexity could be common among many viral taxa including herpesviruses, baculoviruses, circoviruses and others (Moldován, Balázs, *et al.*, 2017; Moldován, Tombácz, *et al.*, 2017; O'Grady, Baddoo and Flemington, 2017; Moldován *et al.*, 2018; Depledge *et al.*, 2019; Tombácz *et al.*, 2019). In my thesis, I discussed the use of LRS for the discovery and annotation of viral transcript isoforms, its capability for the characterization of the gene expression, and the detection of mRNA modifications. Our work on the HSV-1 transcriptome yielded a number of novel 5' and 3' UTR isoforms, multigenic and non-coding RNAs. We show that as a result of these isoforms the viral transcriptome has a highly overlapping nature. Palmer et. al demonstrated that the dislodgement of pre-initiation complexes and the prolonged occlusion of neighbouring promoter regions by paused RNAPs can have a regulatory effect on transcription (Palmer, Egan and Shearwin, 2011). This so-called transcriptional interference was observed *in vivo (Cullen, Lomedico and Ju, 1984; Martens, Laprade and Winston, 2004; Hu et al., 2007)* and *in vitro (Proudfoot, 1986).* We hypothesize that overlapping viral transcripts could have a similar regulatory effect on their neighbouring genes. Longer or shorter UTR regions compared to the main isoform's UTR can modulate the post-transcriptional processing of the RNA through the presence or absence of cis-acting elements (Matoulkova *et al.*, 2012), while uORFs can alter translation (Young and Wek, 2016; Lin *et al.*, 2019). We discovered 5' truncated transcripts with truncated versions of previously annotated ORFs. The function of these possibly protein-coding RNAs needs further investigation; however, we cannot exclude that they represent mere transcriptional noise. Several novel splice isoforms were also annotated, with some of the introns altering the coding sequence. In my thesis I emphasized the importance of filtering of the data, to eliminate possible artefacts resulting from the RT or PCR (Cocquet *et al.*, 2006; Balázs *et al.*, 2019), and the use of multiple library preparation techniques including direct RNA sequencing, which in theory lacks these tendencies.

The quantitative analysis of viral transcriptomes is essential for the understanding the viral life cycle, and virus-host interactions, for which NGS technologies prove to be a crucial tool (Chen *et al.*, 2013; Harkness, Kader and DeLuca, 2014; Peng *et al.*, 2014). In my thesis, I point out the superiority of LRS over

NGS for the characterization of transcript isoforms and present our study on the quantitation of BoHV-1 mRNAs. Using non-amplified cDNA libraries we could characterize the gene expression of the virus, which resembles a similar three-phased pattern to the canonical alphaherpesvirus gene expression derived from HSV-1. We show that TSSs and TESs with a cis-regulating element are in general highly expressed, while those lacking one are generally less expressed. We also point out that the presence of a TATA box will facilitate the expression of 5' UTR isoforms, resulting in higher expression rates. We characterized the expression of the circ gene, which is expressed at equal levels as the canonical early transcripts, suggesting its potential function in the early phases of the viral life cycle.

Information on viral RNA modifications is scarce and focuses mainly on the methylation of the genomes of RNA viruses (Lavi and Shatkin, 1975; Sommer *et al.*, 1976; Tirumuru *et al.*, 2016). To broaden our understanding of the mRNA modification of viruses we used ONTs direct RNA sequencing coupled with signal-level analysis. In theory, the analysis of signal levels produced by the sequencing of RNA molecules can suggest the presence of methylated bases, rendering conventional modification mapping technologies unnecessary. However, they require a ground truth model trained with known methylated positions, which is difficult to produce. At the moment only one such model, for the detection of 5^{m}C is available through the Tombo software suite, which according to our experiments is prone to produce false-positives. We detected abundant methylation of cytosines across the transcriptome of AcMNPV, especially in C and G-rich contexts, which is in concordance with previous studies (Yang *et al.*, 2017). We also detected a potential signal for methylation, the UUA$\underline{\text{C}}$CG, however further studies are needed to validate this finding. Yang et al. demonstrated that 5^{m}Cs are facilitating mRNA export through the ALYREF adaptor in mammalian cells, while Boyne et al showed that the same gene plays a role in the export of viral mRNAs in Kaposi's sarcoma-associated herpesvirus (Boyne, Colgan and Whitehouse, 2008). ALYREF is also present in invertebrates (Shi *et al.*, 2017). We hypothesize that the extensive methylation of Cs on the viral transcripts plays a role in their nuclear export and posttranscriptional modification.

8. Conclusions

In my thesis, I present the possibilities of using long-read sequencing in the structural, expression and modification analysis of viral transcriptomes. Our results regarding the HSV-1 transcriptome added significantly to the knowledge regarding viral RNA isoforms. We manage to validate our results using multiple sequencing platforms. We showed the potential of using the ONT's MinION sequencing of non-amplified direct cDNA libraries of the BoHV-1 transcriptome to quantify gene expression at not just the gene but at the transcript isoform level. These results show a good correlation with previous studies. At the same time, we used direct RNA sequencing to shed light on the epitranscriptome of the AcMNPV, by detecting and mapping 5^{m}C modifications. I outlined the possible strengths and weaknesses of both short- and long-read sequencing technologies, and the areas where they can complement each other.

9. Acknowledgements

The study of the AcMNPV transcriptome presented in the thesis was supported by the National Research Development and Innovation Office in the form of the New National Excellence Programme [UNKP-18-3-I-SZTE-75] scholarship.

I would like to thank Ernő Duda Jr. from the SOLVO Biotechnology Inc. for supplying the AcMNPV and the Sf9 cell line for the study. I would also like to thank Florencia Meyer from the Department of Biochemistry & Molecular Biology, Entomology & Plant Pathology of Mississippi State University for supplying the BoHV-1 for the gene expression study.

I am also grateful for the opportunity of having been able to work with the Pacific Biosciences data, which would not have been possible without the support of Dr Donald Sharon and Prof Michael Snyder from the Department of Genetics at the University of Stanford.

I would like to express my gratitude to all my colleagues at the Department of Medical Biology for their constant help and support. I would like to say special thanks to Marianna Ábrahám and Csilla Magyarné Papdi for their assistance in the lab, to Dr Zsolt Csabai, who did the vast majority of the wet lab work for the presented studies, to Dr Dóra Tombácz, who performed the PacBio sequencing experiments and also helped with the preparations of the ONT sequencing, to Dr Zsolt Balázs, for the development of the LoRTIA toolkit, and for his assistance during the analysis and for brainstorming ideas, and finally to Prof Dr Zsolt Boldogkői, who planned the studies and supervised the writing of the thesis.

10. References

Adams, M. *et al.* (1991) ‘Complementary DNA sequencing: expressed sequence tags and human genome project’, *Science*, pp. 1651–1656. doi: 10.1126/science.2047873.

Adams, M. D. *et al.* (1995) ‘Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence’, *Nature*, 377(6547 Suppl), pp. 3–174.

Amort, T. *et al.* (2017) ‘Distinct 5-methylcytosine profiles in poly(A) RNA from mouse embryonic stem cells and brain’, *Genome biology*, 18(1), p. 1.

Aschenbrenner, J. *et al.* (2018) ‘Engineering of a DNA Polymerase for Direct m A Sequencing’, *Angewandte Chemie* , 57(2), pp. 417–421.

Ayub, M. *et al.* (2013) ‘Nanopore-based identification of individual nucleotides for direct RNA sequencing’, *Nano letters*, 13(12), pp. 6144–6150.

Balázs, Z. *et al.* (2017) ‘Long-Read Sequencing of Human Cytomegalovirus Transcriptome Reveals RNA Isoforms Carrying Distinct Coding Potentials’, *Scientific reports*, 7(1), p. 15989.

Balázs, Z. *et al.* (2019) ‘Template-switching artifacts resemble alternative polyadenylation’, *BMC genomics*, 20(1), p. 824.

Baltimore, D. (1971) ‘Expression of animal virus genomes’, *Bacteriological reviews*, 35(3), pp. 235–241.

Batterson, W. and Roizman, B. (1983) ‘Characterization of the herpes simplex virion-associated factor responsible for the induction of alpha genes’, *Journal of virology*, 46(2), pp. 371–377.

Blissard, G. W. and Rohrmann, G. F. (1990) ‘Baculovirus Diversity and Molecular Biology’, *Annual Review of Entomology*, pp. 127–155. doi: 10.1146/annurev.en.35.010190.001015.

Boldogkői, Z. *et al.* (2018) ‘Transcriptome-wide analysis of a baculovirus using nanopore sequencing’, *Scientific data*, 5, p. 180276.

Boldogkői, Z. *et al.* (2019) ‘Long-Read Sequencing – A Powerful Tool in Viral Transcriptome Research’, *Trends in Microbiology*, pp. 578–592. doi: 10.1016/j.tim.2019.01.010.

Boyne, J. R., Colgan, K. J. and Whitehouse, A. (2008) ‘Recruitment of the Complete hTREX Complex Is Required for Kaposi’s Sarcoma–Associated Herpesvirus Intronless mRNA Nuclear Export and Virus Replication’, *PLoS Pathogens*, p. e1000194. doi: 10.1371/journal.ppat.1000194.

Branton, D. *et al.* (2008) ‘The potential and challenges of nanopore sequencing’, *Nature biotechnology*, 26(10), pp. 1146–1153.

Brosnan, C. A. and Voinnet, O. (2009) ‘The long and the short of noncoding RNAs’, *Current opinion in cell biology*, 21(3), pp. 416–425.

Campadelli-Fiume, G. *et al.* (2000) ‘The novel receptors that mediate the entry of herpes simplex viruses and animal alphaherpesviruses into cells’, *Reviews in medical virology*, 10(5), pp. 305–319.

Chen, Y.-R. *et al.* (2013) ‘The transcriptome of the baculovirus Autographa californica multiple nucleopolyhedrovirus in Trichoplusia ni cells’, *Journal of virology*, 87(11), pp. 6391–6405.

Cocquet, J. *et al.* (2006) ‘Reverse transcriptase template switching and false alternative transcripts’, *Genomics*, 88(1), pp. 127–131.

Coupland, P. *et al.* (2012) ‘Direct sequencing of small genomes on the Pacific Biosciences RS without library preparation’, *BioTechniques*, 53(6), pp. 365–372.

Crick, F. (1970) ‘Central Dogma of Molecular Biology’, *Nature*, pp. 561–563. doi: 10.1038/227561a0.

Crick, F. H. (1958) ‘On protein synthesis’, *Symposia of the Society for Experimental Biology*, 12, pp. 138–163.

Crooks, G. E. *et al.* (2004) ‘WebLogo: a sequence logo generator’, *Genome research*, 14(6), pp. 1188–1190.

Csabai, Z. *et al.* (2019) ‘Analysis of the Complete Genome Sequence of a Novel, Pseudorabies Virus Strain Isolated in Southeast Europe’, *The Canadian journal of infectious diseases & medical microbiology = Journal canadien des maladies infectieuses et de la microbiologie medicale / AMMI Canada*, 2019, p. 1806842.

Cullen, B. R., Lomedico, P. T. and Ju, G. (1984) ‘Transcriptional interference in avian retroviruses--implications for the promoter insertion model of leukaemogenesis’, *Nature*, 307(5948), pp. 241–245.

Delatte, B. *et al.* (2016) ‘RNA biochemistry. Transcriptome-wide distribution and function of RNA hydroxymethylcytosine’, *Science*, 351(6270), pp. 282–285.

Delhon, G. *et al.* (2003) ‘Genome of bovine herpesvirus 5’, *Journal of virology*, 77(19), pp. 10339–10347.

Depledge, D. P. *et al.* (2019) ‘Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen’, *Nature communications*, 10(1), p. 754.

Djavadian, R., Hayes, M. and Johannsen, E. (2018) ‘CAGE-seq analysis of Epstein-Barr virus lytic gene transcription: 3 kinetic classes from 2 mechanisms’, *PLoS pathogens*, 14(6), p. e1007114.

Djebali, S. *et al.* (2012) ‘Landscape of transcription in human cells’, *Nature*, 489(7414), pp. 101–108.

Döhner, K. *et al.* (2002) ‘Function of dynein and dynactin in herpes simplex virus capsid transport’, *Molecular biology of the cell*, 13(8), pp. 2795–2809.

Du, T. *et al.* (2015) ‘Patterns of accumulation of miRNAs encoded by herpes simplex virus during productive infection, latency, and on reactivation', *Proceedings of the National Academy of Sciences*, pp. E49–E55. doi: 10.1073/pnas.1422657112.

El-Mayet, F. S. *et al.* (2019) 'Progesterone increases the incidence of bovine herpesvirus 1 reactivation from latency and stimulates productive infection', *Virus research*, 276, p. 197803.

Emrich, S. J. *et al.* (2007) 'Gene discovery and annotation using LCM-454 transcriptome sequencing', *Genome research*, 17(1), pp. 69–73.

Fraefel, C. *et al.* (1993) 'Immediate-early transcription over covalently joined genome ends of bovine herpesvirus 1: the circ gene', *Journal of virology*, 67(3), pp. 1328–1333.

Fraefel, C., Ackermann, M. and Schwyzer, M. (1994) 'Identification of the bovine herpesvirus 1 circ protein, a myristylated and virion-associated polypeptide which is not essential for virus replication in cell culture', *Journal of virology*, 68(12), pp. 8082–8088.

Fukue, Y. *et al.* (2004) 'Core promoter elements of eukaryotic genes have a highly distinctive mechanical property', *Nucleic acids research*, 32(19), pp. 5834–5840.

Garalde, D. R. *et al.* (2018) 'Highly parallel direct RNA sequencing on an array of nanopores', *Nature methods*, 15(3), pp. 201–206.

Garrity, D. B., Chang, M.-J. and Blissard, G. W. (1997) 'Late Promoter Selection in the Baculovirusgp64 Envelope Fusion ProteinGene', *Virology*, pp. 167–181. doi: 10.1006/viro.1997.8540.

Gerdts, V. *et al.* (2000) 'Pseudorabies Virus Expressing Bovine Herpesvirus 1 Glycoprotein B Exhibits Altered Neurotropism and Increased Neurovirulence', *Journal of Virology*, pp. 817–827. doi: 10.1128/jvi.74.2.817-827.2000.

Gokhale, N. S. *et al.* (2016) 'N6-Methyladenosine in Flaviviridae Viral RNA Genomes Regulates Infection', *Cell host & microbe*, 20(5), pp. 654–665.

Grabherr, M. G. *et al.* (2011) 'Full-length transcriptome assembly from RNA-Seq data without a reference genome', *Nature biotechnology*, 29(7), pp. 644–652.

Guzowski, J. F. and Wagner, E. K. (1993) 'Mutational analysis of the herpes simplex virus type 1 strict late UL38 promoter/leader reveals two regions critical in transcriptional regulation', *Journal of virology*, 67(9), pp. 5098–5108.

Hanson, L., Dishon, A. and Kotler, M. (2011) 'Herpesviruses that infect fish', *Viruses*, 3(11), pp. 2160–2191.

Harkness, J. M., Kader, M. and DeLuca, N. A. (2014) 'Transcription of the herpes simplex virus 1 genome during productive and quiescent infection of neuronal and nonneuronal cells', *Journal of virology*, 88(12), pp. 6847–6861.

Hauenschild, R. *et al.* (2015) 'The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent', *Nucleic acids research*, 43(20), pp. 9950–9964.

Hinkley, S. *et al.* (2000) 'A vhs-like activity of bovine herpesvirus-1', *Archives of* *virology*, 145(10), pp. 2027–2046.

Ho, C. K., Gong, C. and Shuman, S. (2001) ‘RNA triphosphatase component of the mRNA capping apparatus of Paramecium bursaria Chlorella virus 1’, *Journal of virology*, 75(4), pp. 1744–1750.

Hoff, G. L. and Hoff, D. M. (1984) ‘Herpesviruses of Reptiles’, *Diseases of Amphibians and Reptiles*, pp. 159–167. doi: 10.1007/978-1-4615-9391-1_12.

Honess, R. W. and Roizman, B. (1975) ‘Regulation of herpesvirus macromolecular synthesis: sequential transition of polypeptide synthesis requires functional viral polypeptides’, *Proceedings of the National Academy of Sciences of the United States of America*, 72(4), pp. 1276–1280.

Huang, C. J. *et al.* (1996) ‘The herpes simplex virus type 1 VP5 promoter contains a cis-acting element near the cap site which interacts with a cellular protein’, *Journal of virology*, 70(3), pp. 1898–1904.

Hu, B. *et al.* (2016) ‘Functional prediction of differentially expressed lncRNAs in HSV-1 infected human foreskin fibroblasts’, *Virology journal*, 13, p. 137.

Hubbard, K. S. *et al.* (2013) ‘Longitudinal RNA sequencing of the deep transcriptome during neurogenesis of cortical glutamatergic neurons from murine ESCs’, *F1000Research*, 2, p. 35.

Hussain, S. *et al.* (2013) ‘NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs’, *Cell reports*, 4(2), pp. 255–261.

Hussey, G. S. (2019) ‘Key Determinants in the Pathogenesis of Equine Herpesvirus 1 and 4 Infections’, *Veterinary pathology*, 56(5), pp. 656–659.

Hu, X. *et al.* (2007) ‘Transcriptional interference among the murine beta-like globin genes’, *Blood*, 109(5), pp. 2210–2216.

Hu, Y.-C. (2005) ‘Baculovirus as a highly efficient expression vector in insect and mammalian cells’, *Acta pharmacologica Sinica*, 26(4), pp. 405–416.

Javahery, R. *et al.* (1994) ‘DNA sequence requirements for transcriptional initiator activity in mammalian cells’, *Molecular and cellular biology*, 14(1), pp. 116–127.

Kaleta, E. F. (1990) ‘Herpesviruses of birds - a review’, *Avian Pathology*, pp. 193–211. doi: 10.1080/03079459008418673.

Kogan, P. H., Chen, X. and Blissard, G. W. (1995) ‘Overlapping TATA-dependent and TATA-independent early promoter activities in the baculovirus gp64 envelope fusion protein gene’, *Journal of virology*, 69(3), pp. 1452–1461.

Kumar, A. *et al.* (2017) ‘The impact of RNA sequence library construction protocols on transcriptomic profiling of leukemia’, *BMC genomics*. BioMed Central, 18(1), pp. 1–13.

Kwok, S. *et al.* (1990) ‘Effects of primer-template mismatches on the polymerase chain reaction: human immunodeficiency virus type 1 model studies’, *Nucleic acids research*, 18(4), pp. 999–1005.

Lavi, S. and Shatkin, A. J. (1975) ‘Methylated simian virus 40-specific RNA from nuclei and cytoplasm of infected BSC-1 cells’, *Proceedings of the National Academy of Sciences of the United States of America*, 72(6), pp. 2012–2016.

Lichinchi, G. *et al.* (2016) ‘Dynamics of the human and viral m6A RNA methylomes during HIV-1 infection of T cells’, *Nature Microbiology*. doi: 10.1038/nmicrobiol.2016.11.

Li, H. (2018) ‘Minimap2: pairwise alignment for nucleotide sequences’, *Bioinformatics* , 34(18), pp. 3094–3100.

Lim, C. Y. (2004) ‘The MTE, a new core promoter element for transcription by RNA polymerase II’, *Genes & Development*, pp. 1606–1617. doi: 10.1101/gad.1193404.

Lin, Y. *et al.* (2019) ‘Impacts of uORF codon identity and position on translation regulation’, *Nucleic acids research*, 47(17), pp. 9358–9367.

Liu, H. *et al.* (2019) ‘Accurate detection of mA RNA modifications in native RNA sequences’, *Nature communications*, 10(1), p. 4079.

Li, Y. *et al.* (1995) ‘Characterization of cell-binding properties of bovine herpesvirus 1 glycoproteins B, C, and D: identification of a dual cell-binding function of gB’, *Journal of virology*, 69(8), pp. 4758–4768.

Li, Y. *et al.* (1996) ‘Glycoprotein Bb, the N-terminal subunit of bovine herpesvirus 1 gB, can bind to heparan sulfate on the surfaces of Madin-Darby bovine kidney cells’, *Journal of virology*, 70(3), pp. 2032–2037.

Looker, K. J. *et al.* (2015) ‘Global and Regional Estimates of Prevalent and Incident Herpes Simplex Virus Type 1 Infections in 2012’, *PLOS ONE*, p. e0140765. doi: 10.1371/journal.pone.0140765.

Macdonald, S. J. *et al.* (2012) ‘Genome Sequence of Herpes Simplex Virus 1 Strain KOS’, *Journal of Virology*, pp. 6371–6372. doi: 10.1128/jvi.00646-12.

Martens, J. A., Laprade, L. and Winston, F. (2004) ‘Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene’, *Nature*, 429(6991), pp. 571–574.

Mathieu-Daude, F. (1996) ‘DNA rehybridization during PCR: the “Cot effect” and its consequences’, *Nucleic Acids Research*, pp. 2080–2086. doi: 10.1093/nar/24.11.2080.

Matoulkova, E. *et al.* (2012) ‘The role of the 3’ untranslated region in post-transcriptional regulation of protein expression in mammalian cells’, *RNA Biology*, pp. 563–576. doi: 10.4161/rna.20231.

McCarthy, A. (2010) ‘Third generation DNA sequencing: pacific biosciences’ single molecule real time technology’, *Chemistry & biology*, 17(7), pp. 675–676.

Meller, A., Nivon, L. and Branton, D. (2001) ‘Voltage-driven DNA translocations through a nanopore’, *Physical review letters*, 86(15), pp. 3435–3438.

Moldován, N., Balázs, Z., *et al.* (2017) ‘Multi-platform analysis reveals a complex transcriptome architecture of a circovirus’, *Virus research*, 237, pp. 37–46.

Moldován, N., Tombácz, D., *et al.* (2017) ‘Multi-Platform Sequencing Approach Reveals a Novel Transcriptome Profile in Pseudorabies Virus’, *Frontiers in microbiology*, 8, p. 2708.

Moldován, N. *et al.* (2018) ‘Third-generation Sequencing Reveals Extensive Polycistronism and Transcriptional Overlapping in a Baculovirus’, *Scientific reports*, 8(1), p. 8604.

Morris, K. V. and Mattick, J. S. (2014) ‘The rise of regulatory RNA’, *Nature Reviews Genetics*, pp. 423–437. doi: 10.1038/nrg3722.

Moss, B. *et al.* (1977) ‘5’-Terminal and internal methylated nucleosides in herpes simplex virus type 1 mRNA’, *Journal of virology*, 23(2), pp. 234–239.

Muylkens, B. *et al.* (2007) ‘Bovine herpesvirus 1 infection and infectious bovine rhinotracheitis’, *Veterinary Research*, pp. 181–209. doi: 10.1051/vetres:2006059.

Nguyen, M. and Haenni, A.-L. (2003) ‘Expression strategies of ambisense viruses’, *Virus research*, 93(2), pp. 141–150.

Odelberg, S. J. *et al.* (1995) ‘Template-switching during DNA synthesis byThermus aquaticusDNA polymerase I’, *Nucleic Acids Research*, pp. 2049–2057. doi: 10.1093/nar/23.11.2049.

d’Offay, J. M., Fulton, R. W. and Eberle, R. (2013) ‘Complete genome sequence of the NVSL BoHV-1.1 Cooper reference strain’, *Archives of Virology*, pp. 1109–1113. doi: 10.1007/s00705-012-1574-6.

O’Grady, T., Baddoo, M. and Flemington, E. K. (2017) ‘Analysis of EBV Transcription Using High-Throughput RNA Sequencing’, *Methods in molecular biology* , 1532, pp. 105–121.

Oláh, P. *et al.* (2015) ‘Characterization of pseudorabies virus transcriptome by Illumina sequencing’, *BMC microbiology*, 15, p. 130.

Pääbo, S., Irwin, D. M. and Wilson, A. C. (1990) ‘DNA damage promotes jumping between templates during enzymatic amplification’, *The Journal of biological chemistry*, 265(8), pp. 4718–4721.

Palmer, A. C., Egan, J. B. and Shearwin, K. E. (2011) ‘Transcriptional interference by RNA polymerase pausing and dislodgement of transcription factors’, *Transcription*, 2(1), pp. 9–14.

Peng, X. *et al.* (2014) ‘Deep sequencing of HIV-infected cells: insights into nascent transcription and host-directed therapy’, *Journal of virology*, 88(16), pp. 8768–8782.

Pokhriyal, M. *et al.* (2018) ‘Three newly identified Immediate Early Genes of Bovine herpesvirus 1 lack the characteristic Octamer binding motif- 1’, *Scientific reports*, 8(1), p. 11441.

Polz, M. F. and Cavanaugh, C. M. (1998) ‘Bias in template-to-product ratios in multitemplate PCR’, *Applied and environmental microbiology*, 64(10), pp. 3724–3730.

Pomeranz, L. E., Reynolds, A. E. and Hengartner, C. J. (2005) ‘Molecular biology of pseudorabies virus: impact on neurovirology and veterinary medicine’, *Microbiology* *and molecular biology reviews: MMBR*, 69(3), pp. 462–500.

Prazsák, I. *et al.* (2018) ‘Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus’, *BMC genomics*, 19(1), p. 873.

Proudfoot, N. J. (1986) ‘Transcriptional interference and termination between duplicated alpha-globin gene constructs suggests a novel mechanism for gene regulation’, *Nature*, 322(6079), pp. 562–565.

Proudfoot, N. J. (2016) ‘Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut’, *Science*, 352(6291), p. aad9926.

Qiu, X. *et al.* (2001) ‘Evaluation of PCR-Generated Chimeras, Mutations, and Heteroduplexes with 16S rRNA Gene-Based Cloning’, *Applied and Environmental Microbiology*, pp. 880–887. doi: 10.1128/aem.67.2.880-887.2001.

Rajcáni, J., Andrea, V. and Ingeborg, R. (2004) ‘Peculiarities of herpes simplex virus (HSV) transcription: an overview’, *Virus genes*, 28(3), pp. 293–310.

Ransohoff, J. D., Wei, Y. and Khavari, P. A. (2018) ‘The functions and unique features of long intergenic non-coding RNA’, *Nature reviews. Molecular cell biology*, 19(3), pp. 143–157.

Reiss, J. *et al.* (1990) ‘The effect of replication errors on the mismatch analysis of PCR-amplified DNA’, *Nucleic acids research*, 18(4), pp. 973–978.

Rohrmann, G. F. (2019) *Baculovirus Molecular Biology*. Bethesda (MD): National Center for Biotechnology Information (US).

Roundtree, I. A. *et al.* (2017) ‘Dynamic RNA Modifications in Gene Expression Regulation’, *Cell*, pp. 1187–1200. doi: 10.1016/j.cell.2017.05.045.

Saydam, O. *et al.* (2004) ‘Transactivator protein BICP0 of bovine herpesvirus 1 (BHV-1) is blocked by prostaglandin D2 (PGD2), which points to a mechanism for PGD2-mediated inhibition of BHV-1 replication’, *Journal of virology*, 78(8), pp. 3805–3810.

Saydam, O. *et al.* (2006) ‘Host cell targets of immediate-early protein BICP22 of bovine herpesvirus 1’, *Veterinary microbiology*, 113(3-4), pp. 185–192.

Schaefer, M. *et al.* (2008) ‘RNA cytosine methylation analysis by bisulfite sequencing’, *Nucleic Acids Research*, pp. e12–e12. doi: 10.1093/nar/gkn954.

Schaefer, M., Kapoor, U. and Jantsch, M. F. (2017) ‘Understanding RNA modifications: the promises and technological bottlenecks of the “epitranscriptome”’, *Open Biology*, p. 170077. doi: 10.1098/rsob.170077.

Schmidt, W. M. and Mueller, M. W. (1999) ‘CapSelect: a highly sensitive method for 5’ CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs’, *Nucleic acids research*, 27(21), p. e31.

Shi, M. *et al.* (2017) ‘ALYREF mainly binds to the 5′ and the 3′ regions of the mRNA in vivo’, *Nucleic Acids Research*, pp. 9640–9653. doi: 10.1093/nar/gkx597.

Shiraki, T. *et al.* (2003) ‘Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage’, *Proceedings of the National Academy of Sciences of the United States of America*, 100(26), pp. 15776–15781.

Sinclair, W. R. *et al.* (2017) ‘Profiling Cytidine Acetylation with Specific Affinity and Reactivity’, *ACS chemical biology*, 12(12), pp. 2922–2926.

Sommer, S. *et al.* (1976) ‘The methylation of adenovirus-specific nuclear and cytoplasmic RNA’, *Nucleic acids research*, 3(3), pp. 749–765.

Stoiber, M. *et al.* (2016) ‘De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing’. doi: 10.1101/094672.

Szűcs, A. *et al.* (2017) ‘Long-Read Sequencing Reveals a GC Pressure during the Evolution of Porcine Endogenous Retrovirus’, *Genome announcements*, 5(40). doi: 10.1128/genomeA.01040-17.

Tao, X. Y. *et al.* (2013) ‘The Autographa californica Multiple Nucleopolyhedrovirus ORF78 Is Essential for Budded Virus Production and General Occlusion Body Formation’, *Journal of Virology*, pp. 8441–8450. doi: 10.1128/jvi.01290-13.

Thorvaldsdóttir, H., Robinson, J. T. and Mesirov, J. P. (2013) ‘Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration’, *Briefings in bioinformatics*, 14(2), pp. 178–192.

Tirumuru, N. *et al.* (2016) ‘N(6)-methyladenosine of HIV-1 RNA regulates viral infection and HIV-1 Gag protein expression’, *eLife*, 5. doi: 10.7554/eLife.15528.

Tombácz, D. *et al.* (2014) ‘Strain Kaplan of Pseudorabies Virus Genome Sequenced by PacBio Single-Molecule Real-Time Sequencing Technology’, *Genome announcements*, 2(4). doi: 10.1128/genomeA.00628-14.

Tombácz, D. *et al.* (2016) ‘Full-Length Isoform Sequencing Reveals Novel Transcripts and Substantial Transcriptional Overlaps in a Herpesvirus’, *PloS one*, 11(9), p. e0162868.

Tombácz, D. *et al.* (2017) ‘Long-Read Isoform Sequencing Reveals a Hidden Complexity of the Transcriptional Landscape of Herpes Simplex Virus Type 1’, *Frontiers in microbiology*, 8, p. 1079.

Tombácz, D. *et al.* (2018) ‘Transcriptome-wide survey of pseudorabies virus using next- and third-generation sequencing platforms’, *Scientific data*, 5, p. 180119.

Tombácz, D. *et al.* (2019) ‘Multiple Long-Read Sequencing Survey of Herpes Simplex Virus Dynamic Transcriptome’, *Frontiers in genetics*, 10, p. 834.

Trapnell, C. *et al.* (2010) ‘Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation’, *Nature biotechnology*, 28(5), pp. 511–515.

Tserovski, L. *et al.* (2016) ‘High-throughput sequencing for 1-methyladenosine (m(1)A) mapping in RNA’, *Methods* , 107, pp. 110–121.

Velculescu, V. E. *et al.* (1995) ‘Serial analysis of gene expression’, *Science*, 270(5235), pp. 484–487.

Wang, E. T. *et al.* (2008) ‘Alternative isoform regulation in human tissue transcriptomes’, *Nature*, pp. 470–476. doi: 10.1038/nature07509.

Wang, Z., Gerstein, M. and Snyder, M. (2009) ‘RNA-Seq: a revolutionary tool for transcriptomics’, *Nature Reviews Genetics*, pp. 57–63. doi: 10.1038/nrg2484.

Wu, Z. *et al.* (2019) ‘NormExpression: An R Package to Normalize Gene Expression Data Using Evaluated Methods’, *Frontiers in Genetics*. doi: 10.3389/fgene.2019.00400.

Wyler, E. *et al.* (2017) ‘Widespread activation of antisense transcription of the host genome during herpes simplex virus 1 infection’, *Genome biology*, 18(1), p. 209.

