 doi:10.20944/preprints202012.0616.v1

Article

An Integrated Sequencing Approach for Updating of Pseudorabies Virus Transcriptome

Gábor Torma [1], Dóra Tombácz [1,2], Zsolt Csabai [1], Dániel Göbhardter [1], Zoltán Deim [3], 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, California, USA

[3] Department of Biotechnology, Faculty of Science and Informatics, University of Szeged, Szeged, Hungary

* Correspondence: boldogkoi.zsolt@med.u-szeged.hu

Abstract: In the last couple of years, the implementation of long-read sequencing (LRS) technologies for transcriptome profiling has uncovered an extreme complexity of viral gene expression. In this study, we carried out a systematic analysis on the pseudorabies virus transcriptome by combining our current data obtained by using Pacific Biosciences Sequel and Oxford Nanopore Technologies MinION sequencings with our earlier data generated by other LRS and short-read sequencing techniques. As a result, we identified a number of novel genes, transcripts, and transcript isoforms, including splice and length variants, and also confirmed earlier annotated RNA molecules. One of the major findings of this study is the discovery of a large number of 5′-truncated putative mRNAs embedded into larger host mRNAs. A large fraction of these RNA molecules contain in-frame ORFs, which may encode N-terminally truncated polypeptides. These study demonstrates that the PRV transcriptome is much more complex than previously appreciated.

Keywords: pseudorabies virus; herpesvirus; transcriptome; Pacific Biosciences, nanopore sequencing; long-read sequencing

1. Introduction

Pseudorabies virus (PRV; also called as Suid herpesvirus 1), an important veterinary pathogen belongs to the subfamily Alpherpesvirinae of the family Herpesviridae. Its closest relatives are the bovine alphaherpesvirus type 1(BoHV-1) [1] and the varicella zoster virus (VZV) [2]. PRV causes Aujeszky's disease [3] in swine, but other mammalian animals, such as dog, cat, sheep, cattle, and raccoon are also susceptible to the virus. Human and horse are resistant to PRV infection. Like other herpesviruses, PRV is an enveloped virus with a nucleocapsid containing a large (~143,000 kbp, depending on the given strain), linear, double-stranded DNA molecule [4]. Besides the lytic cycle, PRV can also enter latency mainly in the trigeminal ganglia of the infected animals [5].

PRV is a model organism for studying the molecular pathogenesis of herpesviruses, including the mechanism of neurotropism [6,7] and the regulation of gene expression. PRV is also served as a model for the development of genetically engineered vaccines [8]. Additionally, this virus is the most popular multisynaptic tracer of neural circuits [9–12]. PRV was genetically modified in order to restrict its spread exclusively in a retrograde manner [10,13,14], to decrease its virulence [15], or to contain fluorescent protein genes for enhancing detectability of virally infected cells [16], or fluorescent activity markers, which allow the simultaneous monitoring of the activity of multiple neurons using optical methods [17]. Finally, PRV has also been utilized

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 December 2020 doi:10.20944/preprints202012.0616.v1

as a vector for gene delivery to cardiac muscle cells [18] and to embryonic spinal cord graft [19], as well as an oncolytic agent [20].

The transcriptional cascade of herpesviruses was originally defined by assessing of viral RNAs and proteins using inhibitors of protein synthesis (cycloheximide) or DNA replication (phosphonoacetic acid). The immediate-early (IE) genes can be expressed in the absence of *de novo* viral protein synthesis. The *ie180* is the only PRV IE gene, and it encodes a transcription activator [21]. Most of the E genes specifies enzymes needed for the synthesis is viral DNA. The L genes code for the structural elements of the virion, including capsid and spike proteins. Late genes were further delineated as leaky late (L1) and true late (L2) based upon whether they start to be expressed before or after the genome replication, respectively. Viral transcriptomes have been profiled by real-time RT-PCR [22] or RNA sequencing using both Illumina-based short-read sequencing (SRS) [23,24], and long-read sequencing (LRS), including three different platforms from Pacific Biosciences (PacBio) [25–27], Oxford Nanopore Technologies (ONT) [27], and Loop Genomics [28]. ONT approaches included amplified cDNA , direct cDNA (dcDNA) [28] and native (direct) RNA (dRNA) sequencings [29].

Previously, we reported the characterization of PRV transcriptome using SRS based on Illumina platform [24] and LRS based on PacBio RS II [26,30] and ONT MinION [31] platforms. In this current study, we carried out PacBio Sequel and ONT MinION sequencings using novel library preparation approaches. The number of sequencing reads was significantly increased, which allowed the discovery of novel genes, transcripts and transcript isoforms, and the confirmation of our earlier data. Additionally, PacBio Sequel and ONT dRNA sequencings generated very long reads, which helped the identification of novel ultra-long polygenic viral RNAs. In this work, we combined our novel data with the earlier data obtained using different sequencing techniques, which helped not only the annotation of novel transcripts, but also to put a few low-abundance transcripts into the 'uncertain' category.

2. Results

2.1. Analysis of PRV transcriptome using sequencing data obtained in this and in our earlier studies.

In this work, we carried out the transcriptome profiling of pseudorabies virus using novel and formerly published sequencing data. Two LRS platforms were applied for the generation of novel data: PacBio Sequel and ONT MinION. Oligo(dT) primers were used for the reverse transcription (RT) in PacBio sequencing, and oligo(dT) and random primers were used for the RT in ONT sequencing. The criterion for the acceptance of transcripts was the identification of their transcription start sites (TSSs) and transcription end sites (TESs) by at least two independent techniques using LoRTIA software suit developed in our laboratory [32]. For accepting splice sites, embedded transcripts and short 5'-untranslated region (5'-UTR) isoforms, we applied an even more stringent criterion: identification by at least two independent techniques plus by direct RNA (dRNA) sequencing.

Compared to our earlier publications, in this work, we used novel sequencing chemistries for dRNA and dcDNA sequencings, the latest ONT-guppy basecaller instead of the earlier used Albacore, as well as Minimap2 long-read mapper instead of GMAP. It is well-known that RT, PCR and other processes can produce false TESs, TSSs and splice sites [33] due to RNA degradation, false priming and template switching at the repetitive sequences of the transcripts. LoRTIA software was used to eliminate spurious products and also for checking the quality of sequencing adapters and poly(A) sequences.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 December 2020 doi:10.20944/preprints202012.0616.v1

Direct RNA sequencing generates less false transcripts, but it produces incomplete reads since 15-30 bp sequences always lack from the 5'-termini, and in many cases poly(A) tails are also missing. Another disadvantage of dRNA-Seq is its low throughput compare to the cDNA-Seq techniques. Direct cDNA sequencing circumvents these problems, and it also eliminates non-specific events related to PCR amplification. Additionally, dcDNA-Seq produces longer reads than the amplified sequencing techniques and it yields higher coverage than the dRNA-Seq technique. Direct RNA sequencing has become the golden standard by now because it is considered to be error-free (excluding the problems with the transcript termini). However, according to our experience, dRNA-Seq technique has its own specific biases for the generation of false transcripts, because several LoRTIA transcripts obtained using this technique were unreproducible by other techniques. Thus, the distinct techniques have different strengths and limitations, which underlines the importance for the use of multiplatform approaches in transcriptome research.

In total, 13 ONT, 29 PacBio and 2 Illumina samples were used for the analysis of the PRV transcriptome**. Figure 1** shows the workflow of experiments and bioinformatic analyses**. Table S1** shows the details of transcription reads and coverages obtained by the various techniques. The detailed sequencing statistics is shown in **Table 1.**

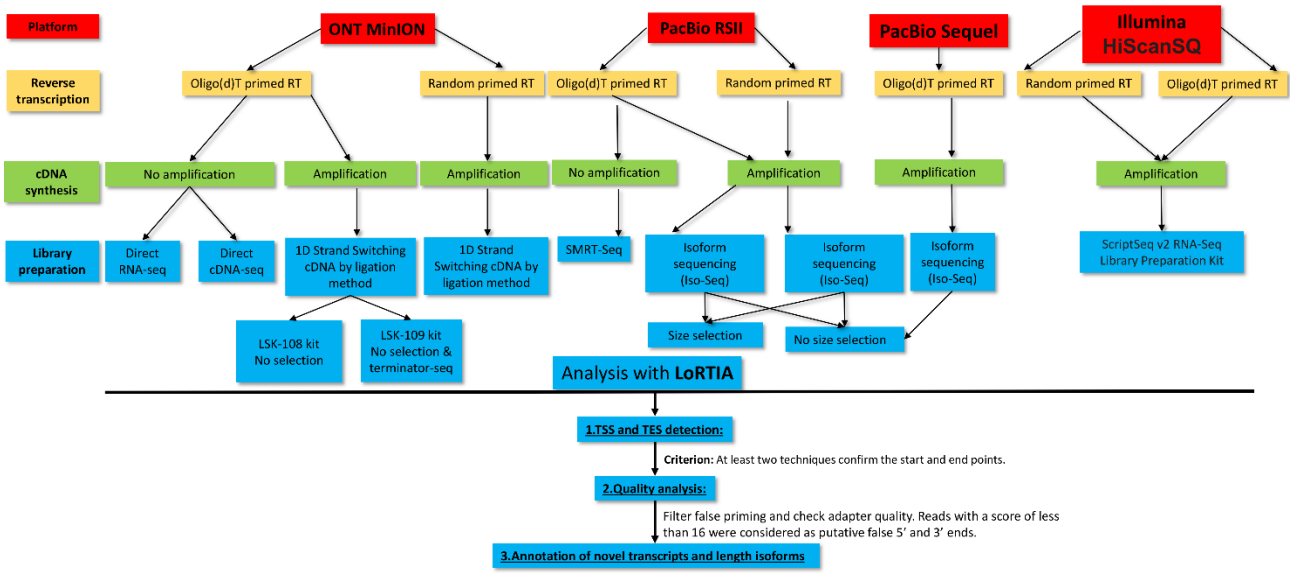

Figure 1. ***Workflow of the PacBio, MinION and Illumina sequencing.*** *The LoRTIA program identified the TSS and TES positions in ONT cDNA, dcDNA, terminator-seq, PacBio RSII random, IsoSeq and Sequel samples. LoRTIA software suit also helped in the validation of TESs and introns in dRNA-Seq samples. The Illumina data were used in the validation of certain low-abundance transcripts and splice sites, and in the identification of transcription readthroughs.*

