HIGH-THROUGHPUT TRANSCRIPTOMIC ANALYSIS OF PSEUDORABIES VIRUS

Ph.D. Thesis

Péter Oláh Msc

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

Supervisor: prof. Zsolt Boldogkői

Szeged
2017

LIST OF PUBLICATIONS

Publications directly related to the subject of the thesis:

I. **Oláh P**, Tombácz D, Póka N, Csabai Z, Prazsák I, Boldogkői Z. Characterization of pseudorabies virus transcriptome by Illumina sequencing.BMC Microbiology. 2015;15:130. doi:10.1186/s12866-015-0470-0. **IF:2.581 (2015)**

II. Tombácz, D., Csabai, Z., **Oláh, P**., Havelda, Z., Sharon, D., Snyder, M., & Boldogkői, Z. (2015). Characterization of Novel Transcripts in Pseudorabies Virus. Viruses, 7(5), 2727–2744. http://doi.org/10.3390/v7052727 **IF:3.437 (2015)**

III. Tombácz D, Sharon D, **Oláh P**, Csabai Z, Snyder M, Boldogkői Z. Strain Kaplan of Pseudorabies Virus Genome Sequenced by PacBio Single-Molecule Real Time Sequencing Technology. Genome Announcements. 2014;2(4):e00628-14. doi:10.1128/genomeA.00628-14.

Publications indirectly related to the subject of the thesis:

IV. Tombácz D, Csabai Z, **Oláh P**, et al. Full-Length Isoform Sequencing Reveals Novel Transcripts and Substantial Transcriptional Overlaps in a Herpesvirus. PLoS ONE. 2016;11(9):e0162868. doi:10.1371/journal.pone.0162868. **IF: 3.54 (2015)**

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

CTO	**Close To Ori-L (gene)**
dsDNA	**double-stranded DNA**
DSE	**downstream sequence element**
E	**Early (gene expression)**
E/L	**Early/Late (gene expression)**
EBV	**Epstein-Barr Virus**
gDNA	**genomic DNA**
HCMV	**Human cytomegalovirus**
HSV-1	**Herpes simplex virus 1**
IE	**Immediate-early (gene expression)**
IR	**Inverted repeat**
IRES	**Internal Ribosome Entry Site**
IRS	**internal repeat sequence**
Ka	**strain Kaplan of Pseudorabies virus**
KSHV	**Kaposi's Sarcoma Herpes Virus**
L	**Late (gene expression)**
lncRNA	**long non-coding RNA**
miRNA	**microRNA**
p.i.	**Post infection**
PA-peak	**signal of mRNA 3' end in PolyA-sequencing**
PAP	**PolyA polymerase**
PAS	**polyA signal**
PA-Seq	**high-throughput 3'-end RNA sequencing**
PK-15	**Porcine kidney 15 cell line**
polyA+	**polyadenylated RNA**
PRV	**Pseudorabies virus**

qRT-PCR	**quantitative real-time polymerase chain reaction**
RNA-Seq	**high-throughput RNA sequencing**
RPKM	**Reads per kilobase per million**
ssDNA	**single-stranded DNA**
TRS	**Terminal repeat sequence**
TSS	**transcriptional start site**
USE	**upstream sequence element**
VZV	**Varicella zoster virus**
ZMW	**Zero-Mode Waveguide**

1. INTRODUCTION

1.1 Aujeszky's disease virus – Pseudorabies

Pseudorabies virus (PRV, also called Aujeszky's disease virus or Suid herpesvirus 1) is a neurotropic alphaherpesvirus, member of the genus *Varicellovirus* of family *Herpesviridae* and subfamily *Alphaherpesvirinae*. The virus was first characterized by Hungarian veterinarian Aladár Aujeszky in 1902 [1] as an infectious agent causing rabies-like symptoms in dogs, cattle and swine. The only natural reservoir host of the virus is swine, where infection is fatal in piglets, and causes respiratory disease in adult animals, as it is shown to inhibit the functions of alveolar macrophages [2, 3]. While the virus infects a wide range of mammals, human and tailless monkeys do not contract the pathogen, which makes it an important model organism for studying neuronal tracing [4, 5, 6] and viral transcriptional regulation [7, 8].

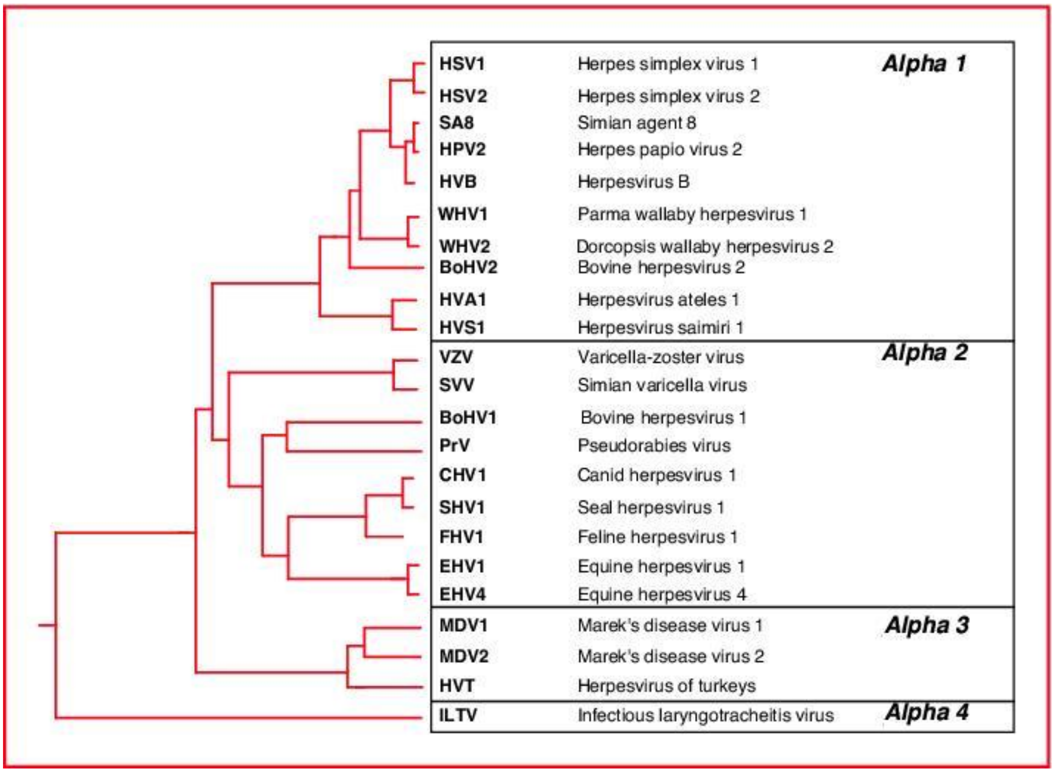

HSV1 Herpes simplex virus 1
Alpha 1
HSV2 Herpes simplex virus 2
SA8 Simian agent 8
HPV2 Herpes papio virus 2
HVB Herpesvirus B
WHV1 Parma wallaby herpesvirus 1
WHV2 Dorcopsis wallaby herpesvirus 2
BoHV2 Bovine herpesvirus 2
HVA1 Herpesvirus ateles 1
HVS1 Herpesvirus saimiri 1
VZV Varicella-zoster virus
Alpha 2
SVV Simian varicella virus
BoHV1 Bovine herpesvirus 1
PrV Pseudorabies virus
CHV1 Canid herpesvirus 1
SHV1 Seal herpesvirus 1
FHV1 Feline herpesvirus 1
EHV1 Equine herpesvirus 1
EHV4 Equine herpesvirus 4
MDV1 Marek's disease virus 1
Alpha 3
MDV2 Marek's disease virus 2
HVT Herpesvirus of turkeys
ILTV Infectious laryngotracheitis virus
Alpha 4

***Fig.1.** Phylogenetic tree of alphaherpesviruses (from [9])*

1.2 Virion structre

Herpesviruses share the same virion structure and basic genome composition, with the closest relatives of PRV being Bovine- and Equine herpesviruses, *Varicella zoster*, and *Herpes simplex 1* and *2*. (Fig1). The 143kbp-long dsDNA genome is contained within the icosahedral capsid consisting of 162 capsomers, surrounded by a tegument layer and the viral envelope acquired from the host cell lipid membrane, and spiked with viral glycoproteins (Fig. 2).

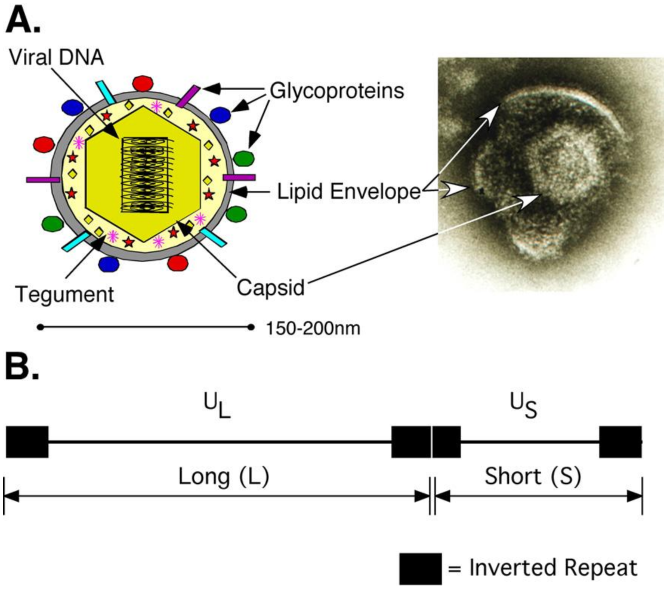


Fig.2. *A: Virion structure of PRV. B: Basic genome composition of PRV.(from [10])*

1.3 Genome composition

The organization of protein-coding genes is also shared among herpesvirus families, with genes forming polycistronic clusters which occur in various permutations across species. Based on the widely-referenced PRV composite genome, created from six PRV strains [11], ~70 protein-coding genes were recognized, arranged in several overlapping gene clusters. Two unique genomic sequences are designated Unique Short (U_S) and Unique Long (U_L), while two copies of the major Inverted Repeat (IR) form the Internal and Terminal Repeat

Sequences (IRS and TRS), flanking the US region in opposite orientations. The extreme GC-content of PRV (73.6%) is also a unique characteristic, and provides methodological benefits, as well as drawbacks during sequencing studies.

1.4 Viral life cycle

The duality of lytic (productive) and latent infectious cycles is also a common trait shared by herpesviruses. In the case of PRV, latency is established in the trigeminal ganglia of the pig host, resulting in persistent infection, and subsequent reactivation may occur in response to stress or immune suppression. PRV infection is initiated by binding to the host cell surface through the interaction of viral glycoproteins gC and gD with cellular receptors nectin 1 and 2, CD55, heparan sulfate and PILR-alpha [12, 13, 14]. The presence of multiple interactors is in accordance with the broad host range of the virus. The fusion of the viral envelope and cellular plasma membrane is mediated by glycoproteins gB, gH and gL. While gD is required for PRV penetration, it is not essential for cell-to-cell spread of the virus [15]. Transportation of the capsid to the nucleus is facilitated by the attachment to the dynein motor molecule and its movement along microtubules. Following nuclear entry, the cellular machinery is utilized to initiate viral transcription and translation. DNA replication is thought to switch rapidly from theta structure to the rolling-circle mechanism, producing concatameric genome copies [16]. In the following, highly coordinated process, DNA is cleaved into monomeric, linear form, and loaded into the assembled viral capsids through the pUL6 portal protein [17, 19]; at least six viral proteins are known to interact in this process. Primary envelopment occurs during nuclear egress, followed by de-envelopment, and the attachment of tegument proteins to the bare capsids in the cytoplasm. Viral particles then go through secondary envelopment on the cytoplasmic face of a Golgi-derived specialized compartment, which then also facilitates the transport of mature virions to the cell surface [18, 20].

1.5 Gene expression cascade

One main advantage of viral gene expression studies is the compact nature of the examined genomes, which, due to evolutionary pressure, function as tightly-regulated machineries with highly compressed and optimized genetic components and program. 40 protein-coding genes are shared between all alpha-beta- and gammaherpesviruses, encoded in the U_L region. These core genes coordinate the major viral functions of infection, capsid assembly, replication and egress. The cascade-like gene expression during lytic alphaherpesviral infections can be divided into four main temporal categories: immediate-early (IE), early (E), early/late (E/L), and late (L) genes. Although recent RT-qPCR experiments have shown that such a classification is simplified, and the borders between temporal gene classes are more continuous [21], this system still provides a broad overview of transcriptional kinetics. The IE class of genes are the first to be transcribed during infection and act as transactivators for the remaining classes, which require transcription factors for expression, and are sensitive to protein translation inhibitors, such as cycloheximide. While in the most well-studied relative of PRV, HSV-1, there are five IE genes (*icp0, icp4, icp22, icp27* and *icp47*) controlling gene expression and inhibiting antigen presentation [22, 23, 24], in PRV *ie180* (homolog of *icp4*) is the sole IE-class transactivator, affecting the activity of several viral promoters. PRV genes early protein 0 (*ep0*, homolog of *icp0*) and *ul54* (*icp27* homolog) are expressed as E-class genes, along with *us1* (*icp22* homolog), one of the most abundantly expressed mRNAs of the virus. During the productive infectious cycle, *ie180* is expressed by 40 mins post infection (p.i.), and protein synthesis lasts until ~2.5 h p.i. The E-class genes appear at ~1h p.i, with peak levels ~3-4h, and primarily encode nucleotide metabolism and DNA replication-related protein products. L-class genes appear ~2.5h p.i, and mainly produce structural and scaffolding proteins [16].

When latency is established in neurons, the viral genome is thought to be almost completely silenced, with the exception of the latency-associated transcript (LAT), encoded on the opposite strand to the *ep0* locus, with the second exon of the Long Latency Transcript (LLT) being similarly antisense to the *ie180* gene [26, 27], therefore possibly blocking their expression through antisense activity. The LAT promoter (LAP) is also shown to be highly neuron-specific, and inhibit apoptotic processes, thus ensuring long term latency of the virus [28].