Xi, H. *et al.* (2007) ‘Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1’, *Genome Research*, pp. 798–806. doi: 10.1101/gr.5754707.

Yang, X. *et al.* (2017) ‘5-methylcytosine promotes mRNA export - NSUN2 as the methyltransferase and ALYREF as an mC reader’, *Cell research*, 27(5), pp. 606–625.

Young, M. D. *et al.* (2012) ‘Differential Expression for RNA Sequencing (RNA-Seq) Data: Mapping, Summarization, Statistical Analysis, and Experimental Design’, *Bioinformatics for High Throughput Sequencing*, pp. 169–190. doi: 10.1007/978-1-4614-0782-9_10.

Young, S. K. and Wek, R. C. (2016) ‘Upstream Open Reading Frames Differentially Regulate Gene-specific Translation in the Integrated Stress Response’, *The Journal of biological chemistry*, 291(33), pp. 16927–16935.

Zhao, B. S., Roundtree, I. A. and He, C. (2018) ‘Publisher Correction: Post-transcriptional gene regulation by mRNA modifications’, *Nature Reviews Molecular Cell Biology*, pp. 808–808. doi: 10.1038/s41580-018-0075-1.

Zhu, J. *et al.* (1999) ‘Identification of a novel 0.7-kb polyadenylated transcript in the LAT promoter region of HSV-1 that is strain specific and may contribute to virulence’, *Virology*, 265(2), pp. 296–307.

11. Copies of publications upon which the thesis was based

I.

SCIENTIFIC REPORTS

OPEN

Third-generation Sequencing Reveals Extensive Polycistronism and Transcriptional Overlapping in a Baculovirus

Received: 21 February 2018
Accepted: 22 May 2018
Published online: 05 June 2018

Norbert Moldován[1], Dóra Tombácz[1], Attila Szűcs[1], Zsolt Csabai[1], Zsolt Balázs[1], Emese Kis[2], Judit Molnár[2] & Zsolt Boldogkői[1]

The *Autographa californica multiple nucleopolyhedrovirus* (AcMNPV) is an insect-pathogen baculovirus. In this study, we applied the Oxford Nanopore Technologies platform for the analysis of the polyadenylated fraction of the viral transcriptome using both cDNA and direct RNA sequencing methods. We identified and annotated altogether 132 novel transcripts and transcript isoforms, including 4 coding and 4 non-coding RNA molecules, 47 length variants, 5 splice isoforms, as well as 23 polycistronic and 49 complex transcripts. All of the identified novel protein-coding genes were 5′-truncated forms of longer host genes. In this work, we demonstrated that in the case of transcript start site isoforms, the promoters and the initiator sequence of the longer and shorter variants belong to the same kinetic class. Long-read sequencing also revealed a complex meshwork of transcriptional overlaps, the function of which needs to be clarified. Additionally, we developed bioinformatics methods to improve the transcript annotation and to eliminate the non-specific transcription reads generated by template switching and false priming.

The *Autographa californica multiple nucleopolyhedrovirus* (AcMNPV) is an insect virus belonging to the *Alphabaculovirus* genus of *Baculoviridae* family[1]. The lifecycle of AcMNPV includes two distinct forms: the budded virus (BV) is released from the infected cells first, while the occlusion-derived virus (ODV) is released later. ODV is responsible for the primary infection, while the BV infects the host cells during the secondary infection[2].

The AcMNPV is a double-stranded DNA virus with a 133 kilobase-pair (kbp)-long circular genome encompassing 156 closely-spaced open reading frames (ORFs). The viral genome is complex with respect to the transcription[3]. The baculovirus genes are expressed in three distinct phases: early (E), late (L) and very late (VL); the early genes can be subdivided into immediate early (IE) and delayed early (DE) genes[2]. The promoters of IE and E genes commonly harbour a canonical TATA motif that are recognized by the host RNA polymerase II, and their transcription starts at an early initiator CAGT sequence[4]. On the other hand, the L and VL transcripts tend to bind to a late initiator sequence (LIS) harbouring a TAAG motif recognized by viral RNA polymerase (RNAP), which starts the transcription from the second nucleotide of the motif[3,5]. The E transcripts contain the consensus AAUAAA or AUUAAA polyadenylation (PA) signal (PAS), and their PA tail formation is carried out by the nuclear polyadenylation machinery of the host cell. On the other hand, it is assumed that the L and VL transcripts do not require the consensus PAS. Using an *in vitro* system, Jin and Guarino[6] demonstrated that the viral RNAP enzyme lacking the carboxy-terminal domain, an essential part in the recruitment of the polyadenylation apparatus, can perform a non-templated addition of adenosines after terminating at a T-rich sequences. Baculoviruses are commonly used as gene delivery vectors in insect cell systems[7] for the expression of recombinant proteins[8] or biopesticides[9].

Next-generation sequencing (NGS) techniques have proved to be useful for discovering novel genes and characterizing their expressions[10–12]. While these platforms are highly accurate and produce a massive amount of output data, they are inefficient for identifying polycistronic and complex transcripts, and for distinguishing between RNA length and splice isoforms[13] and between overlapping transcripts.

[1]Department of Medical Biology, Faculty of Medicine, University of Szeged, Szeged, 6720, Hungary. [2]Solvo Biotechnology, Szeged, 6720, Hungary. Correspondence and requests for materials should be addressed to Z.B. (email: boldogkoi.zsolt@med.u-szeged.hu)

Third generation sequencing (TGS) techniques can surmount these shortcomings by their ability to determine the full-length sequence of the RNA molecules[14–16] using cDNA sequencing (cDNA-Seq) or direct (d) RNA sequencing (dRNA-Seq). The Oxford Nanopore Technologies (ONT) MinION platform is based on the passage of unlabelled DNA or RNA molecules through a protein channel present on a synthetic membrane, which is regulated by a motor protein[17]. The passage of the nucleotides causes changes in the ionic flow through the nanopore, which can be associated with the specific nucleotide[18]. Theoretically, the ONT technology has no upper limit regarding the read length, but currently falls short of its competitors, especially compared to the Pacific Biosciences (PacBio) platform, in terms of the precision of base identification[19]. However, in the case of well-annotated genomes and high data coverage, this technology is optimal for global transcriptome analysis. Both NGS and TGS technologies use reverse transcription (RT) and PCR for library preparation, which can produce false products (false splice sites as well as false 5′ and 3′ ends) through template switching[20,21], or through false priming[22]. The dRNA-Seq represents a useful approach for the elimination of these artefacts. However, besides the relatively high false INDEL and base substitution ratio, dRNA sequencing is afflicted by two other innate flaws, which should be taken into account when searching for novel transcript variants. In our previous study[16], we observed that the 5′ ends of direct RNA sequencing reads were on average 23 bp shorter than the actual transcription start sites (TSSs). We assume that this could be the result of the premature release of the RNA strand by the motor protein, causing a rapid transition of the RNA molecule through the nanopore, and thus cause a perturbation of the base calling near the 5′ end. Our other observation was that the PA tails were miscalled as CT-rich regions[16] for which the reason may be that the dRNA-Seq base caller algorithm is perturbed by the presence of a DNA adaptor ligated to the PA sequences.

Our research group investigated various viruses using PacBio sequencing alone[23–25] and using a multiplatform (PacBio, ONT and Illumina) sequencing approach[16]. The advantage of the PacBio platform over ONT sequencing was the very high accuracy, while the ONT was found to be superior in identifying transcripts with sizes ranging from 200 to 800 bp.

Previous studies of the AcMNPV transcriptome using microarray[26] and real-time RT-PCR[27] analysis focused mainly on the expression dynamics, while a study using 5′ and 3′ rapid amplification of cDNA ends (RACE) and Illumina sequencing of 5′ capped RNA described many of the potential TSSs and transcription end sites (TESs)[3]. Despite the high precision of these methods, they are unable to uncover the transcriptome complexity in its entirety, they fall short especially in the identification of transcript isoforms, multigenic RNA molecules and transcript overlaps.

In this study, we used the ONT MinION long-read real-time cDNA sequencing for confirming the 5′ and 3′ ends of already known AcMNPV transcripts with base-pair precision, and for identifying novel RNA molecules, including putative protein-coding and non-coding transcripts, TSS and TES isoforms, as well as polycistronic and complex transcripts. Additionally, we also applied a dRNA-Seq technique in order to exclude the artefacts produced by PCR and RT, and for confirming the existence of the RNA molecules identified by cDNA-Seq.

Materials and Methods

Cells and viral infection. The baculovirus AcMNPV, used in this study expresses the *lacZ* gene, which was inserted in the promoter region of the *polh* gene (βgal-AcMNPV). This virus was propagated on the Sf9 cell line (kindly provided by Ernő Duda Jr., Solvo Biotechnology, Hungary). Cells were cultivated in 200 ml of GIBCO Sf-900 II SFM insect cell medium (Thermo Fisher Scientific) in a Corning spinner flask (Merck) set to 70 rpm at 26 °C, and were infected with a viral titre of 2 multiplicity of infection (MOI = plaque-forming units per cell). A five ml sample was measured and centrifuged at 2,000 rpm at 4 °C at nine consecutive time points (0 h, 1 h, 2 h, 4 h, 6 h, 16 h, 24 h, 48 h and 72 h), followed by washing with PBS and centrifuged again. Pellets were stored at −80 °C until use. The samples were mixed for the sequencing analysis.

RNA purification. Total RNA was isolated using the Nucleospin RNA Kit (Macherey-Nagel) according to the manufacturer's guidance. In short, infected cells were collected by centrifugation and the cell membrane was disrupted by the addition of lysis buffer (derived from the kit). Genomic DNA was digested by treatment with RNase-free rDNase solution (supplied with the kit). Samples were eluted in a total volume of 50 μl nuclease free water. To eliminate residual DNA contamination, samples were treated with TURBO DNA-free Kit (Thermo Fisher Scientific). The RNA concentration was measured using a Qubit 2.0 Fluorometer, using the Qubit RNA BR Assay Kit (Thermo Fisher Scientific). The poly(A)+ RNA fraction was isolated from the samples using the Oligotex mRNA Mini Kit (Qiagen). An mRNA mix was prepared by using 10 μl from each time point. RNA samples were stored at −80 °C until use.

Oxford Nanopore MinION sequencing. *The 'strand switching cDNA by ligation' approach* The cDNA library was prepared using the Ligation Sequencing kit (SQK-LSK108; Oxford Nanopore Technologies) following the 1D strand switching cDNA by ligation protocol. Briefly: single-stranded (ss)cDNA synthesis was carried out using 50 ng poly(A)+ RNA, SuperScript IV Reverse Transcriptase (Thermo Fisher Scientific) and anchored adapter-primer with (VN)T_{20} nucleotides (nts; supplied by the kit). A 5′ adapter sequence with three O-methyl-guanine RNA bases was added for the facilitation of strand switching. PCR was carried out using Kapa HiFi DNA polymerase (Kapa Biosystems) and the primers supplied in the kit. End repair was conducted using NEBNext End repair/dA-tailing Module (New England Biolabs) followed by adapter ligation using adapters (supplied in the kit) and NEB Blunt/TA Ligase Master Mix (New England Biolabs). The cDNA sample was purified between each step using Agencourt AMPure XP magnetic beads (Beckman Coulter) and the library concentration was determined using a Qubit 2.0 Fluorometer (through use of the Qubit (ds)DNA HS Assay Kit (Thermo Fisher Scientific). Samples were loaded on R9.4 SpotON Flow Cells, and base calling was performed using Albacore v1.2.6.

The direct RNA sequencing approach Libraries were prepared using the Direct RNA Sequencing Kit (SQK-RNA001; Oxford Nanopore Technologies). The first strand cDNA was synthesized by SuperScript IV Reverse Transcriptase (Thermo Fisher Scientific) using an RT adapter with T_{10} nts and the mRNA mix. Adapters (supplied by the kit) were ligated using T4 DNA ligase (New England Biolabs). The RNA-DNA hybrid was purified between each step by using Agencourt AMPure XP magnetic beads (Beckman Coulter), and then treated with RNaseOUT Recombinant Ribonuclease Inhibitor (Thermo Fisher Scientific). Sample concentration was determined using a Qubit 2.0 Fluorometer and the Qubit DNA HS Assay Kit (Thermo Fisher Scientific). Libraries were loaded on R9.4 SpotON Flow Cells. Albacore software (v1.2.6) was used for base calling.

Transcript annotation, visualisation and *in silico* analysis. Reads of both sequencing approaches were aligned to the circularized genome of AcMNPV strain E2 (GeneBank accession: KM667940.1) and the host cell genome (*Spodoptera frugiperda* isolate Sf9; BioProject accession: PRJNA380964) using GMAP v2017-04-24[28]. For the annotation of TSSs, AcMNPV alignments were analysed using the Smith-Waterman algorithm, with a match cost of +2, a mismatch cost of −3, a gap open cost of −3, and a gap extension cost of −2. The last 16 nucleotides of the MinION 5′ adapter were aligned to a window of −10 nucleotides (nt) upstream and +30 nt downstream of the first mapped nucleotide of the read. Reads with a score of less than 17 were considered as putative false 5′ ends caused by strand switching or non-specific priming. A position was considered the 5′ end of a read if the last four nucleotides of the MinION 5′ adapter were detected in a window of −2 upstream to +3 downstream of the first mapped nucleotide. The use of this 5-nucleotide window is necessary due to the varying numbers of base called G letters, caused by the sequencer's homopolymer error, and the homology uncertainty of the mapper. A 5′ end position was considered TSS if the number of reads starting at this position was significantly higher than at other nucleotides in the region surrounding this start position. For this the Poisson-probability (Poisson [k_0; λ]) of k_0 read starting at a given nucleotide in the −50 nt to +50 nt window from each local maximum was calculated with $\lambda = \frac{\sum_{i=-50}^{50} k_i}{101}$.

The 5′ ends of the longest low-abundance reads were individually inspected using Integrative Genomics Viewer (IGV)[29]. Poly(A) tails were defined as soft-clipped homopolymer A or homopolymer T stretches of at least 15 nts, or in some cases for the ONT direct RNA sequencing data soft-clipped CT-rich or GA-rich regions. The latter is required because of the base caller error of the dRNA-Seq[16]. The last mapped nt upstream of a poly(A) tail was considered as TES if at least 10 mapped reads ended in the given position, except when the TES was discovered using dRNA-Seq, where in the absence of any known bias resulting in a false TES, a single mapped read was deemed confirmatory. To avoid non-specific priming, reads with three or more mapped A letters on their 3′ end were discarded from the cDNA dataset. Reads with poly(A) tails on both of their ends were discarded, except for complex transcripts, for which the previously annotated TES was considered. Reads with a larger than 10 nt difference in their 5′ or 3′ ends were considered novel length isoforms (L: longer 5′ UTR, S: shorter 5′ UTR, AT: alternative 3′ termination). Short length isoforms harbouring a truncated version of the known open reading frame (ORF) were considered novel putative coding transcripts, and designated as '0.5'. Short transcripts without an ORF and with a TES upstream of the coding transcript's 3′ end were designated novel 3′-truncated (TR) non-coding transcripts. Long reads spanning at least two known transcripts with different directions were named complex transcripts (C). We assume that these complex transcripts start at the closest upstream annotated TSSs. Splice junctions were accepted if the intron boundary consensus sequences (GT and AG) were present in at least two sequencing reads, or were confirmed either by dRNA-Seq or by PCR analysis. Promoters and initiation sites were discovered using MEME[29]. Possible protein products were predicted by aligning ORFs of putative non-coding RNAs to online databases using the BLASTP suit[30] with an expected threshold of 10. Reads were visualized using the Geneious software suite[31] and IGV.

PCR analysis. PCR analysis and polyacrylamide gel electrophoresis was performed for validating antisense transcripts and splicing events. SuperScript III Reverse Transcriptase (Life Technologies) enzyme, 70 ng of total RNA and gene specific primers were used for the cDNA synthesis, according to the manufacturer's instructions. A noRT control was used for testing the potential DNA contamination. The cDNA samples were amplified using the Applied Biosystem's Veriti Thermal Cycler with KAPA HiFi PCR Kit (KAPA Biosystems) according to the manufacturer's recommendations. The running conditions were as follows: 3 min at 95 °C for initial denaturation, followed by 35 cycles at 98 °C for 20 s (denaturation), at 63 °C for 20 s (annealing), and at 72 °C for 2 min (extension). Final elongation was set at 72 °C for 5 min. Primers used in this study are outlined in Supplementary Table S1. For amplicon separation and visualization, a 12% polyacrylamide gel was prepared. Lanes were loaded with either GeneRuler Ultra Low Range DNA Ladder (Thermo Fisher Scientific), or the samples. Staining was performed with GelRed (Biotium).

Short-read sequencing data acquisition and analysis. In order to compare short-read sequencing data of the AcMNPV transcriptome to our own long-read sequencing data, the single end Illumina sequencing reads deposited in the Sequence Read Archive under accession SRA057390 were retrieved. The reads were aligned to the reference genome of AcMNPV strain E2 (GeneBank accession: KM667940.1) using TopHat v. 2.1.1[32]. Reads were visualized using IGV.

Results

Analysis of the AcMNPV transcriptome with long-read sequencing. In this study, we carried out ONT MinION cDNA and direct RNA sequencing analysis (Fig. 1) of the AcMNPV transcriptome (Fig. 2). The cDNA-Seq yielded 324,677 reads of which 103,133 mapped to the AcMNPV genome (the rest was mapped to the transcripts of the host cell), with an average read length of 1,053 bp and an average genome coverage of 510. The dRNA-Seq technique yielded 6,482 reads, 2,430 mapping to the viral genome, with an average read length

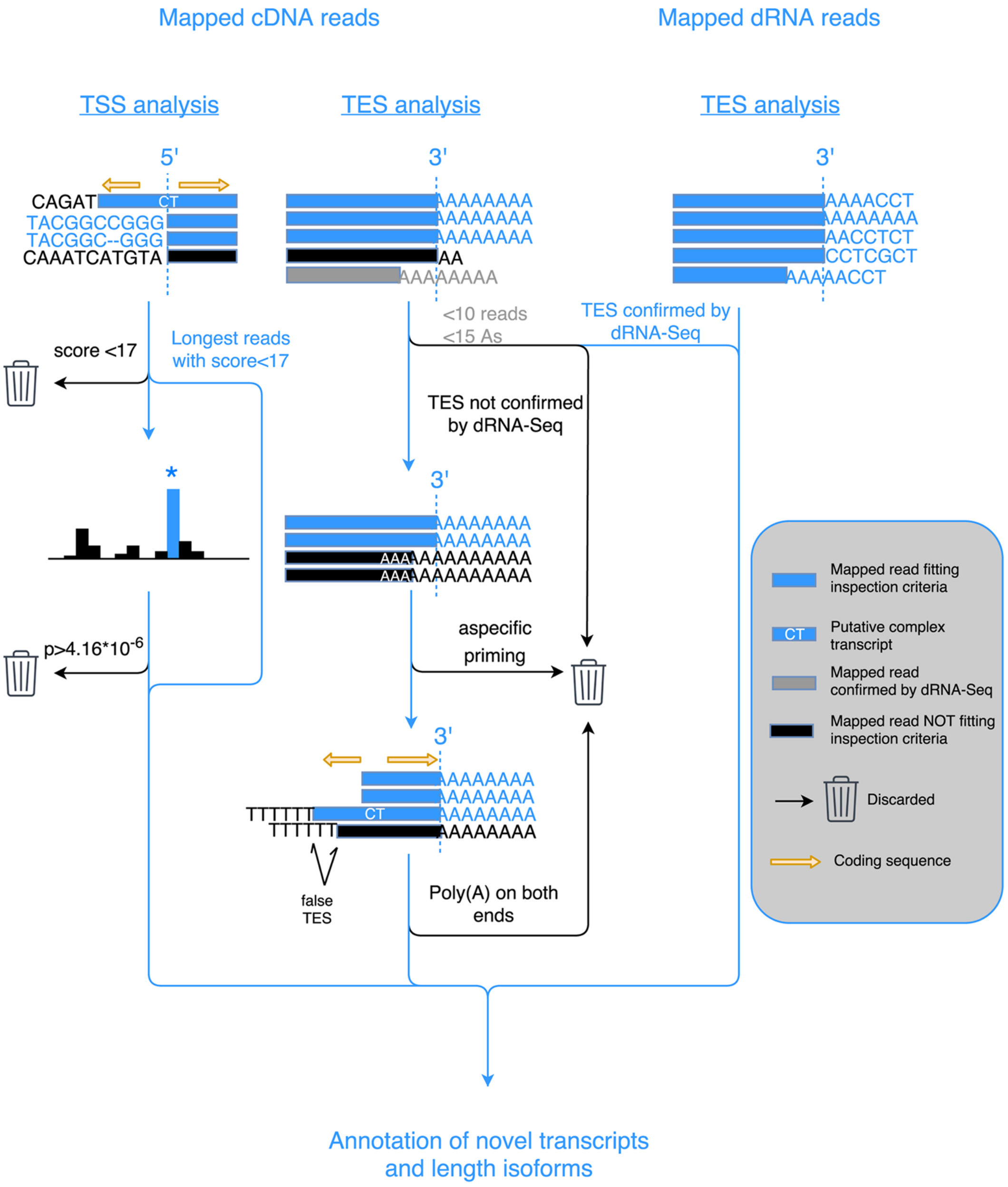


Figure 1. The schematic representation of the workflow used to annotate novel transcripts. Only the cDNA-sequencing reads were used for identifying TSSs. The read start distribution was obtained by only counting reads which contained the 5′ adaptor sequence. Nucleotide positions where the read start distribution was significantly different from a Poisson distribution were annotated as TSSs. In the case of complex transcripts, if no TSS could be determined, the closest upstream TSS was assumed to be the TSS. Both dRNA and cDNA reads were used to identify TESs. In case of the dRNA-Seq, support by one polyA-containing read was sufficient to annotate a TES, while in the case of cDNA-Seq, at least 10 reads were required to support a given TES, and also reads that could have arisen through non-specific priming, and reads containing poly(A) sequences on both ends were discarded from the TES analysis. The complex transcripts, supported by reads with double polyA-tails, were annotated, and their orientation was determined according to the polyA-tail, which passed the TES annotation criteria.

of 614 bp and an average genome coverage of 10. In this work, we detected and annotated altogether 132 novel RNA molecules, including 80 full-length transcripts and 46 transcripts with undetermined TSSs (Supplementary Table S2). We identified five novel splice isoforms (Supplementary Table S3). Additionally, we determined the TSSs of 64 and the TESs of 113 earlier reported transcripts with base-pair precision. The MinION cDNA sequencing technique allows the precise identification of TSSs, although similarly to other techniques, it is also afflicted by sample degradation. Another issue arises when more than three nucleotides of the MinION strand switching oligonucleotide is found on the viral genome, allowing both template switching and false priming. These events resulted in a total of 154 probably false 5′ ends, of which 47 belong to novel transcripts or transcript isoforms,

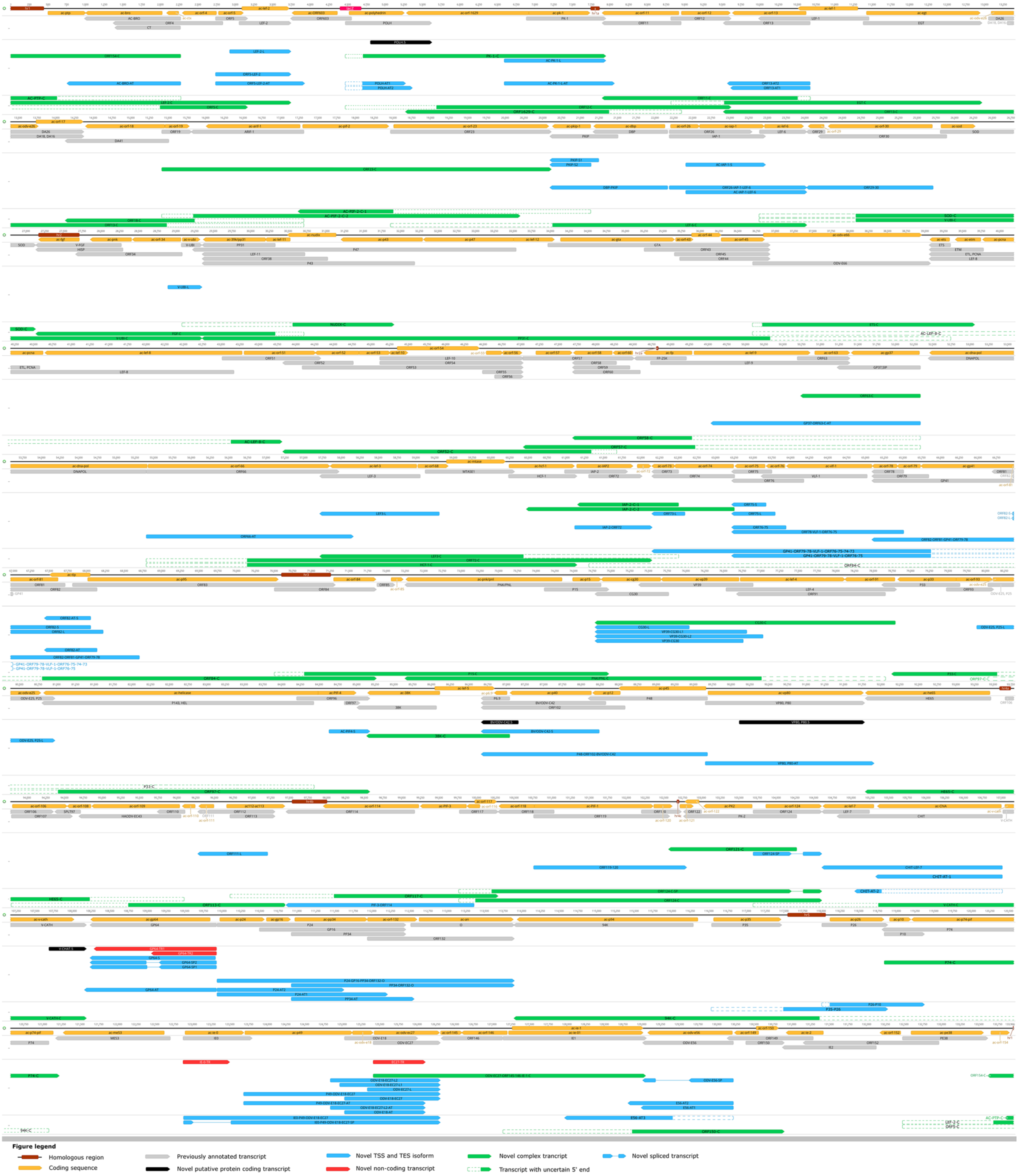


Figure legend
Homologous region
Coding sequence
Previously annotated transcript
Novel putative protein coding transcript
Novel TSS and TES isoform
Novel non-coding transcript
Novel complex transcript
Transcript with uncertain 5' end
Novel spliced transcript

Figure 2. Location of the previously and newly annotated transcripts on the linear view of the AcMNPV genome. Colour code: brown rectangles: homologous regions; yellow arrow-rectangles: coding sequences; grey arrow-rectangles: previously annotated transcripts; black arrow-rectangles: novel putative protein coding transcripts; blue arrow-rectangles: novel TSS and TES isoforms and novel polycistronic transcripts; red arrow-rectangles: novel non-coding transcripts; green arrow-rectangles: novel complex transcripts, purple rectangle: lacZ gene inserted in to the genome. Transcripts with undetermined TSSs were hypothesized to start at the closest upstream TSS in the same orientation, and their missing segment is marked by dashed contours.

including 38 complex transcripts, 6 polycistronic transcripts, and 3 transcripts with alternative termination. The 5′ ends of these transcripts were annotated as undetermined TSSs. False 5′ ends, marked by the absence of the MinION strand switching oligonucleotide, show a much higher variation around the TSS than real TSSs (Fig. 3, panel a), which is probably caused by the frequent template switching events, and demonstrates the utility of our workflow for distinguishing between false and real TSSs. It has been previously shown[33] that the TSSs of the transcripts vary with a few nucleotides. In our annotation, we chose a position as TSS, where significantly more reads

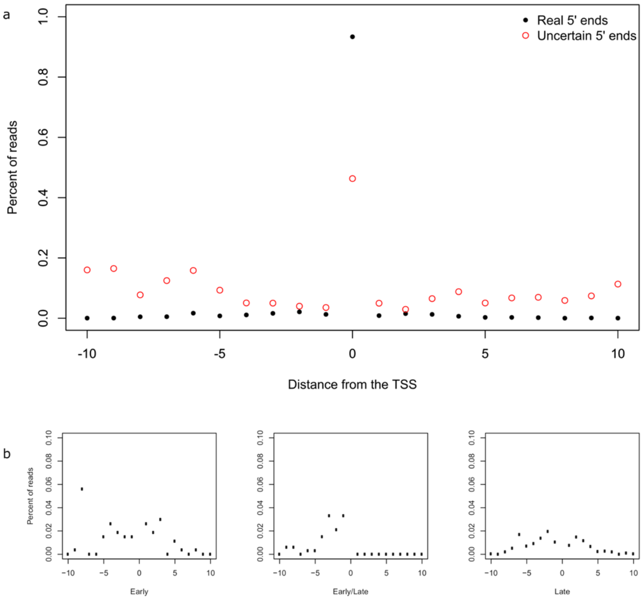


Figure 3. TSS variations. Panel a shows the frequency of reads starting in the vicinity of TSSs for real 5′ ends (black dots) and in case of uncertain 5′ ends (red circles). Panel b shows the frequency of reads starting in the vicinity of real TSSs of temporally different expression. The salient values of the TSS positions are not shown.

started than expected in its surrounding 101 bp region. Out of the 655 positions where reads with real 5′ ends started, 101 were accepted as TSSs ($p < 4.16 * 10^{-6}$, Bonferroni 0.05/101/119), of which 79 belonged to novel transcripts. The remaining positions were considered to be the result of TSS variation and RNA degradation. Template switching and false priming errors contribute to the artefactual 3′ ends of the reads, where the oligo(dT) primers, hybridize with homologous stretches of the transcripts, generally with a much lower affinity. Out of the 23,261 3′ end positions, we found 496 where the upstream genomic region contained at least three A letters, which were removed from further analysis. We discovered 45 novel positions with no evidence of template switching or false priming, thus these were considered as novel TESs.

Analysis of the viral transcriptome using short-read sequencing. In order to compare short- and long-read sequencing of the viral transcriptome, we retrieved and aligned the Illumina sequencing data produced by Chen and coworkers[3]. In total 188,025,259 reads mapped to the AcMNPV genome.

Novel putative protein-coding genes. The putative protein-coding genes identified in this work are all 5′-truncated forms of previously annotated genes. These transcripts share common PA signals with their host genes. In total, we discovered four novel putative mRNAs, three of these starting from the second nucleotide of the canonical LIS, TAAG (the underlined A letter is the first nucleotide of the transcripts). BV/ODV-C42.5, V-CHAT.5, and P80.5 feature the same initiator as their non-truncated forms implying their late transcription. This is indicated by the absence of canonical TATA promoters in these transcripts. The TSS of POLH.5 is located in a ACAGG motif, which is similar to the previously described arthropod initiator element ACAGT[34,35] except that it is not preceded by a canonical TATA motif. The presence of this initiator indicates that an 819 bp long 5′ truncated POLH may be transcribed by the host RNAP. Regarding their TESs, all novel putative protein-coding transcripts contain one or more PASs upstream their 3′ ends, starting at an average distance of 16.2 bp.

Novel non-coding transcripts. In this study, we identified four novel non-coding transcripts, all being 3′ truncated forms of already known mRNAs. BLASTP analyses of the transcripts' ORFs show no homology with known proteins. These transcripts start from the same promoter (all from the LIS) as their host genes, as in GP64-TR1, GP64-TR2, IE-0-TR and EC27-TR transcripts. These RNA molecules have the same TSS as their full length transcripts, but they lack stop codons and thus ORFs. The gene *gp64* is expressed at both early and late stages of the viral infection, which is indicated by the presence of two TATA promoters upstream it's TSS. Unlike GP64-TR2, GP64-TR1 has no canonical PAS, instead the transcript ends at a homopolymer T comprised of four nucleotides. This can act as a polyadenylation signal for the viral RNA polymerase[7].