Table 1. Sequencing reads and coverages obtained by using different techniques

Host read	Sample	Viral read	Mapped viral read length	Coverage
PacBio Sequel	117,079	13,292	1,553	143
PacBio RS II amplified	462,202	116,905	1,255	58
PacBio RS II non-amplified	176,919	52,012	1,282	76
PacBio-random primers	112,081	28,364	932	30

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 December 2020 doi:10.20944/preprints202012.0616.v1

MinION-amplified oligo(d)T	4,273,446	1,385,284	517	658
MinION-non-amplified oligo(d)T	4,907,412	3,451,129	909	1,118
MinION-random	5,144,609	231,500	341	178

Using the LoRTIA suit and applying the stringent criteria, we identified overall 465 TSSs and 57 TESs. Using these latter TSSs and TESs, LoRTIA annotated altogether 619 transcripts of which 410 are novel (**Figure 2** and **Table S2).** Eighty-two long transcripts were annotated manually, due to the uncertain TSSs of these transcripts.

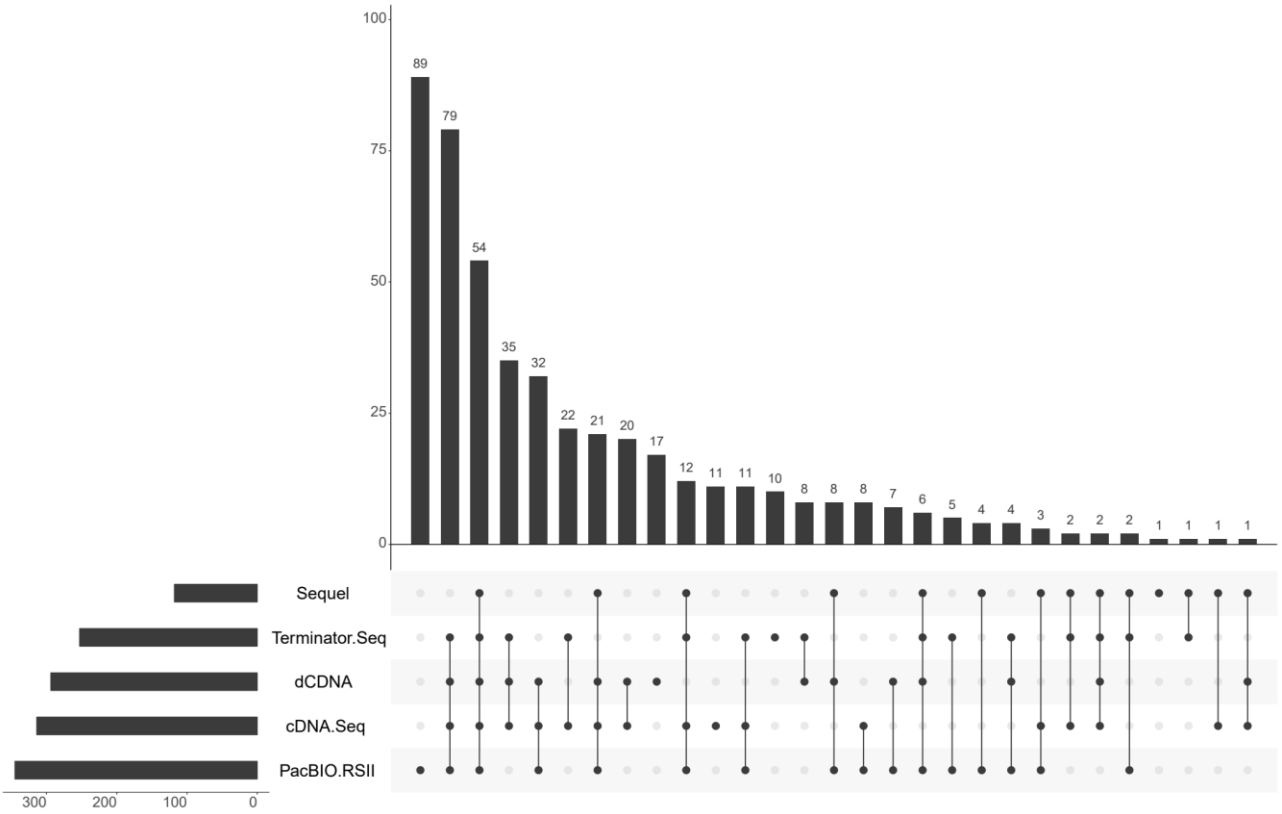

Figure 2. Transcript characteristics. *The upset plot indicates the quantitative distribution of the transcripts identified by using ONT and PacBio techniques.*

The cap-selection and the Terminator-based techniques in principle are able to eliminate transcripts without Caps. However, the 5'-end can be lost in the next steps of processing, therefore these techniques do not provide absolute guarantee for the elimination of transcripts with false TSSs. Nevertheless, they are useful for the enrichment of full-length RNA molecules.

Strain Kaplan of PRV (PRV-Ka) is a well characterized laboratory strain, which may have a reduced set of transcripts due to its multiple passages on cultured cells. Therefore, we also included strain MdBio of PRV (PRV-MdBio) [34], a recently characterized field isolate strain, into the analysis in order to be able to assemble a near complete map of PRV transcripts. PRV-MdBio was used for the dcDNA-Seq analysis. **Figure 3.** illustrate the updated PRV transcriptome, while **Figure S1** indicates the relative abundance of these transcripts.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 December 2020 doi:10.20944/preprints202012.0616.v1