1.6 Transcriptional interference in viruses

The term “transcriptional interference” (TI) is the mechanism by which one transcriptional process interferes with another in cis, and implies the collision of the RNA polymerase (RNAP) machineries [29, 30]. TI has been observed in a wide range of model organisms, including *S. cerevisiae*, *E. coli* and also higher-order species [31, 32, 33, 34]. It is hypothesized that the ubiquitous non-coding transcription observed in most branches of life, while not producing translational products, may in fact serve to coordinate the expression of coding genes by blocking their overlapping promoters and regulatory elements [31]. Both elongating and pausing polymerases form obstacles for transcription, thus slowing or hindering mRNA expression of downstream or oppositely oriented gene clusters. As the human genome is reported to be transcribed in >90% in specific cases [35], the issue of TI is becoming more and more of a hot topic in life sciences. On the basis of this phenomenon, the Transcriptional Interference Network (TIN) hypothesis was put forward [7], which formulates that gene expression machineries may act as an independent self-regulatory system, and that viruses, with their condensed and relatively simple genomes may serve as ideal candidates for modeling such interference networks. The main interactions that the TIN hypothesis considers are between the following overlapping gene clusters: tandemly overlapping genes (waterfall model), divergently or convergently overlapping genes, with 5' or 3' UTR overlaps (seesaw model and extension), and promoter competition model, for bidirectional promoters [7].

1.7 Short-and long-read sequencing technologies

1.7.1 Illumina sequencing by synthesis

With molecular biology stepping into the post-genomic era, the use of so-called “next-generation” -or 2nd generation- DNA sequencing is becoming a routine procedure in a wide range of investigations [36, 37, 38, 39], also setting up the path for new and even more powerful methods, such as single-molecule sequencing or nanopore technologies [42, 43]. The methodological benefits include single-base resolution and customizable coverage per base, hypothesis-free discovery of novel genomic elements and transcripts, and flexibility of sample preparation, ensuring that the range of applications may constantly grow, similarly to those of PCR techniques. Since the initial introduction of the Roche 454 pyrosequencing platform in 2005, the advancements in read length, cost per base and ease of library preparation have granted wider popularity for the Illumina sequencing-by-synthesis approach (originally introduced in 2006) [40]. Sequencing by synthesis utilizes the ligation of universal adapter sequences to fragmented DNA or cDNA molecules, by which fragments are bound to washable slides (flow cells). During sequencing, reversible dye terminators [41] react with the immobilized DNA fragments, and are washed out after imaging by a fluorescent microscope, in order to enable up to 300 repeated cycles (and sequenced bases per fragment). Sequencing is also possible in a paired-end manner, where the given fragments are read from each end, thus providing more useful sequence information, in which case the sequencing “blind spot” (the insert) between the paired ends is set to an approximate length during the fragmentation procedure. Sample multiplexing is achieved by the use of indexed primers, where 6-bp “barcode” sequences are read before each DNA fragment to serve demultiplexing computationally. The resolution of homopolymer stretches, a common challenge in genome reconstruction, is resolved by the application of dye terminators, so that the sequencing reaction can only be extended by one base per cycle. >99.9% accuracy is ensured by the robust chemistry and high copy numbers of the sequenced fragments. However, the relatively shorter read length (>2x300bp) limits the applications in structural genomics, the study of gene expression and splicing kinetics, or metagenomics, where large fragment sizes or complete mRNA sequences are more desirable.

1.7.2 Pacific Biosciences RS II

Pacific Biosciences (PacBio) applies a single-molecule real-time sequencing approach, which makes its RS II platform presently the most popular so-called “third generation” sequencer [40]. It is distinguished from other platforms by the ability to read unprocessed, unamplified DNA fragments, and that the sequencing process is continuously monitored by CCD cameras rather than creating snapshots, such as in a fluorescent microscope. DNA or cDNA strands are loaded onto sample plates which are arrays of 100 nm wells (Zero Mode Waveguide, ZMW), with immobilised DNA polymerase molecules on the bottom. The DNA strands are then processed by the polymerase in the presence of fluorescently-labeled nucleotides. Since random sequences of single-molecule reactions have to be recorded in motion, fluorescent background noise is minimized by ZMWs that are illuminated in *trans* through apertures smaller than the wavelength of illuminating light, thus labeled nucleotides are excited by epifluorescence, with minimal background lighting (Fig.3.). Although the accuracy of a single read is only ~80%, the error rate is compensated by the circularization of molecules, which are then read through several times (Circular Consensus Sequencing, CCS) [48], whereby the random read errors cancel each other out and provide a consensus accuracy >99%. In this manner, single read lengths can reach ~60 kb, which can be used e.g. in scaffolding, while consensus read lengths reach up to ~5 kb. Besides longer read lengths, the technology can be used to decipher epigenetic modifications of DNA, as the time required for the incorporation of each labeled nucleotide by the polymerase is proportional to the type of modification on the template strand (polymerase stalling). This information is currently used to characterize 5-methyl-cytosine (5-mC) patterns in eukaryotes and 4-methyl-cytosine/ 6-methyl-adenine in microbial DNA [47].

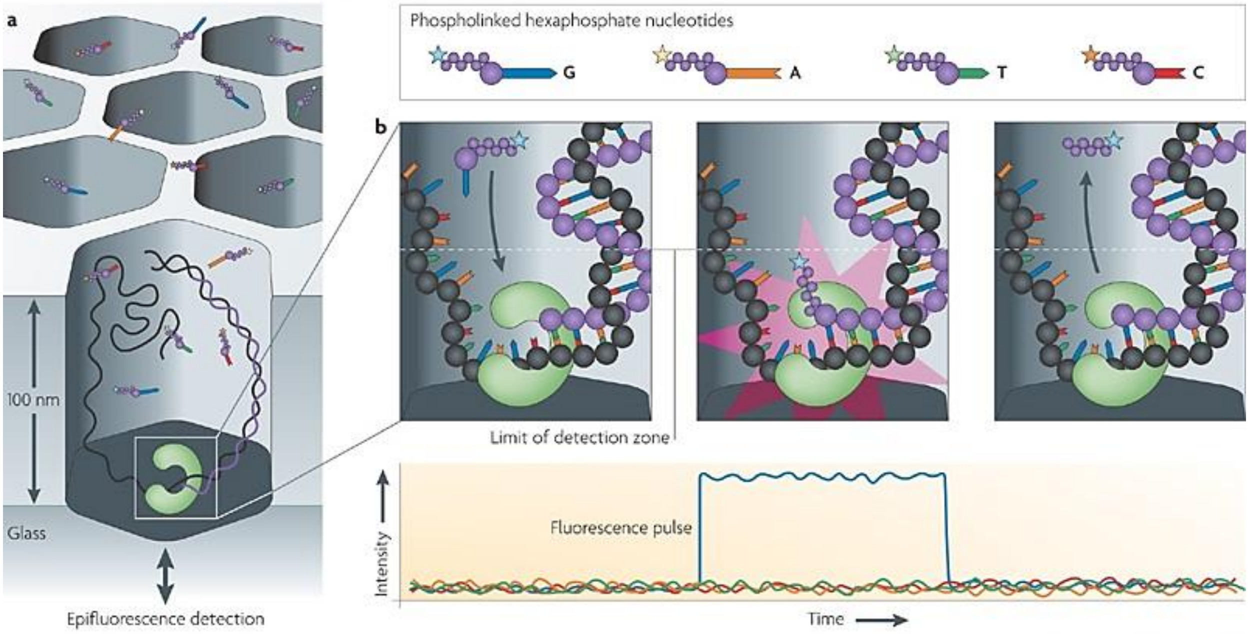


Fig.3. Overview of Pacific Biosciences Single-Molecule Real-Time Sequencing technology (from [45])

1.8 Bioinformatic analyses of sequencing data

The cost- and time-efficient generation of such a wealth of information on the genetic content of various organisms poses a great challenge on the side of digestion and interpretation of the data. With next-generation sequencing technologies, sophisticated algorithms became necessary for data analysis, and nowadays detailed pipelines should be integral parts of study design. Primary sequence analysis consists of the read-out of fluorescent images from the instruments, together with basic quality control, score assignments [44], and usually conversion of results to the popular fastq format, in which each sequencing read maintains its unique identifier, sequence readout and per-base quality score. Secondary analysis consists of demultiplexing barcodes, trimming and discarding of low-quality sequences, followed by either *de novo* genome or transcriptome assembly, or mapping of reads to a known reference genome, in case one is available for the model organism. The resulting sam/bam (or cram) alignment files [46] store the mapping quality and read alignment information, and serve as a basis for the wide variety of tertiary analyses. The most common applications include genotyping, binding site identification and differential expression analysis, which fields now have fairly established guidelines and standard procedures, while in the areas of metagenomics, molecular fingerprinting, network analysis, and integrative genomics, development of robust algorithms continues actively, with less dominant solutions available.

1.9 High-throughput sequencing methods in herpesvirus research

While the use of these high-throughput technologies has obvious advantages for viruses, possessing some of the smallest genomes possible, there are also notable difficulties, e.g. in sample purification, host-pathogen nucleotide ratios, or dealing with extreme GC-content and repetitive sequences. During the last decade, herpesvirus research has benefited tremendously from the application of 2nd-generation sequencing, providing high-resolution genetic maps for several key species [49, 51, 52], epigenetic patterns [50, 53] and transcriptome studies [55, 56, 57, 63], among others. Comparative genomics aides in the functional annotation of conserved elements, along with molecular fingerprinting and diagnostic potentials, previously also limited by the lack of whole-genome sequences for various strains of important pathogens. In viral studies, RNA-sequencing (RNA-Seq) mostly involves the simultaneous survey of host and viral transcripts. While providing insight into host-pathogen gene expression networks, RNA-Seq also expands the views on viral non-coding RNAs. A novel cluster of 5 microRNAs have been shown to be expressed in PRV from the Long Latency Transcript (LLT) intron during latency in neuronal ganglia, with possible roles in viral gene silencing [58], while the same locus gave rise to 11 miRNAs in epithelial cells [59]. Further miRNA discoveries have been reported from HSV-1 and 2, HCMV, EBV and KSHV, where miRNAs were predicted to target both viral and cellular mRNAs, potentially creating a more favorable environment for virus replication by affecting immune evasion and the latent-lytic cycle [60]. The disregulation of host miRNAs during infection also hints at the interplay between these newly discovered RNAs. On the other hand, long non-coding RNAs (lncRNAs) are also detected in most members of *Herpesviridae*. The role of lncRNAs in the latent life cycle has been a long-established fact [61], however, it is of note that several highly abundant lncRNAs have been detected by high-throughput methods in lytic infections, which often account for ~50% of total viral transcript quantities. Examples include human cytomegalovirus (HCMV) *RNA2.7* [56] and the Kaposi's sarcoma herpesvirus (KSHV) *PAN* lncRNA, which is hypothesized to modulate gene expression through binding of modified chromatin [62]. The cataloging of various truncated and spliced isoforms and often extremely variable 5' and 3' UTR sequences is also of interest, especially considering their potential functional roles arising from the condensed, streamlined genome structure, where the efficient encoding of a high number of genes is of key importance. The steadily declining cost per base of sequencing technologies also enables the discovery of novel rare transcripts and underrepresented splice isoforms through ultra-deep sequencing, which are increasingly attributed regulatory functions, instead of being regarded as mere transcriptional noise.

2. AIMS

1. *De novo* assembly of the PRV strain Ka complete genome using long-read sequencing, in order to provide an accurate reference of the strain used in our studies in place of the previously used composite reference.
2. Generating the first single-base resolution transcriptome map of PRV Ka from a mixed-timepoint infection, using short-read sequencing in order to characterize potential new transcripts and splice isoforms.
3. Independent validation and detailed characterization of novel transcripts in multi-time-point samples.
4. Analysis of potential transcriptional interference events in the viral genome in support of the TIN hypothesis.

Remark: The present thesis focuses primarily on the bioinformatical analysis and sequence categorization aspect of the results, which was the main contribution of the author to the source publications, drawing on the genomic DNA sequencing results from publication III, complete transcriptome results from publication I, and specific results on the CTO non-coding RNA from publication II.

3. MATERIALS AND METHODS

3.1 Virus, cells and infection

Immortalized Porcine Kidney PK-15 epithelial cells were used for the propagation of strain Kaplan of PRV. PK-15 cells were cultivated in Dulbecco's modified Eagle medium supplemented with 5% fetal bovine serum (Gibco Invitrogen) with 80 µg gentamycin/ml at 37 °C, 5% CO2 in filter-capped flasks. The following virus stock was prepared for the experiments: semi-confluent PK-15 cells in rapid growth were infected at a multiplicity of infection (MOI) of 0.1 plaque-forming unit (pfu)/cell, and incubation lasted until a complete cytopathic effect was observed. The infected cells were frozen and thawed three times, followed by low-speed centrifugation at 10,000g, 20 min. The supernatant was concentrated and further purified by ultracentrifugation after removal of cell debris, across a 30% sugar cushion at 24,000 rpm for 1h, using a Sorvall AH-628 rotor. The number of cells in a culture flask was 5×10^6. A high MOI (10 pfu/cell) was used for the infection of PK-15 cells in order to generate samples for transcriptome studies. Infected cells were incubated for 1h, followed by removal of the virus suspension and washing with phosphate-buffered saline (PBS). After the addition of new medium to the cells, they were incubated for 1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 or 24h p.i. For the control population, mock-infected cells, treated in the same way as the infected cells, were used.

3.2 Viral DNA extraction

PK-15 cell monolayers were infected at MOI=10pfu/cell, and cultivated at 37 °C until a cytopathic effect was observed. Culture medium was collected and centrifuged at 4,000 rpm for 10 min using a Sorvall GS-3 rotor. Viral particles were sedimented on a 30% sucrose cushion by ultracentrifugation at 24,000 rpm for 1h using a Sorvall AH-628 rotor. The sedimented virus was resuspended in sodium Tris-EDTA buffer. 100 ug/ml Proteinase-K and 0.5% sodium dodecyl sulfate (SDS) was added, the lysate was incubated at 37 °C for 1h, followed by phenol-chloroform extraction.