The IE-0 is the non-spliced version of the IE-1 transcript, and encompasses the *ie0* gene of the virus[36], but is expressed during the late phase. IE-0-TR has the same TSS as IE-0, and harbours a PAS 17 bp upstream of its TES. The promoter consensus sequences of the novel 3′ truncated transcripts are identical to those of their full length variants, albeit the presence of four T letters adjacent to the TES of GP64-TR2 suggests a late transcription.

TSS and TES isoforms. Transcriptional start site isoforms differ in the length of their 5′UTRs from each other. Using long-read sequencing, we could identify 23 novel TSS variants. Twenty-one of these start at the LIS (Supplementary Fig. S1), like their previously annotated isoforms, except two transcripts, ORF73 and EC27,

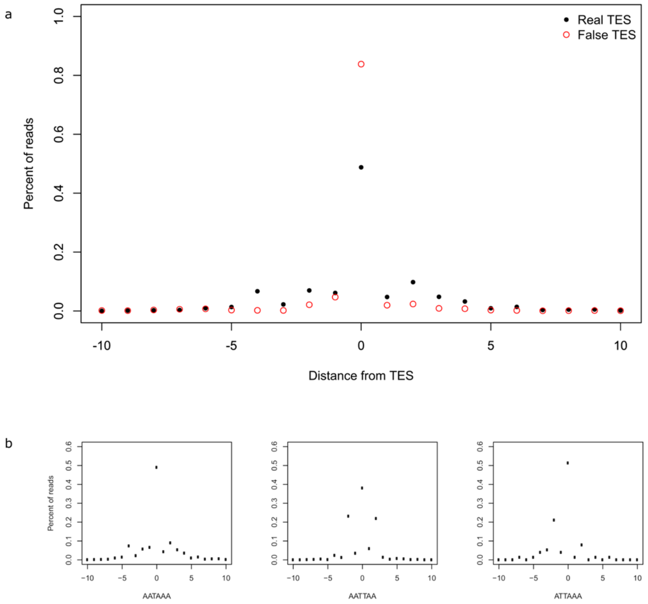


Figure 4. TES variations. Panel a shows the frequency of reads ending in the vicinity of TESs for real 3′ ends (black dots) and in case of uncertain 3′ ends (red circles). Panel b shows the frequency of reads ending in the vicinity of real TESs for the first three most common PASs.

which initiate at a consensus TAAG, but their longer isoform OFR73-L and EC27-L is missing this sequence or a canonical TATA motif. ORF111 was previously characterized as an early gene[3], but its longer isoform, ORF111-L starts at the LIS, and contains two upstream TATA boxes at 24-bp and 41-bp distance. Additionally, LEF-3 and its longer isoform LEF-3-L both initiate at a slightly modified version of the arthropod initiator: CATT and CAAT respectively. We found that five out of the late TSS isoforms contain canonical TATA motifs at an average of 12.4 bp upstream their TSS, three of which (ORF75-L, ORF82-L and VP39-CG30-L1) are adjacent to the LIS. The shorter isoform of GP64, designated GP64-S, contains both a late initiator and a TATA box, which is consistent with the early and the late transcription of its full length isoform. A comparison of the TSS variations of the three kinetic classes shows that early and early/late mRNAs tend to vary more around their most abundant start sites than late transcripts, ($SD_E = 0.014$, $SD_{E/L} = 0.01$, $SD_L = 0.006$, Fig. 3, panel b), which is to be expected due to the LIS-dependent initiation of late transcripts. Transcripts with uncertain 5′ end were labelled as starting at the closest upstream TSS, because we assume that they are controlled by the corresponding upstream promoters.

Transcriptional end site isoforms differ in the 3′ UTRs compared to the formerly annotated transcripts. In this study, we detected twenty-four TES variants. Twelve of these have canonical PASs, while six of the transcripts are terminated at a T-rich region with an average T count of 2.5 (Supplementary Table S4). This and the presence of a LIS at their TSS suggest that they are probably transcribed by the viral RNAP. The same may be true for AC-PK-1-L-AT, the single TES isoform out of the nine without a canonical PAS, but starting at a LIS and terminating at a region composed of 3 T letters. We found that the real 3′ ends show a greater variation around the TES compared to false read endings (Fig. 4 panel a). We also demonstrated that the variance in the TES locations depends on the sequence composition of the upstream PAS (Fig. 4 panel b).

Novel splice isoforms. We identified five novel splice isoforms and confirmed the existence of three previously described spliced transcripts, all with a consensus GT at the splice donor site and AG at the splice acceptor site (Supplementary Table S3). A novel splice variant of ODV-E56, designated ODV-E56-SP, was confirmed by dRNA-Seq, while the other novel splice isoforms were confirmed using PCR analysis followed by PAGE (Supplementary Fig. S2). We detected all of the novel splice sites in the Illumina short-read sequencing datasets, but they represented less than 0.1% of the reads in the given region, which is below the detection threshold of 5 RPKM used by Chen *et al.*[3]. The low abundance of these spliced transcripts is also indicated by the weak bands corresponding for the spliced amplicons in lane A and B of the PAGE. Splicing causes frame-shift in ODV-E56 and GP64-SP2 resulting in putative 3′ truncated, 191 and 261 amino acid long peptides. Further investigations are needed in order to reveal the functions of these transcripts, if they have any.

Polycistronic transcripts. This study revealed extensive polycistronism in the AcMNPV transcriptome. Altogether, twelve bicistronic, four tricistronic, three tetracistronic, three pentacistronic, and a single septacistronic transcripts have been detected by long-read sequencing. Ten of these transcripts share their TSSs with the first genes of the polycistronic unit that are also expressed as monocistronic RNA molecules. Five of the polycistronic transcripts contain novel TSSs. These transcripts contain LISs and are terminated at an average distance of 23.3 bp downstream of a canonical PAS. The ORF29-30 starts at a CAGT motif and contains a PAS 21 bp upstream its TES. Nineteen of the detected polycistronic transcripts have the same TES as the most downstream gene expressed as monocistronic transcripts.

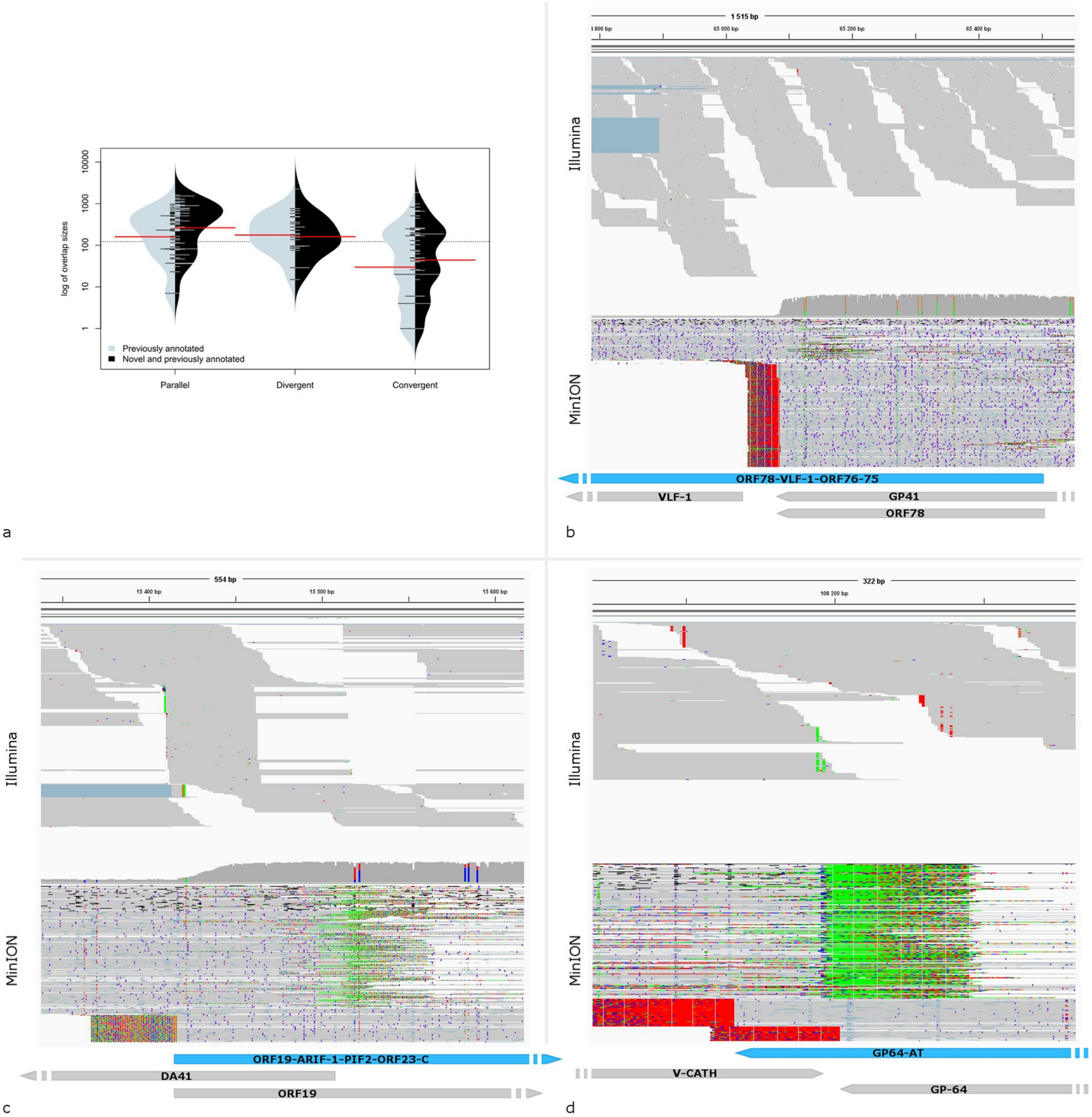


Figure 5. Size distributions and examples for the transcriptional overlaps. Panel a shows the comparison of size distributions of transcriptional overlaps of previously annotated transcripts (grey) versus previously annotated and novel transcripts with a certain TSS (black). Colour code: The red lines represent the mean values. Panel b Parallel; Panel c Divergent, and Panel d Convergent overlaps of previously annotated and novel transcripts. Reads from the Illumina and MinION cDNA sequencing were visualized using IGV, and their schematic representation. Grey arrow-rectangles represent previously annotated transcripts, while blue arrow-rectangles represent novel transcripts. The annotations were cut to fit the representation, indicated by a non-continuous annotation on one side.

Complex transcripts. The complex transcripts are special polycistronic RNAs with genes standing in opposite orientations. In this work, we identified forty-nine complex transcripts, ten of which start at a LIS, and three of them have a canonical TATA promoter at an average distance of just 2.6 bp upstream. One isoform of ORF124-C contains the same splice site as ORF124-SP, and thus has been tagged ORF124-C-SP. Four of the complex transcripts overlap HR1, a homologous region of the AcMNPV genome, which serves as replication initiation site[37]. We detected four complex transcripts that overlap the hr1, a putative origin of replication of AcMNPV.

Novel transcriptional overlaps. The viral transcripts can overlap in parallel (tail-to-head), convergent (tail-to-tail) or divergent (head-to-head) manners. Our study revealed a complex meshwork of transcriptional read-throughs in AcMNPV. We discovered 72 parallel, 37 convergent and 8 divergent overlaps of the previously annotated and novel transcripts, doubling the number (105 previously annotated and 117 novel) and increasing the average size (228.59 bp and 414.7 bp respectively) of transcriptional overlaps (Supplementary Table S5 and Fig. 5). We identified 98 additional overlaps (34 parallel, 31 divergent, 33 convergent) where one of the TSSs of the partner transcripts was only predicted. Comparison of the short-read sequencing data to our long-read sequencing data revealed that the overlapping reads between adjacent transcripts can be found in both datasets, however especially in the case of the parallel overlaps (Fig. 5b), the identification of the TSSs in the overlapping region using short reads is nearly impossible.

Discussion

In this study, we used the Oxford Nanopore Technologies' MinION long-read sequencing platform for the characterization of the transcriptional landscape of a baculovirus. Our investigations revealed a much higher complexity of the viral transcriptome than it had been formerly described. Using both cDNA and direct RNA sequencing methods, we have identified 132 novel transcripts and transcript isoforms, most of which overlap already known transcripts. The longer TSS and TES variants either form a new overlap or they increase the extent of the overlaps formed by the shorter transcript isoforms. No complex transcripts have been described in baculoviruses before this study. These RNA molecules together with the polycistronic transcripts represent very long transcriptional overlaps. Polycistronism is typical in the bacterial genes, however, this phenomenon is extremely rare in eukaryotic organisms. Some eukaryotic viruses solved the problem of reading multiple messages from a single RNA molecule using various mechanisms, such as leaky ribosomal scanning (in retroviruses[38] and papillomaviruses[39]), ribosomal frameshifting (retroviruses[40]) or utilizing internal ribosome entry sites (in picornaviruses[41]). However, no such mechanisms have been described in baculoviruses so far, therefore only the most upstream genes of the RNA molecules will be translated. To address the question of multiple ORFs translating from polycistronic mRNAs there is a need for further studies, such as ribosome profiling[42]. The question can thus be raised about the role of the multigenic RNA molecules in the viral pathogenesis, if there is any. Theoretically, it is possible that these long transcripts represent transcriptional noise. It is also possible these RNA molecules are precursors for smaller regulatory molecules or that they have post-transcriptional function. We can speculate whether these molecules are mere by-products of a transcriptional read-through mechanism whose function is to regulate gene expression through the collision and/or competition between the transcription machineries of adjacent and distal genes. Indeed, the phenomenon of transcriptional interference between the convergent gene pairs have been described in other organsisms[43,44]. These interactions may form a genome-wide meshwork for the regulation of gene expression designated as transcriptional interference network[45]. We also identified four complex transcripts overlapping one of the replication origins (hr1) of the baculovirus, which suggest a potential interaction between the transcription and replication machineries. In this study, we also demonstrated that the TSS isoforms have the same types of promoters and initiator regions, suggesting that the expression kinetics of the longer and the shorter transcripts of a given gene are the same. In contrast, a different promoter or initiator motif may mean an uncoupled transcription of the mRNA isoforms, which could be the result of random processes, but can serve regulatory purposes, such as differential gene expression. We annotated promoters and ORFs for the novel 5′-truncated mRNAs, which therefore may encode protein molecules, but we cannot exclude that they are long non-coding RNAs, or are precursors for short non-coding RNAs. Additionally, we demonstrated the significance of template switching and false priming on transcript isoform annotation using the ONT long-read sequencing technology. While it was formerly considered that these effects can play a role in false TES annotation in the transcripts containing homopolymer-A stretches[22] upstream their 3′ ends, our results suggest that the presence of homology between the 5′-end adapter and the transcripts can also result in a number of false transcription reads.

Data availability. The sequencing data and the transcriptome assembly have been uploaded to the European Nucleotide Archive under the project accession number PRJEB24943.

References

1. The Complete DNA Sequence of Autographa californica Nuclear Polyhedrosis Virus. *Virology* **202**, 586–605 (1994).
2. Rohrmann, G. F. *Baculovirus Molecular Biology*. (National Center for Biotechnology Information (US) 2013).
3. Chen, Y.-R. *et al*. The transcriptome of the baculovirus Autographa californica multiple nucleopolyhedrovirus in Trichoplusia ni cells. *J. Virol.* **87**, 6391–405 (2013).
4. Kogan, P. H., Chen, X. & Blissard, G. W. Overlapping TATA-dependent and TATA-independent early promoter activities in the baculovirus gp64 envelope fusion protein gene. *J. Virol.* **69**, 1452–61 (1995).
5. Garrity, D. B., Chang, M.-J. & Blissard, G. W. Late Promoter Selection in the Baculovirusgp64 Envelope Fusion ProteinGene. *Virology* **231**, 167–181 (1997).
6. Jin, J. & Guarino, L. A. 3′-end formation of baculovirus late RNAs. *J. Virol.* **74**, 8930–7 (2000).
7. Kost, T. A., Condreay, J. P. & Ames, R. S. Baculovirus gene delivery: a flexible assay development tool. *Curr. Gene Ther.* **10**, 168–73 (2010).
8. Hu, Y. Baculovirus as a highly efficient expression vector in insect and mammalian cells. *Acta Pharmacol. Sin.* **26**, 405–416 (2005).
9. Haase, S., Sciocco-Cap, A. & Romanowski, V. Baculovirus insecticides in Latin America: historical overview, current status and future perspectives. *Viruses* **7**, 2230–67 (2015).
10. Kukurba, K. R. & Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring Harb. Protoc.* **2015**, 951–69 (2015).
11. Tombácz, D. *et al*. Characterization of the Dynamic Transcriptome of a Herpesvirus with Long-read Single Molecule Real-Time Sequencing. *Sci. Rep.* **7**, 43751 (2017).
12. Oláh, P. *et al*. Characterization of pseudorabies virus transcriptome by Illumina sequencing. *BMC Microbiol.* **15**, 130 (2015).
13. Liu, L. *et al*. Comparison of Next-Generation Sequencing Systems. *J. Biomed. Biotechnol.* **2012**, 1–11 (2012).
14. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016).
15. Križanović, K., Echchiki, A., Roux, J. & Šikić, M. Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics*, https://doi.org/10.1093/bioinformatics/btx668 (2017).
16. Moldován, N. *et al*. Multi-Platform Sequencing Approach Reveals a Novel Transcriptome Profile in Pseudorabies Virus. *Front. Microbiol.* **8**, 2708 (2018).

17. Clarke, J. *et al*. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4**, 265–270 (2009).
18. Manrao, E. A. *et al*. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nat. Biotechnol.* **30**, 349–353 (2012).
19. Laver, T. *et al*. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* **3**, 1–8 (2015).
20. Luo, G. X. & Taylor, J. Template switching by reverse transcriptase during DNA synthesis. *J. Virol.* **64**, 4321–8 (1990).
21. Cocquet, J., Chong, A., Zhang, G. & Veitia, R. A. Reverse transcriptase template switching and false alternative transcripts. *Genomics* **88** (2006).
22. Kuo, R. I. *et al*. Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics* **18**, 323 (2017).
23. Tombácz, D. *et al*. Long-Read Isoform Sequencing Reveals a Hidden Complexity of the Transcriptional Landscape of Herpes Simplex Virus Type 1. *Front. Microbiol.* **8**, 1079 (2017).
24. Balázs, Z. *et al*. Long-Read Sequencing of Human Cytomegalovirus Transcriptome Reveals RNA Isoforms Carrying Distinct Coding Potentials. *Sci. Rep.* **7**, 15989 (2017).
25. Tombácz, D. *et al*. Full-Length Isoform Sequencing Reveals Novel Transcripts and Substantial Transcriptional Overlaps in a Herpesvirus. *PLoS One* **11**, e0162868 (2016).
26. Smith, I. Misleading messengers? Interpreting baculovirus transcriptional array profiles. *J. Virol.* **81**, 7819-20-1 (2007).
27. Jiang, S. S. *et al*. Temporal transcription program of recombinant Autographa californica multiple nucleopolyhedrosis virus. *J. Virol.* **80**, 8989–99 (2006).
28. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–75 (2005).
29. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).
30. Altschul, S. F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–402 (1997).
31. Kearse, M. *et al*. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–9 (2012).
32. Kim, D. *et al*. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
33. Amman, F. *et al*. TSSAR: TSS annotation regime for dRNA-seq data. *BMC Bioinformatics* **15**, 89 (2014).
34. Xing, K. *et al*. Analysis and prediction of baculovirus promoter sequences. *Virus Res.* **113**, 64–71 (2005).
35. Cherbas, L. & Cherbas, P. The arthropod initiator: the capsite consensus plays an important role in transcription. *Insect Biochem. Mol. Biol.* **23**, 81–90 (1993).
36. Stewart, T. M., Huijskens, I., Willis, L. G. & Theilmann, D. A. The Autographa californica multiple nucleopolyhedrovirus ie0-ie1 gene complex is essential for wild-type virus replication, but either IE0 or IE1 can support virus growth. *J. Virol.* **79**, 4619–29 (2005).
37. Leisy, D. J., Rasmussen, C., Kim, H.-T. & Rohrmann, G. F. The Autographa californica Nuclear Polyhedrosis Virus Homologous Region 1a: Identical Sequences Are Essential for DNA Replication Activity and Transcriptional Enhancer Function. *Virology* **208**, 742–752 (1995).
38. Schwartz, S., Felber, B. K. & Pavlakis, G. N. Mechanism of translation of monocistronic and multicistronic human immunodeficiency virus type 1 mRNAs. *Mol. Cell. Biol.* **12**, 207–19 (1992).
39. Stacey, S. N. *et al*. Leaky scanning is the predominant mechanism for translation of human papillomavirus type 16 E7 oncoprotein from E6/E7 bicistronic mRNA. *J. Virol.* **74**, 7284–97 (2000).
40. Jacks, T. *et al*. Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature* **331**, 280–283 (1988).
41. Pelletier, J. & Sonenberg, N. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature* **334**, 320–325 (1988).
42. McGlincy, N. J. & Ingolia, N. T. Transcriptome-wide measurement of translation by ribosome profiling. *Methods* **126**, 112–129 (2017).
43. Prescott, E. M. & Proudfoot, N. J. Transcriptional collision between convergent genes in budding yeast. *Proc. Natl. Acad. Sci. USA* **99**, 8796–801 (2002).
44. Greger, I. H., Demarchi, F., Giacca, M. & Proudfoot, N. J. Transcriptional interference perturbs the binding of Sp1 to the HIV-1 promoter. *Nucleic Acids Res.* **26**, 1294–301 (1998).
45. Boldogkői, Z. Transcriptional interference networks coordinate the expression of functionally related genes clustered in the same genomic loci. *Front. Genet.* **3**, 122 (2012).

Acknowledgements

We would like to thank Ernő Duda Jr. for supplying the virus and the cell line used in this study. This work was supported by the Swiss-Hungarian Cooperation Programme [SH/7/2/8] to Z.B.o. and the Nation's Talents Programme [NTP-NFTÖ-17B] scholarship of the Hungarian Ministry of Human Capacities awarded to Z.B.a. The work was also supported by the Bolyai János Scholarship of the Hungarian Academy of Sciences to D.T.

Author Contributions

N.M. carried out the ONT MinION dRNA sequencing, analysed the data, participated in the sequence alignment and drafted and wrote the manuscript. D.T. carried out the ONT MinION cDNA sequencing, participated in the design of the study, and took part in drafting the manuscript. A.S. participated in the sequence alignment and carried out the *in silico* analysis. Z.C. prepared the RNA, DNA and cDNA samples and participated in the MinION sequencing. E.K. and M.J. carried out the virus infection and propagated the cells. Z.Ba. performed statistical analysis and revised the manuscript. Z.B.o. conceived, designed and coordinated the study and wrote the manuscript. All authors have read and approved the final version of the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-26955-8.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

II.

Cell Press
REVIEWS

Review

Long-Read Sequencing – A Powerful Tool in Viral Transcriptome Research

Zsolt Boldogkői,[1,*] Norbert Moldován,[1] Zsolt Balázs,[1] Michael Snyder,[2] and Dóra Tombácz[1]

Long-read sequencing (LRS) has become increasingly popular due to its strengths in *de novo* assembly and in resolving complex DNA regions as well as in determining full-length RNA molecules. Two important LRS technologies have been developed during the past few years, including single-molecule, real-time sequencing by Pacific Biosciences, and nanopore sequencing by Oxford Nanopore Technologies. Although current LRS methods produce lower coverage, and are more error prone than short-read sequencing, these methods continue to be superior in identifying transcript isoforms including multispliced RNAs and transcript-length variants as well as overlapping transcripts and alternative polycistronic RNA molecules. Viruses have small, compact genomes and therefore these organisms are ideal subjects for transcriptome analysis with the relatively low-throughput LRS techniques. Recent LRS studies have multiplied the number of previously known transcripts and have revealed complex networks of transcriptional overlaps in the examined viruses.

Transcriptome Research on Viruses

Viruses represent a diverse class of microorganisms with polyphyletic origin. Compared to cellular organisms, even the largest DNA viruses have small genomes with closely-spaced genes. This feature makes viruses excellent model systems in molecular biology to explore the general principles of genetic regulation and transcriptome organization.

Alternative splicing increases the coding potential of the genome through the production of multiple RNA and protein molecules from a single gene. Similarly, alternative transcription initiation and termination also contribute to the genomic complexity. Polycistronism is a common phenomenon in bacteria and in their viruses but it is extremely rare in eukaryotes. The reason for this is that, in prokaryotes, the Shine–Dalgarno sequences allow the translation of each gene in the mRNA [1]. Nevertheless, in eukaryotes only the most upstream gene of a polygenic transcript is translated because of the Cap-dependent initiation system. Some small RNA viruses have evolved miscellaneous strategies to solve the problem of translation of multiple (generally two) proteins from a single transcript, which includes the utilization of an internal ribosome entry site (IRES), or mechanisms to bypass the 5′-proximal AUG to enable downstream initiation, such as the leaky ribosomal scanning mechanisms and ribosomal frameshifting [2]. Nonetheless, in the majority of DNA viruses no such mechanisms have been described so far. The canonical termination sequences are not always efficient in stopping the RNA polymerase (RNP); therefore, transcription is continued until the next termination site is reached, which results in transcriptional readthrough (TRT) producing readthrough (rt)RNAs. Transcriptional overlaps (TOs), in most cases produced by TRT, have been shown to represent a common phenomenon in diverse organisms [3]. The latest studies have also shown an intricate meshwork of TRTs and TOs in various viruses.

Highlights

Long-read sequencing (LRS) has revolutionized genomics and transcriptomics. These third-generation approaches have a relatively low throughput compared to short-read sequencing, but they can solve problems that used to be a challenge for earlier techniques.

The PacBio and ONT sequencing are able to read full-length transcripts and allow the direct study of base modifications on both DNA and RNA molecules. Nanopore technology is able to sequence RNA directly.

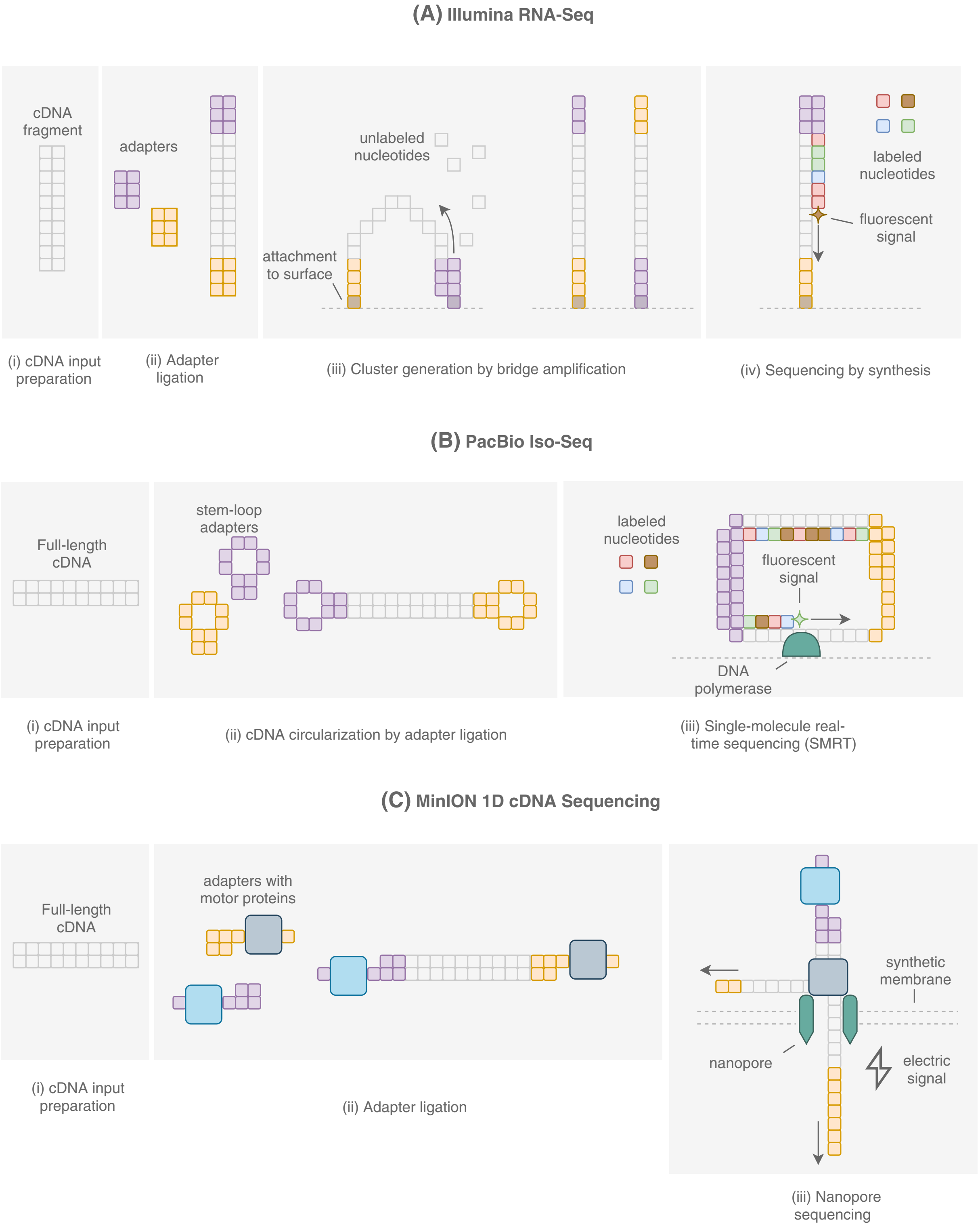
LRS has revealed a much more complex viral transcriptome. Among other capabilities, these techniques allow the discrimination between multispliced transcript variants, RNA length isoforms, embedded RNAs, and polycistronic RNA molecules.

The viral genomes express a highly complex pattern of transcriptional overlaps, the function of which continues to remain unknown.

[1]Department of Medical Biology, Faculty of Medicine, University of Szeged, Szeged, Hungary
[2]Department of Genetics, School of Medicine, Stanford University, 300 Pasteur Drive, Stanford, CA 94305-5120, USA

*Correspondence:
boldogkoi.zsolt@med.u-szeged.hu
(Z. Boldogkői).

https://doi.org/10.1016/j.tim.2019.01.010


(A) Illumina RNA-Seq
cDNA fragment
adapters
unlabeled nucleotides
attachment to surface
labeled nucleotides
fluorescent signal
(i) cDNA input preparation
(ii) Adapter ligation
(iii) Cluster generation by bridge amplification
(iv) Sequencing by synthesis
(B) PacBio Iso-Seq
Full-length cDNA
stem-loop adapters
labeled nucleotides
fluorescent signal
DNA polymerase
(i) cDNA input preparation
(ii) cDNA circularization by adapter ligation
(iii) Single-molecule real-time sequencing (SMRT)
(C) MinION 1D cDNA Sequencing
Full-length cDNA
adapters with motor proteins
synthetic membrane
nanopore
electric signal
(i) cDNA input preparation
(ii) Adapter ligation
(iii) Nanopore sequencing

Trends in Microbiology

(See figure legend at the bottom of the next page.)