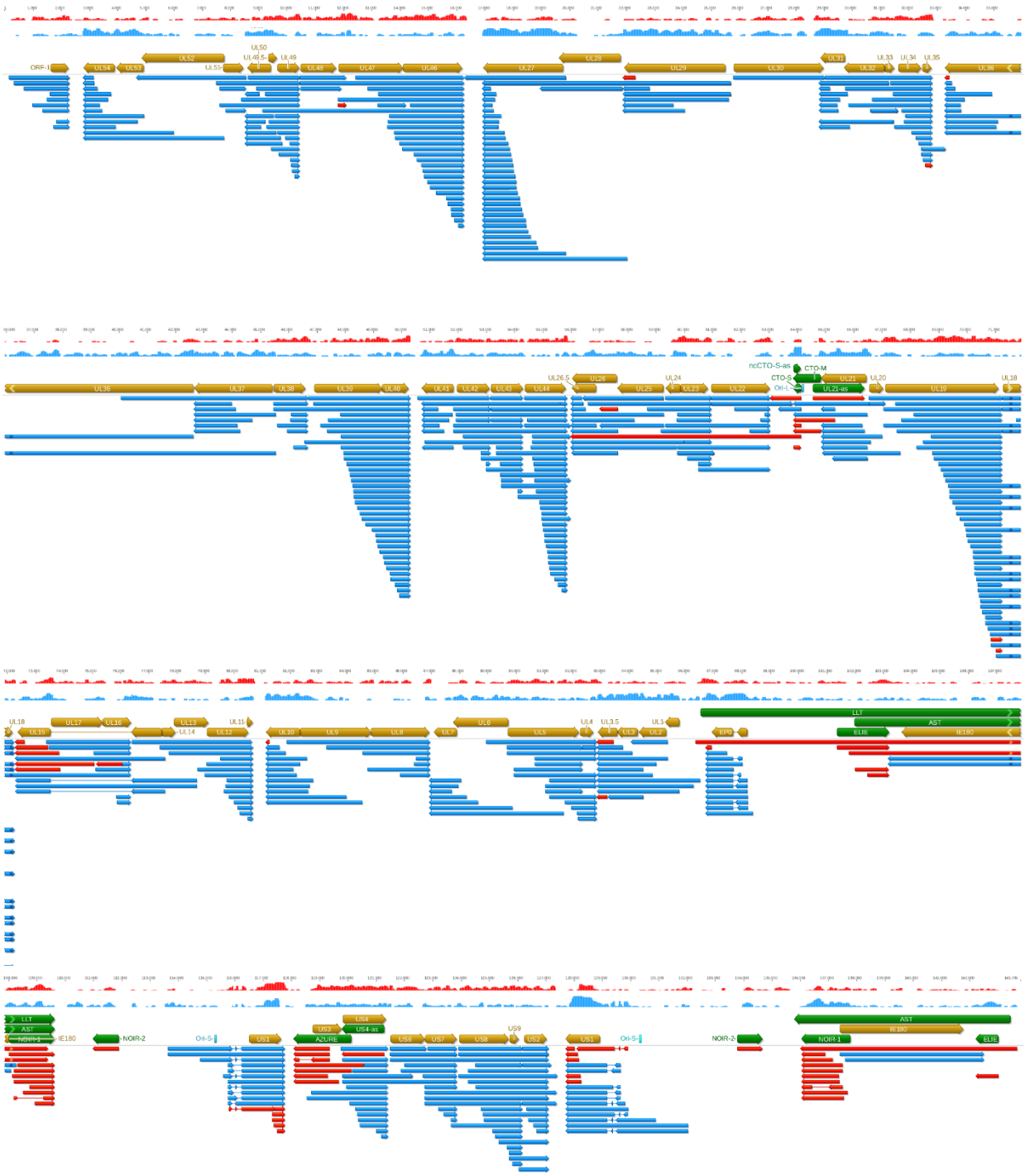

Figure 3. ***The updated PRV transcriptome.*** *PRV transcriptome contains those transcripts which were identified by the integrated approach using novel and earlier short-and long-read sequencing datasets. The light brown arrows represent the open reading frames of genes; the green arrows show the non-coding RNA genes; the blue arrows are the mRNAs, while the red arrows illustrate the non-coding transcripts. The putative embedded mRNAs are considered here as real mRNAs, but this has to be verified by proteomic studies. Complex transcripts are also colored by blue, although, it is unknown whether they really function as mRNAs, especially those ones of which the firs gene stands in opposite direction, that is non translated. LLT and AST transcripts were detected in latently infected neurons, and by real-time RT PCR in lytic infection, but not in our sequencing experiments.*

2.2. Novel putative embedded genes

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 December 2020 doi:10.20944/preprints202012.0616.v1

In this part of our work, we detected a total of 206 embedded transcripts (of which 189 have not been published before) lacking a large part of their 5′-region including the ATG of the canonical open reading frame (ORF) compared to the main mRNAs into which they are embedded. These embedded ORFs share their stop codons with the canonical ORFs but contain one or more in-frame ATGs downstream of the main ORFs. These shorter ORFs potentially encode N-terminally truncated polypeptides. 103 out of the 206 transcripts contain their own ORFs, whereas 103 are 5′-UTR (TSS) variants of the transcripts with the same ORFs. Shorter embedded transcripts without in-frame ORFs are considered to be non-coding. Our results demonstrated that intra-genomic transcription initiation appears to be extremely frequent in the PRV genome. Mono- and bicistronic RNA molecules share the same TSS in most cases, which indicates that these alternative transcription initiations are functional and not the result of mere transcriptional noise (**Figure 4, Table S3**).

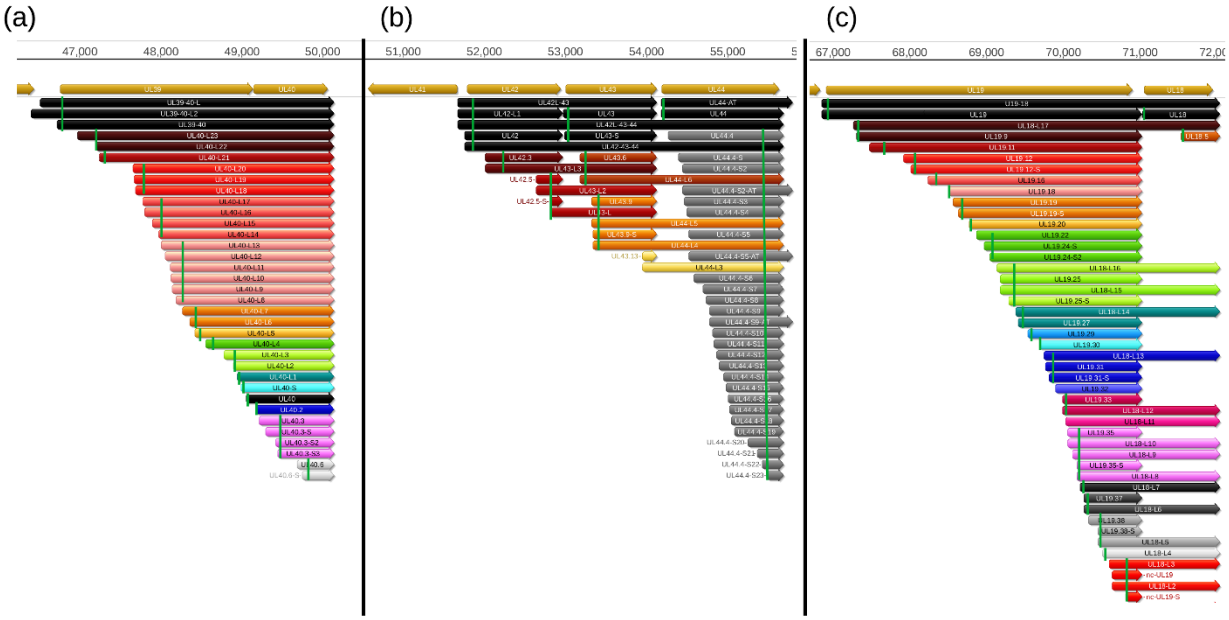

Figure 4. Embedded mRNAs. *The following genomic regions are selected for the illustration of the embedded mRNAs: (a) ul39-ul40, (b) ul42-43-44, (c) ul19-18. Arrows with the same color represent transcripts containing the same ORFs but distinct TSSs or TESs. The rectangular green lines indicate the fist in-frame ATGs within the transcripts. The "nc" letters at the end of the names indicate the lack of the stop codons.*

The stepped structure of the TSSs of multiple embedded RNA molecules can also be seen on the transcription reads (**Figure S2**), which further confirms that they are not artifacts generated by the annotation software. Furthermore, the TSSs of these transcripts were confirmed by dRNA sequencing in each case. Intriguingly, the genes with a substantial intragenic TSS variety contain a large number of in-frame ATGs and practically no ATGs can be found in the other two reading frames, which suggests that these embedded ORFs might indeed be functional.

Together, we identified a surprisingly large number of embedded RNAs, which if translated would significantly increase the complexity of PRV proteome.

2.3. Non-coding transcript

In this study, we annotated novel non-coding RNAs (ncRNAs), which all belong to the heterogeneous group of long non-coding RNAs (lncRNAs) (**Table S4, Figure 5**). TRL and TRS transcripts are ncRNAs at the unique

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 December 2020 doi:10.20944/preprints202012.0616.v1

long and unique short genomic region of PRV, respectively. Similarly to the embedded mRNAs, these non-coding transcripts are also co-terminal with the canonical mRNAs at their TESs but they lack a large segment at the 5'-terminal and do not contain in-frame ORFs due to the absence of ATG at this reading frame. In this work, we identified 13 novel TRLs and 7 novel TRSs. However, we could only confirm a few of the earlier published [31] putative non-coding low-abundance NCL transcripts, which are co-terminal with the mRNAs at their TSSs, but lack a certain part of their 3'-ends including the stop codons. Thus premature transcription termination might be resulted by transcriptional noise, which produces varying 3'-termini, therefore they do not meet the stringent criteria of transcripts with well-defined termini. In this study, we only annotated two such transcripts encoded within the *ul47* and *ul26* genes.

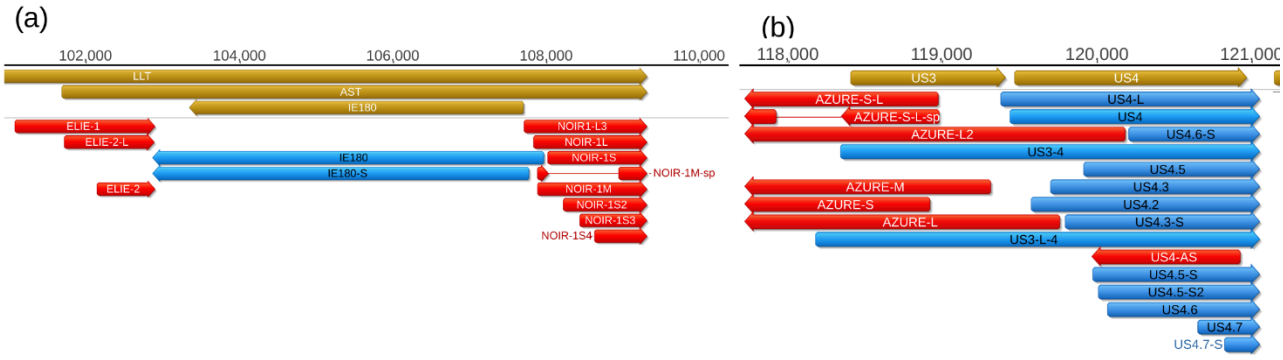

Figure 5. Coding and non-coding RNA molecules at the ie180-us4 genomic region. *A high density of non-coding transcription can be observed at this genomic region. Color code: light brown: coding, or non-coding genes, blue: mRNAs, red: ncRNAs.*

The *ul15* gene of alpha-herpesviruses has a unique structure in the sense that *ul16* and *ul17* genes are embraced within its ORF in an opposite polarity. We have earlier reported [26] that the downstream segment of *ul15* gene can also be separately transcribed and the resulting RNA molecule contains a novel out-of-frame ORF (termed fORF15). In this work, we detected novel isoforms of these transcripts.