3.3 TotalRNA extraction

RNA was extracted from samples of each individual time point of infection by using the NucleoSpin RNA II Kit (Macherey-Nagel GmbH and Co. KG). Following centrifugation and cell lysis with buffer containing chaotropic ions, the nucleic acids were purified on silica column. DNA was removed by RNase-free DNase solution (supplied with the NucleoSpin RNA II Kit). Finally, the RNAs were eluted from the column in RNase-free water (supplied with the kit). To eliminate the residual DNA contamination, all RNA samples were treated by an additional digestion with Turbo DNase (Ambion Inc.). The concentrations of the RNA samples were measured by spectrophotometric analysis with a BioPhotometer Plus instrument (Eppendorf) and Qbit fluorometer (Thermo Fisher Scientific). RNA samples were stored at −80°C until further use.

3.4 PacBio RS II gDNA prepration and sequencing

SMRTbell template libraries were prepared from DNA using standard protocols for 6-kb and 20-kb library preparation. Sequencing was performed in five single-molecule real-time (SMRT) cells with P5 DNA polymerase and C3 chemistry (P5-C3) yielding a total of 78,111 reads and an extremely high coverage (1,200x) throughout the genome. Sequencing and library preparation were carried out in the Department of Genetics, Stanford University.

3.5 Illumina cDNA library preparation and sequencing

Strand-specific total RNA libraries were prepared for paired-end, 2x100bp sequencing by using the Illumina-compatible ScriptSeq v2 RNA-Seq Library Preparation Kit (Epicenter). For polyA-sequencing, a single-end library was constructed through the use of custom anchored adaptor-primer oligonucleotides with an oligo(VN)T-10 primer sequence, in which a two-nucleotide anchor is followed by 10 T nucleotides. Anchored primers compensate for the loss in throughput due to the high fraction of reads containing solely adenine bases, as in the case of conventional oligo(dT) primers. Sequencing was performed on an Illumina HiScanSQ platform at the Genomic Medicine and Bioinformatic Core Facilty of the University of Debrecen.

3.6 RT-qPCR analysis of alternative splicing

In order to validate splicing events, two sets of primers were applied, with lengths from 19 to 23 nucleotides, approximately 100 bp upstream and downstream of the given splice site (Table 1). 5 µl solutions were prepared for reverse transcription reactions, containing 0.02 µg of total RNA, 2 pmol of the gene-specific primer, 0.25 µl of dNTP mix, 1 µl of 5× First-Strand Buffer, 0.25 µl (50 units/µl) of SuperScript III Reverse Transcriptase (Invitrogen) and 1 U of RNAsin (Applied Biosystems Inc.). RT mixes were incubated at 55 °C for 60 min. The reaction was stopped at 70 °C for 15 min. No-RT control reactions (RT reactions without Superscript III enzyme) were run to test the potential viral DNA contamination by conventional PCR. For RT-qPCR reactions, RNA samples with no detectable DNA contamination were used. Real-time quantitative PCR experiments were carried out for each sample in triplicate, on a Rotor-Gene 6000 cycler (Corbett Life Science). RT-qPCR reactions were done in 20-µl mixtures containing 7 µl of ×10 dilution cDNA, 10 µl of ABsolute qPCR SYBR Green Mix (Thermo Fisher Scientific), 1.5 µl of forward and 1.5 µl of reverse primers (10 µM each). The running conditions were as follows: 15 min at 95 °C, 30 cycles of 94 °C for 25 s (denaturation), 60 °C for 25 s (annealing), and 72 °C for 6 s (extension). Products were visualized on 12 % polyacrylamide gel stained with Gel Red dye, gel images were acquired using a ProteinSimple AlphaImager HV gel documentation system.

EP0 SPLICE FW 1	tgtcaaacagcgcatcgacgagg
EP0 SPLICE REV 1	cacaagttctgtctggactgcatcca
EP0 SPLICE FW 2	Gatgttgtccacgacggcctc
EP0 SPLICE REV 1	cacaagttctgtctggactgcatcca
EP0 SPLICE FW 3	Ctggcggttcatcccgtgctc
EP0 SPLICE REV 1	cacaagttctgtctggactgcatcca

***Table 1:** EP0 splice site validation primer sequences*

3.7 Northern Blot Analysis

Total RNA was isolated from PK-15 cells by TRIzol (Life Technologies) extraction according to the manufacturer's instructions for conventional Northern blot experiments. Samples were denatured in loading buffer for 5 min at 65 °C. Extracted RNA samples (10 ug) were fractionated in formaldehyde/1.2% agarose gel, transferred to a Nytran N membrane (Schleicher & Schuell BioScience, Dassel, Germany) by a capillary method and fixed by ultraviolet cross-linking. The membrane was probed by using the random primed PCR product and the total viral DNA with the DecaLabel DNA Labeling Kit (Fermentas, Vilnius, Lithuania). PCR reactions were carried out with AccuPrime GC-Rich DNA Polymerase (Life Technologies) according to the manufacturer's recommendations. The oligonucleotide probe was labeled with [α-32P]CTP. Filter prehybridization was carried out in 50% formamide, 0.5% SDS, 5× SSPE, 5× Denhardt's solution and 20 µg/mL sheared, denatured salmon sperm. The probe was heated for 1 min at 95 °C. Overnight hybridization was carried out at 68 °C. Finally, the hybridization membranes were washed in 2× SSC 0.1% SDS at 68 °C for ones 10 min, 0.5× SSC, 0.1% SDS at 68 °C for 10 min, 0.1× SSC 0.1% SDS at 68 °C for 10 min.

3.8 Data analysis

Sequence analysis was carried out using a range of open-source tools and custom in-house scripts incorporated into an analysis pipeline tailored to the specific questions of our studies, as detailed in section 4.1.

3.9 Data availability

The complete genome sequence of strain Kaplan of pseudorabies virus was assigned DDBJ/EMBL/GenBank accession no. KJ717942. Raw read data from PA-Seq and RNA-Seq experiments are deposited in the European Nucleotide Archive under accession code PRJEB9526.

4. RESULTS AND DISCUSSION

The whole-genome sequence of PRV strain Ka was determined by cutting-edge long-read sequencing in order to facilitate accurate transcriptome mapping and further epigenetic studies of the virus. Illumina short-read, high-throughput sequencing was used for the first time on lytic-infection PRV samples in order to create a single-base resolution transcriptome map, along with the detailed polyadenylation landscape of the virus by PA-Seq. In accord with expectations, most of the viral genome was transcribed, with the exception of several small intergenic repetitive sequences, and loci in the large internal and terminal repeats. Among the findings are a novel polyadenylated lncRNA near the OriL origin of replication, and the single-base resolution mapping of 3′ UTRs across the viral genome. A number of genes exhibited alternative polyadenylation sites, while previously described splice sites were confirmed and expanded with a novel alternative splicing event in the key regulator *ep0* gene. As PRV mRNAs mainly form polycistronic clusters, it is also of note that we identified several genes possessing distinct PA sites that were previously thought of as polycistronic.

4.1 Bioinformatic analysis pipeline

4.1.1 Data quality assessment

For PacBio RS II sequencing, reads spanning several kilobases were obtained using the 6kb and 20kb library preparation protocols. Individual reads with >25% error rate were discarded from further analysis, however high accuracy was achieved by the >1000x coverage of the genome, through ~78.000 reads.

For Illumina sequencing, both the RNA integrity measurements during the sample preparation, FastQC [64] quality metrics, and the low signal-to-noise ratio in the 1 kb region surrounding the PA peaks during the analysis indicated high library quality. Sequencing of the total RNA isolates of infected cells yielded a data set of～208 million 100 bp paired-end reads for the random hexamer-primed library, with a mean insert size ~ 300bp. 1.3 million reads aligned to the viral genome version KJ717942.1, and the majority of the remaining sequences aligned to the host organism genome Sus scrofa 10.2. PA-Seq resulted in～103 million single- end, 50 bp reads, with a higher ratio of 10 million reads aligning to the PRV reference.

4.1.2 Long-read genome assembly

Raw reads of PRV gDNA were retrieved in PacBio's information-rich bax.h5 format and were subjected to the SmrtAnalysis pipeline for filtering, followed by small genome de novo assembly using the HGAP and Quiver algorithms. [65]. The HGAP (Hierarchical Genome Assembly Process) algorithm aligns quality-filtered raw reads by the longest overlapping seeds in order to create high-accuracy consensus reads of several kbp length during pre-assembly. Genome assembly is then carried out by the Celera Assembler, optimized for microbial genome sizes. The Celera Assembler was used for the whole-genome shotgun assembly of the human reference genome by Celera Genomics, and since has been continually developed, until recently. The algorithm is based on Overlap-Layout Consensus *de novo* assembly and inherently works efficiently on long reads (such as those from Sanger sequencing), building unitigs which are then classified as unique (U-unitigs) or repeat, and processed into larger contigs. [66]. In the case of the compact PRV genome, the longest contigs practically encompass the complete genome, naturally without the large Terminal Repeat Sequence. Caution is required, however, especially with regard to the known systematic errors of long-read single-molecule sequencing. While the high error rate of individual raw reads is effectively cancelled out by coverage, due to the random nature of the error, the overestimation of homopolymer guanine stretches in a fraction of the reads often ends up incorporated in the final consensus. These can be easily scanned for as they result in a drop of coverage, and sequence overrepresentation, and the erroneous reads are discarded. The resulting consensus is polished by the Quiver algorithm, which can use a preexisting reference genome in order to fill out possible gaps or structural arrangement- in the case of PRV, the placement of the large TRS sequence was guided by Quiver, with manual inspection of repeat-boundary spanning anchor sequences, and further cross-validation by short-read cDNA fragments located over the repeat boundaries, which confirmed placement and orientation of the TRS.

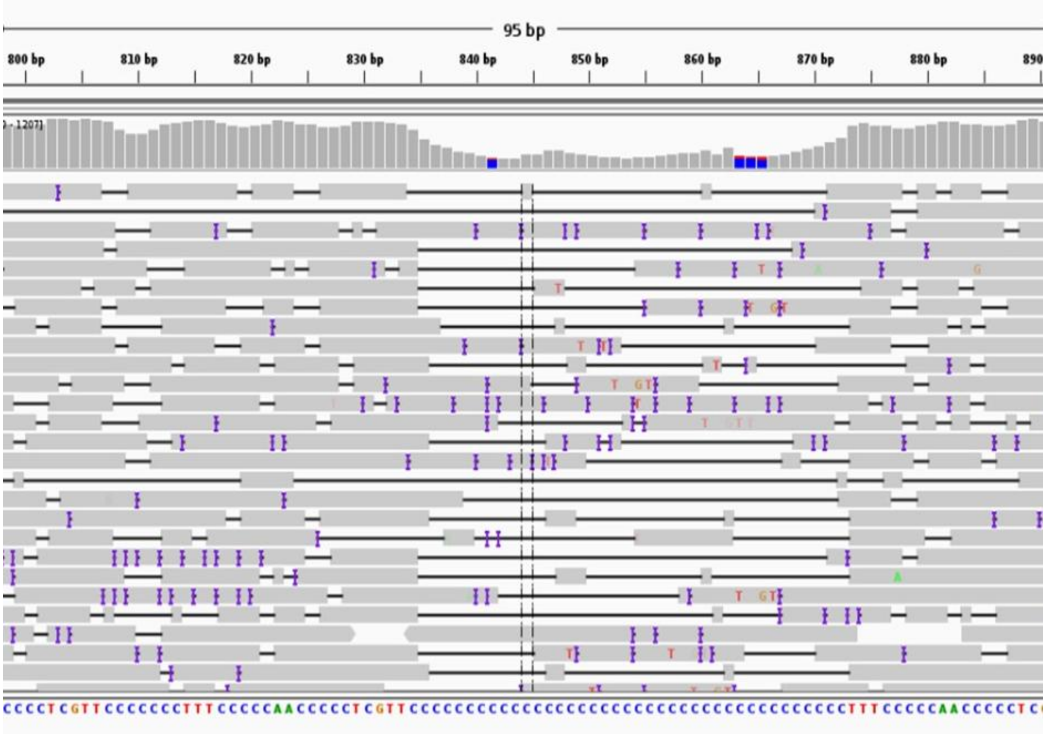

95 bp
800 bp
810 bp
820 bp
830 bp
840 bp
850 bp
860 bp
870 bp
880 bp
890
CCCCTCGTTCCCCCCCTTTCCCCCAACCCCCTCGTTCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCTTTCCCCCAACCCCCTC

Fig.4. *Alignment of PacBio genomic DNA reads to the de novo reference, illustrating a systematic sequencing error in guanine-homopolymer stretches, which was corrected during genome finalization, and the random-error nature of PacBio sequencing. Top bar: genomic coordinates; histogram: per-base read coverage; gray lines: individual single-pass reads, with insertions (blue), and deletions (black horizontal lines). It is shown that only a small fraction of individual reads contain the excess G (or C) homopolymer stretches, and an accurate consensus can be constructed by filtering for these errors.*

4.1.3 Short-read transcriptome analysis

The quality-filtered short reads were mapped to the respective host and viral genomes using Tophat [67] splicing-aware aligner for the random-hexamer primed library and Bowtie2 [68] for the PA-Seq library. In both approaches, a primary alignment is created with the Burrows-Wheeler alignment algorithm, as implemented in Bowtie2. End-to-end alignment with default seed lengths of 20 proved efficient for mapping, with deletions and insertions (indels) of 5-5 bases considered, although such long indels were rarely observed. A notable exception was a longer variant of 12 bp in the *ul27* gene, which occurred in ~50% of sequenced viral gDNA fragments, and as such accounted for the greatest genomic variation, with its functional characterization awaiting further studies. In the case of random hexamer samples, splice site detection followed, by realignment of reads, using adjustments for insert sizes from 50bp to 5000bp. Splice sites were considered for further analysis, and included in the resulting list of junctions, if coverage was above 10 bases, with anchor sizes >5bp, and at maximum two mismatches present in a given set of anchors. PA-Seq peak detection was carried out using HOMER [70] in strand-specific mode, adjusting for the peculiarities of oligo(dT) primed samples, in that the peak slopes detection optimized for 5' CAGE-sequencing analysis was effectively reversed. Peak categories were assigned according to the following criteria: the presence of at least 2 consecutive adenine mismatches in at least 10 independent, non-overlapping reads at the PA site, and the presence or absence of a PAS in the 50 bp region upstream from the PA site. Visualization and annotation were undertaken in a variety of tools including the Artemis Genome Browser v15.0.0 [69], IGV v2.2 [70], Circos [71] and R [72]. GC bias in the alignments was inspected by using the Bioconductor R package. The prediction of canonical and non-canonical PAS was carried out using PolyApred support vector machine-based algorithm [73].