The next-generation short-read sequencing (SRS) technology, released in the mid-2000s, has revolutionized genomic and transcriptomic sciences due to its massively parallel nature, which has enabled sequencing of millions of DNA fragments simultaneously at a relatively low cost. The Illumina platform is by far the most widely applied SRS technique. The enormous number of reads generated by SRS enabled the sequencing of entire genomes of various organisms at an unprecedented speed. The currently running genome programs are mainly based on the SRS approach [4,5]. This technique has also been extensively used to study the transcriptomes of various organisms [6,7]. The third-generation LRS technology emerged in 2011, when Pacific Biosciences (PacBio) commercialized the single-molecule real-time (SMRT®) technology [8]. Currently, two LRS technologies are in use: the PacBio and the Oxford Nanopore Technology (ONT) platforms. MinION, the first prototype of ONT was released in 2014 [9]. Both LRS techniques are based on the development of novel biochemistry, which enables the direct capture of long DNA sequences or cDNAs from full-length transcripts. ONT has also developed a method for sequencing native RNAs [10] and, since Helicos [11] has withdrawn from the market, nanopore sequencing is the only commercially available direct (d)RNA sequencing method. SRS, however, is still outstanding for producing high-quality, deep-coverage datasets. This technique is more cost-effective and has a lower per-base error rate than the LRS approaches [12,13]. Complex genomic regions, including sequences with a high GC content, as well as repetitive sequences, cannot be efficiently resolved by SRS. The short read length also makes computations difficult or impossible for the determination of exon connectivity and for the identification of transcript isoforms, such as multispliced RNAs as well as transcription start site (TSS) and transcription end site (TES) variants [14]. Furthermore, alternative polycistronism and TOs, especially between embedded RNA (eRNA) molecules, pose challenges for the SRS platforms. These challenges can be overcome by the LRS technology since it is able to provide full contig information about transcripts.

In recent years, LRS has been widely utilized in the analysis of the genomes of various organisms including prokaryotic [15] and eukaryotic [16] species as well as viruses [17–22]. Nevertheless, current LRS techniques are only able to characterize small genomes and transcriptomes in high depth due to the comparatively low throughput. Viral transcriptomes used to be investigated by traditional techniques, including Northern blotting [23,24], quantitative PCR [25,26], RACE analysis [27], and microarray studies [28,29]. The introduction of Illumina sequencing [30–32] to virus research has led to a significant progress in the discovery and precise annotation of viral transcripts. Currently, the global transcriptome of several viruses belonging to different families has been analyzed by using various techniques of PacBio and ONT platforms [33–37]. Our review aims to provide an overview of the potentials and limitations of LRS methods, to present the transcriptome diversity that has been detected by long-read sequencing, and to discuss future paths of viral genomics opened up by this technology.

LRS

Similar to the Illumina approach, PacBio also adopts a sequencing-by-synthesis strategy, but while Illumina detects augmented signals from amplified DNA fragments, the PacBio technique captures a single DNA molecule (Figure 1 and Table 1). The PacBio SMRT® sequencing utilizes

Figure 1. Comparison of the Various Sequencing Platforms. (A) Illumina sequencing uses cDNA fragments, each of which is amplified multiple times to form a cluster. Sequencing is based on the fluorescent signal of the incorporated nucleotides. The emitted fluorescence from each cluster is strong enough to generate a specific signal in the detector. (B) During PacBio sequencing, full-length cDNA is circularized by stem-loop adapters and loaded into ZMWs, where the immobilized polymerase incorporates fluorescently labeled nucleotides. ZMWs are capable of detecting the fluorescent signal of a single nucleotide incorporation. The polymerase is able to take multiple passes on the cDNA template; thus, the consensus reads are more accurate than the raw reads. (C) The motor proteins ligated to the cDNA templates bind to nanopores on a synthetic membrane in MinION sequencing. The motor proteins ratchet one strand of the DNA through the nanopore, whereas a detector is measuring the potential changes on the two sides of the membrane. The electric signal is specific to the nucleotides passing through the membrane.

Table 1. Comparison of the Various Sequencing Platforms.

		Illumina		Pacific Biosciences		Oxford Nanopore Technologies	
		HiSeq	MiSeq	RSII	Sequel	MinION 1D	dRNA-Seq
Required amount of input material (ng)		1–50	1–50	–	1000	1000	500–775
Mapped read length (bp)	Mean	92	175	2088	13 800	1503.50	968
	Median	–	–	1720	–	1439	713
	Standard deviation	–	–	1438.14	–	969.18	–
	Maximum	101	250	8006	–	9345	21 866
Average percentage of mapped reads		84.1	84.4	90.29	–	69.27	96.5
Substitutions per base		0.0053	0.0142	0.0212	0.005	0.0754	0.024
INDELs per base		7.2×10^{-6}	–	0.0231	0.001	0.0856	0.0435
Sample type		gDNA	gDNA	cDNA	gDNA	cDNA	RNA
Mapping software used in the study		BWA-MEM	BWA-MEM	GMAP	–	GMAP	Minimap2
Refs		[41]	[41]	[42]	[43]	[42]	[44]

zero-mode waveguides (ZMW) [38] for single-molecule analysis, which allows the detection of fluorescent signals emitted during the incorporation of labeled nucleotides. A single DNA polymerase molecule, fixed at the bottom of a ZMW, reads the circularized template multiple times. When a nucleotide is incorporated in the growing DNA strand, the fluorescent tag is cleaved off, and it gets out of the observation area. The base-call is made by the detection of the fluorescent signal of the nucleotide incorporated within the ZMW [39]. The accuracy of the obtained consensus sequence (reads of inserts, ROI) depends upon the number of polymerase passes around the circular template [12]. Sequel, the newest PacBio platform, launched in 2015, has a capacity sevenfold greater than the former RS II platform [40]. Additionally, the Sequel system has a considerably decreased loading bias compared to RS II; therefore, it does not require size-selection[i]. SMRT® sequencing generates subreads, thereby resulting in multiple base coverage in a given base, which leads to increased precision. The base-calling accuracy depends on the read length and on the movie length. The PacBio Isoform sequencing (Iso-Seq®) allows the generation of full-length cDNA sequences without the need for contig assembly, and thus it is suitable for the confident characterization of the full complement of transcript isoforms across an entire transcriptome or within the targeted genes.

The nanopore technology is based on monitoring the transit of DNA or RNA molecules through a protein pore; it measures variations in electric currents produced by the nucleotides that are threaded through the nanopores aided by a molecular motor protein. Nanopore sequencing is able to determine very long nucleic acid sequences [45]. ONT 1D sequencing has low accuracy (approximately 85%) [46]. The 1D2 technology improves this accuracy by ligating an adapter to the end of the reads, which increases the probability that a strand and its complementary strand pass through a pore consecutively. The base-calling algorithm creates a consensus of the two reads, and it has an average quality of over 95%. ONT can also identify the nucleotide modifications of RNA molecules by native RNA sequencing [10]. The advantages of nanopore sequencing over the PacBio platform are the longer read length, the higher throughput, and the lower costs

[42]. The two LRS techniques are prone to similar errors, such as homopolymer bias and indel errors. High sequencing error rate makes accurate DNA sequencing difficult, including *de novo* sequencing and variant calling. Sequencing errors, however, do not represent a major obstacle in transcriptome research if well-annotated genome sequences to which the transcript reads can be aligned are available. The lower throughput compared to the SRS approach means that LRS can only characterize abundant transcripts and that technical by-products are more difficult to filter out from LRS data. Ligation and template switching are common causes of such artifacts. Template switching is caused by the release of the template strand by the polymerase molecule during synthesis followed by binding to another template that shares homology with the original template and can occur at both the reverse-transcription (RT) [47] and the PCR [48] steps. The advantage of dRNA sequencing of ONT is that it is free from RT and PCR artifacts. The shortcomings of the current dRNA technique are its demand for the starting material (at least 500 ng PolyA$^+$ RNA), very low throughput, and that the produced reads lack short sequences at both the 5′ and 3′ termini [35]; therefore it does not resolve isoforms with base-pair precision. The cDNA sequencing methods require slightly less starting material (at least 250 ng PolyA+ RNA without PCR or 200 ng amplicon after PCR using ONT protocols or 2 ng total RNA using the Iso-Seq® protocol) and have a higher throughput, although not as high as SRS techniques. To date, the majority of LRS protocols have focused on full-length polyA-selected RNAs [49]. Cap-selection of RNA molecules can also be used to enrich full-length RNA molecules [50].

SRS can Complement LRS

It has been demonstrated that SRS coupled with the so-called synthetic long-read sequencing method (SLR-Seq) can represent an alternative approach for full-length characterization of transcripts [51] at the cost of reducing the sequencing yields. SLR-Seq is also afflicted by SRS biases, such as poor characterization of GC-rich regions. LRS is definitely superior to SRS regarding isoform detection and the differential quantitation of isoforms; SRS still offers many advantages and could be used alongside LRS techniques [52]. For example, ChIP-Seq [53] and ribosome profiling [54,55] are methods which provide valuable information and are not well suited for current LRS technologies. SRS can also be used to improve LRS not only through error correction (Box 1)

Box 1. Technology Corner: The Bioinformatic Challenges of LRS

Owing to the higher error rate but greater length of long reads, different tools are needed to analyze LRS and SRS data. The preprocessing of the reads is platform-specific. SMRT Link from PacBio creates accurate consensus reads and also assembles consensus isoforms which can be mapped to the genome or can be analyzed further without the need for a genome sequence [90]. Such consensus isoforms are usually highly accurate, and no further error correction is necessary. The processing of nanopore reads is less standardized. Guppy has recently been declared to be the recommended base caller by ONT. For genome sequencing, the next step would be error correction either by using short reads or based solely on the nanopore sequencing [91–93]. For transcriptome sequencing, however, error correction does not appear to be beneficial prior to isoform identification as it may interfere with both the qualitative and the quantitative analysis [94]. A novel method which uses rolling-circle amplification to produce concatemers of cDNA molecules (R2C2) greatly improves the quality of nanopore reads while preserving the benefits of single-molecule sequencing [95]. Minimap2 is used for the alignment of reads from both sequencing technologies [96]. Recently, a number of software programs have been developed for the task of isoform discovery. Mandalorion was designed for isoform detection in 2D reads [97]; however, ONT no longer supports 2D technology, the new version of the pipeline currently only accepts highly accurate R2C2 reads. FLAIR focuses on the splice isoforms and requires an annotation of splice sites [98]. Pinfish[ii] bases isoform discovery on the clustering of reads to define median exon boundaries, whereas LoRTIA[iii] identifies and filters transcript features (TSS, intron, and TES), then constructs isoforms based on these features. At the moment, all of these tools require an available genome sequence to determine the exact nucleotide sequence of the transcripts, but, except for FLAIR, they do not require an existing annotation of transcript features such as splice sites Therefore they can be applied to the investigation of viral transcriptomes that have not been studied before. Due to the numerous challenges in the analysis of LRS data and the relative novelty of these tools, artifacts may be common and should be filtered out by inspection or, ideally, by quality-control programs, such as SQANTI, a pipeline that characterizes and filters isoforms based on an annotation [99].

but also through the precise characterization of transcript features, which can be helpful in isoform identification. PRO seq [56] identifies TSSs, whereas technologies such as 3′READS+ [57] characterize TESs with higher sensitivity and specificity than is possible with current LRS technologies. Concurrently, LRS can be used for the precise quantitative analysis of the viral transcriptome at the isoform level [58]. Using a non-amplified Iso-Seq® technique, the results of the kinetic categorization of PRV transcripts have been in agreement with earlier observations obtained by real-time RT-PCR analysis [26]. The ability of LRS methods to sequence PCR-free cDNA or RNA provides more accurate quantitation that is devoid of amplification bias. However, due to the low throughput of such LRS methods, SRS is still more efficient in characterizing host transcription. When combined with microfluidic technologies provided by 10x Genomics, LRS can differentiate between transcript isoforms whereas SRS can characterize gene expression at the level of a single cell [59], resulting in a more specific analysis of the viral–host interactions. Using LRS and SRS coupled with other techniques can eschew the deficiencies of each approach and opens the possibility of a wider analysis of the viral transcriptome.

Novel Viral Transcripts Identified by LRS

LRS techniques have already been used in the investigation of the transcriptomes of various viruses, including herpesviruses [33–37], baculoviruses [60], retroviruses [61], circoviruses [62], and poxviruses [22,63]. The application of these techniques has identified a much greater complexity of viral transcriptomes compared to earlier approaches. Most of the newly discovered transcripts belong to categories which are difficult to study with SRS and other techniques, such as embedded messenger RNAs (emRNAs), polygenic transcripts, overlapping transcripts, and RNA isoforms including splice, TSS, and TES variants. LRS studies have also revealed noncoding RNAs (ncRNAs), such as antisense RNAs (asRNAs), intergenic RNAs (iRNAs), and embedded noncoding RNAs (encRNAs). Special classes of transcripts termed near-replication-origin RNAs (nroRNAs) and nro-like transcripts have also been recently described [64,65].

Messenger RNAs

LRS techniques are especially efficient in identifying eRNAs sharing a common TSS or TES with the longer host transcript. An emRNA containing an in-frame open reading frame (ORF) within the coding region of the longer host gene has the potential to specify an N-terminally truncated polypeptide. LRS studies have multiplied the number of these truncated mRNAs in each subfamily of herpesviruses [35–37,65], and in a baculovirus [60].

Noncoding Transcripts

It used to be believed that DNA and protein molecules play a fundamental role in the control of cell functions, whereas RNAs were considered to have only subsidiary roles. Nonetheless, recent transcriptomics studies have revealed a huge variety of ncRNAs with a wide range of functions, including epigenetic, transcriptional, and post-transcriptional regulation of gene expression [66]. These transcripts are classified below according to their locations and orientations.

Intergenic Transcripts

These RNA molecules are located between two coding sequences without or with no significant overlap with the adjacent genes. LRS studies have identified several iRNAs in viruses, although these techniques are not superior to SRS in the detection of these types of transcript.

Embedded Noncoding Transcripts

The encRNAs can be mapped within either mRNAs or ncRNAs. The most typical mRNA-overlapping encRNAs are the 5′- and 3′-truncated transcripts. Generally, encRNAs have a common TSS or TES with other ncRNAs with which they overlap.

Antisense Transcripts

The asRNAs either completely or partially overlap the mRNAs in an antiparallel manner. These transcripts can either be controlled by their own promoters or they can be the result of transcriptional overlaps between neighboring or distal convergent or divergent genes (Figure 2). In this latter case, only the overlapping part of the transcript is antisense.

Replication-Associated RNAs, a Novel Class of Transcripts

The replication-associated RNAs (raRNAs) are mapped in close vicinity to the replication origins (Oris) of viral DNA [33,64,65]. LRS techniques have detected several such RNA molecules, including nroRNAs in herpesviruses [31,36,37], and nro-like transcripts in baculoviruses [60] and circoviruses [62]. Six types of replication-associated transcript can be distinguished in terms of their coding potency and position to the Ori: (i) mRNAs that do not overlap the Ori; (ii) ncRNAs that do not overlap the Ori; (iii) ncRNAs that do overlap the Ori; (iv) mRNA isoforms with very long overlapping alternative TES; (v) mRNA isoforms with very long overlapping alternative TSS; and (vi) mRNA with Ori overlapping ORF (Figure 3).

Readthrough RNAs

The readthrough RNAs (rtRNAs) are produced by occasional TRT of coding or noncoding genes due to the inefficient recognition of transcriptional termination sequences by the RNP molecules. The rtRNAs are TES variants if they have an exact termination site and contain the same ORF as the shorter transcript isoform. Complex transcripts are also the results of TRTs, whereas the polycistronic RNAs can only be considered as rtRNAs if the upstream genes also have their own transcription termination signals. Studies using quantitative RT-PCR [26] and Illumina sequencing [31, 67] have demonstrated a pervasive, genome-wide expression of asRNAs in herpesviruses. It is possible that these transcripts lack poly(A) tails and that is why they are undetected by oligo (dT)-primed sequencing techniques. These molecules may be expressed at a low abundance and/or have a short half-life due to the lack of polyadenylation.

Transcript Isoforms

Transcript isoforms include the length and splice variants of mRNAs and ncRNAs.

Transcription Start-Site Isoforms

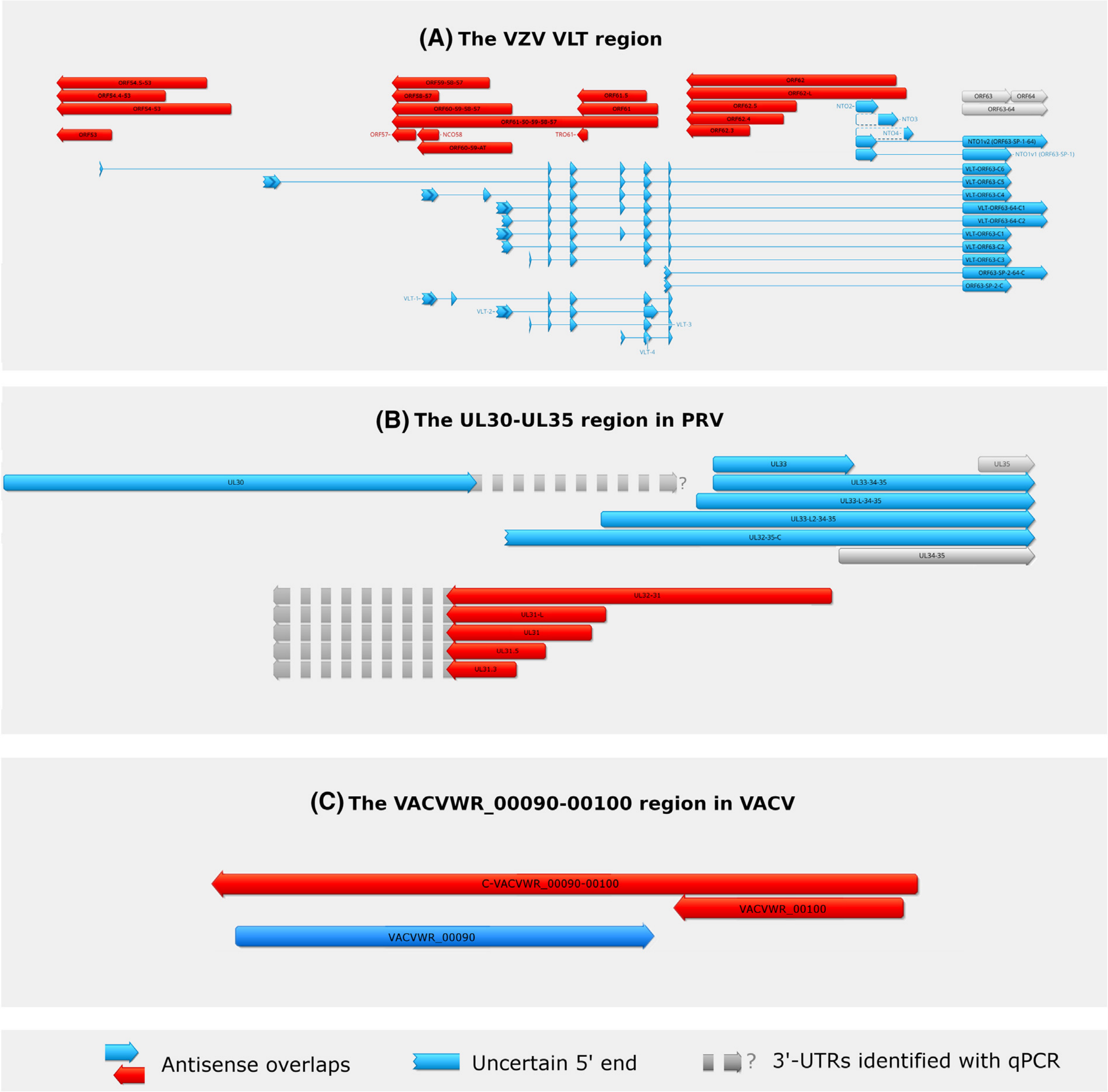
Genes can be controlled by alternative promoters which express TSS isoforms [68]. Multiple promoter-controlled genes can be differentially expressed throughout the viral life cycle [69]. Additionally, TSS isoforms can have diverse functions as the various 5′-UTR (untranslated region) structures can control the translation in a differential manner (as reviewed in [70]).

Transcription End-Site Isoforms

In certain cases, the progression of RNPs does not stop at the transcription termination sequences, which leads to longer TES variants [71,72]. On many occasions, these longer molecules overlap the transcripts generated by the downstream genes. SRS has been utilized to identify alternative polyadenylation [73]. LRS provides additional information about the various 5′- and 3′-UTR combinations utilized by the various transcript isoforms. It is important, however, to filter the putative polyadenylation sites for signs of internal priming, as it has been shown that adenine-rich regions may appear as false polyadenylation sites [74].

Splice Isoforms

The PacBio Iso-Seq® technique is especially suitable for the detection of novel splice sites [75–77]. Furthermore, alternatively processed multispliced transcripts can be reliably identified only by the LRS techniques; hence, they are able to sequence full-length transcripts and thus to map exon connectivity. Splicing events are relatively rare in alphaherpesviruses, baculoviruses,


(A) The VZV VLT region

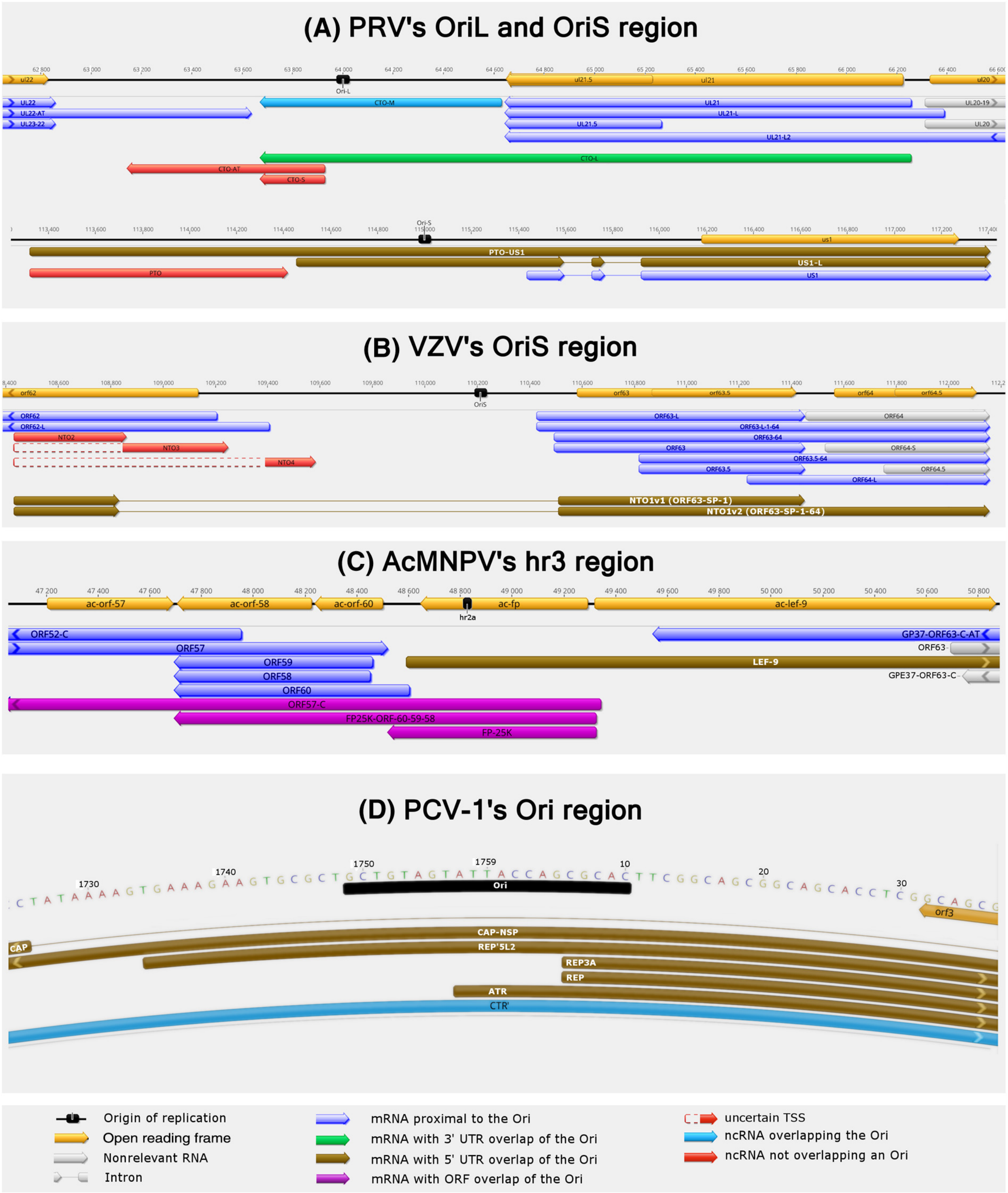
(B) The UL30-UL35 region in PRV

(C) The VACVWR_00090-00100 region in VACV

Trends in Microbiology

Figure 2. Antisense Transcripts in Two Herpesviruses. (A) The very long splice isoforms of ORF63, ORF63-64, and the VLT$_{lyt}$ RNAs (blue arrows) of varicella-zoster virus overlap several transcripts and transcript isoforms (red arrows) in antisense orientation in an almost 18-kbp long region of the viral genome. The NTO3 and NTO4 antisense transcripts are probably controlled by the same promoter as the NTO1v1 transcript. The unprocessed version of these noncoding transcripts overlaps with several mRNAs in a convergent or divergent manner, thereby generating antisense parts of the RNA molecules. The arrows illustrate exons, while the lines between the arrows are introns. (B) The various RNAs and transcript isoforms produced from the *ul30-35* genomic region of pseudorabies virus (PRV) overlap each other either in convergent (e.g., *ul30* and *ul31*) or in divergent (e.g., *ul32-31* and *ul33*) manners, thereby producing antisense sequences on the RNA molecules. The longer 3′ UTR versions of some transcripts detected by qPCR result in more extended antisense overlaps (gray broken arrows). (C) Convergent overlapping antisense transcripts of the vaccinia virus transcriptome from the VACVWR-00090-00100 genomic region.

and orthomyxoviruses [36,60,78], but they are common in beta- and gammaherpesviruses, retroviruses, and hepadnaviruses [34,61,67,79], whereas there is no splicing in poxviruses at all [80]. As a result of the application of LRS techniques, the number of new splice sites and splice



Trends in Microbiology

Figure 3. Replication-Associated Transcripts. A large variety of raRNAs have been evolved in various viruses. In herpesviruses, the ncoRNAs can be located near the replication origin, or they can overlap it (panels A and B). These transcripts can be protein-coding (e.g., US1) or noncoding (e.g., cto-s, pto, and NTO2-4), or, alternatively, the noncoding parts of long mRNA isoforms (e.g., PTO-US1, US1-L, and NTO1). The baculovirus hr1 region represents an additional transcript type, which overlaps the Ori with its protein-coding part (ORF; panel C). All transcripts of the porcine circovirus type 1 overlap the viral origin of replication (panel D).

isoforms has been radically increased. Nonetheless, thorough filtering is advised, as template-switching and ligation can both introduce many chimeric cDNA products that may resemble splicing. Common filters require the presence of consensus splice sequences (usually GT/AG, GC/AG, or AT/AC) and/or the absence of short homologous sequences that could facilitate template switching. Even if very strict criteria are used for the identification of splice variants, analysis of individual transcripts by, for example, Northern blot, is needed to confirm their real existence because the RT or the sequencing protocols can produce splice artifacts [47].

Multigenic Transcripts

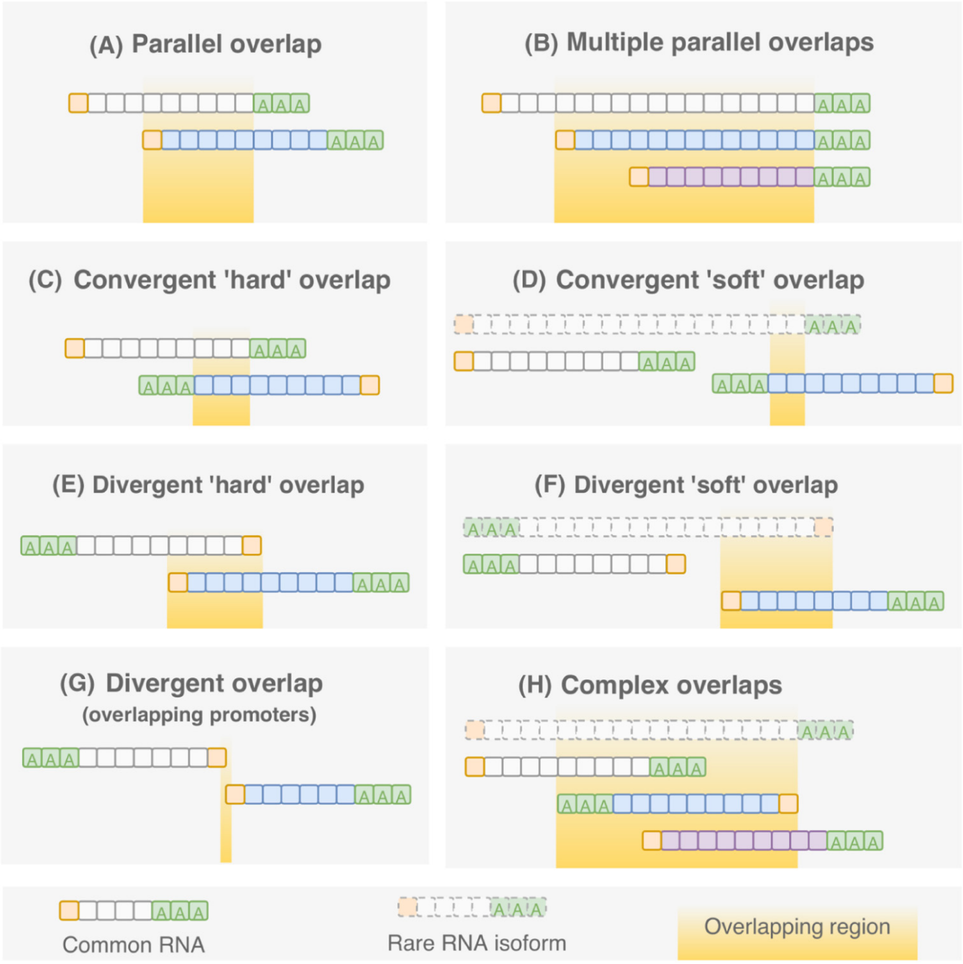
Multigenic transcripts include polycistronic and complex RNA molecules. Polycistronic transcripts contain two or more genes arranged in tandem array, whereas, in complex transcripts, at least two genes stand in an opposite, convergent or divergent, orientation. Polycistronism is common in viruses; however, the translation of internal genes from these RNA molecules is very rare in large DNA viruses. Two well-described exceptions are the translation of the ORF72-71 and the ORF35-36-37 transcripts of Kaposi sarcoma-associated herpesvirus (KSHV), where the expression of the downstream genes is facilitated by an IRES sequence or an upstream (u)ORF, respectively [81,82]. What could be the function of these RNA molecules, if not translation? LRS characterizes transcriptome diversity with a much higher sensitivity than is possible by SRS. Therefore, LRS studies of viruses with complex transcriptomes, such as DNA viruses, and some RNA viruses, will deepen our understanding of the general aspects of genetic regulation.

Transcriptional Overlaps

Transcripts can overlap with each other in a convergent (tail-to-tail), divergent (head-to-head), or a parallel (tail-to-head) manner (Figure 4). In many cases, TRTs produce overlapping transcripts (‘soft’/alternative overlaps), but TOs can also be produced without readthrough (‘hard’ TOs), for example, in *ul30-31* genes of alphaherpesviruses (Figure 2) [33–37]. This gene pair also produces longer TES variants with uncertain termination. The alternative promoter usage can also produce divergent soft TOs, but in, for example, PRV, most divergent genes overlap each other in a hard manner. This is also the case in the tandem gene clusters of herpesviruses in which the adjacent genes overlap in a tail-to-head manner. LRS studies revealed an intricate meshwork of TOs in every examined virus family [34,36,37,60,62]. The question can be raised whether this TO complexity is functional, or not, and if it is, what could this function be.