Additionally, we also describe here novel NOIR-1variants. Transcripts of NOIR-1 family are co-terminal with the antisense transcript (AST) and long latency transcript (LLT) of PRV. The AZURE transcripts overlap the *us3* gene (and the long TSS variant partly the *us4* gene) in an antisense polarity. We also detected novel AZURE isoforms, including a spliced transcript. We also identified novel splice isoforms of the NOIR-1 transcript family.

Antisense RNAs (asRNAs) were detected by both SRS and LRS in PRV. These transcripts are produced either from distinct promoters, such as the LLT, AST and AZURE transcripts (**Figure 5**), or they can also be the product of transcriptional read-thorough between convergent genes, or transcriptional overlaps between divergently oriented genes. Additionally, we re-annotated the exact TSS position of US4-as antisense transcript.

The AST and LLT transcripts are included in our transcript list, however – as they are expressed in latency – we did not detect these transcripts in out lytic infection experiment. These transcripts are likely to be expressed in a very low abundance in lytic infection, because we could detect the using real-time RT-PCR [22]. Nonetheless, we could detect the NOIR-1 transcript family, which shares the TES with the latency transcript.

2.4. Replication-origin associated transcripts

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 December 2020 doi:10.20944/preprints202012.0616.v1

Replication origin-associated RNAs (raRNAs) have been described in several viruses, including herpesviruses [35]. First the CTO family of PRV raRNAs have been described [30], which was followed by the discovery of the PTO transcripts [26,31] (**Figure 6**). In this study, we report the identification of 5 additional transcripts at this genomic region, including a novel TES isoform of COT-S (CTO-S-AT2) and a TSS variant of CTO-M (CTO-M/L), a very long complex RNA molecule (CTO-S-cx) and two other transcripts running with opposite polarity with respect of CTOs (CTO-as and UL21-as). While the CTO-S is among the most abundant if not the most abundant PRV transcript, the novel RNA molecules are expressed in a lower copy numbers. We also detected a splice variant of PTO-US1.

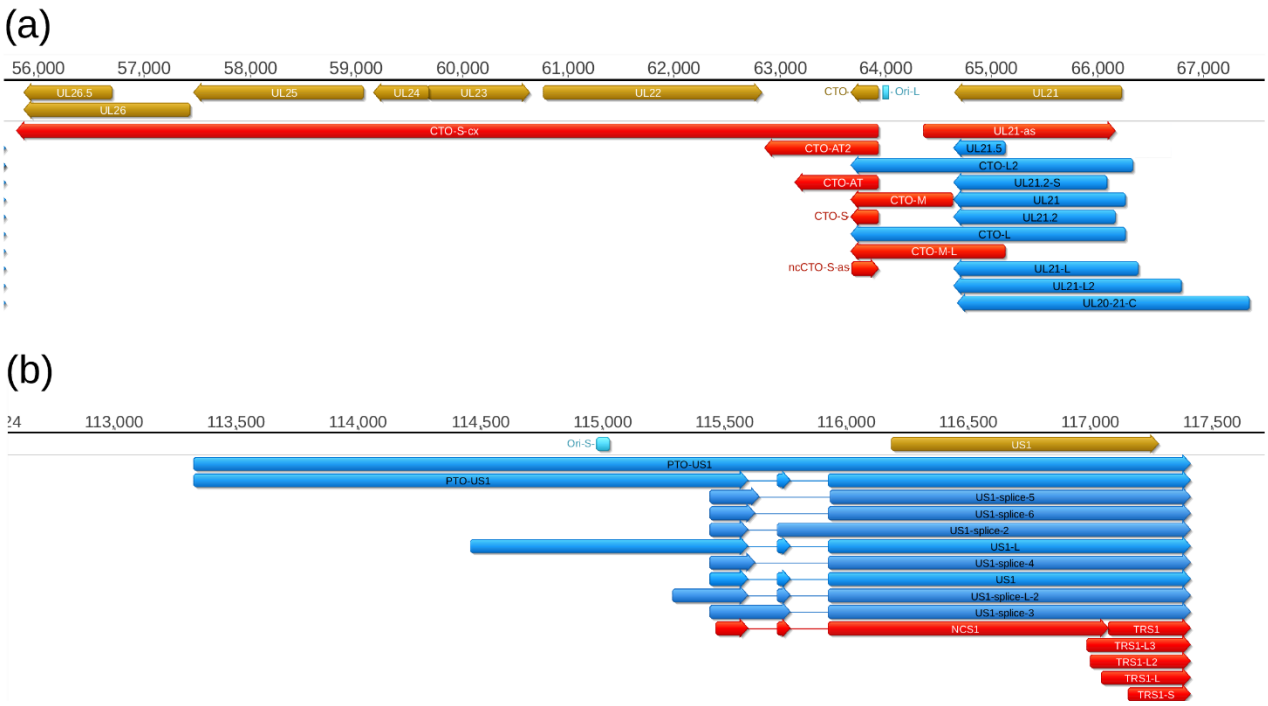

Figure 6. Replication origin-associated transcripts at the (a) Ori-L (a) and (b) the Ori-S genomic regions. *These RNA molecules have been described in several viruses, including herpesviruses. Except the CTO-S family, these transcripts overlap the replication origins through either their 3′-UTRs (CTO-L) or their 5′-UTR (PTO-US1). Both the raRNAs and the transcript of adjacent genes are overlapped by antisense RNAs of which some are controlled by separate promoters. Color code: light brown: coding, or non-coding genes, blue: mRNAs, red: ncRNAs*

2.5. TSS and TES isoforms

In this study, an even higher diversity of TSSs and TESs are described compare to the earlier reports [31]. Altogether, we identified 24 shorter (not including the embedded mRNAs) and 166 longer novel TSS isoforms, as well as 22 novel TES variants (**Table S5**). Due to the stringent criteria used for the annotations, the number of short TSS isoforms are likely higher what we publish here. Transcripts with longer 5′-UTR sequences than the canonical transcript are termed by adding an 'L' letter to the end of the name, whereas the shorter variants are designated by adding an 'S' letter to the original names. Transcripts with longer 3′-UTRs are designated by adding 'AT' letters to the original names. The base content at the 5′- and 3′-termini of the PRV transcripts is illustrated in **Figure 7**. The GGG sequence was the most common triplet at the 5′-end, whereas AU-rich sequences are found upstream and GU-rich sequences are common downstream of the poly(A) signal.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 December 2020 doi:10.20944/preprints202012.0616.v1

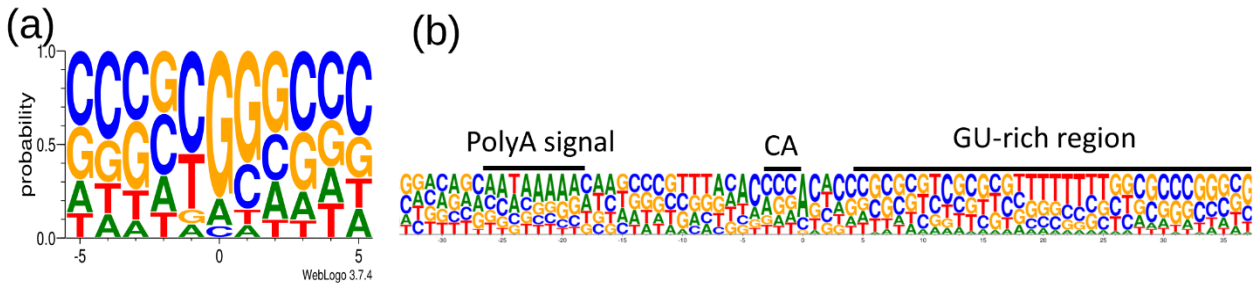

Figure 7. *The base content at the 5′- and 3′-termini of the PRV transcripts.* The image is generated using weblogo 3.0

2.6. Novel splice isoforms

Splicing in alphaherpesvirus transcripts is much less common than in other subfamilies of herpesviruses. A few PRV transcript has been described previously [31]. In this work, we identified altogether 9 novel introns (**Table 2**) in the transcripts EP0, US1, AZURE and NOIR-1, PTO-US1 (**Table S6**). Introns were only accepted if their splice donor and acceptor sequences contained the canonical GT/AT or GC/AG sequences. Each splice site was validated by dRNA sequencing.

Table 2. Splice junction sites of the PRV transcriptome.