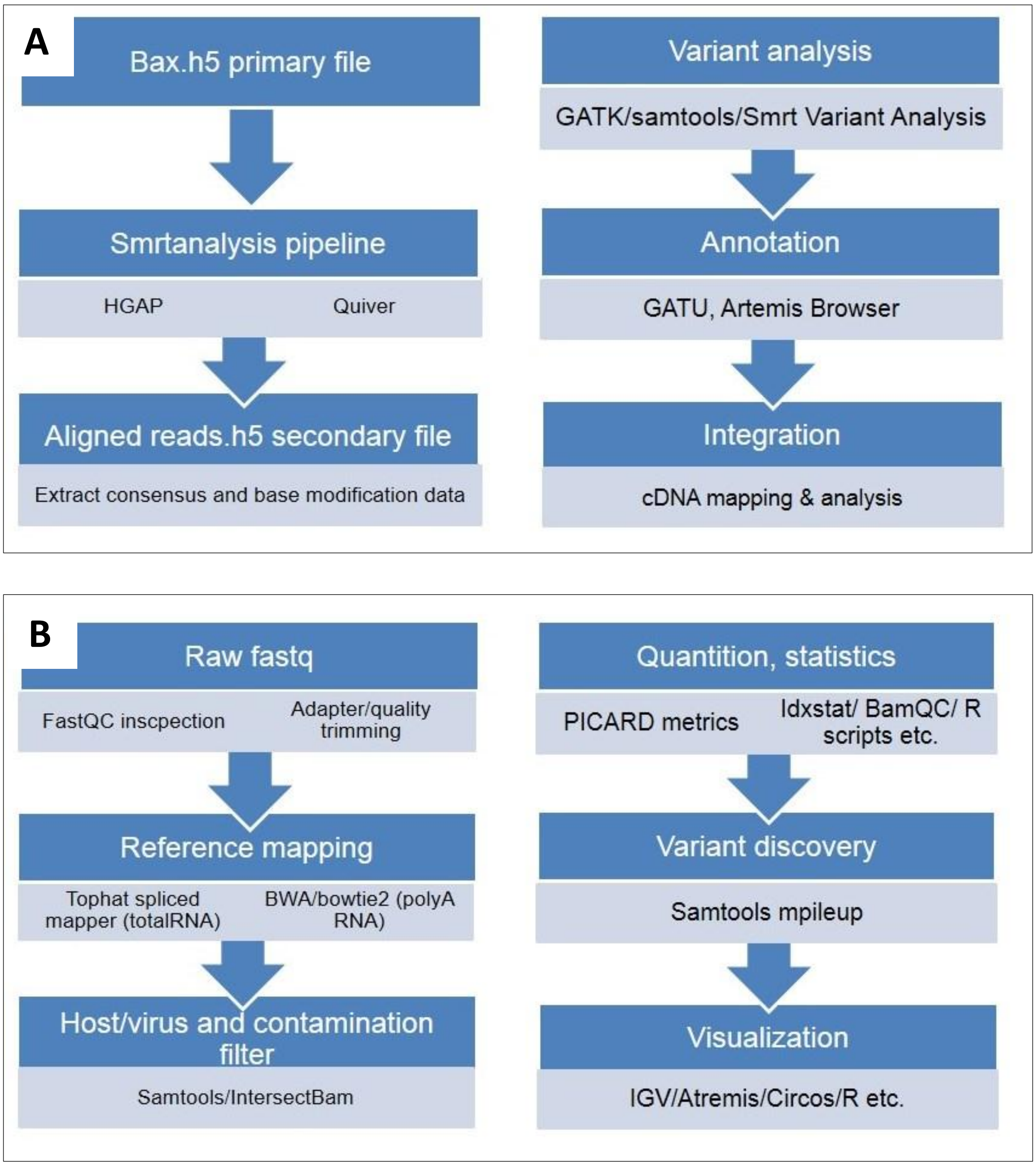


Fig.5. *A: Overview of long-read genome assembly. B: Overview of short-read sequencing data analysis*

4.2 Correlation between random-hexamer and PA-Seq results

The overall expression values of genes were estimated in RPKM (Reads per Kilobase per Million) values. This normalization method takes into account the length of the transcripts, so the read count from these loci are not skewed, and also normalizes counts by the library size (or total read count) factor. In PA-Seq, as reads are concentrated on the 3' ends of transcripts, the mean peak width value (in base pairs) was taken into account instead of transcript length. The highest differences arise from the UL7-9 cluster, where, although random hexamer reads are present, these form disconnected and biased fragments, and PA-Seq peaks better define the transcripts. On the other hand, 3' transcript ends are not well defined in the US region, where conservative polyA signals are sparsely found, which is reflected in the ratio of coverages from the different libraries. Among the genes with highest expression (*us1* and the CTO lncRNA) differences are also attributable to the extremely high coverage over very short transcripts, at which the RPKM normalization does not perform optimally. Altogether there is significant correlation of the expression values between the two approaches.

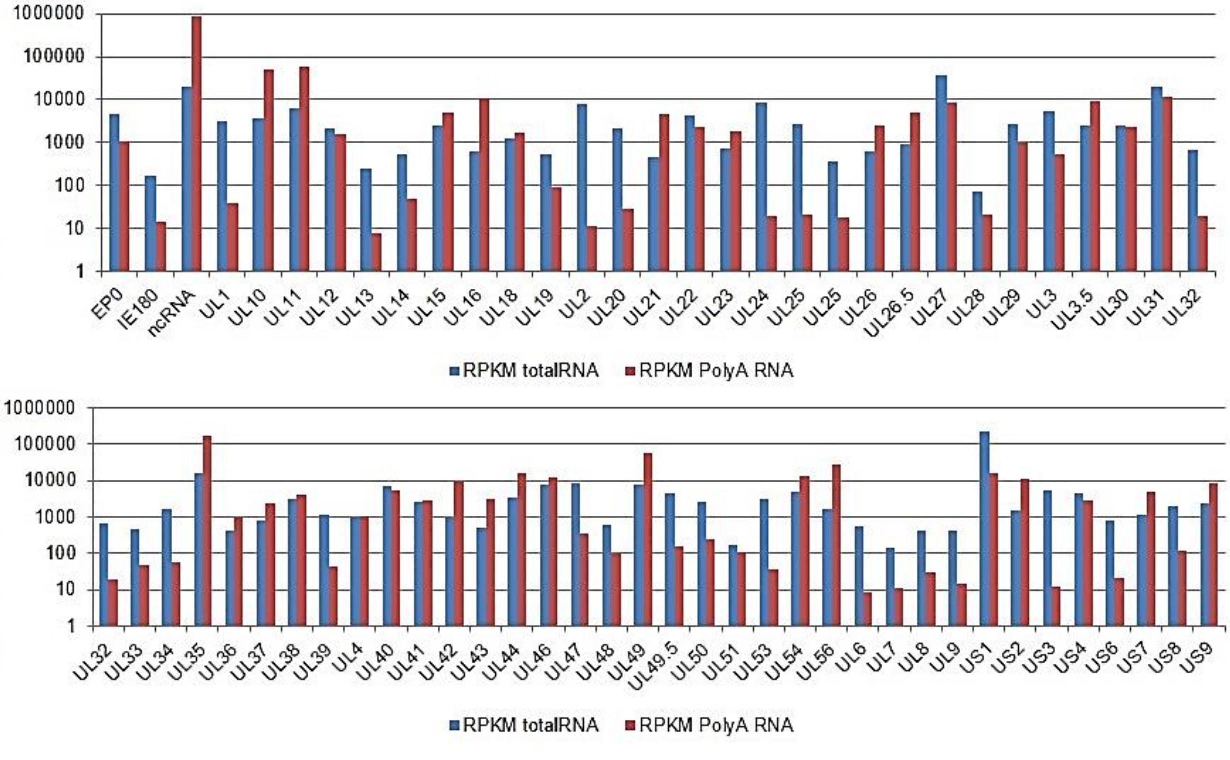


Fig.6. *Overall agreement between random-hexamer primed and anchored oligo(dT)-primed (PA-Seq) short-read sequencing libraries. Y-axis: gene expression values in Reads Per Kilobase per Million (correlation between libraries: r=0.724).*

4.3 PRV strain Ka genome map

Although PRV is a widely-studied organism among herpesviruses, and the complete viral genome of strain Kaplan had previously been sequenced [8], the available draft genomes are mostly poorly annotated, and contain several discrepancies, mainly in low-complexity regions. We carried out DNA sequencing in order to assess intra-strain heterogeneity, and to gather preliminary data for future epigenetic studies. Furthermore, the most well-annotated genome to date (NC006151.1) is a composite of six different strains, and as such could not directly be used in transcriptomic studies.

The especially long read lengths enabled the construction of highly overlapping contigs for assembly. The complete genome consists of 143,423bp, with 73.59% GC-content; sequence identity compared to other PRV strains available in GenBank ranges between 97-99%. The extent of intra-strain variability was lower than expected, with well-defined variable base positions only outside of protein-coding sequences, and occurring quite infrequently. An intriguing source of heterogeneity was present in a 12bp semi-palindromic repeat in the *ul27* gene, being absent in ~50% of the viral genome copies. Further studies ruled out the possibility of technical error, and showed that the repeat is specific to certain mutant strains of Ka. Protein-coding genes were predicted, and existing, matching annotations lifted by the GATU tool [75]. Manual annotation was used on genomic features such as replication origins and repeat motifs. Annotation of a previously unknown noncoding RNA Close to OriL (CTO), a novel splice site of the *ep0* gene, and new isoforms of 11 protein-coding genes were based on our short-read RNA sequencing results. Annotation of PRV miRNAs was created on the basis of previously published precursor miRNAs found in strains NIA-3 and Ea [58, 59], and as such were included in GenBank annotation for the first time.

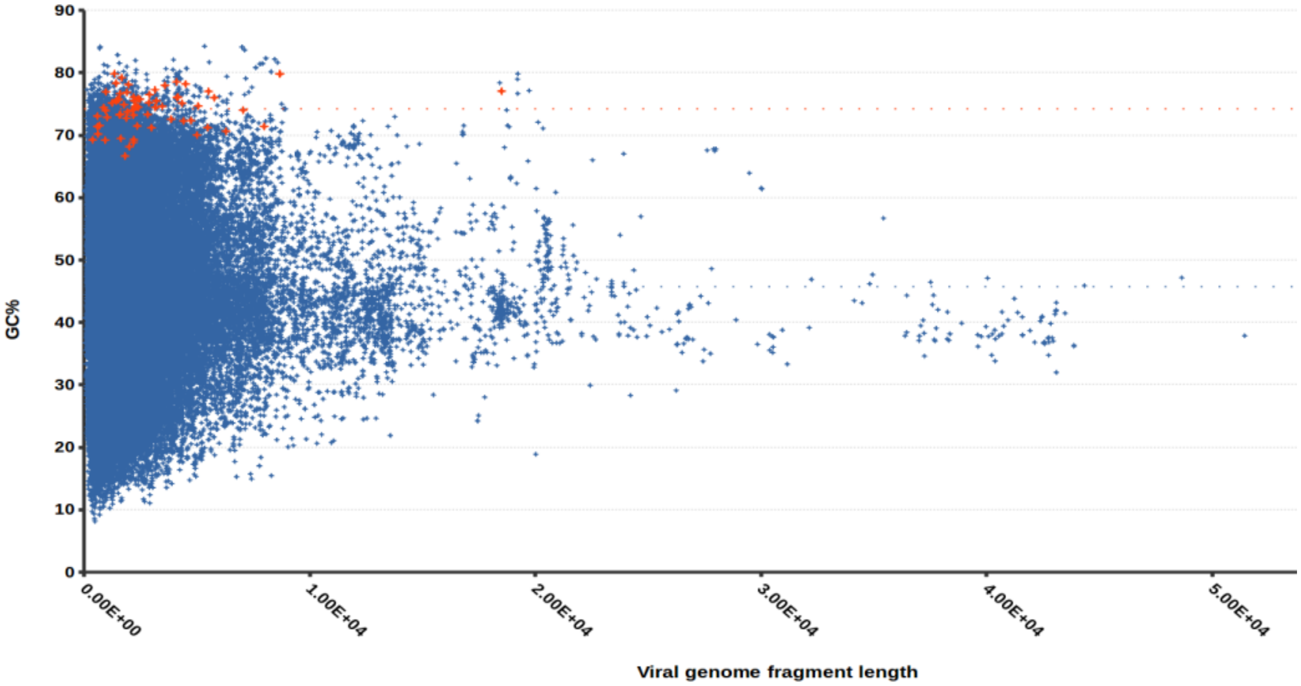


Fig.7. *The GC content of PRV complete genome and fragments (red) plotted against all available RefSeq viral sequences (blue).*