Transcriptional Interference Networks

It has been earlier proposed that the role of TOs is to carry out transcription-level interference (TI) between the herpesvirus genes in order to control gene expression [83]. It is supposed that TIs are organized into a system (transcription interference network, TIN) that may coordinate the viral life cycle in a spatiotemporal manner through the physical interaction of the transcriptional apparatuses. transcriptional interference network (TIN) might represent an additional level of gene regulation, which could have coevolved with the transcription factor-dependent regulation of gene expression [84]. TI is considered to be a system-level property because every viral gene produces transcripts that overlap with other transcripts in a variety of ways; therefore, the effect of a change in the expression of a gene can spread throughout the entire genome. It is hypothesized that these interactions form a self-regulatory network and result in a strictly timed alteration of the ON/OFF states of genes along the whole viral DNA. TIN is supposed to coregulate the closely spaced genes through synchronization and/or negative synchronization of the transcriptions, thereby resulting in a well-defined temporal pattern of genome-wide gene expressions. TIN may also be able to reduce the transcriptional noise, that is, it cooperates with the transcription factor-based system for suppressing the expression of genes whose products are not needed at a given stage of the viral life cycle.



Trends in Microbiology

Figure 4. Transcript Overlaps. Viral RNA molecules can form various types of overlap with respect to orientation and length of transcripts. The prototypic organization of the transcriptome of herpesviruses and baculoviruses is that tandem genes express overlapping transcripts with common 3′-termini. The transcripts with 'soft' overlaps have shorter non-overlapping variants. Complex overlaps span at least two complete genes.

Transcriptional and Replication Interference Networks

The expression of raRNAs is supposed to facilitate the regulation of replication initiation and the orientation of replication fork progression through either a collision between the replication and transcription machineries, or by unwinding the DNA strands by the RNP near the Ori. Since the identified nroRNAs are all polyadenylated, they are likely to have a function as RNA molecules beside being the byproducts of a regulatory mechanism. Moreover, an overall decrease in the transcriptional activity in individual herpesvirus genes has been observed following the onset of DNA replication [85], which also suggests an interplay between the two machineries. The interactions between the RNA and DNA synthesis apparatus have been supposed to form a transcription and replication interference network (TRIN) that controls global gene expression and replication in a cooperative manner [64].

Transcriptional overlaps and the overlaps of raRNAs with the replication origins may represent a novel level of genetic regulation. Nevertheless, further investigations are needed to elucidate the role of these RNA molecules.

Upstream ORFs

Ribosome footprint analysis has detected hundreds of translationally active short uORFs in the human cytomegalovirus (HCMV) and KSHV genomes [54,55]. It has also been shown that transcript isoforms differ from each other in the presence or absence of uORFs [35,37]. Some of these short peptides may be functional; it is more likely, however, that their main role is the regulation of translation reinitiation at downstream ORFs. The uORFs permit translation reinitiation of a downstream gene because they are very short, and thus the initiation factors have not yet dissociated. When translation is initiated at an uORF upstream of the HCMV UL4 gene, the mRNA structure stalls the ribosome and it completely inhibits the downstream translation from that RNA molecule [86]. The translation of KSHV gene ORF36 has been shown to be regulated by uORFs upstream of the ORF35 coding sequence through a termination–reinitiation mechanism [81]. This herpesvirus has reversed the host strategy to repress translation by uORFs: it allows the expression of a downstream gene. It has also been shown that RNA isoforms can increase the coding capacity of HCMV genes due to the alternative presence or absence of uORFs [35,37].

Detection of RNA Editing and Modification

G mismatches of the mapped reads can be caused either by natural variation of the viral population or by A-to-I RNA hyperediting. The two cases can be easily distinguished by using LRS techniques when the edited RNA overlaps other transcripts that are unmodified. LRS has been used to discover a hyperediting event on a varicella-zoster virus ncRNA, the NTO3 [65]. Several types of RNA modification have been described in viral transcriptomes with functions including mRNA maturation [87] and evasion of the host's immune response [88,89]. The ability to sequence full-length native RNA molecules simplifies the problem of assorting the detected modified nucleotides among the overlapping transcripts.

Concluding Remarks

In the past few years, LRS techniques have become essential in genome and transcriptome research, and their application is expected to be dramatically increased in the near future. The current LRS techniques produce a lower sequencing coverage than the SRS approaches; therefore, small-genome organisms are ideal subjects for these third-generation sequencing techniques. LRS platforms are able to determine full-length transcripts; thus, they are superior for identifying long, multigenic and multispliced transcripts as well as TSS and TES isoforms. TOs are also easier to study with these techniques. All in all, LRS approaches have multiplied the number of transcripts in every examined viral species. Considering that many viral transcriptomes are poorly annotated, LRS may greatly increase our understanding of the gene expression of most viruses with complex transcriptomes. Direct RNA sequencing is also useful for the examination of viruses with simple transcriptomes [e.g., (+)ssRNA viruses] as it can detect RNA modifications which would not be detectable by other methods. The extremely complex network of TOs suggests a sophisticated interplay between the adjacent and distal genes through physical interaction between the transcriptional apparatuses. Additionally, the discovery of nroRNAs and nro-like transcripts raises the possibility of the interaction between the replication and transcription machineries in order to regulate gene expression and DNA synthesis in a cooperative manner (see Outstanding Questions).

LRS technology is versatile and is rapidly evolving. Its capability of distinguishing between RNA isoforms by reading full-length RNAs and detecting RNA modifications makes it a competent tool for viral transcriptomics.

Outstanding Questions

How could LRS be better exploited for the identification of novel transcripts and transcript isoforms, and how could the potential artefacts of these techniques be eliminated?

While ONT sequencing is superior to the PacBio sequencing in terms of cost-effectiveness, in throughput, and in read length, it possesses a major disadvantage regarding a high error rate. Will ONT be able to solve this problem in the future? Alternatively, will PacBio be able to keep pace with ONT in continuing to improve the above parameters?

10x Genomics has developed an add-on for the short-read sequencing-based Illumina system, which provides long-range genome information. Will this technology outcompete the PacBio and ONT from the market?

The function of most viral noncoding transcripts remains unknown. How can these functions be characterized and harnessed for controlling the virulence of the various viruses?

Can upstream ORFs be utilized for the control of viral gene expression?

What could be the function of RNA editing in some noncoding transcripts?

Do the extensive transcriptional overlaps represent a mere economization of the small viral genomes, or do they represent a novel regulatory layer based on the interference between the transcriptional machineries of adjacent and distal genes?

What could be the function of the 'near replication origin' transcripts? Do they regulate the replication on a collision-based manner? Does the replication regulate the transcription?

Acknowledgments

Z.Bo. was supported by the National Research, Development and Innovation Office (NKFIH) K 128247 grant, M.S. by the National Institutes of Health, Centers of Excellence in Genomic Science Center for Personal Dynamic Regulomes (No. 5P50HG00773502), and D.T. by the National Research, Development and Innovation Office (NKFIH) FK 128252 grant.

Resources

[i]www.pacb.com/wp-content/uploads/Clark-PAG-2017-Full-Length-cDNA-Sequencing-on-the-PacBio-Sequel_Platform.pdf
[ii]https://github.com/nanoporetech/pinfish
[iii]https://github.com/zsolt-balazs/LoRTIA

References

1. Shine J. and Dalgarno L. (1975) Determinant of cistron specificity in bacterial ribosomes. *Nature* 254, 34–38.
2. Firth A.E. and Brierley I. (2012) Non-canonical translation in RNA viruses. *J. Gen. Virol.* 93, 1385–1409.
3. Yuan C. et al. (2017) It is imperative to establish a pellucid definition of chimeric RNA and to clear up a lot of confusion in the relevant research. *Int. J. Mol. Sci.* 18, E714.
4. Weimer B.C. (2017) 100K Pathogen genome project. *Genome Announc.* 5, e00594-17.
5. Turnbull C. et al. (2018) The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ* k1687, 361.
6. Wang E.T. et al. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476.
7. Goodwin S. et al. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351.
8. McCarthy A. (2010) Third generation DNA sequencing: Pacific Biosciences' single molecule real time technology. *Chem. Biol.* 17, 675–676.
9. Ip C.L.C. et al. (2015) MinION Analysis and Reference Consortium: phase 1 data release and analysis. *F1000Research* 4, 1075.
10. Garalde D.R. et al. (2018) Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* 15, 201–206.
11. Ozsolak F. et al. (2009) Direct RNA sequencing. *Nature* 461, 814–818.
12. Rhoads A. and Au K.F. (2015) PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* 13, 278–289.
13. Quick J. et al. (2014) A reference bacterial genome dataset generated on the MinION^{TM} portable single-molecule nanopore sequencer. *Gigascience* 3, 22 (2047-217X-3-22).
14. Steijger T. et al. (2013) Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* 10, 1177–1184.
15. Chin C.-S. et al. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10, 563–569.
16. Pendleton M. et al. (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* 12, 780–786.
17. Sz⊠cs A. et al. (2017) Long-read sequencing reveals a GC pressure during the evolution of porcine endogenous retrovirus. *Genome Announc.* 5, e01040-17.
18. Tombácz D. et al. (2014) Strain Kaplan of pseudorabies virus genome sequenced by PacBio single-molecule real-time sequencing technology. *Genome Announc.* 2, e00628-14.
19. Nakano K. et al. (2017) Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Hum. Cell* 30, 149–161.
20. Dilernia D.A. et al. (2015) Multiplexed highly-accurate DNA sequencing of closely-related HIV-1 variants using continuous long reads from single molecule, real-time sequencing. *Nucleic Acids Res.* 43, e129.
21. Bull R.A. et al. (2016) A method for near full-length amplification and sequencing for six hepatitis C virus genotypes. *BMC Genomics* 17, 247.
22. Prazsák I. et al. (2018) Full genome sequence of the Western Reserve strain of vaccinia virus determined by third-generation sequencing. *Genome Announc.* 6, e01570-17.
23. Mankertz J. et al. (1998) Transcription analysis of porcine circovirus (PCV). *Virus Genes* 16, 267–276.
24. Farrell M.J. et al. (1991) Herpes simplex virus latency-associated transcript is a stable intron. *Proc. Natl. Acad. Sci. U. S. A.* 88, 790–794.
25. Nagel M.A. et al. (2011) Varicella-zoster virus transcriptome in latently infected human ganglia. *J. Virol.* 85, 2276–2287.
26. Tombácz D. et al. (2009) Whole-genome analysis of pseudorabies virus gene expression by real-time quantitative RT-PCR assay. *BMC Genomics* 10, 491.
27. Sadler R.H. and Raab-Traub N. (1995) Structural analyses of the Epstein–Barr virus BamHI A transcripts. *J. Virol.* 69, 1132–1141.
28. Aguilar J.S. et al. (2006) Quantitative comparison of the HSV-1 and HSV-2 transcriptomes using DNA microarray analysis. *Virology* 348, 233–241.
29. Lacaze P. et al. (2011) Temporal profiling of the coding and noncoding murine cytomegalovirus transcriptomes. *J. Virol.* 85, 6065–6076.
30. Oláh P. et al. (2015) Characterization of pseudorabies virus transcriptome by Illumina sequencing. *BMC Microbiol.* 15, 130.
31. O'Grady T. et al. (2014) Global bidirectional transcription of the Epstein–Barr virus genome during reactivation. *J. Virol.* 88, 1604–1616.
32. Chen Y.-R. et al. (2013) The transcriptome of the baculovirus *Autographa californica* multiple nucleopolyhedrovirus in *Trichoplusia ni* cells. *J. Virol.* 87, 6391–6405.
33. Tombácz D. et al. (2016) Full-length isoform sequencing reveals novel transcripts and substantial transcriptional overlaps in a herpesvirus. *PLoS One* 11, e0162868.
34. O'Grady T. et al. (2016) Global transcript structure resolution of high gene density genomes through multi-platform data integration. *Nucleic Acids Res.* 44, e145.
35. Moldován N. et al. (2017) Multi-platform sequencing approach reveals a novel transcriptome profile in pseudorabies virus. *Front. Microbiol.* 8, 2708.
36. Tombácz D. et al. (2017) Long-read isoform sequencing reveals a hidden complexity of the transcriptional landscape of herpes simplex virus type 1. *Front. Microbiol.* 8, 1079.
37. Balázs Z. et al. (2017) Long-read sequencing of human cytomegalovirus transcriptome reveals RNA isoforms carrying distinct coding potentials. *Sci. Rep.* 7, 15989.
38. Levene M.J. et al. (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* 299, 682–686.
39. Eid J. et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138.
40. Lin H.-H. and Liao Y.-C. (2015) Evaluation and validation of assembling corrected PacBio long reads for microbial genome completion via hybrid approaches. *PLoS One* 10, e0144305.
41. Schirmer M. et al. (2016) Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinf.* 17, 125.
42. Weirather J.L. et al. (2017) Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* 6, 100.
43. Hebert P.D.N. et al. (2018) A sequel to Sanger: amplicon sequencing that scales. *BMC Genomics* 19, 219.
44. Workman R.E. et al. (2018) Nanopore native RNA sequencing of a human poly(A) transcriptome. *bioRxiv* Published online November 9, 2018. https://doi.org/10.1101/459529.

45. Jain M. et al. (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36, 338–345.
46. Lu H. et al. (2016) Oxford Nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics* 14, 265–279.
47. Cocquet J. et al. (2006) Reverse transcriptase template switching and false alternative transcripts. *Genomics* 88, 127–131.
48. Kebschull J.M. and Zador A.M. (2015) Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res.* 43, e143.
49. Zhu Y.Y. et al. (2001) Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* 30, 892–897.
50. Edery I. et al. (1995) An efficient strategy to isolate full-length cDNAs based on an mRNA cap retention procedure (CAPture). *Mol. Cell. Biol.* 15, 3363–3371.
51. Tilgner H. et al. (2015) Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* 33, 736–742.
52. Depledge D.P. et al. (2019) Going the distance: optimizing RNA-Seq strategies for transcriptomic analysis of complex viral genomes. *J. Virol.* 93, e01342-18.
53. Park P.J. (2009) ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10, 669–680.
54. Stern-Ginossar N. et al. (2012) Decoding human cytomegalovirus. *Science* 338, 1088–1093.
55. Arias C. et al. (2014) KSHV 2.0: a comprehensive annotation of the Kaposi's sarcoma-associated herpesvirus genome using next-generation sequencing reveals novel genomic and functional features. *PLoS Pathog.* 10, e1003847.
56. Mahat D.B. et al. (2016) Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* 11, 1455–1476.
57. Zheng D. et al. (2016) 3′READS+, a sensitive and accurate method for 3′ end sequencing of polyadenylated RNA. *RNA* 22, 1631–1639.
58. Tombácz D. et al. (2017) Characterization of the dynamic transcriptome of a herpesvirus with long-read single molecule real-time sequencing. *Sci. Rep.* 7, 43751.
59. Gupta I. et al. (2018) Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.* 36, 1197–1202.
60. Moldován N. et al. (2018) Third-generation sequencing reveals extensive polycistronism and transcriptional overlapping in a baculovirus. *Sci. Rep.* 8, 8604.
61. Moldován N. et al. (2018) Multiplatform next-generation sequencing identifies novel RNA molecules and transcript isoforms of the endogenous retrovirus isolated from cultured cells. *FEMS Microbiol. Lett.* 365 (5), fny013.
62. Moldován N. et al. (2017) Multi-platform analysis reveals a complex transcriptome architecture of a circovirus. *Virus Res.* 237, 37–46.
63. Tombácz D. et al. (2018) Dynamic transcriptome profiling dataset of vaccinia virus obtained from long-read sequencing techniques. *Gigascience* 7, giy139.
64. Tombácz D. et al. (2015) Characterization of novel transcripts in pseudorabies virus. *Viruses* 7, 2727–2744.
65. Prazsák I. et al. (2018) Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus. *BMC Genomics* 19, 873.
66. Palazzo A.F. and Lee E.S. (2015) Non-coding RNA: what is functional and what is junk? *Front. Genet.* 6, 2.
67. Gatherer D. et al. (2011) High-resolution human cytomegalovirus transcriptome. *Proc. Natl. Acad. Sci. U. S. A.* 108, 19755–19760.
68. Pal S. et al. (2011) Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Res.* 21, 1260–1272.
69. Isomura H. et al. (2008) Noncanonical TATA sequence in the UL44 late promoter of human cytomegalovirus is required for the accumulation of late viral transcripts. *J. Virol.* 82, 1638–1646.
70. Leppek K. et al. (2018) Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat. Rev. Mol. Cell Biol.* 19, 158–174.
71. Mainguy G. et al. (2007) Extensive polycistronism and antisense transcription in the mammalian Hox clusters. *PLoS One* 2, e356.
72. Gullerova M. and Proudfoot N.J. (2008) Cohesin complex promotes transcriptional termination between convergent genes in *S. pombe*. *Cell* 132, 983–995.
73. Majerciak V. et al. (2013) A viral genome landscape of RNA polyadenylation from KSHV latent to lytic infection. *PLoS Pathog.* 9, e1003749.
74. Nam D.K. et al. (2002) Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proc. Natl. Acad. Sci. U. S. A.* 99, 6152–6156.
75. Abdel-Ghany S.E. et al. (2016) A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.* 7, 11706.
76. Gonzalez-Garay M.L. (2016) *Introduction to Isoform Sequencing Using Pacific Biosciences Technology (Iso-Seq)*. Springer, pp. 141–160.
77. Wang B. et al. (2016) Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* 7, 11708.
78. Shih S.R. et al. (1995) The choice of alternative 5′ splice sites in influenza virus M1 mRNA is regulated by the viral polymerase complex. *Proc. Natl. Acad. Sci. U. S. A.* 92, 6324–6328.
79. Sommer G. and Heise T. (2008) Posttranscriptional control of HBV gene expression. *Front. Biosci.* (13), 5533–5547.
80. Moss B. (1991) Vaccinia virus: a tool for research and vaccine development. *Science* 252, 1662–1667.
81. Kronstad L.M. et al. (2013) Dual short upstream open reading frames control translation of a herpesviral polycistronic mRNA. *PLoS Pathog.* 9, e1003156.
82. Low W. et al. (2001) Internal ribosome entry site regulates translation of Kaposi's sarcoma-associated herpesvirus FLICE inhibitory protein. *J. Virol.* 75, 2938–2945.
83. Boldogkői Z. (2012) Transcriptional interference networks coordinate the expression of functionally related genes clustered in the same genomic loci. *Front. Genet.* 3, 122.
84. Nasser W. et al. (2002) Transcriptional regulation of fis operon involves a module of multiple coupled promoters. *EMBO J.* 21, 715–724.
85. Takács I.F. et al. (2013) The ICP22 protein selectively modifies the transcription of different kinetic classes of pseudorabies virus genes. *BMC Mol. Biol.* 14, 2.
86. Geballe A.P. and Mocarski E.S. (1988) Translational control of cytomegalovirus gene expression is mediated by upstream AUG codons. *J. Virol.* 62, 3334–3340.
87. Sommer S. et al. (1976) The methylation of adenovirus-specific nuclear and cytoplasmic RNA. *Nucleic Acids Res.* 3, 749–765.
88. Anderson B.R. et al. (2011) Nucleoside modifications in RNA limit activation of 2′-5′-oligoadenylate synthetase and increase resistance to cleavage by RNase L. *Nucleic Acids Res.* 39, 9329–9338.
89. Karikó K. et al. (2005) Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA. *Immunity* 23, 165–175.
90. Workman R.E. et al. (2018) Single-molecule, full-length transcript sequencing provides insight into the extreme metabolism of the ruby-throated hummingbird *Archilochus colubris*. *Gigascience* 7, giy009.
91. Madoui M.-A. et al. (2015) Genome assembly using nanopore-guided long and error-free DNA reads. *BMC Genomics* 16, 327.
92. Salmela L. and Rivals E. (2014) LoRDEC: accurate and efficient long read error correction. *Bioinformatics* 30, 3506–3514.
93. Koren S. et al. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736.
94. Lima L.I.S. de et al. (2018) Comparative assessment of long-read error-correction software applied to RNA-sequencing data. *bioRxiv* Published online November 23, 2018. https://doi.org/10.1101/476622.

95. Volden R. et al. (2018) Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc. Natl. Acad. Sci. U. S. A.* 115, 9726–9731.
96. Li H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100.
97. Byrne A. et al. (2017) Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* 8, 16027.
98. Tang A.D. et al. (2018) Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *bioRxiv* Published online September 6, 2018. https://doi.org/10.1101/410183.
99. Tardaguila M. et al. (2017) SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* 28, 396–441 Published online August 21, 2017. https://doi.org/10.1101/gr.222976.117.

III.

ORIGINAL RESEARCH
published: 24 September 2019
doi: 10.3389/fgene.2019.00834

Multiple Long-Read Sequencing Survey of Herpes Simplex Virus Dynamic Transcriptome

*Dóra Tombácz[1], Norbert Moldován[1], Zsolt Balázs[1], Gábor Gulyás[1], Zsolt Csabai[1], Miklós Boldogkői[1], Michael Snyder[2] and Zsolt Boldogkői[1]**

[1] Department of Medical Biology, Faculty of Medicine, University of Szeged, Szeged, Hungary, [2] Department of Genetics, School of Medicine, Stanford University, Stanford, CA, United States

Long-read sequencing (LRS) has become increasingly important in RNA research due to its strength in resolving complex transcriptomic architectures. In this regard, currently two LRS platforms have demonstrated adequate performance: the Single Molecule Real-Time Sequencing by Pacific Biosciences (PacBio) and the nanopore sequencing by Oxford Nanopore Technologies (ONT). Even though these techniques produce lower coverage and are more error prone than short-read sequencing, they continue to be more successful in identifying polycistronic RNAs, transcript isoforms including splice and transcript end variants, as well as transcript overlaps. Recent reports have successfully applied LRS for the investigation of the transcriptome of viruses belonging to various families. These studies have substantially increased the number of previously known viral RNA molecules. In this work, we used the Sequel and MinION technique from PacBio and ONT, respectively, to characterize the lytic transcriptome of the herpes simplex virus type 1 (HSV-1). In most samples, we analyzed the poly(A) fraction of the transcriptome, but we also performed random oligonucleotide-based sequencing. Besides cDNA sequencing, we also carried out native RNA sequencing. Our investigations identified more than 2,300 previously undetected transcripts, including coding, and non-coding RNAs, multi-splice transcripts, as well as polycistronic and complex transcripts. Furthermore, we found previously unsubstantiated transcriptional start sites, polyadenylation sites, and splice sites. A large number of novel transcriptional overlaps were also detected. Random-primed sequencing revealed that each convergent gene pair produces non-polyadenylated read-through RNAs overlapping the partner genes. Furthermore, we identified novel replication-associated transcripts overlapping the HSV-1 replication origins, and novel LAT variants with very long 5' regions, which are co-terminal with the LAT-0.7kb transcript. Overall, our results demonstrated that the HSV-1 transcripts form an extremely complex pattern of overlaps, and that entire viral genome is transcriptionally active. In most viral genes, if not in all, both DNA strands are expressed.

Keywords: herpesviruses, herpes simplex virus, long-read sequencing, direct RNA sequencing, Pacific Biosciences, Oxford Nanopore Technologies, transcript isoforms

OPEN ACCESS

Edited by:
Ishaan Gupta,
Indian Institute of Science Education and Research, India

Reviewed by:
Milind B. Ratnaparkhe,
ICAR Indian Institute of Soybean Research, India
Daniel Pearce Depledge,
New York University, United States

Correspondence:
Zsolt Boldogkői
boldogkoi.zsolt@med.u-szeged.hu

Specialty section:
This article was submitted to Genomic Assay Technology, a section of the journal Frontiers in Genetics

Received: *31 January 2019*
Accepted: *13 August 2019*
Published: *24 September 2019*

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

INTRODUCTION

Next-generation short-read sequencing (SRS) technology has revolutionized the research fields of genomics and transcriptomics due to its capacity of sequencing a large number of nucleic acid fragments simultaneously at a relatively low cost (Mortazavi et al., 2008; Wang et al., 2009; Djebali et al., 2012). However, SRS technologies have inherent limitations both in genome and transcriptome analyses. This approach does not perform adequately in mapping repetitive elements and GC-rich DNA sequences, or in discriminating paralogous sequences. In transcriptome research, SRS techniques have difficulties in identifying multi-spliced transcripts, overlapping transcripts, transcription start site (TSS), and transcription end site (TES) isoforms, as well as multigenic RNA molecules.

Long-read sequencing (LRS) techniques can resolve these obstacles. The LRS technology is able to read full-length RNA molecules, therefore it is ideal for application in the analysis of complex transcriptomic profiles. Currently two techniques are available in the market, the California-based Pacific Biosciences (PacBio) and the British Oxford Nanopore Technologies (ONT) platforms. The PacBio approach is based on single-molecule real-time (SMRT) technology, while the ONT platform utilizes the nanopore sequencing concept. Both techniques have already been applied for the structural and dynamic transcriptomic analysis of various organisms (Byrne et al., 2017; Chen et al., 2017; Cheng et al., 2017; Li et al., 2018; Nudelman et al., 2018; Wen et al., 2018; Zhang et al., 2018; Jiang et al., 2019; Zhao et al., 2019), including viruses (Boldogkői et al., 2019b), such as herpesviruses (Tombácz et al., 2015; O'Grady et al., 2016; Tombácz et al., 2016; Balázs et al., 2017a; Balázs et al., 2017b; Moldován et al., 2017b; Tombácz et al., 2017b; Tombácz et al., 2017a; Tombácz et al., 2018b; Depledge et al., 2019), poxviruses (Tombácz et al., 2018a), baculoviruses (Moldován et al., 2018b), retroviruses (Moldován et al., 2018a), coronaviruses (Viehweger et al., 2019), and circoviruses (Moldován et al., 2017a). Additionally, the ONT technology is capable of sequencing DNA and RNA in its native form, allowing epigenetic and epitranscriptomic analysis (Wongsurawat et al., 2018; Liu et al., 2019; Shah et al., 2019).

Herpes simplex virus type 1 (HSV-1) is a human pathogenic virus belonging to the *Alphaherpesvirinae* subfamily of the *Herpesviridae* family. Its closest relatives are the HSV-2, the Varicella-zoster virus (VZV), and the animal pathogen pseudorabies virus (PRV). The most common symptom of HSV-1 infection is cold sores, which can recur from latency causing blisters primarily on the lips. HSV-1 may cause acute encephalitis in immunocompromised patients. The ability of herpesviruses to establish lifelong latency within the host organism significantly contributes to their evolutionary success: according to WHO's estimates, more than 3.7 billion people under the age of 50 are infected with HSV-1 worldwide (Looker et al., 2015).

HSV-1 has a 152-kbp linear double-stranded DNA genome that is composed of unique and repeat regions. Both the long (UL) and the short (US) unique regions are flanked by inverted repeats (IRLs and IRSs, respectively) (Macdonald et al., 2012). The viral genome is transcribed by the host RNA polymerase in a cascade-like manner producing three kinetic classes of transcripts and proteins: immediate-early (IE), early (E), and late (L) (Harkness et al., 2014). IE genes encode transcription factors required for the expression of E and L genes. E genes mainly code for proteins playing a role in DNA synthesis, whereas L genes specify structural elements of the virus. Earlier studies and *in silico* annotations have revealed 89 mRNAs, 10 non-coding (nc)RNAs (Rajčáni et al., 2004; McGeoch et al., 2006; Macdonald et al., 2012; Lim, 2013; Hu et al., 2016), and 18 microRNAs (Du et al., 2015). Our recent study (Tombácz et al., 2017b) based on PacBio RS II sequencing has identified additional 142 transcripts and transcript isoforms, including ncRNAs. The detection and the kinetic characterization of HSV-1 transcriptome face an important challenge because of the overlapping and polycistronic nature of the viral transcripts. Polycistronic transcription units are different from those of bacterial operons, in that the downstream genes on multigenic transcripts are untranslated because herpesvirus mRNAs use cap-dependent translation initiation (Merrick, 2004). The majority of herpesvirus transcripts are organized into tandem gene clusters generating overlapping transcripts with co-terminal TESs. The *ul41-44* genomic region of HSV-1 does not follow this rule, since these genes are primarily expressed as monocistronic RNA molecules. Our earlier study has revealed that these genes also produce low-abundance bi- and polycistronic transcripts. Alternatively, many HSV-1 genes, which were believed to be exclusively expressed as parts of multigenic RNAs, have also been shown to specify low-abundance monocistronic transcripts (Tombácz et al., 2017b).

SRS technologies have become useful tools for the analysis of transcriptomes. However, conventionally applied SRS platforms cannot reliably distinguish between multi-spliced transcript isoforms, and TSS variants, as well as between embedded transcripts and their host RNAs, etc. Additionally, SRS, even if applied in conjunction with auxiliary techniques such as RACE analysis, has limitations in detecting multigenic transcripts, including polycistronic RNAs and complex transcripts (cxRNAs; containing genes standing in opposite orientations). LRS is able to circumvent these problems. Both PacBio and ONT approaches are capable of reading cDNAs generated from full-length transcripts in a single sequencing run and permit mapping of TSSs and TESs with base-pair precision. The most important disadvantage of LRS compared to SRS techniques is lower coverage. In PacBio sequencing, if any errors occur in raw reads, they are easily corrected thanks to the very high consensus accuracy of this technique (Miyamoto et al., 2014). Thus, it is only a widespread myth that SMRT sequencing is too error prone to be used for precise sequence analysis. The precision of basecalling is substantially lower for ONT platform than that of PacBio, but the former technique is far more cost-effective, and yields both higher throughput and longer reads. The high error rate of the ONT technique can be circumvented by obtaining high sequence coverage. Nonetheless, this latter problem is not critical in transcriptome research if the genome sequence of the examined organism has already been annotated.

A diverse collection of methods and approaches have already been employed for the investigation of herpesvirus transcriptomes, including *in silico* detection of open reading frames (ORFs) and cis-regulatory motifs, Northern-blot analysis (Costa et al., 1984; Sedlackova et al., 2008), S1 nuclease mapping (McKnight, 1980; Rixon and Clements, 1982), primer extension (Perng et al., 2002; Naito et al., 2005), real-time reverse transcription-PCR (RT^2-PCR) analysis (Tombácz et al., 2009), microarrays (Stingley et al., 2000), Illumina sequencing (Harkness et al., 2014; Oláh et al., 2015), PacBio RS II (O'Grady et al., 2016; Tombácz et al., 2017b), and Sequel sequencing, as well as ONT MinION cDNA and direct RNA sequencing (Boldogkői et al., 2018; Prazsák et al., 2018; Depledge et al., 2019).

In this study, we report the application of PacBio Sequel and ONT MinION long-read sequencing technologies for the characterization of the HSV-1 lytic transcriptome. We used an amplified isoform sequencing (Iso-Seq) protocol of PacBio that was based on PCR amplification of cDNAs prior to sequencing. We used both cDNA and direct (d)RNA sequencing on the ONT platform. Additionally, we applied Cap-selection for ONT sequencing. In order to identify non-polyadenylated transcripts, we also applied random oligonucleotide primer-based RT in addition to the oligo(dT)-priming. Furthermore, the latter technique is more efficient for the mapping of the TSSs, and it is useful for the validation of novel RNA molecules. Our intentions of using novel LRS techniques were to analyze the dynamic viral transcriptome, to generate a higher number of sequencing reads, and to identify novel transcripts that had been undetected in our earlier PacBio RS II-based approach. Furthermore, in this report, we also reanalyzed our earlier results that were obtained using a single-platform method (Tombácz et al., 2017b).

MATERIALS AND METHODS

Cells and Viral Infection

The strain KOS of HSV-1 was propagated on an immortalized kidney epithelial cell line (Vero) isolated from the African green monkey (*Chlorocebus sabaeus*). Vero cells were cultivated in Dulbecco's modified Eagle medium supplemented with 10% fetal bovine serum (Gibco Invitrogen) and 100 μl penicillin–streptomycin 10K/10K mixture (Lonza)/ml and 5% CO_2 at 37°C. The viral stocks were prepared by infecting rapidly-growing semi-confluent Vero cells at a multiplicity of infection (MOI) of 1 plaque-forming unit (pfu)/cell, followed by incubation until a complete cytopathic effect was observed. The infected cells were then frozen and thawed three times. The cells were then centrifuged at 10,000 ×g for 15 min using low-speed centrifugation. For the sequencing studies, cells were infected with MOI = 1, incubated for 1 h. This was followed by removal of the virus suspension and a PBS washing step. Next, the cells were supplied with a fresh culture medium and were then incubated for 1, 2, 4, 6, 8, 10, 12, or 24 h.