Transcript name	TSS	TES	Exon 1 position	Exon 2 position	Strand
EP0-sp2	96421	97899	96421-97389	97480-97875	-
NOIR-1M-splice	107880	109304	107880-10801	108932-109304	+
US1-sp-2	115437	117407	115437-115591	115713-117407	+
US1-sp-3	115437	117407	115437-115638	115931-117407	+
US1-sp-4	115437	117407	115437-115621	115922-117407	+
US1-sp-5	115437	117407	115437-115632	115922-117407	+
US1-sp-6	115437	117407	115437-115765	115922-117407	+
AZURE-S-L-sp	117716	118967	117716-117917	118340-118967	-
PTO-US1-sp	113321	117407	115713-115765	115922-117407	+

2.7. Novel multigenic transcripts

In this work, we identified 87 novel polycistronic transcripts of which 59 are bicistronic, 19 are tricistronic, 8 are tetracistronic, and 1 is pentacistronic (**Table S7)**. The pentacistronic transcript (UL49.5-49-48-47-46) is the longest among them (7,130 bps long). We also detected 24 novel complex transcripts, in which at least one of the genes stands in an opposite direction relative the other genes. The longest complex transcript is the CTO-S-cx (8,135 bps), that comprises the CTO-S, *ul22*, *ul23*, *ul24*, *ul25*, and *ul26* genes of which *ul22* stands in opposite polarity relative to the others. We could not annotate the TSSs in 9 out of the 24 complex transcripts precisely, because only a few reads were obtained and because they were detected by dRNA sequencing that generates transcription reads lacking 15-20 base pairs from the 5′-ends. Although, both PacBio and ONT platforms are able to read extremely long DNA stretches, it is difficult to obtain very long RNA and cDNA reads for which the reason could be the fragility of RNA molecules, the imperfectness of RT reaction, and the preference of PCR producing for short amplicons.

2.8. Transcriptional overlaps

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 December 2020 doi:10.20944/preprints202012.0616.v1

This study revealed an even higher complexity of transcriptional overlaps than in earlier reports [31]. Transcripts can overlap with each other in a parallel, divergent or convergent manners. Parallel and convergent overlaps are produced by transcriptional readthroughs from tandem and convergent genes, whereas divergent overlap are formed by the overlap of 5'-termini of RNA molecules.

Tandem genes of herpesviruses are organized into gene clusters producing co-terminal transcripts as such: 'abcd', 'bcd', 'cd' and 'd' of which, according to our current knowledge, only the most upstream genes are translated. The phenomenon of parallel overlaps has been long well-known. The high amount of embedded RNAs reported in this study makes more complex of the overlapping transcription patterns. Our earlier studies revealed [26,31] that divergent genes also exhibit a great extent of transcriptional overlaps especially through the very long alternative 5'-UTRs. This study revealed an even higher intricacy of this type of overlaps. Convergent genes are in most cases separated by relative long intergenic regions composed mainly of repetitive sequences (the only exceptions are the ul7/8, ul30/31 and ul50/51 gene pairs of which TESs are located within each other). In this systematic analysis using both SRS and LRS data, we demonstrated that occasional transcriptional readthroughs occurs in every convergent gene pairs thereby producing antisense segments in the resulted read-through RNA molecules. An intriguing phenomenon was also observed, namely that some genes produce long TES isoform transcripts of which 3'-UTR sequences span the intergenic region but are terminated at exactly the TESs of the partner convergent genes (in UL27-AT, UL35-AT, UL44-AT, CTO-S-AT, US2-AT, **Figure 3**). The existence of complex transcripts may indicate an interaction between distal viral genes at the level of transcription.

3. Discussion

The advantage of LRS over the SRS in transcriptomics is that it can produce full-length transcripts and therefore is more valuable for the assembly of viral transcriptome than SRS, but this latter technique generates higher data coverage. Furthermore, the distinct LRS methods have different limitations and strengths. Therefore, the combination of the various approaches is advantageous for addressing the complexity of transcriptome architectures. We have developed a pipeline for the analysis of long-read RNA sequencing data. The LoRTIA software suit proved to be useful for the identification of transcript and transcript isoforms and for the exclusion of potential erroneous signals that may arise as a result of RNA degradation or during RT, PCR and other processes.

This study delineated that the PRV transcriptome is more complex than previously anticipated. Here, we identified an unexpectedly large number of potential novel genes, which are embedded into larger host genes, and encode 5'-truncated RNA molecules many of them containing a large number of in-frame ATGs (but practically no out-of-frame ATGs in any of the other two reading frames), which may allow the translation of the truncated transcripts in a large diversity. The N-terminal truncated version of the putative proteins may have an altered effector function [36], or may play an until now uncovered role in the herpesvirus pathogenicity. The smallest embedded RNAs do not contain in-frame ATGs therefore they are considered to be non-coding RNAs. The large number of embedded transcripts suggest a novel aspect of genomic organization, which may be not restricted to the herpesviruses, but represent a general phenomenon. Detection of these putative overlapping polypeptides would basically redefine the herpesvirus proteome. Functional clarification of these transcripts and polypeptides would go far beyond the significance of this phenomenon in herpesviruses.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 December 2020 doi:10.20944/preprints202012.0616.v1

Intriguingly, the members of ELIE and NOIR-1 transcript families overlap the 5′- and 3′-ends of LLT/AST transcripts, respectively, but do not overlap the IE180 transcript. This form of organization suggests the ELIE and NOIR-1 transcripts do somehow interfere with transcription of the *llt* and *ast* genes, but not with the *ie180* gene during lytic infection and thereby suppressing the transcription of the two non-coding latency genes, which are normally expressed in latency.

Earlier we have described a number of raRNAs in a variety of viruses [35]. The most complex pattern of these transcripts was described in PRV [26,30,31]. PRV raRNA molecules include ncRNAs, longer TSS or TES variants of protein-coding genes and complex transcripts. One of the proposed functions of these transcripts is to provide a replication-transcription interference mechanism that controls the initiation and the orientation of DNA replication [37]. However, since these transcripts contain poly(A) sequences, we assume that they are not a mere by-products of the above putative mechanism but also play until now unascertained roles as RNA molecules. It would be especially intriguing to study the potential effect of the novel ncCTO-S-as on the very abundant CTO-s. Since ntCTO-S-as contains poly(A) sequences, it is likely functional as an RNA molecule, and not a mere result of the operation of a transcription interference-based mechanism.

In this and previous studies, we demonstrated the existence of a vast diversity of transcription initiation from practically each herpesvirus gene. The significance of this phenomenon is currently unknown. One possibility of using multiple promoters is the differential transcription regulation of gene expression throughout the viral life cycle. Another, not necessarily exclusive explanation for the alternative TSS usage is that only the longer 5′-UTR variants contain upstream ORFs (uORFs), but not the shorter ones thereby providing additional coding potentials. The uORFs has been shown to play a role in the control of translation in multiple ways [38]. We have described such isoform variation first in human cytomegalovirus before [39]. Additionally, long 5′-UTRs overlap with the adjacent or even more distal genes which may lead to transcriptional interference between gene expressions as suggested before [37]. The potential function of alternative TES usage is also unknown. This phenomenon might provide an alternative regulation of transcription and/or translation.

Polycistronism is typical in prokaryotic genes but not in eukaryotes. The genetic organization of eukaryotic viruses, including herpesviruses, resemble to that of bacteria in that they also express polycistronic mRNA molecules. However, polycistronism serve different functions in the two groups of organisms since at this moment we do not have evidence for the translation of the downstream genes on a polycistronic RNA molecule in the viruses. One of the functions of this transcription system may be the (down)regulation of transcription of downstream genes by the upstream genes through interference of transcription machineries [37]. It is unknown whether complex transcripts are translated or instead they may function as a lncRNAs, or may be both.

Application of LRS techniques for transcriptome profiling revealed that the viral genomes are transcribed well beyond of gene boundaries thereby generating an intricate meshwork of transcriptional overlaps. The functionality of this phenomenon was initially regarded with skepticism. However, mounting evidence suggests that overlapping transcription fulfills regulatory purposes and provides novel strategies for the coordination of gene expression [40].

4. Materials and Methods

4.1. Cells and viruses

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 December 2020 doi:10.20944/preprints202012.0616.v1

PK-15 porcine kidney epithelial cell line (ATCC® CCL-33™) was used for the propagation of strains Kaplan (PRV-Ka) and MdBio (PRV-MdBio) of pseudorabies virus. Cells were cultivated in DMEM (Gibco/Thermo Fisher Scientific), supplemented with 5% fetal bovine serum (Gibco/Thermo Fisher Scientific) and 80 μg of gentamycin per ml (Gibco/Thermo Fisher Scientific) at 37°C in the presence of 5% CO_2. For the preparation of virus stock solution, cells were infected with 0.1 multiplicity of infection [MOI = plaque-forming units (pfu)/cell]. Viral infection was allowed to progress until complete cytopathic effect was observed. It was followed by three successive cycles of freezing and thawing of infected cells in order to release of viruses from the cells. In all of the previous experiments of which data were used here, PRV-Ka was grown in PK-15 cells using the same cultivation conditions.

4.2. RNA isolation

For the extraction of total RNAs, the NucleoSpin® RNA II kit and NucleoSpin® RNA kit (both from Macherey-Nagel) were used for SRS and LRS sequencing, respectively. In short, cells were collected by centrifugation, then the lysis was carried out by incubation in a chaotropic ion containing solution, which inactivates the RNase enzyme. DNA and RNA molecules bind to the silica membrane. Samples were handled with DNase I solution (supplied in the kit) to eliminate DNA. Total RNAs were eluted from the membrane in RNase-free water. To remove the potential remaining DNA contamination, samples were handled with Ambion® TURBO DNA-free™ Kit. RNA samples were kept at −80 °C until use. The polyA(+) fraction of the total RNAs were purified using the Oligotex mRNA Mini Kit (Qiagen; "Spin Columns" protocol). Half of the sample was handled with Terminator™ 5′-Phosphate-Dependent Exonuclease (Lucigen), which digests RNA with 5′-monophosphate ends but not RNAs with 5′-triphosphate, 5′-cap or 5′-hydroxyl groups starting from the 5′ end, therefore it helps to characterize 5′-ends of RNAs.