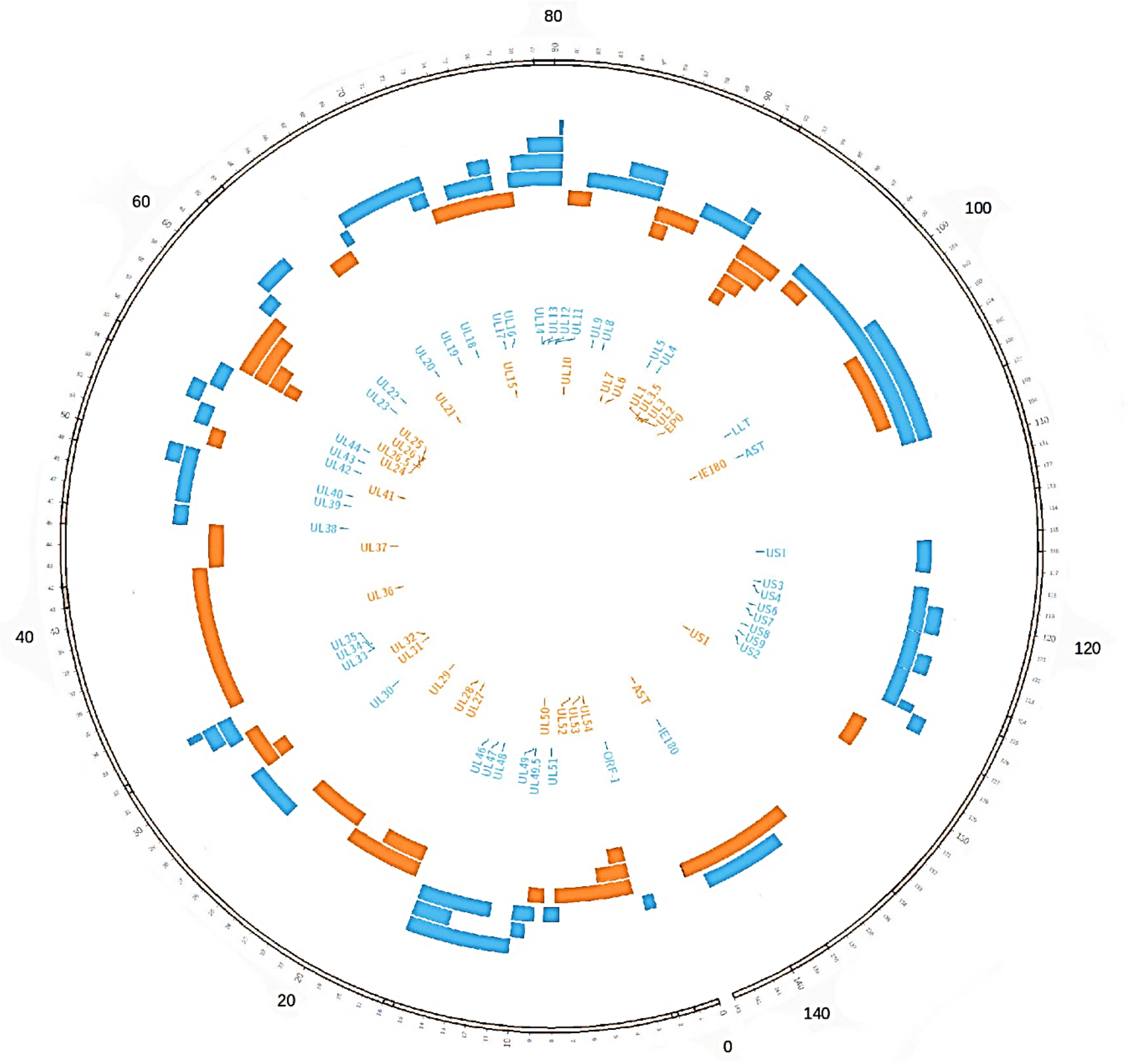


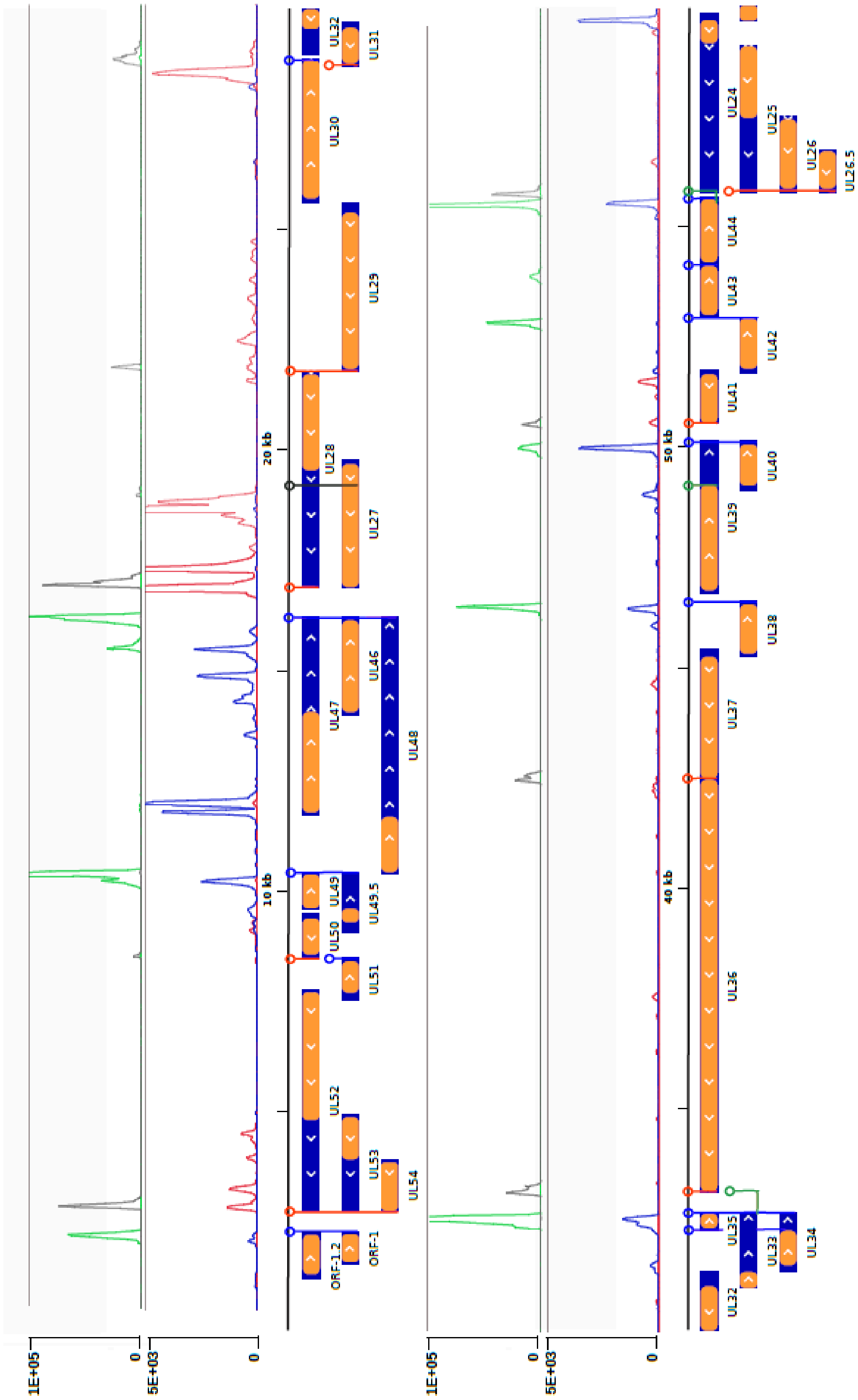
Fig.8. *The genome arrangement and nested gene clusters of PRV strain Ka illustrated on Circos plot. Blue: positive strand protein coding sequences, with latency-associated lncRNAs; orange: negative strand protein coding sequences.*

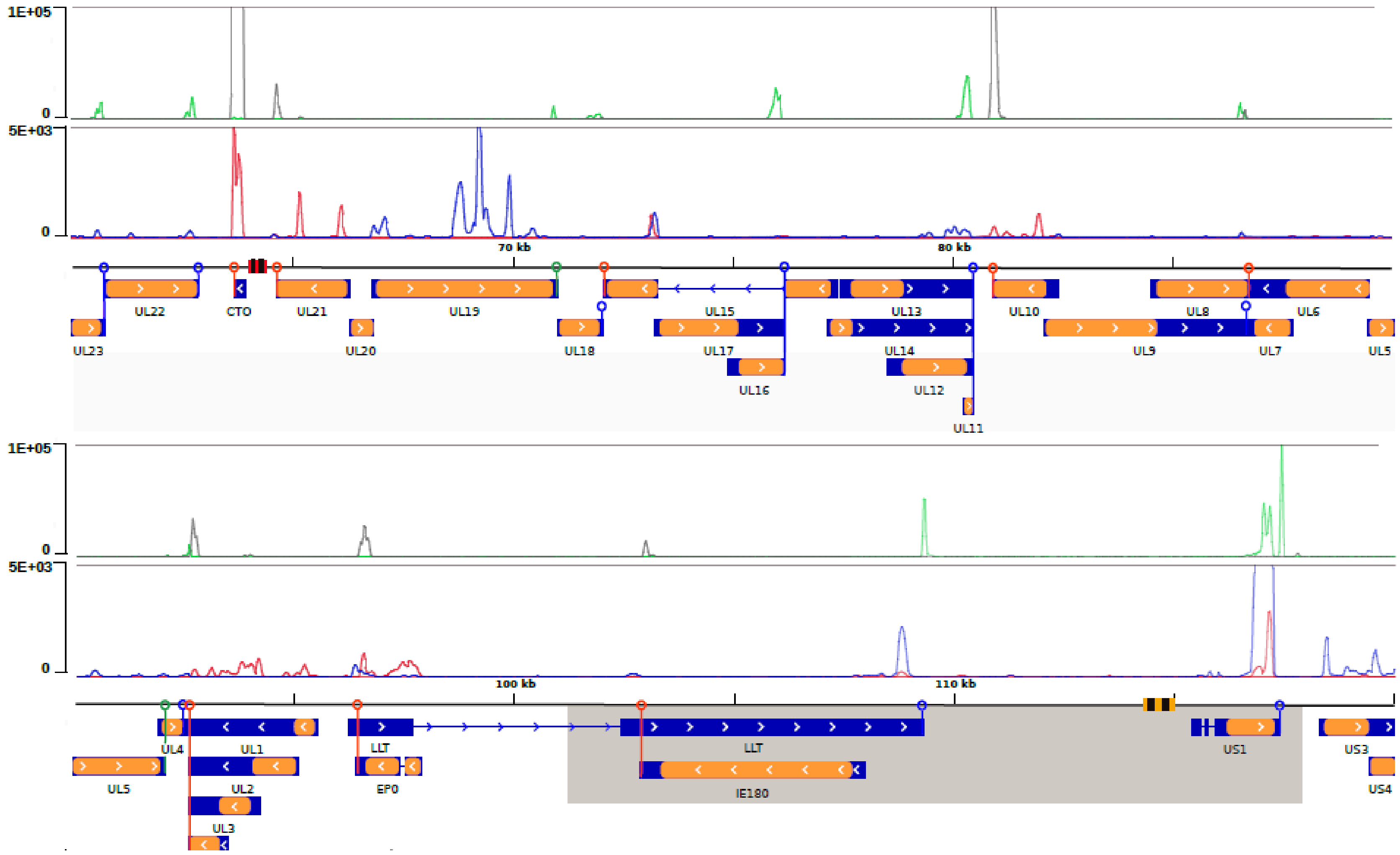
4.4 Assessment of the PRV transcriptome by total RNA sequencing and PA-Seq

For the investigation of the lytic PRV transcriptome, porcine kidney (PK-15) epithelial cells were infected with a high dose (MOI=10 pfu) of PRV strain Ka. Samples were gathered up to 24 h post-infection (p.i.) in order to capture most RNA species during lytic infection for sequencing library preparation. Both random hexamer-primed and oligo(dT)-primed libraries were prepared in order to assess total RNA and mRNA transcripts separately. In our modified polyadenylation sequencing (PA-Seq) protocol [55], total RNA was reverse-transcribed by using custom designed oligo(T10-VN) anchored primers containing standard Illumina strand-specific adaptor sequences. The two-nucleotide anchor sequence ensures the annealing of primers at exactly the PA site of mRNAs, providing considerably fewer reads that contain redundant adenine homopolymer stretches, with more useful sequence information resulting for the given depth of sequencing. PA-peaks occurred on both strands, mainly in accordance with previously existing ORF annotations, and also long non-coding RNAs, including the latency-associated transcript (LAT) and the long-latency transcript (LLT), which have been shown to be expressed during lytic infection in moderate amounts, despite previous expectations.

4.5 PRV transcriptome profiling

As anticipated, nearly the complete viral genome was transcribed, with the exception of short intergenic repetitive sequences, and longer spans of low-complexity sequences in the large internal and terminal repeat regions. Transcripts with the highest cumulative expression in the mixed time-point samples were the *us1* gene (RPKM = 2.32×105, total RNA library), encoding the *icp22* homolog Rsp40 immediate-early transactivator; followed by the novel lncRNA, CTO (RPKM = 1.6×10^6 in the total RNA library) Transcription at insulator sequences was observed only in two convergently oriented gene pairs, *ul44-ul26* and, in a less pronounced fashion, in *ul35-ul36*. The extent to which leaky transcription traverses the intergenic repeat boundaries is 109 bp and 443 bp, respectively, indicating alternative transcript termination. On the other hand, non-transcribed, repetitive regions were markedly present between ORF-1 and *ul54; ul46 and ul27; ul40* and *ul41*; and *ul11* and *ul10*. No expression was detected in these boundary regions, indicating their mechanistic role involved with the transcriptional machinery of the virus.