RNA Isolation

The total RNA samples were purified from cells using the NucleoSpin® RNA kit (**Table 1**) according to the kit's manual and our previously described methods (Boldogkői et al., 2018). The RNA samples were quantified using the Qubit® 2.0 Fluorometer and were stored at -80°C until use. The samples taken from each experiment were then mixed for sequencing. Samples were subjected to ribodepletion for the random primed sequencing, while selection of the poly$(A)^+$ RNA fraction was being carried out for polyA-sequencing. All experiments were performed in accordance with the relevant guidelines and regulations.

Pacific Biosciences RS II and Sequel Platforms—Sequencing of the Polyadenylated RNA Fraction or the Total Transcriptome

The Clontech SMARTer PCR cDNA Synthesis Kit was used for cDNA preparation according to the PacBio Isoform Sequencing (Iso-Seq) protocol. For the analysis of relatively short viral RNAs, the 'No-size selection' method was used and samples were run on the RSII and Sequel platforms, both. The SageELF™ and BluePippin™ Size-Selection Systems (Sage Science) were also used to carry out size-selection for capturing the potential long, rare transcripts. The reverse transcription (RT) reactions were primed by using the oligo(dT) from the SMARTer Kit. However, we also used random primers for a non-size selected sample to detect non-polyadenylated RNAs. The cDNAs were amplified by

TABLE 1 | Summary of the kits used for RNA preparation and quantitation.

Method		Kit	Company
RNA purification	Total RNA extraction	NucleoSpin RNA	Macherey Nagel
	PolyA(+) RNA isolation	Oligotex mRNA Mini Kit	Qiagen
	Ribodepletion	Ribo-ZeroTM Magnetic Kit H/M/R	Epicentre/Illumina
Concentration measurement	Total RNA	Qubit RNA BR Assay Kit	Life Technologies
	PolyA(+) RNA	Qubit RNA HS Assay Kit	
	rRNA depleted RNA		
Elimination of non-capped RNAs	5'-phasopahte-dependent-exonuclease digestion	Terminator™ 5′-Phosphate-Dependent Exonuclease	Epicentre/Lucigen

the KAPA HiFi Enzyme from KAPA Biosystems, according to PacBio's recommendations (Balázs et al., 2017b; Tombácz et al., 2018b). The SMRTbell libraries were generated using PacBio Template Prep Kit 1.0. For binding the DNA polymerase and annealing the sequencing primers, the DNA/Polymerase Binding Kit P6-C4 and v2 primers, as well as the Sequel Sequencing Kit and v3 primers were used for the RSII and Sequel sequencing, respectively. The DNA/Polymerase Binding Kit P6-C4 and v2 primers were used for binding DNA polymerase and for annealing sequencing primers. Whereas, the Sequel Sequencing kit and v3 primers were used for RSII and Sequel sequencing.

The polymerase-template complexes were bound to MagBeads with the PacBio MagBead Binding Kit. Samples were loaded onto the RSII SMRT Cell 8Pac v3 or Sequel SMRT Cell 1M. The movie time was 240 or 360 min *per* SMRT Cell for the RSII, while 600-min movie time was set to the Sequel run.

Oxford Nanopore Minion Platform—cDNA Sequencing Using Oligo(dT) or Random Primers

Regular (No Cap Selection) Protocol

The 1D Strand switching cDNA by ligation protocol (Version: SSE_9011_v108_revS_18Oct2016) from the ONT was used for sequencing HSV-1 cDNAs on the MinION platform. The ONT Ligation Sequencing Kit 1D (SQK-LSK108) was applied for the library preparation using the recommended oligo(dT) primers, or custom-made random oligonucleotides, as well as the SuperScipt IV enzyme for the RTs. The cDNA samples were subjected to PCR reactions with KAPA HiFi DNA Polymerase (Kapa Biosystems) and Ligation Sequencing Kit Primer Mix (part of the 1D Kit). The NEBNext End repair/dA-tailing Module (New England Biolabs) was used for the end repair, whereas the NEB Blunt/TA Ligase Master Mix (New England Biolabs) was utilized for the adapter ligation. The enzymatic steps (e.g.: RT, PCR, and ligation) were carried out in a Veriti Cycler (Applied Biosystems) according to the 1D protocol (Moldován et al., 2018b; Tombácz et al., 2018b). The Agencourt AMPureXP system (Beckman Coulter) was used for the purification of samples after each enzymatic reaction. The quantity of the libraries was checked using the Qubit Fluorometer 2.0 and the Qubit (ds)DNAHS Assay Kit (Life Technologies). The samples were run on R9.4 SpotON Flow Cells from ONT.

Cap Selection Protocol

The TeloPrime Full-Length cDNA Amplification Kit (Lexogen) was used for generating cDNAs from 5' capped polyA$^{(+)}$ RNAs. RT reactions were carried out with oligo(dT) primers (from the kit) or random hexamers (custom made) using the enzyme from the kit. A specific adapter (capturing the 5' cap structure) was ligated to cDNAs (25°C, overnight), then the samples were amplified by PCR using the Enzyme Mix and the Second-Strand Mix from the TeloPrime Kit. The reactions were performed in a Veriti Cycler and the samples were purified on silica membranes (TeloPrime Kit) after the enzymatic reactions. The Qubit 2.0 and the Qubit dsDNA HS quantitation assays (Life Technologies) were used for measuring the concentration of the samples. A quantitative PCR reaction was carried out for checking the specificity of the samples using the Rotor-Gene Q cycler (Qiagen) and the ABsolute qPCR SYBR Green Mix from Thermo Fisher Scientific. A gene-specific primer pair (HSV-1 *us9* gene, custom made) was used for the test amplification. The PCR products were used for ONT library preparation and sequencing. The end-repair and adapter ligation steps were carried out as was described in the 'Regular' protocol, and in our earlier publication (Boldogkői et al., 2018). The ONT R9.4 SpotONFlow Cells were used for sequencing.

Application of Terminator Exonuclease

Some of the non-Cap-selected samples were treated by Terminator exonuclease (Epicentre/Lucigen) in order to reduce the proportion of sequencing reads with incomplete 5'-UTR regions. The protocol has been carried out as recommended by the manufacturer. Briefly, 2 µl of buffer A, 1 µg of total RNA, 0.5 µl of RNaseOUT (Invitrogen), and 1 U of Terminator exonuclease were mixed and incubated at 30°C for 60 min. The same reaction was carried out using buffer B instead of buffer A, after which the two mixtures were pooled.

Oxford Nanopore Minion Platform—Direct RNA Sequencing

The ONT's Direct RNA sequencing (DRS) protocol (version: DRS_9026_v1_revM_15Dec2016) and the ONT Direct RNA Sequencing Kit (SQK-RNA001) were used to examine the transcript isoforms without enzymatic reactions—to avoid the potential biases—as well as to identify possible base modifications alongside the nucleotide sequences. Polyadenylated RNA was extracted from the total RNA samples and it was subjected to DRS library preparation according to the ONT's protocol (Boldogkői et al., 2018). The quantity of the sample was measured by Qubit 2.0 Fluorometer using the Qubit dsDNA HS Assay Kit (both from Life Technologies). The library was run on an ONT R9.4 SpotON Flow Cell. Basecalling was carried out using Albacore (v 2.3.1).

Mapping and Data Analysis

The minimap2 aligner (Li, 2018) was used with options *-ax splice -Y -C5 –cs* for mapping the raw reads to the reference genome (X14112.1), followed by the application of the LoRTIA toolkit (https://github.com/zsolt-balazs/LoRTIA) for the determination of introns, the 5' and 3' ends of transcripts, as well as for detecting the full-length reads. Putative introns were defined as deletions with the consensus flanking sequences (GT/AG, GC/AG, AT/AC). The complete intron lists are available as additional material. We used even stricter criteria: only those splice sites were accepted, which were validated by dRNA-Seq [used in our present work and in Depledge and coworkers' study (Depledge et al., 2019)]. These transcripts all have the canonical splice site: GT/AG and they are abundant (> 100 read in Sequel data).

The 5' adapter and the poly(A) tail sequences were identified at the ends of reads by the LoRTIA toolkit based on Smith-Waterman alignment scores (**Table 2**). If the adapter or poly(A) sequence ended at least three nucleotides (nts) downstream from the start of the alignment, the adapter was discarded, as it could have been placed there by template-switching. Transcript features such as introns, transcriptional start sites (TSS) and transcriptional end sites (TES) were annotated if they were detected in at least two reads and in 0.1% of the local coverage. In order to reduce the effects of RNA degradation, only those TSSs were annotated, which were significant peaks compared to their ±50-nt-long windows according to Poisson distribution. Reads being connected a unique set of transcript features were annotated as transcript isoforms. Low-abundance reads detected in a single experiment were accepted as transcripts if the same TSS and TES were also used by other transcripts. In most cases, those reads were accepted as isoforms, which were detected in at least two independent experiments. The 5′-ends of the long low-abundance reads were checked individually using the Integrative Genome Viewer (IGV; https://software.broadinstitute.org/software/igv/download). The workflow of the data analysis can be found in **Supplementary Figure 1**.

RESULTS

Analysis of the HSV-1 Transcriptome With Full-Length Sequencing

In this work, we report the application of two distinct LRS techniques (the PacBio Sequel and the ONT MinION platforms), and multiple library approaches for the investigation of the HSV-1 lytic transcriptome. We also reutilized our previous PacBio RS II data for the validation of novel transcripts. The PacBio sequencing is based on an amplified Iso-Seq template preparation protocol that utilizes a switching mechanism at the 5' end of the RNA template, and is thereby able to produce complete full-length cDNAs (Zhu et al., 2001). We applied both cDNA and dRNA sequencing for the ONT technique. Additionally, we used Cap-selection for a fraction of samples. A single sample was treated by Terminator exonuclease, which selectively degrades uncapped and non-polyadenylated transcripts. ONT sequencing was also used for the kinetic analysis of HSV-1 gene expressions. Sequencing reads were mapped to the HSV-1 (X14112) genome using the Minimap2 alignment tool (Li, 2018) with default parameters.

Altogether, we obtained 80,061 full-length ROIs mapping to the HSV-1 genome using Sequel sequencing, whereas PacBio RSII platform generated 38,972 ROIs (**Supplementary Table 1**). ONT sequencing produced altogether 1,505,848 sequencing reads mapping to the viral genome. The reason behind the relatively low proportion of the full-length read count within the MinION samples is that this method—compared to PacBio—generates a higher number of 5' truncated reads. We and others have reported in previous publications that the dRNA-Seq method is not optimal for capturing entire transcripts (Moldován et al., 2017b; Moldován et al., 2018b; Workman et al., 2018): we found that short 5' sequences of transcripts and in many cases the polyA-tails were missing from most of the reads. However, a recently published technique utilizing adapter ligation to the 5' end of full-length mRNAs is able to solve this problem (Jiang et al., 2019). Another drawback of native RNA sequencing is its low throughput compared to cDNA sequencing. The advantage of dRNA-Seq is that it is free of false products which are typically produced by RT, PCR, and cDNA sequencing.

Table 3 shows the average read lengths of mapped full-length ROIs and MinION reads in the different samples. A detailed description of reads obtained from all libraries is found in **Supplementary Table 1**.

TABLE 3 | Average mapped read-lengths and transcript lengths.

Technique	Average length of the reads (bp)	Average length of the abundant full-length transcripts (bp)
PacBio RSII *oligo(dT)*	1,369	1,409
PacBio RSII *random*	924	NA
PacBio Sequel	1,923	1,789
ONT MinION 1D *oligo(dT)*	967	1,222
ONT MinION 1D *random*	766	NA
ONT MinION Cap-seq *oligo(dT)*	683	797
ONT MinION dRNA-Seq	823	NA
ONT MinION *Terminator*	873	1,225
ONT MinION *Cap-seq random*	388	NA
ONT MinION time points	826	1,232

The data obtained from the individual p.i. time-points are discussed in ***Supplementary Table 1****.*

TABLE 2 | 5' adapter sequences and settings for adapter detection with the LoRTIA pipeline. The scoring of the Smith-Waterman alignment was set to +2 for matches and -3 for mismatches, gap openings and gap extensions.

Method	Adapter sequence	Score limit	Distance from the start of the alignment
PacBio	AGAGTACATGGG	16	+5/–15
MinION	TGCCATTAGGCCGGG	15	+5/–15
Teloprime	TGGATTGATATGTAATACGACTCACTATAG	20	+5/–30

Cap-selection performed suboptimally in our experiment, because it produced relatively short average sequencing reads. Random RT-priming allowed the analysis of non-polyadenylated transcripts and helped the validation of TSSs and splice sites. Additionally, this technique proved to be superior for identifying the 5'-ends of very long transcripts, including polycistronic and complex RNA molecules. Terminator exonuclease was used for the enrichment of intact TSSs of the transcripts.

The following technical artifacts can be generated by RT and PCR: template switching, and nonspecific binding of oligod(T) or PCR primers. In addition to poly(A) tails, oligo(dT) primers occasionally hybridize to A-rich regions of transcripts and thereby produce false reads. These products were discarded from further analysis, albeit in some cases we were unsure about the non-specificity of the removed reads. We ran altogether 11 parallel sequencing reactions using 8 different techniques for providing independent reads. Additionally, in some cases, the same TSS, TES or splice junctions were found in other transcripts detected within the same sequencing reaction which further enhanced the number of independent sequencing reads. In our earlier publication (Tombácz et al., 2017b), we could not detect all spurious products, therefore, in the present work, we have made a minor correction to our formerly published results.

We used a novel bioinformatics tool (LoRTIA) — developed in our laboratory — for the identification of TSS and TES positions, as well as splice donor and acceptor sites (**Supplementary Figure 1**). This software suite detected a total of 1,677 putative TSSs 162 putative TESs and 379 putative introns (**Supplementary Table 2**). A putative TSS or TES was accepted as real if LoRTIA detected it in at least three independent samples in the case of longer isoforms, and five independent samples in the shorter variants, including 5'-truncated ORF-containing RNAs. The reason for a more stringent selection criterion for the short isoforms is that these can be the result of fragmentation, which is not the case for longer isoforms. These analyses yielded altogether 537 TSSs and 77 TESs. Only those sequencing reads were accepted as transcripts, which contained a TSS and a TES annotated in the above way. This method yielded 667 transcripts (**Supplementary Table 3**). For very long transcripts (≥ 3,000 bp), we applied a different rule: a read was accepted as a transcript if it was longer than all annotated overlapping transcripts even if it was represented in a few copies and had no annotated TSS. A large number of very long transcripts were identified this way in most cases in the Sequel dataset. Thus, altogether 2,250 transcripts were identified in this study (**Supplementary Table 3**). We assume that much more low-abundance and very long transcripts are expressed by the HSV-1 genome than we detected with our very strict criteria. Our dataset is available for further investigations, which can confirm or reject these latter categories of putative transcripts.

For intron identification, we used the following criteria: the candidate intron had to carry one of the canonical splice junction sequences: GT/AG, GC/AG, AT/AC; and it had to be detected by dRNA-Seq and both cDNA-Seqs (PacBio and ONT platforms). Besides introns based on hard evidence, we enlist additional putative introns of which the criterion was their detection by both dRNA-Seq and at least one of the cDNA (PacBio or ONT) sequencings. The third category of introns includes very abundant splice variants and introns on very long transcripts that were exclusively identified using Sequel sequencing in most cases. This study identified a large number of rare variants with deletions, which represented less than 5% of the total isoforms of a certain transcript. These putative splice variants were not accepted as transcripts. Altogether, 182 introns were identified in terms of the above criteria, among which 155 carry canonical GT/AG, 22 GC/AG, and 2 AT/AC splice junction sequences (**Supplementary Table 2**). Our analysis detected 80 transcripts containing one or more of these introns (**Supplementary Table 3**).

In Silico Analysis of Promoters and Poly(A) Signals

In order to detect promoter sequences, we analyzed the -150 to +1 upstream region of the TSSs *in silico* (**Figure 1**). We found that 45% (371) of the TSSs are preceded by a canonical GC box sequence at a mean distance of 66.301nt ($\sigma = 31.205$), 4% (35) by a CAAT box at a mean distance of 113.428nt ($\sigma = 15.471$), and 11% (91) by a TATA box at a mean distance of 30.373nt ($\sigma = 2.058$) (Mackem and Roizman, 1982; Guzowski and Wagner, 1993). Some of the GC boxes may be nonfunctional, since they may be the result of the high GC-content of the viral genome. Earlier studies found a canonical initiator region (INR) ± 5 nt around the TSS of eukaryotic organisms (Lim et al., 2004; Xi et al., 2007). These have been shown to be used during the early viral gene expression, whereas late transcription is initiated from a G-rich sequence (Huang et al., 1996; Lieu and Wagner, 2000). We detected 16 TSSs containing a CAG INR (TSS position underlined) and 89 TSSs having YANW (Y: cytosine/thymine, N: adenine/cytosine/thymine/guanine, W: thymine/adenine, TSS position underlined).

We found that TSSs expressed in every time point are abundant and their INRs exhibit high similarity to canonical eukaryotic INRs, whereas TSSs from late samples are similar to the VP5 promoter (**Figure 1A**). Furthermore, these late TSSs are expressed in low abundance (2.8% of all reads starting in these positions) but their ratio is seven-fold higher than those of early TSSs (**Figure 1B**). We carried out *in silico* analysis of the -50nt region located upstream the TESs and detected 59 possible polyadenylation signals (PASs) at a mean distance of 21.779nt ($\sigma = 5.558$). The number of TESs expressed in both early and late phases is slightly higher than the number of TESs expressed only in the late phase of the viral life cycle (**Figure 1C**). TESs expressed throughout the entire viral replication are characterized by canonical PASs, cleavage signals and GU-rich regions. This is in contrast with TESs expressed only in the late phase, which tend to have no canonical signals for polyadenylation and cleavage (**Figure 1D**). Additionally, these late TESs are low abundance, representing only 0.1% of the reads' 3' ends.

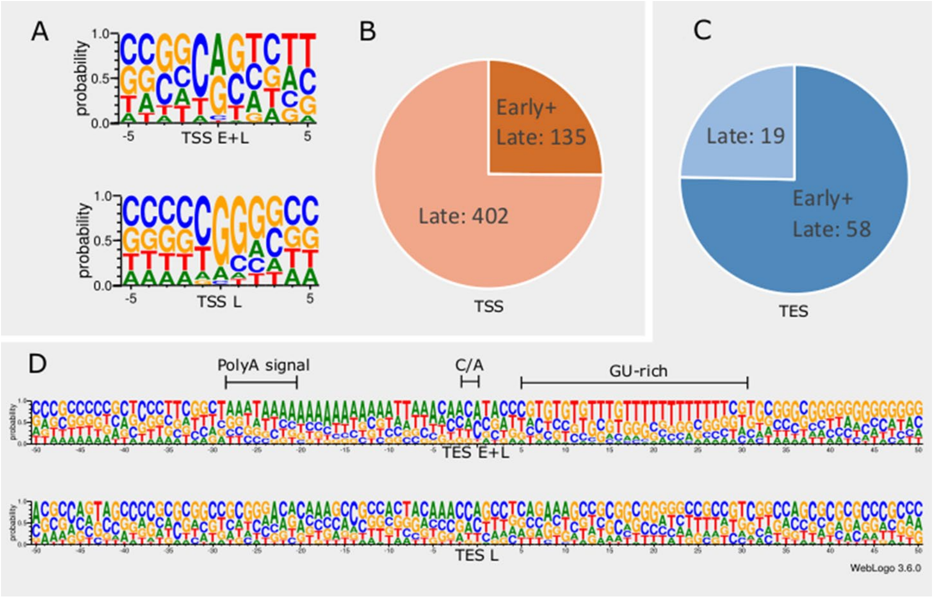


FIGURE 1 | *In silico* analysis of INR and PAS sequences. **(A)** The initiator region (INR) of early samples is similar to the canonical eukaryotic INR sequence, while late INRs show homology with the VP5 promoter. **(B)** The proportion of TSSs present in both early and late or exclusively late time points of infection. **(C)** The proportion of TESs present in both early and late or exclusively late time points of infection. **(D)** The probability of expression of nucleotides in the ±50nt region of TESs throughout the entire infection period compared to those nucleotides that expressed only in late time points. TESs expressed during the entire period of infection (E+L) contain a canonical poly(A) signal, the C/A cleavage site and GU-rich downstream region. Late TESs lack a PAS and the canonical downstream elements, but they contain a GC-rich sequences 15-20nt downstream of the cleavage site.

Novel Putative mRNAs

5’-Truncated transcriptional reads were accepted as transcripts if they were present in at least five independent samples. The first base had to be located within a ±5 window range. Additionally, reads having less than a 5% proportion at the overlapping region were discarded. Present investigations revealed 182 novel 5’-truncated mRNAs (tmRNAs) of HSV-1 (**Supplementary Table 4**), which were all produced from genes embedded in larger host genes of the virus. These 5’-truncated mRNAs are assumed to be generated by alternative transcription initiation from promoters located within larger genes. We could identify promoter modules for only 39 transcripts (we could not identify promoter consensus sequences for several canonical ORFs, too). These transcripts all contain in-frame ORFs. The first in-frame AUG triplet is assumed to encode the translation start codon. Further analyses have to be carried out to verify the coding potential of the ORF-containing tmRNAs. We detected a transcript — termed ‘RL-intron’ (RL2I) — with a TSS identical to that of the TSS of *rl2* gene but with a TES located within the intronic region of this gene. Our BLAST searches resulted in hypothetical proteins predicted to this ORF, but according to our knowledge, no such transcript has been detected until now.

Novel Putative Non-Coding (or Coding) Transcripts

In this part of our study, we detected 18 putative non-coding RNAs, including antisense RNAs (asRNAs, termed as ASTs) and other putative long non-coding RNAs (lncRNAs) (**Table 4**). Furthermore, we validated and determined the base pair-precision termini of the transcripts published earlier by us and

TABLE 4 | Polyadenylated ncRNAs of HSV-1. **(A)** Previously detected and validated ncRNAs; **(B)** Novel ncRNAs. All transcripts are polyadenylated.

Name	Genomic locations	
A		
LAT 0.7 kb - S	7,643	8,393
LAT 0.7 kb	7,643	8,423
AST-1	57,711	59,429
AST-2-L4*	78,315	80,725
AST-2-L3*	78,531	80,725
AST-2 sp	79,792	80,725
AST-2	79,792	80,725
AST-3*	103,152	103,512
AST-4*#	110,816	112,131
LAT 0.7 kb	117,948	118,728
LAT 0.7 kb - S	117,978	118,728
B		
LAT 0.7 kb - ul1-2-3-3.5*	7,643	11,285
LAT 0.7 kb - S2	7,643	8,338
LAT 1.1 kb	7,643	8,733
AST-2-sp2	79,792	80,725
LAT 1.1 kb	118,033	118,728
LAT 0.7 kb - S2	117,638	118,728
LAT 0.7 kb - L*	115,083	118,728
AST-5	141,008	141,629

**unidentified 5’ end # unidentified 3’ end.*

by others. **Supplementary Table 5** shows the potential peptides encoded by the ORFs on these transcripts. Further studies have to confirm whether these ORFs are translated. If so, they are novel protein-coding genes.

(1) Antisense RNAs These transcripts can be controlled by their own promoters or by the promoter of another (mRNA) gene. It has earlier been reported that the 0.7-kb LAT transcript is not expressed in strain KOS of HSV-1 (Zhu et al., 1999). Here we demonstrate that this is not the case, since we were able to detect this transcript. The existence of the shorter LAT-0.7kb-S (Tombácz et al., 2017b) was also confirmed. Additionally, we detected asRNAs being co-terminal with the LAT-0.7 transcripts, but having much longer TSSs. The LAT region and its surrounding genomic sequences are illustrated in **Figure 2A**. Using random oligonucleotide-based LRS techniques, we obtained a large number of antisense-oriented reads, most of them without identified 5'-ends. We also detected antisense transcripts without defined TSSs and TESs within 27 HSV-1 genes (*rl1, rl2, ul1, ul2, ul4, ul5, ul10, ul14, ul15, ul19, ul23, ul29, ul31, ul32, ul36, ul37, ul39, ul42, ul43, ul44, ul49, ul50, ul53, ul54, us4, us5, us8*). The expression level of these asRNAs is low, in most cases only a few reads were detected *per* gene locus. However, a high level of antisense RNA expression was identified within the locus of *ul10* gene. A special class of asRNAs is produced by divergent genes, and read-through RNAs (rtRNAs) generated by transcriptional read-through between convergent gene pairs. These transcripts are mRNAs with long stretches of antisense segments. For example, we detected an antisense transcript originated within the 3' region of *ul4* gene and co-terminated with UL6-7 bicistronic transcript. This RNA molecule contains three splice sites, and can be considered as a very long TSS isoform of the UL6-7 transcript.

(2) Intergenic ncRNAs A ncRNA (termed "intergenic ncRNA"; IGEN-1) located between the *ul26* and *ul27* genes was also identified. This transcript is co-terminal with the UL27-AT RNA, which is a longer TES isoform of UL27 transcript (**Figure 2B**). Another non-coding transcript (IGEN-2) with unidentified transcript ends was detected to be expressed in the outer termini of the HSV-1 unique long region. The potential function of IGEN transcripts remains unclear. A bidirectional, low-level expression from the intergenic region between the *rl2* (icp0) and LAT genes was also observed. These RNA molecules are co-terminal with the LAT-0.7kb transcript and may be parts of the potential RL2-LAT-UL1-2-3 transcript (Tombácz et al., 2017b). Additionally, we detected RNA expression in practically every intergenic region.

(3) Intra-intronic ncRNAs A ncRNA was identified within the intron of the *rl2* gene, which was designated as NCIRL2. This transcript is expressed in a low abundance.

Replication-Associated Transcripts

We identified five replication-associated RNAs (raRNAs) designated OriL-RNA1-2, and OriS-RNA1-3, which overlap the replications origins OriL and OriS, respectively. OriL-RNA1 is a long TSS isoform produced from the *ul30* gene, whereas OriS-RNA2 is a TSS variant of *rs1* (*icp4*) (**Figure 3**). OriL-RNA2 is a transcript without an annotated TES. We suppose that this transcript is the long TSS variant of the *ul29* gene. We were only able to detect certain segments but not the entire OriS-RNA1 described by Voss and Roizman (1988). We also detected a longer TSS isoform of the *us1* gene (US1-L2 = OriS-RNA3) which overlaps the OriS located within the terminal repeat of US region (TRS) (**Figure 3**). Additionally, OriS is also overlapped by a longer 5' variant of the *us12* gene (US12-11-10-L2 = OriS-RNA-4).

TSS and TES Isoforms

The multiplatform system allowed the discovery of novel RNA isoforms and reannotation of the transcript termini published earlier by others and us (Tombácz et al., 2017b; Depledge et al., 2019). The LoRTIA software suit — used for the detection of TSS and TES positions — identified 218 TSS and 56 TES positions (**Supplementary Table 2**). Altogether 53 genes produce at least one TSS isoform, besides the most frequent variants (**Supplementary Table 3**). Fifteen genes were found to produce three different transcript length isoforms (including the most frequent versions). The recent LRS analysis discovered 51 protein-coding and 2 (0.7 kb LAT, and RS1) non-coding transcripts with alternative TSSs. However, a few transcripts with unannotated 5'-ends were also detected (**Supplementary Table 3**). The alternative TSSs may lead to transcriptional overlap or they may enlarge the extent of existing overlaps especially between divergently transcribed genes. Some transcripts (e.g. UL19 and UL10) exhibit an especially high complexity of TSS isoforms (**Figure 4A**). The *ul21* gene produces nine different 5' length variants, the longer ones overlap the divergently oriented *ul22* gene) (**Figure 4B**). Additionally, long TSS isoforms are responsible for the overlaps of each replication origin of HSV-1, which is not the case in PRV, its close relative (Tombácz et al., 2015; Boldogkői et al., 2019a). Many of the longer TSS variants contain upstream ORFs (uORFs), which may carry distinct coding potentials as described by Balázs and colleagues in the human cytomegalovirus (Balázs et al., 2017a). Two novel 3'-UTR variants were also identified in this study.

Novel Splice Sites and Splice Isoforms

In this study, we also used dRNA sequencing, which provides a fundamentally different method from cDNA sequencing and hence can be utilized to filter out spurious splice sites. The splice donor and acceptor sites were also detected by using the LoRTIA tool. Altogether, using different sequencing techniques and bioinformatics analyses, we were able to verify the existence of 5 previously described and 30 novel splice sites. **Table 5** contains the list of introns, which were confirmed by dRNA-Seq (**Figure 5**). By far the most complex splicing pattern was detected in RNAs produced from the *ul41-45* genomic region.

Novel Multigenic Transcripts

Our earlier survey has revealed several novel multigenic RNAs, including polycistronic and complex transcripts (Tombácz

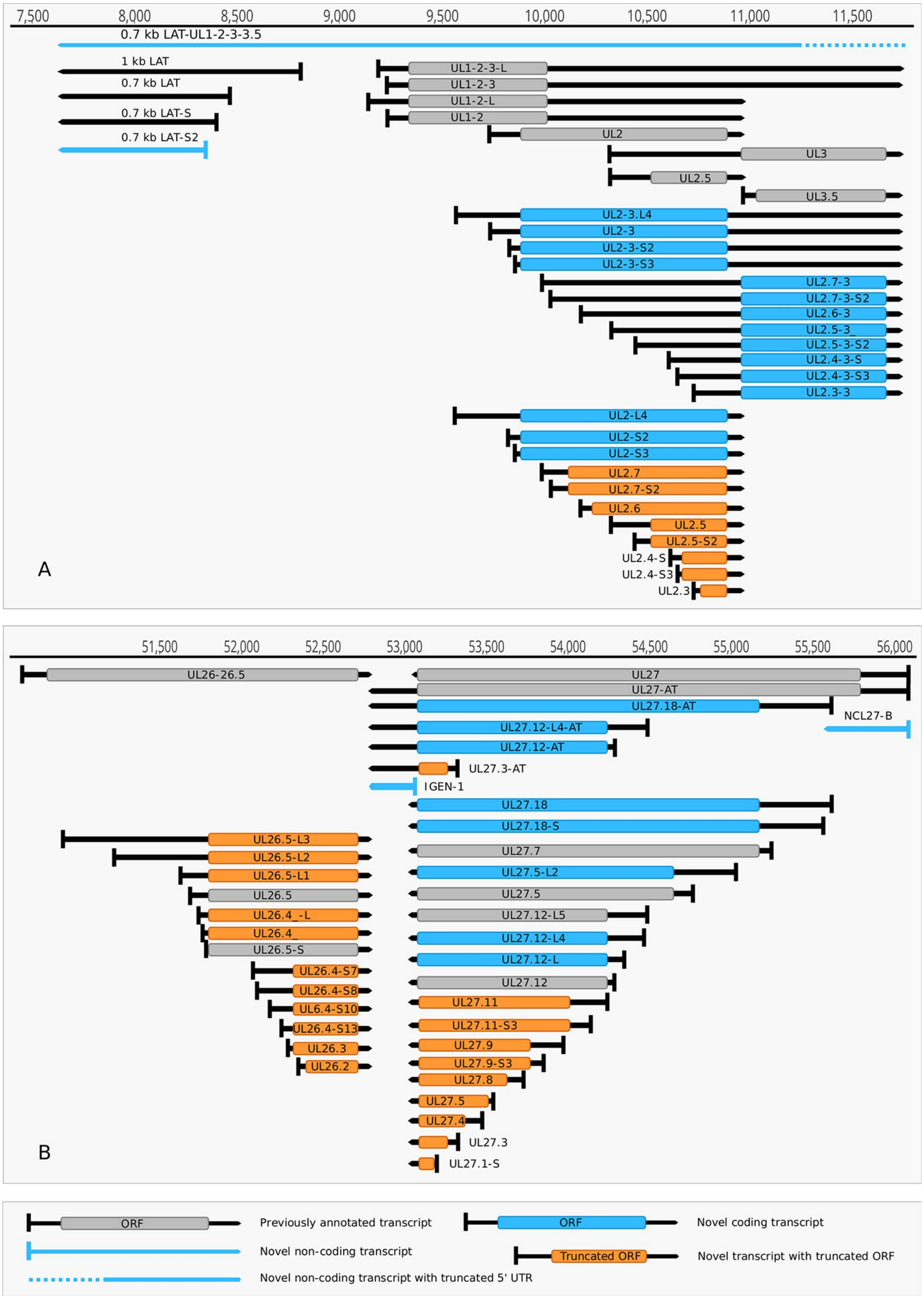


FIGURE 2 | Non-coding HSV-1 RNAs. **(A)** Schematic representation of the LAT region and surroundings. Besides the previously published coding and non-coding transcripts, this figure illustrates the newly discovered shorter TSS version of the 0.7 kb LAT, as well as the oppositely oriented transcript isoforms, which are co-terminal with the 3' ends of the UL2 or UL3 transcripts. **(B)** A novel non-coding transcript designated IGEN-1 is co-terminal with UL27-AT which is a longer TES isoform of UL27. Several other 5' UTR length variants were discovered and annotated in the UL26-UL27 region.