4.3. Pacific Biosciences Isoform sequencing using the Sequel system

4.3.1. Synthesis of cDNAs

The cDNAs were generated from the polyA(+) RNA samples using the Clontech SMARTer PCR cDNA Synthesis Kit according to the PacBio Isoform Sequencing (Iso-Seq) protocol without size selection. The first-strand cDNAs were generated with oligo(dT) primers (part of the Clontech Kit). These samples were amplified using KAPA HiFi Enzyme (Kapa Biosystems), according to the PacBio's recommendations (details in our earlier publication: [41].

4.3.2. SMRTbell template preparation

About 500 ng from the cDNA sample was used to prepare the SMRTbell library using the PacBio DNA Template Prep Kit 1.0 according to the Pacific Biosciences' 2 kb Template Preparation and Sequencing protocol, according to our earlier publication [41]. Briefly, the cDNA ends were repaired then the adapters were ligated to the samples. Finally, the exonuclease treatment was carried out in order to remove the incorrect SMRTbell templates (e.g. with free ends that did not receive an adapter, or contain nicks or other damage) from the library leaving only intact SMRTbell templates. AMPure® PB bead purification steps were performed after each of the enzymatic steps. The SMRTbell library was bound to the P6 DNA polymerase (PacBio) and annealed to v2 primers (PacBio), then this library-polymerase complex was bound to MagBeads with MagBead Binding Kit (PacBio). The total amount of the MagBead-bound complex was loaded onto the SMRT Cell. The MagBead One Cell Per Well protocol was used. One SMRT Cell was run on the Sequel instrument.

4.4. Oxford Nanopore Technologies Nanopore sequencing using the MinION device

4.4.1. Direct RNA sequencing

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 December 2020 doi:10.20944/preprints202012.0616.v1

The Direct RNA sequencing (SQK-RNA002) protocol (Version: DRS_9080_v2_revM_14Aug2019) was used to obtain amplification-free transcriptomic data to remove RT and PCR biases, as well as to explore attributes of native RNA such as modified bases. Five-hundred ng of polyA(+)-tailed RNA was used. The library preparation was carried out according our previous publication [41] with the following modification: Agencourt RNAClean XP beads (Beckman Coulter) was used instead of the RNase OUT (Invitrogen)-treated Agencourt XP beads (Beckman Coulter).

4.4.2. Direct cDNA sequencing

Non-amplified cDNA libraries were prepared from the poly(A)+ fraction of RNAs from the MDBIO strain using the ONT's Direct cDNA Sequencing Kit (SQK-DCS109; DCS_9090_v109_revJ_14Aug2019) according to the manufacturer's protocol. In brief, the Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific) with SSP and VN primers (supplied in the kit) were used for the synthesis of first cDNA strand from 100 ng of poly(A)+ RNA. Next, the potential RNA contamination was eliminated using RNase Cocktail Enzyme Mix (Thermo Fisher Scientific). This step was followed by the second strand synthesis using LongAmp Taq Master Mix (New England Biolabs). Double-stranded cDNA ends were repaired using NEBNext End repair /dA-tailing Module (New England Biolabs) then the sequencing adapter ligation was carried out with the NEB Blunt /TA Ligase Master Mix (New England Biolabs).

4.4.3. Amplified cDNA sequencing

- ONT's ligation-based sequencing protocol (SQK-LSK109; Version: GDE_9063_v109_revU_14Aug2019)

The ONT's LSK109 protocol was used for sequencing the polyA-selected oligo(dT)-primed, rRNA-depleted random-primed or Terminator™-handled oligo(dT)-primed samples. The generation of cDNA was conducted according to our previous publications [31,41] using oligo(dT) or random primers. The DNA repair was carried out according to the SQK-LSK109 protocol. Briefly, the NEBNext FFPE DNA Repair Mix and NEBNext Ultra II End repair / dA-tailing Module reagents (all from New England Biolabs) were mixed with the samples, then the mixtures were incubated at 20°C for 5min and at 65°C for 5min. This step was followed by the adapter ligation: the NEBNext Quick T4 DNA Ligase (New England Biolabs), the Ligation Buffer and Adapter Mix (both from ONT's Kit) were mixed with the cDNA samples and incubated for 10min at room temperature. Samples were purified using the AMPure XP magnetic beads (Beckman Coulter) after each enzymatic step.

- 1D Strand switching cDNA by ligation method (Version: SSE_9011_v108_revS_18Oct2016) and the ONT Ligation Sequencing Kit 1D (SQK-LSK108)

This protocol was used to analyze the random primed cDNA libraries. In short, ribodepleted RNA fraction was used to generate cDNA samples, first it was mixed with dNTPs (10 mM, Thermo Scientific) and random primers (ordered from IDT DNA) and then the mixtures were incubated at 65 °C for 5min. After this step, the DTT and buffer form the SuperScipt IV Reverse Transcriptase kit (Life Technologies), RNase OUT enzyme (Life Technologies) and strand-switching oligo with three O-methyl-guanine RNA bases (PCR_Sw_mod_3G; Bio Basic, Canada) were added and the mixtures were heated to 42 °C for 2min. SuperScript IV Reverse Transcriptase enzyme (200 unit) was mixed into the samples. The generation of the first cDNA strand was conducted at 50 °C for 10min, then the strand switching step was carried out at 42 °C for 10min. For the inactivation of the enzymes the samples were heated to 80 °C for 10min. Samples were amplified using the KAPA HiFi DNA Polymerase (Kapa Biosystems) and Ligation Sequencing Kit Primer Mix (provided by the 1D Kit). The NEBNext End repair / dA-tailing Module (New England Biolabs) was applied to repair cDNA ends, while NEB Blunt/TA Ligase Master Mix (New England Biolabs) was used to ligate the adapters (supplied by the kit).

4.5. Previous sequencing techniques of which data are used in this study

The methods of these approaches have been described in our earlier publications (**Table 3**).

Table 3. List of sequencing techniques used in our previous studies.

Technique	Publication

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 December 2020 doi:10.20944/preprints202012.0616.v1

Illumina HiScan	Oláh et al., 2015 BMC Microbiology [24]
Pacific Biosciences RSII	Tombácz et al., 2016 Plos One [26]
Oxford Nanopore Technologies	Moldován et al., 2018 Frontiers in Microbiology [31]
All: data report, detailed protocols	Tombácz et al., 2018 Scientific Data [41]

4.6. Measurement of nucleic acid quality and quantity

4.6.1. RNA

The Qubit RNA BR Assay Kit (Invitrogen) was used for the total RNA measurement, while the Qubit RNA HS Assay Kit (Invitrogen) was applied to check the quantity of the poly(A)+ and rRNA-depleted RNA fractions. The final concentrations of the RNA samples were determined by Qubit® 4.

4.6.2. cDNA

The concentrations of the cDNA samples and sequencing ready libraries were measured by using the qubit dsDNA HS Assay Kit (Invitrogen).

The RNA quality was assessed with the Agilent 2100 Bioanalyzer (for PacBio sequencing) or Agilent 4150 TapeStation System (for MinION sequencing) and RIN scores above 9.6 were used for cDNA production.

4.7. Pre-processing and data analysis

The processing of the MinION raw data was conducted with the Guppy basecaller v. 3.6.1. with --qscore_filtering. Reads with a Q-score greater than 7 were aligned to the viral genome (NCBI nucleotide accession: KJ717942.1 [42] using the Minimap2 mapper [43]. The PacBio dataset was also mapped with Minimap2 to the same reference genome. Transcription reads were visualized using Genious 11.1.5 software (https://www.geneious.com). The upset plot was visualized by the UpSetR program [44].

The LoRTIA (https://github.com/zsolt-balazs/LoRTIA) software package (v.0.9.9) was used for the detection and annotation of transcripts and transcript isoforms, as was described earlier [28] (**Table 4**). Briefly, the sequencing adapters and the homopolymer A sequences were checked by the LoRTIA toolkit for the identification of TSS and TES, respectively. To eliminate random TSS and TES (which can be caused by RNA degradation), the putative start and end sites were tested against the Poison distribution (with Bonferroni correction). Putative introns were accepted with the following criterions: (1) have one of the three most frequent splice consensus sequence (GT/AG, GC/AG, AT/AC); (2) are more abundant than 1‰ compared to the local coverage.