1E+05
0
5E+03
0
10 kb
20 kb
ORF-1.2
ORF-1
UL54
UL53
UL52
UL51
UL50
UL49.5
UL49
UL48
UL47
UL46
UL28
UL27
UL29
UL30
UL32
UL31
1E+05
0
5E+03
0
40 kb
50 kb
UL32
UL33
UL34
UL35
UL36
UL37
UL38
UL39
UL40
UL41
UL42
UL43
UL44
UL24
UL25
UL26
UL26.5


1E+05
0
5E+03
0
70 kb
80 kb
UL22
CTO
UL21
UL19
UL15
UL13
UL10
UL8
UL6
UL23
UL20
UL18
UL17
UL14
UL9
UL7
UL5
UL16
UL12
UL11
1E+05
0
5E+03
0
100 kb
110 kb
UL4
UL1
LLT
LLT
US1
US3
UL5
UL2
EP0
IE180
US4
UL3

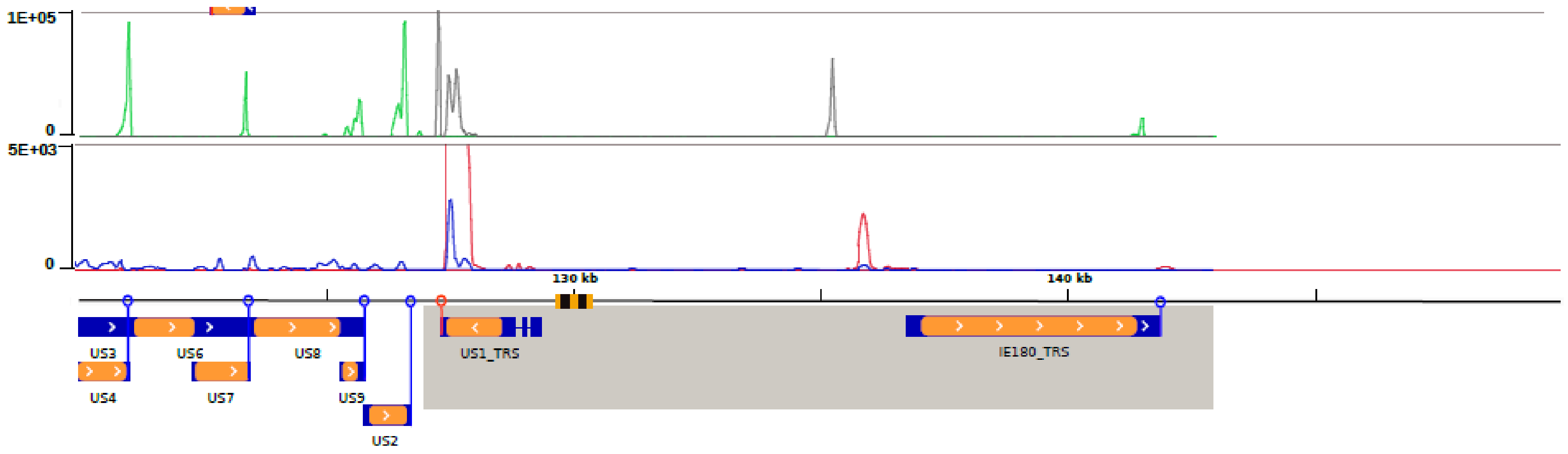


Fig.9. *Transcriptional map of the PRV genome identified by total RNA sequencing and PA-Seq. Genetic map: orange: coding sequences, blue: transcripts, red striped rectangle: OriL palindrome, yellow striped rectangles: OriS palindromes, grey: internal and terminal repeat regions, blue circles: PA site on + strand, red circles: PA site on –strand, green circles: alternative PA site on + strand, black circles: alternative PA site on –strand. Expression levels (in coverage per base): upper box: PA-Seq expression, green: +strand read coverage, black: −strand read coverage, lower box: totalRNA sequencing, blue: +strand coverage, red: −strand coverage*

4.6 Novel RNAs

Through the use of PA-Seq and random hexamer sequencing, transcription of the hypothetical ORF1.2 [75] was evidenced by marked expression in both libraries, also involving 5′ upstream regions. Single-base localization of the transcription start site is complicated by the presence of several repeats in the genomic sequence in the interval 730–960 bp, and thus definitive transcription start sites (TSS) were only assigned by the use of long-read cDNA sequencing in subsequent experiments.

The previously unknown polyA+ non-coding RNA, CTO ("Close to OriL") located between genes *ul21* and *ul22* and spanning 286 bp (Fig.10.) proved to be among the most abundant transcripts, on par with UL1, in all samples. In subsequent studies it was revealed that besides the highly expressed isoform, CTO also possesses considerably longer isoforms, which are the products of various read-throughs, and thus the 286 bp-long transcript variant was designated CTO-S ("CTO-Short"). Other length variants include CTO-L, a readthrough product originating from the *ul21* gene promoter, and an alternative PA site, about 120 nucleotides downstream from the main PAS. A putative CTO-M transcript was decisively detected only in following studies reported in publication IV. As the TSS of CTO-M is near the *ul21* polyA signal, the PAS of the downstream gene might act as a promoter in the opposite direction, as this phenomenon is observed especially in low-expressed transcripts. This alternate usage of the PAS might also be controlled by the flanking regulatory sequence elements, although the unusually high GC-content of PRV makes it difficult to discriminate less conserved motifs, as detailed in section 4.7. An interesting aspect of the novel lncRNA is the genomic context. The predicted promoter of CTO-S is a well conserved TATAA motif within 20-25bp upstream of the TSS, also neighboring an Oct-1 binding site commonly present in the virus, and shown to affect viral DNA replication in adenoviral infections in vivo [76]. In microorganisms, the third-position GC content in a given genome is used with great precision to delineate protein-coding sequences. Indeed, in most microbial and viral species, the third-position GC-content correlates well with the ORFs [77]. However, throughout the span of CTO-S, GC-base composition forms a synchronous peak in all reading frames; an irregular composition which is not present in the genomes of any of PRVs close relatives- bovine, equine, canine herpesviruses, *Herpes simplex* or *Varicella zoster.* PRV strains Becker, Bartha and HeN1 however show >99 % sequence similarity with strain Kaplan in the CTO-S genomic region, and thus express the lncRNA with high probability. Considering that the lncRNA is in close proximity of the OriL, knock-out experiments are difficult to carry out, although the function of the highly expressed transcript may be hypothesized as a regulator of DNA replication. It was also shown in RT-qPCR experiments that all isoforms of CTO share late transcriptional kinetics, which is in agreement with their hypothetical role in replication.

CTO-S RNA was detected by using traditional Northern blot analysis. Due to the very low copy number, we could not detect CTO-L by Northern blot analysis; however, the existence of this transcript was verified by four independent techniques (PacBio PA-Seq, two Illumina RNA-Seq methods and Real-time RT-PCR) in subsequent experiments. Sequence analysis of CTO by using the pre-microRNA hairpin prediction tools miRNAFold [78] and miPred [79] yielded negative results in each case. Moreover, previous studies of the miRNA expression in PRV in both porcine dendritic and epithelial [58, 59] cell lines failed to detect miRNAs from the genomic region of CTO.

A short, 3′-overlapping antisense non-coding transcript (termed SANC) was also detected adjacent to the PA site of the *ul21* gene, near OriL (64558–64674 on reference genome KJ717942.1), with an expression of RPKM = 1.67×10^3 in the total RNA library, the highest non-coding antisense expression in our short-read samples that is not caused by gene cluster overlaps.

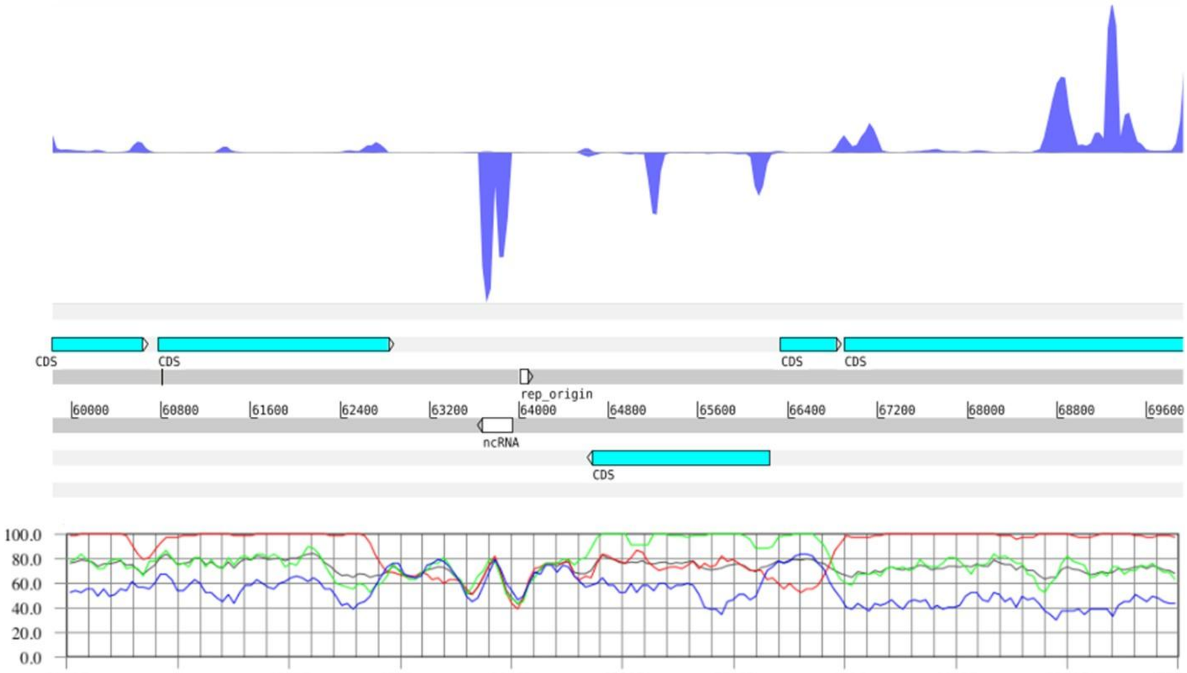


Fig.10. *Expression of the CTO polyA+ lncRNA in the random-hexamer primed library. Blue histogram: read coverage per base position on + and – strands; blue arrows: coding sequences; white arrows: CTO-S RNA and origin of replication; line graph: the third-position GC content per reading frames, showing the irregular GC-composition of the non-coding RNA.*

4.7 Alternative 3' UTR detection

The position of polyA tails on mRNA and polyA+ lncRNA transcripts can be accurately identified between and within gene clusters by using the PA-Seq method, with the additional benefit of a very high coverage of individual genes relative to sequencing depth. Anchored oligo(VN-T10) primers (V=A,G or C) provide greater efficiency, as the length of polyA tails may be well over the length of short reads used in sequencing; thus, with conventional oligo(dT) primers, a higher percentage of reads is lost to only containing adenine bases. In addition, we have detected robust signal from genes that were indistinguishable from background in random-hexamer primed libraries as well as gene-specific RT-PCR experiments. These include genes of the *ul7-ul9* convergently oriented cluster.

The most highly abundant transcripts (CTO, US1, UL31, and UL35) were in accordance with the random hexamer-primed data. The most widely used polyadenylation signal (PAS) in eukaryotic organisms is the AAUAAA canonical motif, which is usually found 10-30 nucleotides upstream of the YA cleavage site, from where the addition of adenines through polyadenine polymerase (PAP) initiates. In line with expectations, the analysis of viral PAS indicated that in ~90% of cases, the strong polyadenylation peaks correspond to the AAUAAA signal., while the second most widely used motif is AUUAAA, in ~10% of all cases. (Fig.11). In eukaryotes, two of the most commonly recognized 3' regulatory motifs are USE (upstream sequence element) and DSE (downstream sequence element), U-rich and U/G-rich sequences, respectively. These are known to be required for the precise cleavage of pre-mRNA, especially in the absence of a conserved PAS [80]. Although it is known that the USE element lies upstream within ~30bp of the cleavage site, while DSE is found >20 nucleotides downstream, the extreme GC-content of PRV complicated the distinction of such motifs from the surrounding sequences, albeit PRV exhibited a number of non-canonical PA sites, mainly in the *ul28* gene and the US region. Human studies reveal that when multiple PA sites are present for a given gene, the most distal tends to use the canonical AAUAAA signal, while the proximal signals vary considerably from this consensus, essentially correlating signal position with usage frequency. Only six genomic positions containing the canonical AATAAA sequence were unused in our PA-Seq samples, 3 motifs residing inside coding sequences (+9072-9077; −52929-52934; −78490-78495) and one motif located directly upstream of the *us3* gene (+118308-118313), and not corresponding to any viral transcript. The remaining signal was present in two copies, (−117738-117743) and (+126996-127001) in the inverted repeats. On the other hand, canonical PAS that were previously considered inactive demonstrated pronounced polyadenylation peaks, providing alternative transcript termination sites in genes *ul35, ul44* and *ul22*, with PAS that were previously considered inactive. The usage frequency of the distal PA site was at least an order of magnitude lower than those of the proximal ones in both cases. Although the PAP is thought cleave at an exact YA sequence motif, a slight wiggle in the nucleotide position of excision was observable in all PA-peaks, with the extent of 5-10 bp; this is in accordance with recent findings in eukaryotes [81, 82]

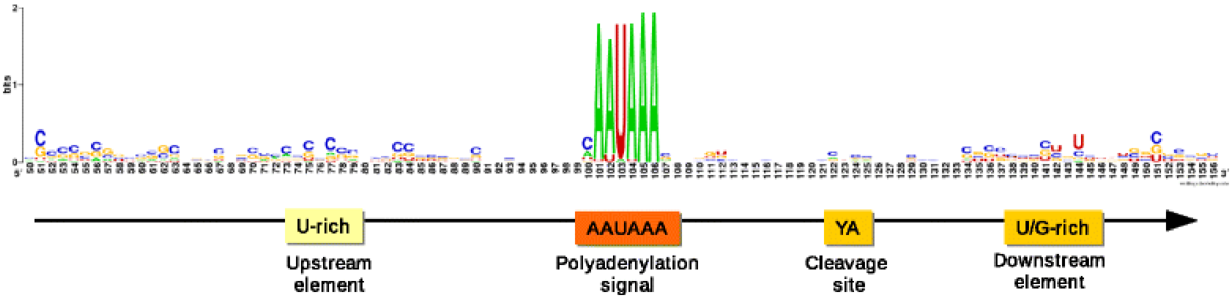


Fig.11. *The organization and nucleotide distribution in the PAS regions. Letter height represents nucleotide frequency in the sequence logo.*

The PA-Seq method additionally revealed polyadenylation peaks in transcripts encoded by upstream genes of tandem gene clusters. These included the polyadenylation of UL19. This transcript has previously been detected in strain Indiana-Funkhauser [83], with the non-canonical PAS ATATAAA; in our PA-Seq samples, we have confirmed the active use of this site in strain Ka. A similar arrangement was found in *ul28*, although no conservative PAS was detectable upstream of the well-defined PA-peak at base position 18960. Though PA peaks within the clusters of the US region were markedly above the background signal and correlated well with the coding sequence boundaries, these signals were several orders of magnitude weaker than the commonly observed polyA peaks, making them difficult to validate. The tandemly oriented UL4 transcript has been hypothesized to be 3′ coterminal with UL5 [11], as the canonical PAS directly downstream of UL5 is inside the UL4 ORF. However, PA-Seq peaks were found at the 3′ ends of both genes, showing that the PAS of UL5 is also active.