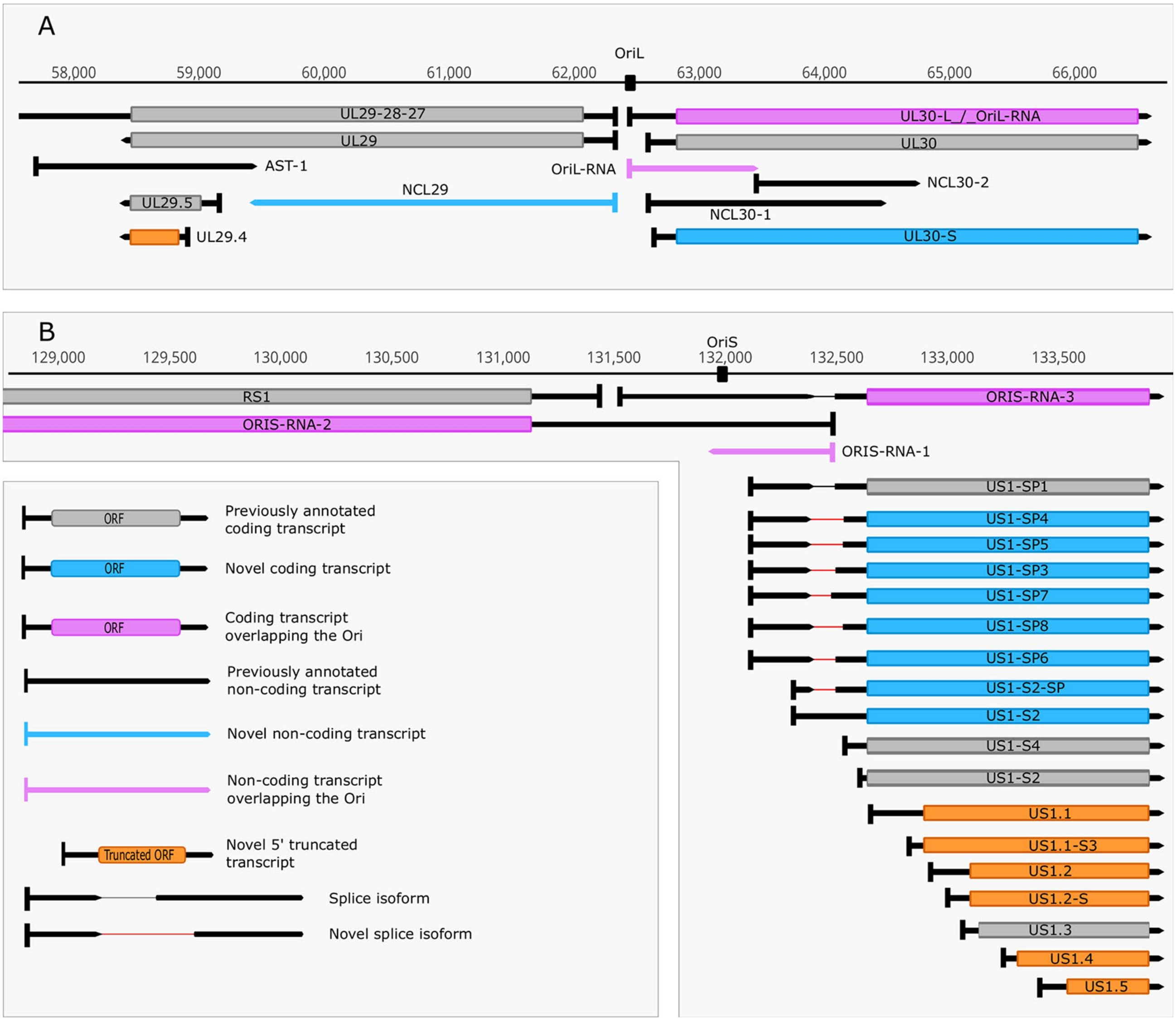


FIGURE 3 | Replication associated transcripts of HSV. **(A)** A novel shorter 5'-UTR isoform of the UL30, and a non-coding transcript sharing the TSS with UL29 but terminating within its ORF was discovered in the vicinity of Ori-L. **(B)** Two isoforms with shorter 5'-UTRs, seven splice isoforms and six novel putative protein-coding transcripts were annotated downstream of Ori-S.

et al., 2017b). In this work, we identified 201 multigenic transcripts containing two or more genes (**Supplementary Table 3**). The cxRNAs are long RNA molecules with at least 2 genes standing in opposite orientation relative to one another. Our intriguing findings are the RL1-RL2 (ICP34,5-ICP0) bicistronic transcript, as well as the 0.7. kb LAT-UL1-2-3-3.5 cxRNA (**Figures 2A, B**). Most of the novel multigenic transcripts are expressed at low levels, which can explain why they had previously gone undetected. In this work, we also identified four novel complex transcripts (0.7 kb LAT-UL1-2-c, UL18-15.5-c, UL20-21-c, US4-3-2-c) with unannotated TSSs (**Figure 2A**). We were able to detect these transcripts by cDNA sequencing and by the reanalysis of a MinION dRNA sequencing dataset (Depledge et al., 2019). Our novel experiments validated previously published cxRNAs. This study demonstrates that full-length overlaps between two divergently-oriented HSV-1 genes are an important source for the cxRNA molecules. The likely reason for the lack of cxRNA TSSs in many cases is that they are very long and low-abundance transcripts. It cannot be excluded with absolute certainty that some of the low-abundance multigenic transcripts are artefacts produced by the template–switch mechanism; other approaches are needed for the validation of their existence one-by-one.

Novel Transcriptional Overlaps

This study revealed an immense complexity of transcriptional overlaps (**Figure 6** and **Table 6**). These overlaps are produced by

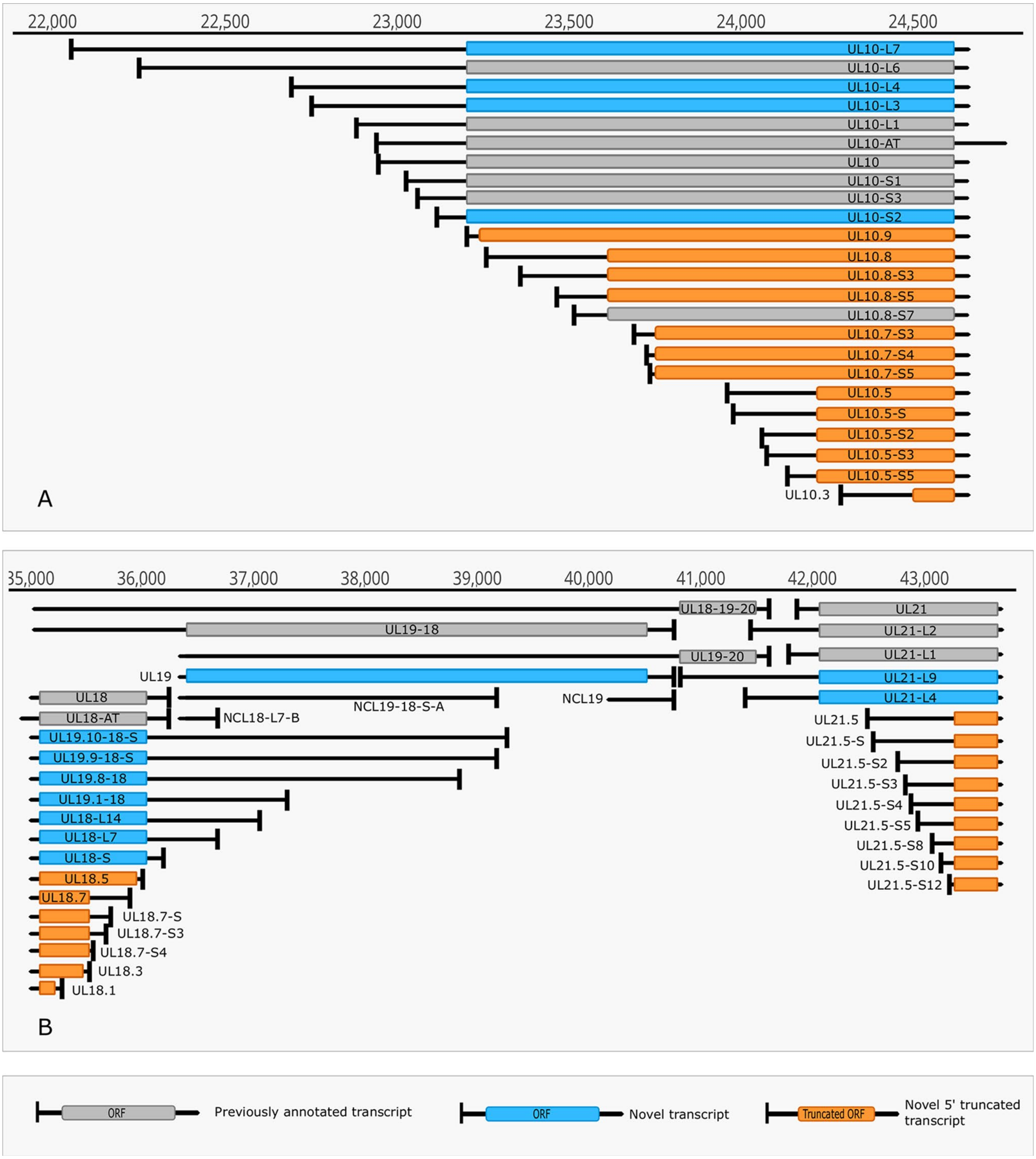


FIGURE 4 | Complexity of TSSs. **(A)** The TSS pattern of UL10 transcript exhibits an especially high complexity. Several TSSs are located downstream from the translation initiation site, resulting in truncated ORFs. RNAs harboring these truncated ORFs may code for N-terminally truncated transcripts or may be non-coding RNA. **(B)** Divergent overlaps between the ul20 and ul21 genes. These overlaps are caused by the high variability in the TSS of UL21.

either transcriptional read-through events between transcripts oriented in parallel [as described in Kara et al. (2019)], or in a convergent manner (thereby generating rtRNAs), or through the use of long TSS isoforms pertaining to one or of both partners of divergently-oriented genes. Transcriptional overlaps can also be produced by antisense transcripts controlled by their own promoters, as seen in LAT transcripts. Besides the 'soft' (alternative) overlaps, adjacent genes can also produce 'hard'

TABLE 5 | The most frequent splice sites of the HSV-1 transcriptome.

Intron start	Intron end	Read count	DNA strand	Splice donor/ acceptor	
2,318	3,082	20	+	GT/AG	
3,750	3,888	6	+	GT/AG	*
3,750	3,885	8	+	GT/AG	
13,449	13,931	37	−	GT/AG	*
30,049	33,634	198	+	GT/AG	
69,593	69,923	12	+	GT/AG	*
69,670	69,923	20	+	GT/AG	*
71,622	71,712	2	−	GC/AG	*
71,622	71,718	6	−	GC/AG	*
71,622	71,724	2	−	GC/AG	*
71,622	71,736	2	−	GC/AG	*
71,622	71,748	4	−	GC/AG	*
91,553	92,535	120	+	GT/AG	
97,724	97,949	228	+	GT/AG	
113,428	113,786	40	+	GT/AG	*
122,483	122,621	7	−	GT/AG	*
122,486	122,621	8	−	GT/AG	*
123,289	124,053	20	−	GT/AG	*
132,373	132,540	74	+	GT/AG	*
132,373	132,506	269	+	GT/AG	*
132,373	132,487	34	+	GT/AG	*
132,373	132,543	2	+	GT/AG	*
132,381	132,518	2,995	−	GT/AG	*
132,386	132,540	11	+	GT/AG	*
132,386	132,506	34	+	GT/AG	*
132,386	132,509	31	+	GT/AG	*
145,646	145,820	66	−	GT/AG	*
145,646	145,860	34	−	GT/AG	*
145,649	145,820	1,077	−	GT/AG	*
145,649	145,860	824	−	GT/AG	*
145,649	145,847	3	−	GT/AG	*
145,671	145,852	23	+	GT/AG	*
145,671	145,873	13	+	GT/AG	*
145,680	145,847	7	−	GT/AG	*
145,683	145,860	53	−	GT/AG	*
145,683	145,847	17	−	GT/AG	*

The newly discovered splice sites are labeled with asterisks.

overlaps when only overlapping transcripts are produced from the same gene pairs. An important novelty of this study is the discovery that practically each convergent gene produces rtRNAs crossing the boundaries of the adjacent genes. Two of the convergent gene pairs (*ul3-ul4* and *ul30-ul*31) form 'hard' transcriptional overlaps, whereas the other gene pairs form 'soft' overlaps. The 'softly' overlapping convergent transcripts are likely to be non-polyadenylated, since we were only able to detect most of them by the random primer-based sequencing technique. The *ul3-ul4* and *ul30-31* gene pairs also express non-polyadenylated rtRNAs that extend beyond their poly(A) sites. Transcriptional read-troughs were detected between each convergent gene pair in most cases from both directions, except in the UL43-44-45/UL48-47-46 cluster (**Figure 6** and **Table 6**). Another important novelty of this study is the discovery of very long TSS variants of divergent transcripts, the 5'-UTRs of which entirely overlap the partner gene. We detected very long transcripts which overlap the following divergent gene clusters: *ul4-5/ul6-7, ul4-5/ul6-7, ul4-5/ul6-7, ul4-5/ul6-7, ul9-8/ul10, ul9-8/ul10, ul14-13-12-11/ul15, ul17/ul15e2, ul20-19-18/ul21, ul20-19-18/ul21, ulL23-22/ul24-25-26, ul29*/OriL/*ul30, ul29*/OriL/*ul30, ul32-31/ul33-34-35, ul37/ul38-39-40, ul41-ul42, ul49.5.49/ul50, ul51/ul52-53-54, us2/US3, us2/us3, us2/us3*. Altogether, our results show that practically every nucleotide of the double-stranded HSV-1 DNA is transcribed.

Kinetics of HSV-1 Transcripts

Cultured Vero cells were incubated with HSV-1 for 1, 2, 4, 6, 8, 10, 12, or 24 h. Altogether, we obtained 1,028,840 viral reads in the kinetic part of the study (**Supplementary Table 1**). The distribution of TSSs and TESs along the HSV-1 genome is illustrated in **Figure 7** (see in detail in **Supplementary Figure 2**) and **Figure 8**. The dynamics of various transcript categories is exemplified in **Figure 9**, including tmRNAs (**panel A**), TSS isoforms (**panel B**), TES isoforms (**panel C**), splice variants (**panel D**), and polycistronic RNAs (**panel E**). Many mono- and polycistronic RNAs and transcript isoforms are differentially expressed throughout the replication cycle of the virus. The cumulative abundance of transcript isoforms in distinct period of HSV infection is depicted in **Supplementary Figure 3**.

DISCUSSION

In the last couple of years, LRS approaches revealed that the viral transcriptome is substantially more complex than previously thought (Boldogkői et al., 2019b). In this study, 2 sequencing platforms (PacBio Sequel and ONT MinION) and 8 library preparation methods were applied for the investigation of the HSV-1 lytic transcriptome, including both poly(A)$^+$ and poly(A)$^-$ RNAs. This research yielded a number of novel transcripts and transcript isoforms. We identified novel tmRNAs embedded into larger host viral genes. All of these short novel transcripts contain in-frame ORFs, but it does not necessarily mean that this coding potency is realized in translation. Indeed, most of the putative tmRNAs are expressed in low abundance (these were not accepted as transcripts), which raises doubts as to whether they code for proteins. These transcripts might have a regulatory role in certain step(s) of gene expression, but we cannot exclude that they represent mere transcriptional noise.

This study also identified a large number of transcript length isoforms varying in their TSSs or TESs. In certain genes, we obtained very high number of TSS isoforms, therefore we did not name them individually. Many of these length variants are expressed in low abundance. It is unknown whether these transcripts have distinct roles, or their function is exactly the same as the high-abundance variants. It is possible that increasing coverage further would reveal that transcripts are initiated from a promoter at each nucleotide within a certain stretch of DNA with varying probabilities. In the human cytomegalovirus and HSV it has been shown that the longer TSS variants may contain uORFs which may have a role in the translational regulation of downstream ORFs, and shorter TSSs, on the other hand, often contain N-terminally truncated ORFs (Stern-Ginossar et al., 2012; Balázs et al., 2017a; Whisnant et al., 2019).

In this work, we also detected novel splice sites and splice isoforms. We applied very strict criteria for the identification

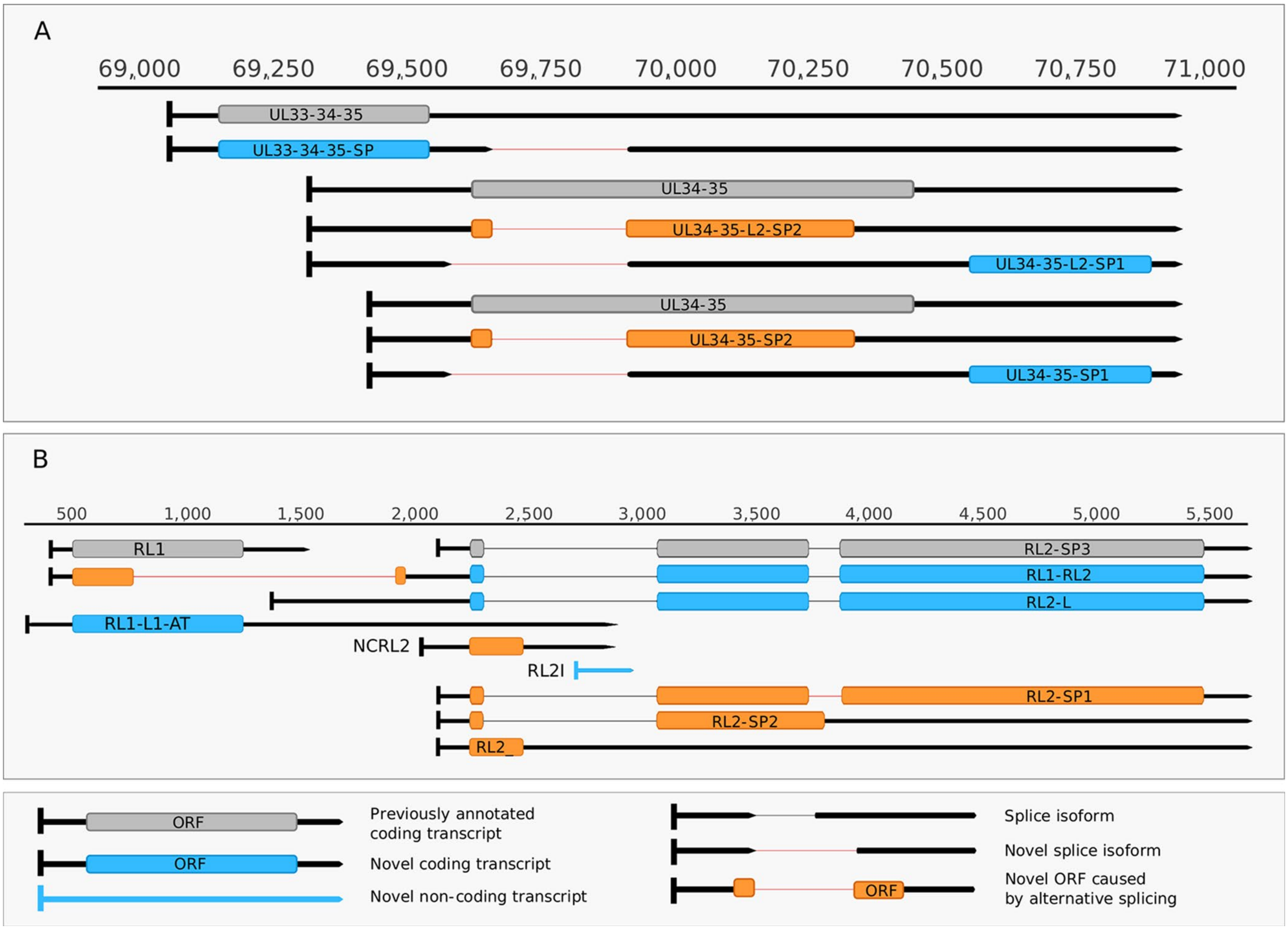


FIGURE 5 | Complexity of spliced transcripts. **(A)** The splice sites of UL34-35 transcript were confirmed by dRNA sequencing. The splicing event leads to frameshifting of the ul34 ORF in UL34-35-SP2 and UL34-35-L-SP2 and in the deletion of the translational initiation site of the ul34 ORF in UL34-35- SP1 UL34-35-L2-SP1. **(B)** The splicing complexity of RL2 and novel non-coding and potentially coding transcripts overlapping RL1 and the 5' UTR of RL2. Alternative splicing in RL2-SP1 produces a frameshift in the second half of the ORF, while alternative splicing in RL2-SP2 and in RL2 results in premature stop codons. Two novel 5'-UTR length isoforms of the RL2 transcripts are shown, one of which may contain a truncated form of the RL1 ORF. This truncation is caused by a splicing event. A novel isoform of the RL1 with longer 5' and 3' UTRs (designated as RS1-L1-AT) was discovered. Another novel putative protein coding transcript is the RS2-ALT, which is co-terminal with RS1-L1-AT.

of introns, therefore, many low-abundance introns have been eliminated. Indeed, after the submission of our manuscript, Tang et al. (2019) have reported the existence of several hundreds of splice sites in HSV-1. Further studies have to decide whether these putative introns are artifacts or they really exist.

Here, we also report the identification of several multigenic RNA molecules including polycistronic and complex transcripts. The existence of cxRNAs, expressed from convergent gene pairs, indicates that transcription does not stop at gene boundaries but occasionally continues across genes standing in opposite directions of one another. The cxRNAs are typically expressed in low amount: however, their abundance is difficult to determine precisely because the amount of long transcripts is significantly underestimated by LRS techniques.

We have also detected pervasive antisense transcript expression throughout the entire viral genome especially with the random primer-based sequencing method. Novel antisense RNAs are typically transcriptional read-through products specified by the promoter of neighboring convergent genes. These normally low-abundance, non-polyadenylated transcription reads contain varying 3'ends. The reason of this phenomenon is the use random nucleotide primers for the RT. The HSV-1 genome also expresses antisense RNAs controlled by their own promoters. For example, we identified a very long 5'-UTR isoform of LAT-0.7 transcript. The LAT RNAs have been shown to play a role in latency (Nicoll et al., 2016). LAT has also been shown to be a source of miRNAs (Lieberman, 2016). Further studies are needed to establish the potential function of LAT expression during the lytic cycle. We also detected novel divergent transcriptional overlaps: in two cases these transcripts appear to be initiated from the 3'-ends of the adjacent genes.

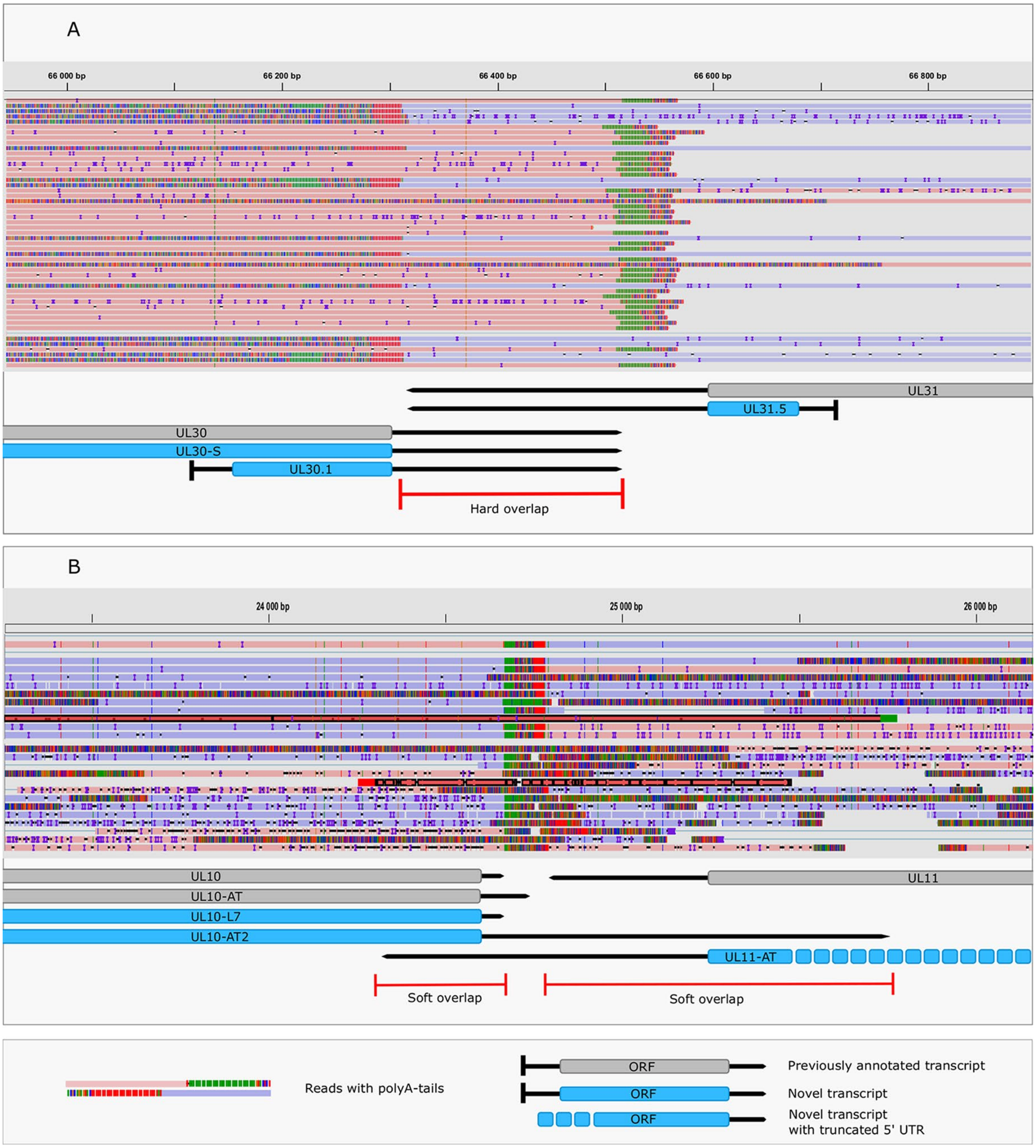


FIGURE 6 | Transcriptional overlaps. **(A)** A hard convergent overlap between the 3'-UTR regions of UL30 and UL31 transcripts shown by sequencing reads and annotations. **(B)** Occasional overlapping events between UL10-AT2 and UL11 and between UL11-AT and UL10 termed "soft convergent overlap". The reads representing UL10-AT2 and UL11-AT are shown in dark red. Reads were visualized using IGV.

In another article, we proposed a potential function for the complex overlapping meshwork formed by transcriptional read-throughs, divergent overlaps, antisense RNAs, as well as polygenic transcripts. We suggest the existence of a novel regulatory layer based on genome-wide interactions between closely located genes through the collision of and competition between their transcriptional machineries (Boldogkői et al., 2019c).

Moreover, we could also identify 2 novel replication-associated transcripts—OriL RNA-1 and OriS RNA-3—overlapping OriL and OriS, respectively. Both raRNAs are long TSS isoforms produced from the neighboring genes, *us1* for OriS, and *ul30* for OriL. Similar transcripts have also been recently described in other alphaherpesviruses (Moldován et al., 2017b; Boldogkői et al., 2018; Prazsák et al., 2018). Intriguingly, since the replication origin is located at different genomic regions of herpesviruses, the sequences

TABLE 6 | Read-through RNAs. **(A)** Novel ncRNAs with unidentified 3' ends; **(B)** Novel ncRNAs with unidentified 5' and 3' ends.

Name	Genomic locations	
A		
rtUL3-4	11,212	12,316
rtUL8-7	17,579	18,659
rtUL16-15L1	30,000	31,607
rtUL51-S-50	107,877	109,169
rtUL51-50	108,179	109,305
rtUL56-55-54-c	114,529	117,080
rtUS2-US1	133,243	135,306
rtUS1-US2	132,127	135,322
rtUS11-10-9	143,185	145,461
rtUS12-11-10-9	143,752	146,102
B		
IGEN-2 (earlier name: ULTN)	6,154	6,608
rtUL4-UL3	11,697	12,500
rtUL7-8	17,931	19,042
rtUL15-18	29,241	35,597
rtUL18-15	34,818	35,068
rtUL21-22	42,780	45,087
rtUL22-21/1	41,950	44,076
rtUL22-21/2	43,654	46,359
rULl26-27	52,662	54,774
rtUL36-35	71,000	71,520
rtUL41-40	89,898	91,274
rtUL40-41	90,900	91,712
AST-3-L	101,939	103,511
AST-3-UL49.5 rtRNA	102,801	103,952

These rtRNAs are probably non-polyadenylated because most of them were detected by random-primed sequencing alone. The genomic locations indicate the mapping of the transcription reads and not the transcript termini. "rt" stands for "read-through", "c" for "complex".

of raRNAs are non-homologous. The function of these transcripts may be the regulation of the initiation of replication fork as in bacterial plasmids (Tomizawa et al., 1981; Masukata and Tomizawa, 1986), or the regulation of replication orientation through a collision-based mechanism, as suggested earlier (Tombácz et al., 2015; Boldogkői et al., 2019a). In the latter case, raRNAs are mere byproducts of a regulatory mechanism, but it does not exclude the possibility that these transcripts have their own functions, which are at least partly different from those of shorter isoforms.

The analysis of the HSV-1 dynamic transcriptome has revealed a temporally differential expression of transcript isoforms, which suggests a function of these forms of diversity.

The prototypic organization of herpesvirus transcripts with respect to the location of genes is as follows (in the case of adjacent genes): abcd, bcd, cd, and d. However, there are some exceptions to this rule, e.g. the *ul41-43* and *ul51-49* regions. Both the regular and the irregular gene clusters exhibit time course differences in their location in mono- and various polycistronic RNAs. Genes are also transcribed in various combinations on RNA molecules but the expression of most genes follows the prototypic organization. All in all, this study identified several novel RNA molecules, and transcript isoforms. Further studies have to be carried out to ascertain the function of these transcripts. The question might be raised as to whether the low-abundance transcripts have any function at all, or whether they are the product of transcriptional noise. These transcripts may also be the by-products of a genome-wide regulatory mechanism discussed above, or they may also be functional.