Table 4. Settings of the LoRTIA software suite for each sample type

Sample	5′ adapter	5′ min score	3′ adapter	3′ min score
PacBio	AGAGTACATGGG	16	AAAAAAAAAAAAAAA	18
MinION cap	default	default	default	default
MinION non-cap	TGCCATTAGGCCGGG	14	AAAAAAAAAAAAAAA	16
MinION dcDNA	GCTGATATTGCTGGG	16	AAAAAAAAAAAAAAA	16

The accepted putative TSSs and TESs were considered as existing if they were detected by at least two different techniques. Potential introns were accepted as real introns if they were present in both dRNA-Seq and at least one of the cDNA-Seq datasets and if they were shorter than 10 kbps. We set a relatively low abundance for

 doi:10.20944/preprints202012.0616.v1

acceptance because it may vary in different cell types. The accepted TSSs, TESs and introns were then assembled into putative transcripts using the Transcript_Annotator software of the LoRTIA suite. Very long unique or low-abundance reads which could not be detected using LoRTIA were annotated manually. These reads were also accepted as putative transcript isoforms if they were longer than any other overlapping RNA molecule. In some cases, the exact TSSs were not annotated. Finally, a read was considered as transcript if it was present in at least three separate samples. Transcript annotation was followed by isoform categorization according to the following principles: the most abundant transcript containing a single ORF was termed canonical monocistronic transcript, whereas isoforms with longer or shorter 5′-UTRs or 3′-UTRs regions than the canonical transcripts were termed TSS or TES isoforms (variants), respectively. Similarly, transcripts with alternative splicing were named splice isoforms. Transcripts with 5′-truncated in-frame ORF were termed as putative mRNAs. Transcripts with multiple non-overlapping ORFs were designated polycistronic, whereas those with ORFs in different orientation were called complex transcripts. Transcripts with no ORFs or ORFs shorter than 30 nts were named non-coding, except if occurred in front of a canonical ORF in a transcript (These small ORFs were termed uORFs).

4.8. Accession codes

The LoRTIA software suite is available on GitHub: https://github.com/zsolt-balazs/LoRTIA

Supplementary Materials

Figure S1.: Sequencing reads are organized in a stepped fashion indicating the separated TSSs of embedded transcripts
Figure S2.: Antisense transcripts near the replication OriL region
Table S1. Sequencing reads and coverages obtained in different techniques.
Table S2. List of PRV transcripts.
Table S3. List of putative embedded genes.
Table S4. List of non-coding transcripts.
Table S5. List of TSS and TES isoforms.
Table S6. Novel Splice isoforms.
Table S7. Polygenic transcripts.

Author Contributions: Conceptualization, Z.B.; methodology, G.T., D.T. and Z.B.; formal analysis, G.T., D.T., D.G., and Z.B.; investigation, G.T., D.T., Z.C., Z.D. and Z.B.; resources, M.S., D.T. and Z.B.; data curation, G.T.; writing—original draft preparation, D.T. and Z.B; writing—review and editing, Z.B.; visualization, G.T. and D.T.; supervision, M.S. and Z.B.; funding acquisition, D.T. and Z.B. All authors have read and agreed to the published version of the manuscript."

Funding: This study was supported by OTKA K 128247 to ZB, OTKA FK 128252 to DT, University of Szeged Open Access Fund 4813. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

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

Ethics declaration: Neither human nor animal experiments were applied in this study.

References

1. Engel, E.A.; Song, R.; Koyuncu, O.O.; Enquist, L.W. Investigating the biology of alpha herpesviruses with MS-based proteomics. *Proteomics* 2015, *15*, 1943–1956.

2. Hogue, I.B.; Scherer, J.; Enquist, L.W. Exocytosis of alphaherpesvirus virions, light particles, and glycoproteins uses constitutive secretory mechanisms. *MBio* **2016**, *7*, doi:10.1128/mBio.00820-16.

3. Aujeszky, A. A contagious disease, not readily distinguishable from rabies, with unknown origin. *Veterinarius*

1902.

4. Pomeranz, L.E.; Reynolds, A.E.; Hengartner, C.J. Molecular biology of pseudorabies virus: impact on neurovirology and veterinary medicine. *Microbiol. Mol. Biol. Rev.* **2005**, *69*, 462–500, doi:10.1128/MMBR.69.3.462-500.2005.

5. Wang, H.H.; Liu, J.; Li, L.T.; Chen, H.C.; Zhang, W.P.; Liu, Z.F. Typical gene expression profile of pseudorabies virus reactivation from latency in swine trigeminal ganglion. *J. Neurovirol.* **2020**, *26*, 687–695, doi:10.1007/s13365-020-00866-9.

6. Sandri-Goldin, R.M. *Alpha Herpesviruses: Molecular and Cellular Biology*; Caister Academic Press, 2006; ISBN 978-1-904455-09-7.

7. Mettenleiter, T.C.; Sobrino, F. *Animal Viruses: Molecular Biology*; Caister Academic Press, 2008; ISBN 978-1-904455-22-6.

8. Feng, Z.; Chen, J.; Liang, W.; Chen, W.; Li, Z.; Chen, Q.; Cai, S. The recombinant pseudorabies virus expressing African swine fever virus CD2v protein is safe and effective in mice. *Virol. J.* **2020**, *17*, doi:10.1186/s12985-020-01450-7.

9. Jasmin, L. Pseudorabies virus as a neuroanatomical tracer. *J. Neurovirol.* 1995, *1*, 326–327.

10. Koyuncu, O.O.; Perlman, D.H.; Enquist, L.W. Efficient retrograde transport of pseudorabies virus within neurons requires local protein synthesis in axons. *Cell Host Microbe* **2013**, *13*, 54–66, doi:10.1016/j.chom.2012.10.021.

11. Card, J.P.; Enquist, L.W. Transneuronal circuit analysis with pseudorabies viruses. *Curr. Protoc. Neurosci.* **2014**, *68*, doi:10.1002/0471142301.ns0105s68.

12. Boldogköi, Z.; Reichart, A.; Tóth, I.E.; Sik, A.; Erdélyi, F.; Medveczky, I.; Llorens-Cortes, C.; Palkovits, M.; Lenkei, Z. Construction of recombinant pseudorabies viruses optimized for labeling and neurochemical characterization of neural circuitry. *Mol. Brain Res.* **2002**, *109*, 105–118, doi:10.1016/S0169-328X(02)00546-6.

13. Kratchmarov, R.; Kramer, T.; Greco, T.M.; Taylor, M.P.; Ch'ng, T.H.; Cristea, I.M.; Enquist, L.W. Glycoproteins gE and gI Are Required for Efficient KIF1A-Dependent Anterograde Axonal Transport of Alphaherpesvirus Particles in Neurons. *J. Virol.* **2013**, *87*, 9431–9440, doi:10.1128/jvi.01317-13.

14. Viney, T.J.; Balint, K.; Hillier, D.; Siegert, S.; Boldogkoi, Z.; Enquist, L.W.; Meister, M.; Cepko, C.L.; Roska, B. Local Retinal Circuits of Melanopsin-Containing Ganglion Cells Identified by Transsynaptic Viral Tracing. *Curr. Biol.* **2007**, *17*, 981–988, doi:10.1016/j.cub.2007.04.058.

15. Szabó, E.; Csáki, A.; Boldogkoi, Z.; Tóth, Z.; Köves, K. Identification of autonomic neuronal chains innervating gingiva and lip. *Auton. Neurosci. Basic Clin.* **2015**, *190*, 10–19, doi:10.1016/j.autneu.2015.03.005.

16. Boldogköi, Z.; Sík, A.; Dénes, Á.; Reichart, A.; Toldi, J.; Gerendai, I.; Kovács, K.J.; Palkovits, M. Novel tracing paradigms - Genetically engineered herpesviruses as tools for mapping functional circuits within the CNS: Present status and future prospects. *Prog. Neurobiol.* 2004, *72*, 417–445.

17. Boldogkoi, Z.; Balint, K.; Awatramani, G.B.; Balya, D.; Busskamp, V.; Viney, T.J.; Lagali, P.S.; Duebel, J.; Pásti, E.; Tombácz, D.; et al. Genetically timed, activity-sensor and rainbow transsynaptic viral tools. *Nat. Methods* **2009**, *6*, 127–130, doi:10.1038/nmeth.1292.

18. Prorok, J.; Kovács, P.P.; Kristf, A.A.; Nagy, N.; Tombácz, D.; Tth, J.S.; Ördg, B.; Jost, N.; Virág, L.; Papp, J.G.; et al. Herpesvirus-mediated delivery of a genetically encoded fluorescent ca 2+ sensor to canine cardiomyocytes. *J.*

Biomed. Biotechnol. **2009**, *2009*, doi:10.1155/2009/361795.

19. Boldogkö, Z.; Szabó, A.; Vrbová, G.; Nógrádi, A. Pseudorabies virus-based gene delivery to rat embryonic spinal cord grafts. *Hum. Gene Ther.* **2002**, *13*, 719–729, doi:10.1089/104303402317322285.

20. Boldogkoi, Z.; Bratincsak, A.; Fodor, I. Evaluation of pseudorabies virus as a gene transfer vector and an oncolytic agent for human tumor cells. *Anticancer Res.* **2002**, *22*, 2153–2159.

21. Cheung, A.K. DNA nucleotide sequence analysis of the immediate-early gene of pseudorabies virus. *Nucleic Acids Res.* **1989**, *17*, 4637–4646, doi:10.1093/nar/17.12.4637.

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

23. Harkness, J.M.; Kader, M.; DeLuca, N.A. Transcription of the Herpes Simplex Virus 1 Genome during Productive and Quiescent Infection of Neuronal and Nonneuronal Cells. *J. Virol.* **2014**, *88*, 6847–6861, doi:10.1128/jvi.00516-14.