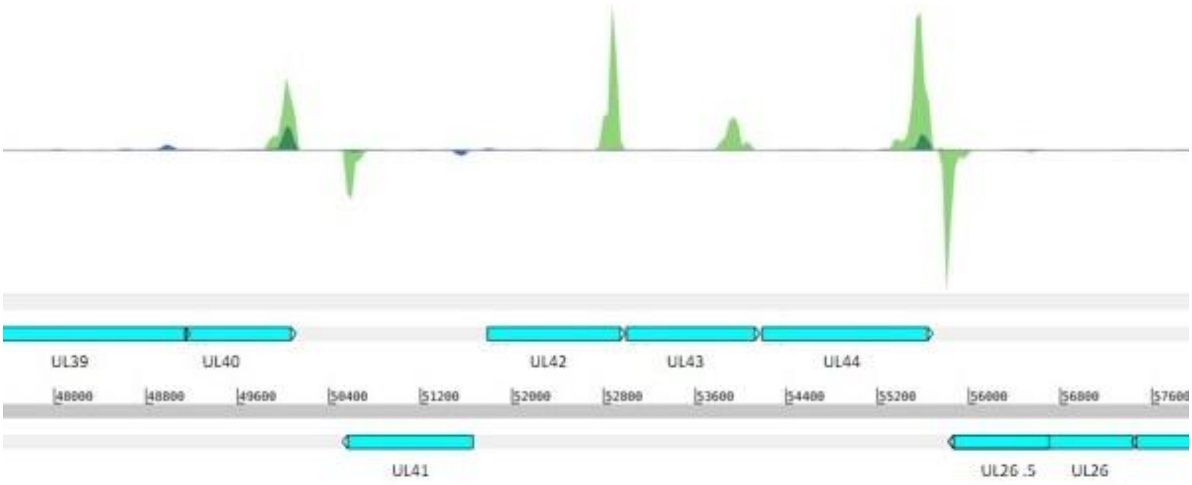


Fig.12. *Alternative polyadenylation peaks in gene clusters. Blue histogram: random hexamer library coverage; green: PA-Seq coverage; blue arrows: coding sequences. The above diagram exemplifies the difference in dynamic range between random-hexamer primed and PA-Seq libraries, as normalization is omitted in the visualization. The 3' ends of transcripts are marked by well-defined spikes, which are easily detected and quantitated by peak analysis software such as HOMER. The gene map gives examples for PA-peak distributions in single genes (ul41), polycistronic genes with 3' coterminal polyadenylation (ul39-ul40, ul25-ul26) and a tandem cluster of monocistronic genes (ul42-ul44).While there is no transcription in the intergenic sequence between UL40 and UL41 mRNAs, the UL44-UL26 boundary is an example where alternative polyadenylation traverses the intergenic repeat, resulting in a secondary PA-peak for UL26 (not shown due to scale)*

Detected splice sites			**Alternative polyadenylation**			**Convergently overlapping gene clusters**		**Divergently overlapping genes**		**Tandem gene clusters**
Gene	**Donor site**	**Acceptor site**	**Gene**	**Alternative polyadenylation signal**	**Coordinate**					
UL15	D −76165	A −73285	UL35	AAUAAA	33133–33138	UL51	UL50	UL52	UL51	UL52-UL54
US1	D +115592	A +115713	UL44	AAUAAA	55768–55773	UL30	UL31, UL32	UL50	UL49.5	UL48-UL46
	D +115766	A +115921	UL22	N/A*	63624	UL33, UL34, UL35	UL36	UL29	UL30	UL31-UL32
US1	D −129158	A −129037	UL19	AUAUAAA	71005–710011	UL44	UL26.5, UL26, UL25, UL24	UL32	UL33	UL33-UL35
	D −128984	A −128829	UL28	N/A*	18960	UL8, UL9	UL6, UL7	UL37	UL38	UL39-UL40
EP0	D −97480	A −97389	UL5	AAUAAA	92065–92070			UL41	UL42	UL24-UL26.5
	D −97528		CTO	N/A*	63538			UL24	UL23	UL17-UL16
								UL21	UL20	UL14-UL11
								UL15	UL14	UL9-UL8
								UL10	UL9	UL7-UL6
								UL6	UL5	UL1-UL3.5
										US3-US4
										US6-US7
										US8-US9

*No prediction available for canonical or non-canonical polyA signal using PolyApred server

***Table 2:** Transcriptional overlaps, splice sites and alternative polyadenylation in the PRV genome*

4.8 Splice sites in the PRV transcriptome

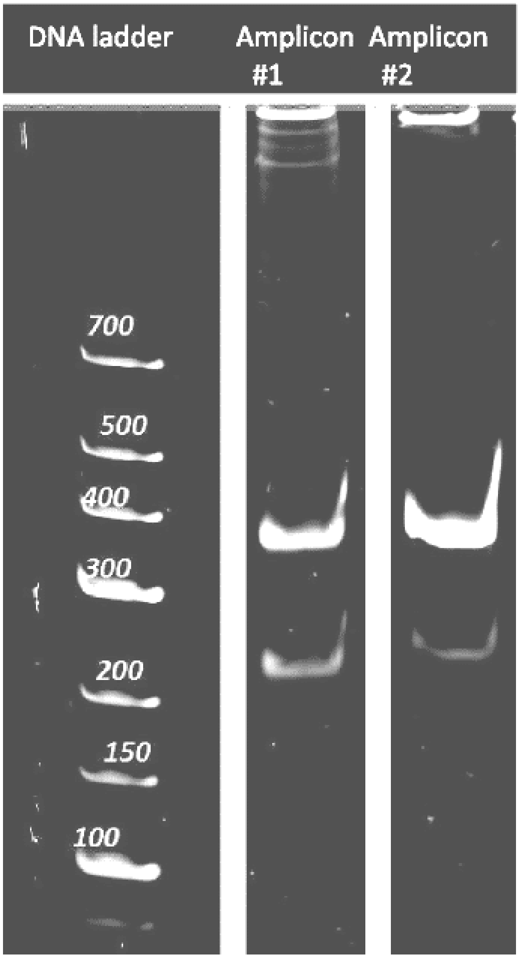


***Fig.12.** RT-PCR validation of the ep0 splice site*

For splice site analysis, total RNA reads were aligned to PRV strain Ka genome KJ717942.1. All possible splice donor and acceptor sites were considered, with a lower bound of at least 10 supporting reads, and an anchor of at least five nucleotides. Through the exclusion of low-coverage junction candidates, artifacts possibly occurring due to mispriming or template switching during amplification steps could be neglected. The initial set of splice acceptor and donor sites contained 97 candidate splice junctions, with 49 sites above the threshold coverage. This set contained several permutations of the US1 3′ UTR splice junctions, which were screened for the presence of short anchor regions and high mismatch ratios within these anchors. After screening for anchors of <5 bases, consistent splice junctions were readily identifiable. The remaining, high-coverage splice sites are denoted as follows: (D + 10000^A + 12000), with D denoting the donor site, A the acceptor site, and +/− the DNA top and bottom strand, respectively, along the coordinates of the splice junction. Splice sites have previously been characterized in the protein-coding region of UL15 [84, 85] (D −76165^A −73285), and in the 3′ UTR of US1 [86] (D +115592^A +115713; D +115766^A +115921), present in both terminal and internal repeats, while one site in the non-coding RNA LLT [87] (D +97765; A +102403) was expressed at an insufficient level for accurate splice site identification. A low percentage of reads also mapped outside the assigned acceptor and donor sites. A novel alternative splice site was characterized in the *ep0* gene [88], the homolog of HSV-1 *icp0*, which is also a spliced gene, but expressed in the immediate-early class in HSV-1. The newly characterized *ep0* alternative splicing consists of two potential donor sites at (D −97480) and (D −97528) and the acceptor site at (A −97389). While the splice junction formed between the acceptor and proximal donor sites conforms to the rule of GT/AG nucleotides comprising ~99 % of junctions in eukaryotic organisms [89], the junction formed with the distal donor site contains GT/CG bases. Experimental validation of the novel splice site has been carried out by RT-PCR, using two primer sets (Table 1) designed approximately 100 bp upstream and downstream of the splice site, followed by polyacrylamide gel electrophoresis. The experiments confirmed the presence of the novel isoform during lytic infection robustly after visualization (Fig.12). The disordered nature of the protein hinders *in silico* predictions of the altered structure of the shorter form, however, the zinc-finger domain of *ep0* is not contained in the intron, indicating that a similar protein function is likely to be retained.

4.9 Transcriptional overlaps and TI

The genome of PRV is an ideal candidate for studying TI, as it is among the larger viral genomes, with >70 coding and non-coding genes forming overlapping gene clusters. In convergently oriented clusters, more extensive overlaps include coding regions of the opposite genes, potentially giving rise to TI between the interacting partners [7]. An example of such a relation is between UL30 and UL31, with a tail-to-tail overlap of 80 nucleotides. Here, the expression of UL30 mRNA exceeds that of UL31, with considerable antisense expression over the latter gene, possibly due to transcriptional read-through from UL30. As anticipated, convergent genes with more than ~45 bp separating their respective PA signals do not demonstrate detectable transcriptional overlap, ranging from UL18-UL15 (45 bp) to UL46-UL27 (632 bp), while convergent gene pairs in closer proximity exhibit longer 3′ UTR overlapping regions. The various overlaps between the viral genes are presented in Table 2. It is the hypothesis behind TINs that the transcription of these neighboring genes might affect each other in a mechanistic fashion, forming self-regulatory networks and a novel layer of complexity in the genome. We assessed the various transcript overlaps, including parallel (tandem), divergent and convergent overlaps (Fig.13, Table 2). Most of the PRV genes are organized into tandem gene clusters producing polycistronic RNAs (Table 2).

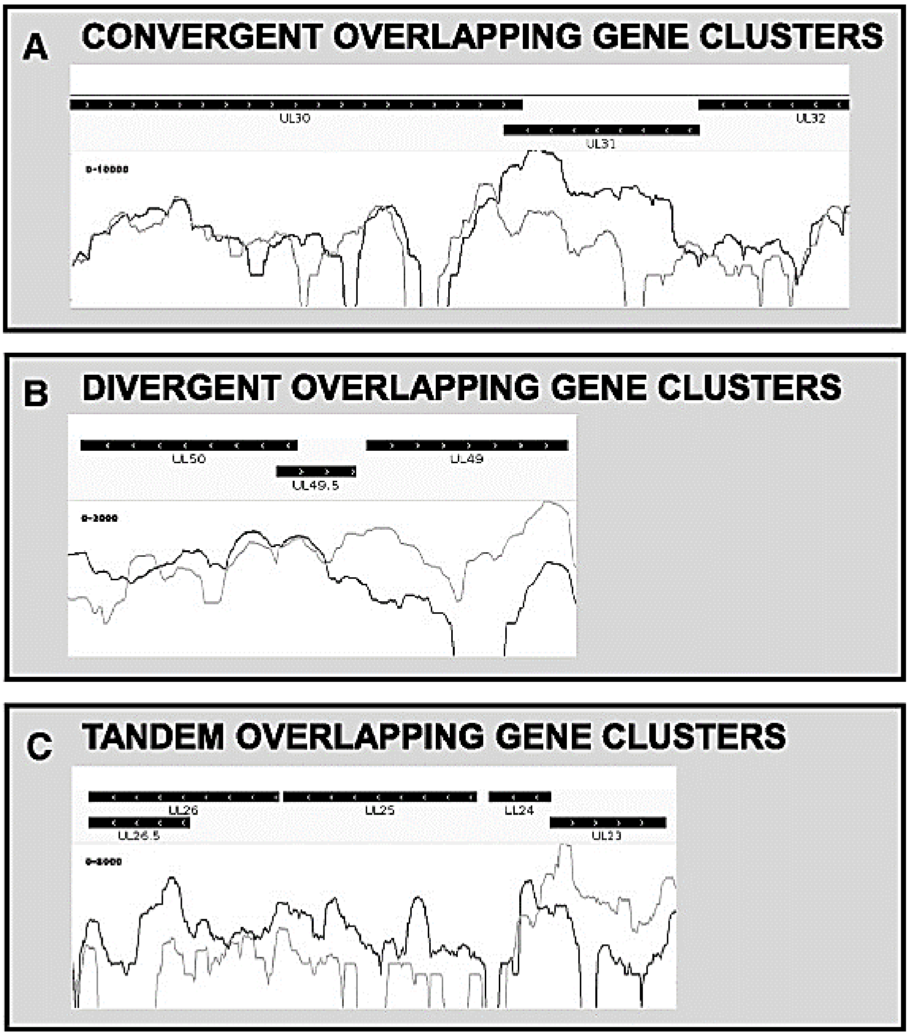


Fig.13. *1: Categories of transcriptional overlaps plotted along the PRV strain Ka genome sequence. Legend from perimeter to center: light brown: coding sequences on both + and – strands; dark brown: non-transcribed, or sparsely*

transcribed repetitive sequences between gene clusters; gray histogram: GC content in averaged windows; red line: 50% GC; colored ribbons: areas of transcriptional overlap corresponding to categories; purple: 3' coterminal tandem overlaps; blue: convergent overlaps; green: divergent overlaps; orange: full overlaps; red: tandem overlap without 3' coterminus

*2: Convergent (**a**), divergent (**b**) and tandem (**c**) overlaps in the PRV genome, as shown by random-hexamer primed samples. Extensive transcriptional overlaps are frequent throughout the condensed viral genome. Black boxes: coding sequences, white arrows: gene orientation, grey line graph: positive strand expression, black line graph: negative strand expression*

An interesting feature of organization is that all of the upstream genes of the clusters end within the downstream genes. Similarly, the divergent gene pairs overlap in every case. Thus, the translation of the downstream genes would be hindered in eukaryotic organisms, a problem for which a proposed mechanism is the wide use of Internal Ribosome Entry Sites (IRES) and ribosome skipping (90). However, the genetic arrangement and the observed, wide-spread read-throughs between genes indicate a basis of inter-gene regulation through TI, which has been observed in a range of model organisms previously. Theoretically, the overlaps between convergent, divergent and within tandem gene clusters may be explained by the restriction of the viral genome length. However, since these overlaps are not too extensive, they probably provide a regulatory mechanism for transcription. The distant convergent genes are separated from each other by repetitive sequences, indicating a mechanism with a likely function for the prevention of transcriptional collisions. In the *ul35* and *ul44* genes, these alternative termination sites traversed intergenic repetitive regions, previously considered to be transcriptional barriers.