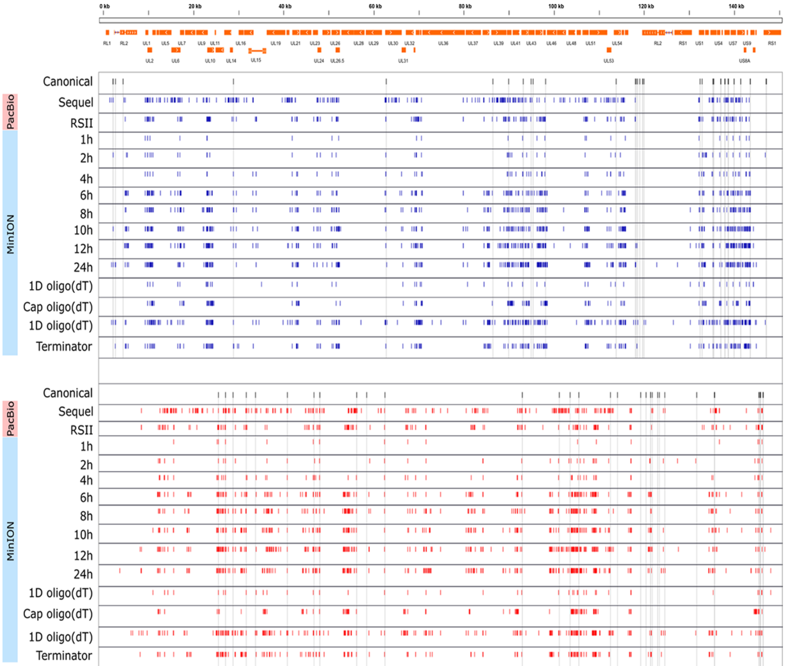


FIGURE 7 | Genome-wide kinetics of the TSSs of HSV-1. The TSSs were determined using the LoRTIA software suite in each sample. Blue dashes represent TSSs on the forward strand, while red dashes represent TSSs on the reverse strand. Black dashes represent previously known TSSs, whereas grey lines starting from the TSS and spanning to the bottom of the figure show the locations of known TSSs in every sample. Orange rectangles represent the ORFs. A higher resolution illustration is presented in **Supplementary Figure 2**.

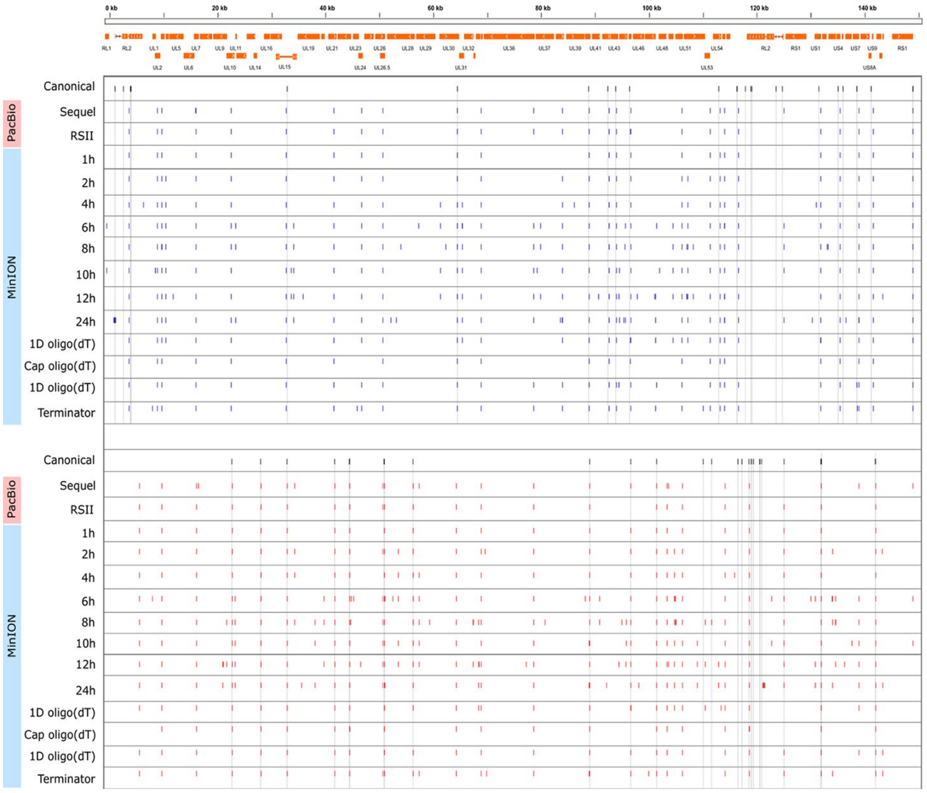


FIGURE 8 | Genome-wide kinetics of the TESs of HSV-1. The TESs were determined using the LoRTIA software suite in each sample. Blue dashes represent TESs on the forward strand, while red dashes represent TESs on the reverse strand. Black dashes represent previously known TESs, whereas grey lines starting from these and spanning to the bottom of the figure show the locations of known TESs in every sample. Orange rectangles represent the ORFs.

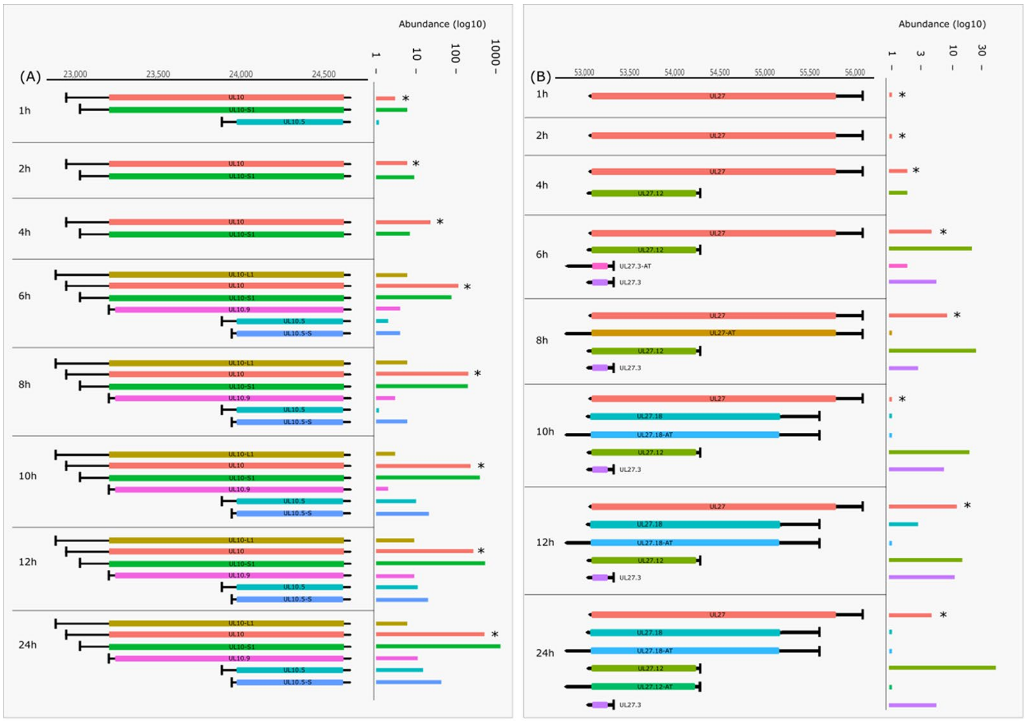


FIGURE 9 | Continued

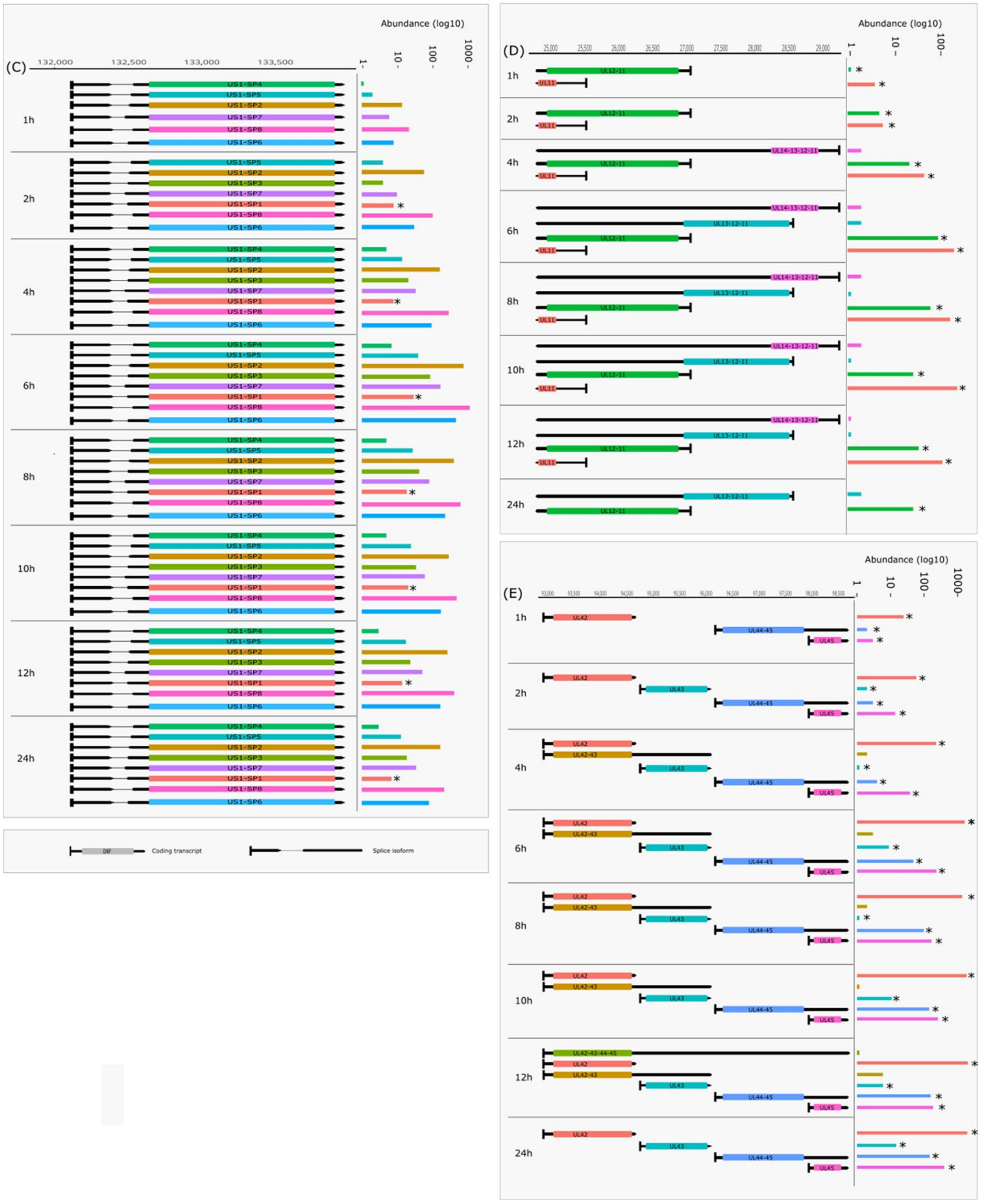


FIGURE 9 | Dynamic HSV-1 transcriptome—examples. The structure of transcript isoforms and of their position on the HSV-1 genome is shown by the annotations, while their abundance in distinct time points of the infection is represented on a log10 scale by bar plots on the right side of the annotation. Transcripts annotated in other works are marked with an asterisk (*). Transcript structures and counts were determined using the LoRTIA software suite. **(A)** The change in abundance of the 5'-UTR and 5' truncated isoforms of UL10. **(B)** Expression of UL27 RNA and its isoforms, including those with alternative termination. **(C)** Transcription kinetics of the US1 splice variants. **(D)** The change in abundance of polycistronic and monocistronic transcripts in the coterminal transcript at the UL11-UL14 region. **(E)** Transcription kinetics abundance of polycistronic and monocistronic transcripts in the UL42-UL45 region. Some of these transcripts are coterminal, while others have alternative terminations.

ACCESSION NUMBER

The PacBio RSII sequencing files and data files have been uploaded to the NCBI GEO repository and can be found with GenBank accession number GSE97785. The alignment files from MinION pooled samples, individual time points and Sequel sequencing have been deposited to the European Nucleotide Archive (ENA) under accession number PRJEB25433. Additional data from other sources utilized in this work for validation of rare transcripts and isoforms are available at the ENA with the study accession code PRJEB27861 (MinION dRNA-seq).

DATA AVAILABILITY

The datasets generated for this study can be found in European Nucleotide Archive, PRJEB25433.

AUTHOR CONTRIBUTIONS

DT designed the experiments, prepared the PacBio and ONT sequencing libraries, performed the PacBio RSII, Sequel and ONT MinION sequencing, analyzed the data, and drafted the manuscript. NM analyzed the dynamic transcriptome data and drafted the manuscript. ZBa adapted the LoRTIA pipeline for the analysis. GG analyzed the PacBio and ONT dataset and maintained the cell cultures. ZC isolated RNAs, generated cDNAs, prepared ONT libraries, and performed ONT MinION sequencing. MB analyzed the PacBio data and made manuscript revisions. MS conceived and designed the experiments. ZBo conceived and designed the experiments, supervised the study, analyzed the data, and wrote the final manuscript. All authors have read and approved the final version of the manuscript.

FUNDING

This study was supported by OTKA K 128247 to ZBo and OTKA FK 128252 to DT. DT was also supported by the Bolyai János Scholarship of the Hungarian Academy of Sciences and by the Eötvös Scholarship of the Hungarian State. The project was also supported by the NIH Centers of Excellence in Genomic Science (CEGS) Center for Personal Dynamic Regulomes [5P50HG00773502] to MS.

ACKNOWLEDGMENTS

We would like to thank Marianna Ábrahám (University of Szeged) for her technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00834/full#supplementary-material

SUPPLEMENTARY FIGURE 1 | Workflow of the data analysis.

SUPPLEMENTARY FIGURE 2 | High resolution TSS kinetics. TSSs and TESs were determined using the LoRTIA software suite in each sample. Blue dashes represent TSSs on the forward strand, while red dashes represent TSSs on the reverse strand. Orange rectangles represent the ORFs.

SUPPLEMENTARY FIGURE 3 | The cumulative abundance of transcript isoforms. Transcript isoforms were annotated and counted in separate stages of the viral infection using the LoRTIA software suite. The names of isoforms annotated in previous works by other methods are in red, whereas the isoforms detected by long-read sequencing are in black.

SUPPLEMENTARY TABLE 1 | Reads' statistics.

SUPPLEMENTARY TABLE 2 | TSSs, TESs and introns.

SUPPLEMENTARY TABLE 3 | (A) Genome coordinates and abundance of transcripts identified by software. TSSs with bold letters were detected in at least 3 independent samples. **(B)** Spliced transcripts with genome coordinates and intron abundances. Abbreviations: HA: highly abundant, A, abundant; LA, low abundance.

SUPPLEMENTARY TABLE 4 | Novel 5'-truncated transcripts with putative coding potential. This table summarizes novel and the previously published embedded mRNAs, as well as their genomic positions. Asterisks indicate transcripts that were also detected in our earlier study (Tombácz et al., 2017b).

SUPPLEMENTARY TABLE 5 | NcRNA_codepot table. The table enlists the transcript start and end positions, the ORF composition, excluding introns for spliced ORFs, the orientation of the ORFs, the size of the ORF and the amino acid sequence of the ORF. Homology of these ORFs was analyzed by aligning them to Non-redundant protein database using the BLASTp suite. Hits with the highest E-score were included in the table.

REFERENCES

Balázs, Z., Tombácz, D., Szűcs, A., Csabai, Z., Megyeri, K., Petrov, A. N., et al. (2017a). Long-read sequencing of human cytomegalovirus transcriptome reveals rna isoforms carrying distinct coding potentials. *Sci. Rep.* 7, 15989. doi: 10.1038/s41598-017-16262-z

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

Boldogkői, Z., Balázs, Z., Moldován, N., Prazsák, I., and Tombácz, D. (2019a). Novel classes of replication-associated transcripts discovered in viruses. *RNA Biol.* 16(2), 166–175, doi: 10.1080/15476286.2018.1564468

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

Boldogkői, Z., Szűcs, A., Balázs, Z., Sharon, D., Snyder, M., and Tombácz, D. (2018). Transcriptomic study of Herpes simplex virus type-1 using full-length sequencing techniques. *Sci. Data* 5, 180266. doi: 10.1038/sdata.2018.266

Boldogkői, Z., Tombácz, D., and Balázs, Z. (2019c). Interactions between the transcription and replication machineries regulate the RNA and DNA synthesis in the herpesviruses. *Virus Genes* 55, 274–279. doi: 10.1007/s11262-019-01643-5

Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., et al. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* 8, 16027. doi: 10.1038/ncomms16027

Chen, S.-Y., Deng, F., Jia, X., Li, C., and Lai, S.-J. (2017). A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. *Sci. Rep.* 7, 7648. doi: 10.1038/s41598-017-08138-z

Cheng, B., Furtado, A., and Henry, R. J. (2017). Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. *Gigascience* 6, 1–13. doi: 10.1093/gigascience/gix086

Costa, R. H., Cohen, G., Eisenberg, R., Long, D., and Wagner, E. (1984). Direct demonstration that the abundant 6-kilobase herpes simplex virus type 1 mRNA mapping between 0.23 and 0.27 map units encodes the major capsid protein VP5. *J. Virol.* 49, 287–292.

Depledge, D. P., Srinivas, K. P., Sadaoka, T., Bready, D., Mori, Y., Placantonakis, D. G., et al. (2019). Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nat. Commun.* 10, 754. doi: 10.1038/s41467-019-08734-9

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., et al. (2012). Landscape of transcription in human cells. *Nature* 489, 101–108. doi: 10.1038/nature11233

Du, T., Han, Z., Zhou, G., Roizman, B., and Roizman, B. (2015). Patterns of accumulation of miRNAs encoded by herpes simplex virus during productive infection, latency, and on reactivation. *Proc. Natl. Acad. Sci.* 112, E49–E55. doi: 10.1073/pnas.1422657112

Guzowski, J. F., and Wagner, E. K. (1993). Mutational analysis of the herpes simplex virus type 1 strict late UL38 promoter/leader reveals two regions critical in transcriptional regulation. *J. Virol.* 67, 5098–5108.

Harkness, J. M., Kader, M., and DeLuca, N. A. (2014). Transcription of the herpes simplex virus 1 genome during productive and quiescent infection of neuronal and nonneuronal cells. *J. Virol.* 88, 6847–6861. doi: 10.1128/JVI.00516-14

Hu, B., Huo, Y., Chen, G., Yang, L., Wu, D., and Zhou, J. (2016). Functional prediction of differentially expressed lncRNAs in HSV-1 infected human foreskin fibroblasts. *Virol. J.* 13, 137. doi: 10.1186/s12985-016-0592-5

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, 1898–1904.

Jiang, F., Zhang, J., Liu, Q., Liu, X., Wang, H., He, J., et al. (2019). Long-read direct RNA sequencing by 5'-Cap capturing reveals the impact of Piwi on the widespread exonization of transposable elements in locusts. *RNA Biol.* 16(7), 950–959, doi: 10.1080/15476286.2019.1602437

Kara, M., O'Grady, T., Feldman, E. R., Feswick, A., Wang, Y., Flemington, E. K., et al. (2019). Gammaherpesvirus readthrough transcription generates a long non-coding RNA that is regulated by antisense miRNAs and correlates with enhanced lytic replication *in vivo. Noncoding RNA* 5, 6. doi: 10.3390/ncrna5010006

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, Y., Fang, C., Fu, Y., Hu, A., Li, C., Zou, C., et al. (2018). A survey of transcriptome complexity in Sus scrofa using single-molecule long-read sequencing. *DNA Res.* 25, 421–437. doi: 10.1093/dnares/dsy014

Lieberman, P. M. (2016). Epigenetics and genetics of viral latency. *Cell Host Microbe* 19, 619–628. doi: 10.1016/j.chom.2016.04.008

Lieu, P. T., and Wagner, E. K. (2000). Two leaky-late HSV-1 promoters differ significantly in structural architecture. *Virology* 272, 191–203. doi: 10.1006/viro.2000.0365

Lim, F. (2013). HSV-1 as a model for emerging gene delivery vehicles. *ISRN Virol.* 2013, 1–12. doi: 10.5402/2013/397243

Lim, C. Y., Santoso, B., Boulay, T., Dong, E., Ohler, U., and Kadonaga, J. T. (2004). The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev.* 18, 1606–1617. doi: 10.1101/gad.1193404

Liu, H., Begik, O., Lucas, M. C., Mason, C. E., Schwartz, S., Mattick, J. S., et al. (2019). Accurate detection of m6A RNA modifications in native RNA sequences. *bioRxiv* 525741. doi: 10.1101/525741

Looker, K. J., Magaret, A. S., May, M. T., Turner, K. M. E., Vickerman, P., Gottlieb, S. L., et al. (2015). Global and regional estimates of prevalent and incident Herpes Simplex Virus Type 1 infections in 2012. *PLoS One* 10, e0140765. doi: 10.1371/journal.pone.0140765

Macdonald, S. J., Mostafa, H. H., Morrison, L. A., and Davido, D. J. (2012). Genome sequence of herpes simplex virus 1 strain KOS. *J. Virol.* 86, 6371–6372. doi: 10.1128/JVI.00646-12

Mackem, S., and Roizman, B. (1982). Structural features of the herpes simplex virus alpha gene 4, 0, and 27 promoter-regulatory sequences which confer alpha regulation on chimeric thymidine kinase genes. *J. Virol.* 44, 939–949.

Masukata, H., and Tomizawa, J. (1986). Control of primer formation for ColE1 plasmid replication: conformational change of the primer transcript. *Cell* 44, 125–136. doi: 10.1016/0092-8674(86)90491-5

McGeoch, D. J., Rixon, F. J., and Davison, A. J. (2006). Topics in herpesvirus genomics and evolution. *Virus Res.* 117, 90–104. doi: 10.1016/j.virusres.2006.01.002

McKnight, S. L. (1980). The nucleotide sequence and transcript map of the herpes simplex virus thymidine kinase gene. *Nucleic Acids Res.* 8, 5949–5964. doi: 10.1093/nar/8.24.5949

Merrick, W. C. (2004). Cap-dependent and cap-independent translation in eukaryotic systems. *Gene* 332, 1–11. doi: 10.1016/j.gene.2004.02.051

Miyamoto, M., Motooka, D., Gotoh, K., Imai, T., Yoshitake, K., Goto, N., et al. (2014). Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. *BMC Genom.* 15, 699. doi: 10.1186/1471-2164-15-699

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

Moldován, N., Szűcs, A., Tombácz, D., Balázs, Z., Csabai, Z., Snyder, M., et al. (2018a). Multiplatform next-generation sequencing identifies novel RNA molecules and transcript isoforms of the endogenous retrovirus isolated from cultured cells. *FEMS Microbiol. Lett.* 365, fny013. doi: 10.1093/femsle/fny013

Moldován, N., Tombácz, D., Szűcs, A., Csabai, Z., Balázs, Z., Kis, E., et al. (2018b). Third-generation sequencing reveals extensive polycistronism and transcriptional overlapping in a baculovirus. *Sci. Rep.* 8, 8604. doi: 10.1038/s41598-018-26955-8

Moldován, N., Tombácz, D., Szűcs, A., Csabai, Z., Snyder, M., and Boldogkői, Z. (2017b). Multi-platform sequencing approach reveals a novel transcriptome profile in pseudorabies virus. *Front. Microbiol.* 8, 2708. doi: 10.3389/fmicb.2017.02708

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Naito, J., Mukerjee, R., Mott, K. R., Kang, W., Osorio, N., Fraser, N. W., et al. (2005). Identification of a protein encoded in the herpes simplex virus type 1 latency associated transcript promoter region. *Virus Res.* 108, 101–110. doi: 10.1016/j.virusres.2004.08.011

Nicoll, M. P., Hann, W., Shivkumar, M., Harman, L. E. R., Connor, V., Coleman, H. M., et al. (2016). The HSV-1 latency-associated transcript functions to repress latent phase lytic gene expression and suppress virus reactivation from latently infected neurons. *PLOS Pathog.* 12, e1005539. doi: 10.1371/journal.ppat.1005539

Nudelman, G., Frasca, A., Kent, B., Sadler, K. C., Sealfon, S. C., Walsh, M. J., et al. (2018). High resolution annotation of zebrafish transcriptome using long-read sequencing. *Genome Res.* 28, 1415–1425. doi: 10.1101/gr.223586.117

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–e145. doi: 10.1093/nar/gkw629

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

Perng, G.-C., Maguen, B., Jin, L., Mott, K. R., Kurylo, J., BenMohamed, L., et al. (2002). A novel herpes simplex virus type 1 transcript (AL-RNA) antisense to the 5' end of the latency-associated transcript produces a protein in infected rabbits. *J. Virol.* 76, 8003–8010. doi: 10.1128/JVI.76.16.8003-8010.2002

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

Rajčáni, J., Andrea, V., and Ingeborg, R. (2004). Peculiarities of Herpes Simplex Virus (HSV) transcription: an overview. *Virus Genes* 28, 293–310. doi: 10.1023/B:VIRU.0000025777.62826.92

Rixon, F. J., and Clements, J. B. (1982). Detailed structural analysis of two spliced HSV-1 immediate-early mRNAs. *Nucleic Acids Res.* 10, 2241–2256. doi: 10.1093/nar/10.7.2241

Sedlackova, L., Perkins, K. D., Lengyel, J., Strain, A. K., van Santen, V. L., and Rice, S. A. (2008). Herpes simplex virus type 1 ICP27 regulates expression of a variant, secreted form of glycoprotein C by an intron retention mechanism. *J. Virol.* 82, 7443–7455. doi: 10.1128/JVI.00388-08

Shah, K., Cao, W., and Ellison, C. E. (2019). Adenine methylation in drosophila is associated with the tissue-specific expression of developmental and regulatory genes. *G3 (Bethesda)* 9 (6), 1893–1900. doi: 10.1534/g3.119.400023

Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V. T. K., Hein, M. Y., Huang, S.-X., et al. (2012). Decoding human cytomegalovirus. *Science* 338, 1088–1093. doi: 10.1126/science.1227919

Stingley, S. W., Ramirez, J. J., Aguilar, S. A., Simmen, K., Sandri-Goldin, R. M., Ghazal, P., et al. (2000). Global analysis of herpes simplex virus type 1 transcription using an oligonucleotide-based DNA microarray. *J. Virol.* 74, 9916–9927. doi: 10.1128/JVI.74.21.9916-9927.2000

Tang, S., Patel, A., and Krause, P. R. (2019). Hidden regulation of herpes simplex virus 1 pre-mRNA splicing and polyadenylation by virally encoded immediate early gene ICP27. *PLOS Pathog.* 15, 1–30. doi: 10.1371/journal.ppat.1007884

Tombácz, D., Balázs, Z., Csabai, Z., Moldován, N., Szűcs, A., Sharon, D., et al. (2017a). Characterization of the dynamic transcriptome of a herpesvirus with long-read single molecule real-time sequencing. *Sci. Rep.* 7, 43751. doi: 10.1038/srep43751

Tombácz, D., Csabai, Z., Oláh, P., Balázs, Z., Likó, I., Zsigmond, L., et al. (2016). Full-length isoform sequencing reveals novel transcripts and substantial transcriptional overlaps in a herpesvirus. *PLoS One* 11, e0162868. doi: 10.1371/journal.pone.0162868

Tombácz, D., Csabai, Z., Oláh, P., Havelda, Z., Sharon, D., Snyder, M., et al. (2015). Characterization of novel transcripts in pseudorabies virus. *Viruses* 7, 2727–2744. doi: 10.3390/v7052727

Tombácz, D., Csabai, Z., Szűcs, A., Balázs, Z., Moldován, N., Sharon, D., et al. (2017b). Long-read isoform sequencing reveals a hidden complexity of the transcriptional landscape of herpes simplex virus type 1. *Front. Microbiol.* 8, 1079. doi: 10.3389/fmicb.2017.01079

Tombácz, D., Prazsák, I., Szűcs, A., Dénes, B., Snyder, M., and Boldogkői, Z. (2018a). Dynamic transcriptome profiling dataset of vaccinia virus obtained from longread sequencing techniques. *Gigascience* 7, giy139. doi: 10.1093/gigascience/giy139

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

Tombácz, D., Tóth, J. S., Petrovszki, P., and Boldogkoi, Z. (2009). Whole-genome analysis of pseudorabies virus gene expression by real-time quantitative RT-PCR assay. *BMC Genom.* 10, 491. doi: 10.1186/1471-2164-10-491

Tomizawa, J., Itoh, T., Selzer, G., and Som, T. (1981). Inhibition of ColE1 RNA primer formation by a plasmid-specified small RNA. *Proc. Natl. Acad. Sci. U.S.A.* 78, 1421–1425. doi: 10.1073/pnas.78.3.1421

Viehweger, A., Krautwurst, S., Lamkiewicz, K., Madhugiri, R., Ziebuhr, J., Hölzer, M., et al. (2019). Direct RNA nanopore sequencing of full-length coron-avirus genomes provides novel insights into structural variants and enables modification analysis. *bioRxiv,* 483693. doi: 10.1101/483693

Voss, J. H., and Roizman, B. (1988). Properties of two 5'-coterminal RNAs transcribed part way and across the S component origin of DNA synthesis of the herpes simplex virus 1 genome. *Proc. Natl. Acad. Sci. U.S.A.* 85, 8454–8458. doi: 10.1073/pnas.85.22.8454

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi: 10.1038/nrg2484

Wen, M., Ng, J. H. J., Zhu, F., Chionh, Y. T., Chia, W. N., Mendenhall, I. H., et al. (2018). Exploring the genome and transcriptome of the cave nectar bat Eonycteris spelaea with PacBio long-read sequencing. *Gigascience* 7, giy116. doi: 10.1093/gigascience/giy116

Whisnant, A. W., Jürges, C. S., Hennig, T., Wyler, E., Prusty, B., Rutkowski, A. J., et al. (2019). Integrative functional genomics decodes herpes simplex virus 1. *bioRxiv* 603654. doi: 10.1101/603654

Wongsurawat, T., Jenjaroenpun, P., Wassenaar, T. M., Wadley, T. D., Wanchai, V., Akel, N. S., et al. (2018). Decoding the epitranscriptional landscape from native RNA sequences. *bioRxiv,* 487819. doi: 10.1101/487819

Workman, R. E., Tang, A., Tang, P. S., Jain, M., Tyson, J. R., Zuzarte, P. C., et al. (2018). Nanopore native RNA sequencing of a human poly(A) transcriptome. *bioRxiv,* 459529. doi: 10.1101/459529

Xi, H., Yu, Y., Fu, Y., Foley, J., Halees, A., and Weng, Z. (2007). Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. *Genome Res.* 17, 798–806. doi: 10.1101/gr.5754707

Zhang, B., Liu, J., Wang, X., and Wei, Z. (2018). Full-length RNA sequencing reveals unique transcriptome composition in bermudagrass. *Plant Physiol. Biochem.* 132, 95–103. doi: 10.1016/j.plaphy.2018.08.039

Zhao, L., Zhang, H., Kohnen, M. V., Prasad, K. V. S. K., Gu, L., and Reddy, A. S. N. (2019). Analysis of transcriptome and epitranscriptome in plants using PacBio Iso-Seq and nanopore-based direct RNA sequencing. *Front. Genet.* 10, 253. doi: 10.3389/fgene.2019.00253

Zhu, J., Kang, W., Marquart, M. E., Hill, J. M., Zheng, X., Block, T. M., et al. (1999). Identification of a Novel 0.7-kb polyadenylated transcript in the LAT promoter region of HSV-1 that is strain specific and may contribute to virulence. *Virology* 265, 296–307. doi: 10.1006/viro.1999.0057

Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R., and Siebert, P. D. (2001). Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* 30, 892–897. doi: 10.2144/01304pf02

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a past collaboration with several of the authors ZB, MS.