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

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

26. Tombácz, D.; Csabai, Z.; Oláh, P.; Balázs, Z.; Likó, I.; Zsigmond, L.; Sharon, D.; Snyder, M.; Boldogkői, Z. Full-Length Isoform Sequencing Reveals Novel Transcripts and Substantial Transcriptional Overlaps in a Herpesvirus. *PLoS One* **2016**, *11*, e0162868, doi:10.1371/journal.pone.0162868.

27. Olasz, F.; Tombácz, D.; Torma, G.; Csabai, Z.; Moldován, N.; Dörmő, Á.; Prazsák, I.; Mészáros, I.; Magyar, T.; Tamás, V.; et al. Short and Long-Read Sequencing Survey of the Dynamic Transcriptomes of African Swine Fever Virus and the Host Cells. *Front. Genet.* **2020**, *11*, doi:10.3389/fgene.2020.00758.

28. Moldován, N.; Torma, G.; Gulyás, G.; Hornyák, Á.; Zádori, Z.; Jefferson, V.A.; Csabai, Z.; Boldogkői, M.; Tombácz, D.; Meyer, F.; et al. Time-course profiling of bovine alphaherpesvirus 1.1 transcriptome using multiplatform sequencing. *Sci. Rep.* **2020**, *10*, doi:10.1038/s41598-020-77520-1.

29. Balázs, Z.; Tombácz, D.; Szűcs, A.; Snyder, M.; Boldogkői, Z. Dual Platform Long-Read RNA-Sequencing Dataset of the Human Cytomegalovirus Lytic Transcriptome. *Front. Genet.* **2018**, *9*, doi:10.3389/fgene.2018.00432.

30. Tombacz, D.; Csabai, Z.; Olâh, P.; Havelda, Z.; Sharon, D.; Snyder, M.; Boldogkôi, Z. Characterization of novel transcripts in pseudorabies virus. *Viruses* **2015**, *7*, 2727–2744, doi:10.3390/v7052727.

31. Moldován, N.; Tombácz, D.; Szűcs, A.; Csabai, Z.; Snyder, M.; Boldogkői, Z. Multi-Platform Sequencing Approach Reveals a Novel Transcriptome Profile in Pseudorabies Virus. *Front. Microbiol.* **2017**, *8*, 2708, doi:10.3389/fmicb.2017.02708.

32. Balázs, Z.; Tombácz, D.; Csabai, Z.; Moldován, N.; Snyder, M.; Boldogkoi, Z. Template-switching artifacts resemble alternative polyadenylation. *BMC Genomics* **2019**, *20*, 1–10, doi:10.1186/s12864-019-6199-7.

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

34. Csabai, Z.; Tombácz, D.; Deim, Z.; Snyder, M.; Boldogkoi, Z. Analysis of the complete genome sequence of a novel, pseudorabies virus strain isolated in Southeast Europe. *Can. J. Infect. Dis. Med. Microbiol.* **2019**, *2019*,

doi:10.1155/2019/1806842.

35. Boldogkői, Z.; Balázs, Z.; Moldován, N.; Prazsák, I.; Tombácz, D. Novel classes of replication-associated transcripts discovered in viruses. *RNA Biol.* 2019, *16*, 166–175.

36. Isomura, H.; Stinski, M.F.; Kudoh, A.; Murata, T.; Nakayama, S.; Sato, Y.; Iwahori, S.; Tsurumi, T. Noncanonical TATA Sequence in the UL44 Late Promoter of Human Cytomegalovirus Is Required for the Accumulation of Late Viral Transcripts. *J. Virol.* **2008**, *82*, 1638–1646, doi:10.1128/jvi.01917-07.

37. Boldogkői, Z.; Tombácz, D.; Balázs, Z. Interactions between the transcription and replication machineries regulate the RNA and DNA synthesis in the herpesviruses. *Virus Genes* 2019, *55*, 274–279.

38. Calvo, S.E.; Pagliarini, D.J.; Mootha, V.K. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 7507–12, doi:10.1073/pnas.0810916106.

39. Balázs, Z.; Tombácz, D.; Szucs, A.; Csabai, Z.; Megyeri, K.; Petrov, A.N.; Snyder, M.; Boldogkoi, Z. Long-Read Sequencing of Human Cytomegalovirus Transcriptome Reveals RNA Isoforms Carrying Distinct Coding Potentials. *Sci. Rep.* **2017**, *7*, doi:10.1038/s41598-017-16262-z.

40. Toledo-Arana, A.; Lasa, I. Advances in bacterial transcriptome understanding: From overlapping transcription to the excludon concept. *Mol. Microbiol.* 2020, *113*, 593–602.

41. Tombácz, D.; Sharon, D.; Szûcs, A.; Moldován, N.; Snyder, M.; Boldogkõi, Z. Transcriptomewide survey of pseudorabies virus using next-A nd third-generation sequencing platforms. *Sci. Data* **2018**, *5*, 1–13, doi:10.1038/sdata.2018.119.

42. Tombácz, D.; Sharon, D.; Oláh, P.; Csabai, Z.; Snyder, M.; Boldogkoi, Z. Strain Kaplan of pseudorabies virus genome sequenced by PacBio single-molecule real-time sequencing technology. *Genome Announc.* **2014**, *2*, 6–7, doi:10.1128/genomeA.00628-14.

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

44. Lex, A.; Gehlenborg, N.; Strobelt, H.; Vuillemot, R.; Pfister, H. UpSet: Visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.* **2014**, *20*, 1983–1992, doi:10.1109/TVCG.2014.2346248.

Figure Legends

Figure 1. ***Workflow of the PacBio, MinION and Illumina sequencing.*** The LoRTIA program identified the TSS and TES positions in ONT cDNA, dcDNA, terminator-seq, PacBio RSII random, IsoSeq and Sequel samples. LoRTIA software suit also helped in the validation of TESs and introns in dRNA-Seq samples. The Illumina data were used in the validation of certain low-abundance transcripts and splice sites, and in the identification of transcription readthroughs.

Figure 2. ***Transcript characteristics.*** The upset plot indicates the quantitative distribution of the transcripts identified by using ONT and PacBio techniques.

Figure 3. ***The updated PRV transcriptome.*** PRV transcriptome contains those transcripts which were identified by the integrated approach using novel and earlier short-and long-read sequencing datasets. The light brown arrows represent the open reading frames of genes; the green arrows show the non-coding RNA genes; the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 December 2020 doi:10.20944/preprints202012.0616.v1

blue arrows are the mRNAs, while the red arrows illustrate the non-coding transcripts. The putative embedded mRNAs are considered here as real mRNAs, but this has to be verified by proteomic studies. Complex transcripts are also colored by blue, although, it is unknown whether they really function as mRNAs, especially those ones of which the firs gene stands in opposite direction, that is non translated. LLT and AST transcripts were detected in latently infected neurons, and by real-time RT PCR in lytic infection, but not in our sequencing experiments.

Figure 4. Embedded mRNAs. The following genomic regions are selected for the illustration of the embedded mRNAs: (a) ul39- ul40, (b) ul42-43-44, (c) ul19-18. Arrows with the same color represent transcripts containing the same ORFs but distinct TSSs or TESs. The rectangular green lines indicate the fist in-frame ATGs within the transcripts. The "nc" letters at the end of the names indicate the lack of the stop codons.

Figure 5. ***Coding and non-coding RNA molecules at the ie180-us4 genomic region.*** A high density of non-coding transcription can be observed at this genomic region. Color code: light brown: coding, or non-coding genes, blue: mRNAs, red: ncRNAs.

Figure 6. ***Replication origin-associated transcripts at the (a) Ori-L* (a) *and (b) the Ori-S genomic regions.*** These RNA molecules have been described in several viruses, including herpesviruses. Except the CTO-S family, these transcripts overlap the replication origins through either their 3′-UTRs (CTO-L) or their 5′-UTR (PTO-US1). Both the raRNAs and the transcript of adjacent genes are overlapped by antisense RNAs of which some are controlled by separate promoters. Color code: light brown: coding, or non-coding genes, blue: mRNAs, red: ncRNAs

Figure 7. ***The base content at the 5′- and 3′-termini of the PRV transcripts.*** The image is generated using weblogo 3.0

Legend for Tables

Table 1. ***Sequencing reads and coverages obtained using different techniques***

Table 2. ***The list of splice isoforms***

Table 3. ***Methods of the approaches described in our earlier publications.***

Table 4. ***Settings of the LoRTIA software suite for each sample type***

Appendix

Figure S1. ***The relative abundance of PRV transcripts.*** Black arrows indicate very high-abundance transcripts, whereas light gray arrows indicate very low-abundance transcripts.

Figure S2. ***Sequencing reads are organized in a stepped fashion indicating the separated TSSs of embedded transcripts***. This distribution pattern of the raw reads indicates that they were not generated by spurious processes or degradation.

Table S1. ***Sequencing reads and coverages obtained in different techniques.***

Table S2. ***List of PRV transcripts.***

Table S3. ***List of putative embedded genes.***

Table S4. ***List of non-coding transcripts.***

Table S5. ***List of TSS and TES isoforms.***

 doi:10.20944/preprints202012.0616.v1

Table S6. *Novel Splice isoforms.*

Table S7. *Polygenic transcripts.*