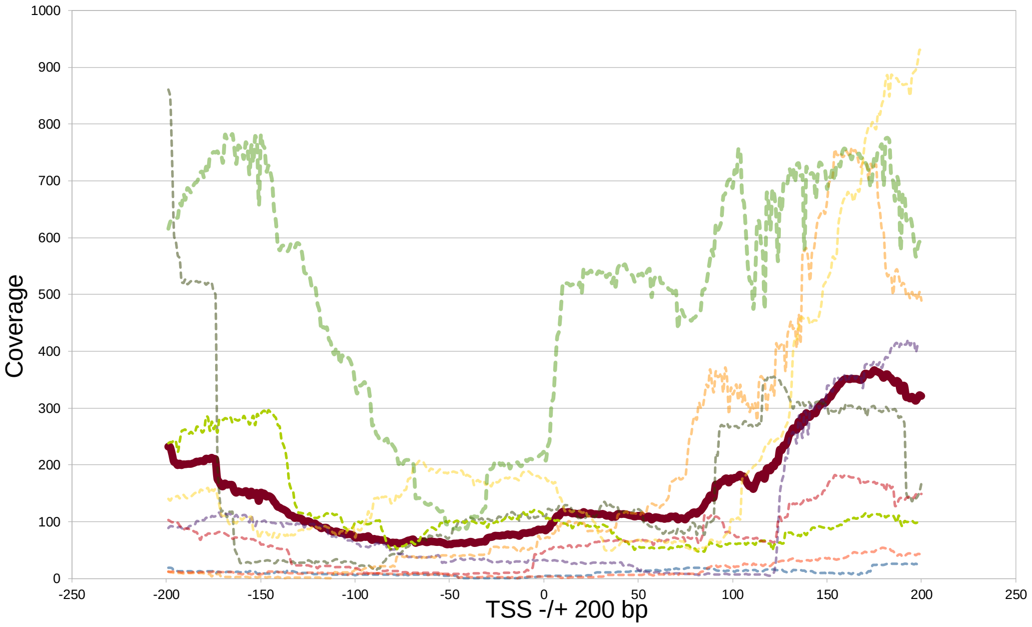


Fig.14. *The distribution of coverage across all bidirectional promoters in the PRV genome, showing extensive transcription in both directions from the promoters. Dotted lines: coverage peaks per individual promoter, red line: mean expression across bidirectional promoters.*

This finding indicates that low-frequency “leaky” transcription occurs more often than anticipated in PRV. Although the function of these PA sites is unknown, it is noteworthy that a highly similar arrangement was present between convergent gene clusters *ul9-ul8* and *ul7-ul6*, with the difference that a strongly repetitive sequence resembling the above-mentioned intergenic repeats in both length and base content, was found within the comparatively long 3′ UTR of UL7. This gave the sole example for a third type of transcriptional boundary between convergent clusters, as illustrated in Fig.15.

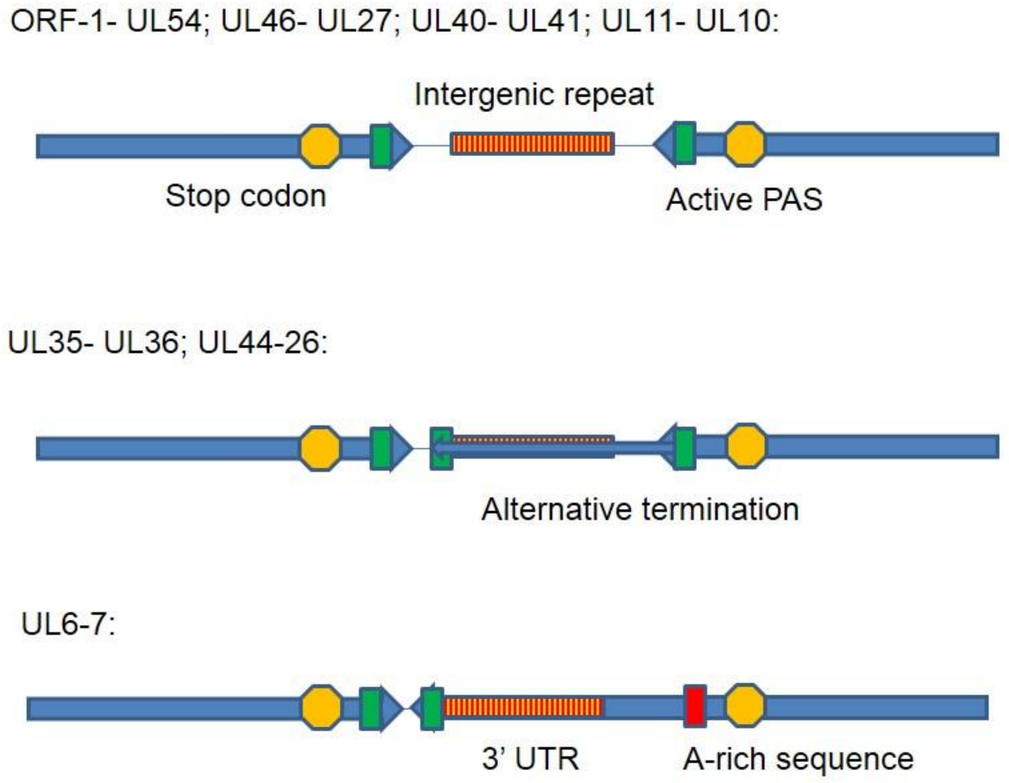


Fig.15. *Types of short intergenic repeats as transcriptional insulators between convergently oriented gene clusters in PRV.*

5. CONCLUSION

5.1 gDNA sequencing

In our series of experiments, the survey of Pseudorabies virus strain Kaplan genomic DNA and total RNA content was carried out for the first time using high-throughput sequencing methods. Through the use of complementary sequencing technologies and independent validation methods, detailed information was generated on the genomic context in which further, more specific experiments were devised to investigate questions of viral transcriptional regulation and the roles of transcriptional interference in the extremely condensed and temporally highly regulated viral genome.

The gDNA sequencing of PRV strain Ka using the cutting-edge PacBio SMRT sequencing technology yielded an updated genetic map of the strain, providing an alternative to the previously widely used reference genome which consisted of mixed sequences of six viral strains. Using the information gathered from transcriptome sequencing, and previously available sources regarding regulatory motifs and microRNAs, rich and up-to-date annotation information was also deposited in the corresponding GenBank record, which, through sequence homology, might be useful in annotating and understanding the frequently isolated strains from farms mainly in Asia, but around the world as well. [91,92,93]

On a technical level, the study has also been among the first which employed the SMRT technology in sequencing a viral genome, and as such, provided opportunities for methodological improvement, notably in the systematic errors arising in high-GC content organisms, and also effective screening and removal of such errors. With these additions, the long-read sequencing technology has proven efficient and economic for the use in viral gDNA studies, as complete genomes can be assembled from a minimal number of contigs and very little cross-validation using the long reads.

5.2 cDNA sequencing

The catalog of PRV transcripts has been updated based on our cDNA sequencing experiments of mixed-timepoint lytic infections on the PK-15 epithelial cell line, providing a detailed view on the transcriptional landscape of the virus, and accurately defining 3' polyA+ RNA boundaries, which are key to the composition of the characteristic, polycistronic gene arrangement in the Herpesviridae family.

5.3 Splice sites

From random-hexamer primed sequencing, data emerged on a splice isoform of the key early transactivator ep0 gene, the functions of which require further investigations, as the disordered nature of the protein prohibits in silico modeling and predictions. Based on amino acid sequence, however, the zinc-finger domain of ep0 is not covered by the spliced intron, indicating retained function of the shorter isoform.

Further splice sites of in the PRV genome have been confirmed and accurately detected in us1 and ul15. The expression of latency-associated transcripts was also detected in our lytic samples at low levels, although the coverage did not enable accurate identification of the LLT splice site. Further, the presence of low-coverage, difficult to validate splice junctions was later confirmed via deep-sequencing isoform analysis reported in publication IV, and led to the identification of several lower-expressed mRNA isoforms and ncRNAs.

5.4 Polyadenylation landscape

Through PA-Seq, overall gene expression and transcript boundaries were assessed using highly efficient anchored primers, resulting in the detection of transcripts in the UL7-UL9 cluster, often missed by other PCR-based methods. The detailed organization of PRV gene clusters was also revealed, serving as a foundation for further evaluation of the transcriptional kinetics and regulatory mechanisms that may shed light on transcriptional interference networks. The usage frequency and distribution of both canonical and non-canonical polyadenylation signals was assessed, with sequence motifs corresponding highly to those identified in eukaryotic model organisms, but also distinct features arising from the compact and overlapping 3' UTR sequences and the polycistronic arrangement.

5.5 Novel transcripts

Recent results in herpesvirus research indicated a wide range of non-coding RNAs in clinically important pathogens such as KSHV, HCMV and EBV (REF). The ncRNAs described in lytic infections were of similarly high expression, often accounting to ~50% of the total transcribed RNA quantity. Further, the screening of PRV samples for miRNAs from both latent infection on neuronal ganglia and lytic phase in epithelial cells gave positive results [58, 59], with pre-miRNAs originating from the LLT intron. As such, the assessment of the totalRNA and polyA+ RNA fractions was both timely and promising, eventually resulting in the characterization of the rather short (268bp), but extremely highly expressed polyA+ lncRNA CTO-S, near the viral origin of replication. Lacking a direct genomic target in either the host or the viral genome, the initial hypothesis for the function of the gene is the regulation of viral DNA replication, a function which is in agreement with its late expression kinetics. Other novel transcripts include the low-expressed short 3'-antisense RNA SANC, also in the OriL region, and validation of the ORF1.2 hypothetical mRNA, along with several new 3'-UTR variations arising from alternative polyadenylation.

5.6 Transcriptional interference

The above findings indicated that key genomic features are present in PRV which may serve as the focus of transcriptional interference studies, including the extensive overlaps between various gene clusters, the location of ncRNAs and short intergenic repetitive sequences, and the newly discovered transcript boundaries within gene clusters. The evaluation of previously studied sites, such as the *ul30-ul31* gene pairs added qualitative information to the results of earlier RT-qPCR results.

ACKNOWLEDGEMENT

I would like to thank all my colleagues and staff of the Department of Medical Biology for their kind support.

Foremost I thank my supervisor, prof. Zsolt Boldogkői for continuous support, motivation and guidance throughout the graduate program, without which any of the presented work would have been impossible.

I am grateful for the expertise and help both in- and outside the laboratory of dr. Dóra Tombácz, Zsolt Csabai, István Prazsák, Marianna Ábrahám, Csilla Magyarné Papdi, Nándor Póka and Kata Filkor, and the administrative help from Gabriella Szilágyiné Teleki.

I would like to honor the memory of our late colleague, dr. Antal Mai.

Last, I give my special thanks and gratitude to my supporting family and my loving wife, Eszter.

Illumina short-read library preparation, sequencing and primary bioinformatic analysis of the samples was performed at the Genomic Medicine and Bioinformatic Core Facility of the University of Debrecen, Faculty of Medicine.

Pacific Biosciences SMRT-sequencing and library preparation were carried out in the Department of Genetics, Stanford University, laboratory of prof. Michael Snyder.

The research was supported by grant (SH/7/2/8) of the Swiss-Hungarian Cooperation Programme to Zsolt Boldogkői, and by CEGS NIH grant 1P50HG007735-01 to Michael Snyder.

REFERENCES

1. Aujeszky A: A contagious disease, not readily distinguishable from rabies, with unknown origin. Veterinarius 1902, 25:387-396.
2. Jun Cao, Korneel Grauwet, Ben Vermeulen, Bert Devriendt, Ping Jiang, Herman Favoreel, Hans Nauwynck, Suppression of NK cell-mediated cytotoxicity against PRRSV-infected porcine alveolar macrophages in vitro, Veterinary Microbiology, Volume 164, Issues 3–4, 28 June 2013, Pages 261-269, ISSN 0378-1135, http://dx.doi.org/10.1016/j.vetmic.2013.03.001.
3. Matthias J. Deruelle, Céline Van den Broeke, Hans J. Nauwynck, Thomas C. Mettenleiter, Herman W. Favoreel, Pseudorabies virus US3- and UL49.5-dependent and -independent downregulation of MHC I cell surface expression in different cell types, Virology, Volume 395, Issue 2, 20 December 2009, Pages 172-181, ISSN 0042-6822, http://dx.doi.org/10.1016/j.virol.2009.09.019.
4. Boldogkői Z, Reichart A, Tóth IE, Sík A, Erdélyi F et al: Construction of recombinant pseudorabies viruses optimized for labeling and neurochemical characterization of neural circuitry. Mol Brain Res 2002, 109 (1-2): 105-118.
5. Hao Y, Tian X-B, Liu C, Xiang H-B. Retrograde tracing of medial vestibular nuclei connections to the kidney in mice. *International Journal of Clinical and Experimental Pathology*. 2014;7(8):5348-5354.
6. Boldogkői Z, Bálint K, Awatramani GB, Balya D, Busskamp V, Viney TJ, et al. Genetically timed, activity sensor and rainbow transsynaptic viral tools. Nat Methods. 2009;6:127–30. doi: 10.1038/nmeth.1292.
7. Boldogköi Z. Transcriptional interference networks coordinate the expression of functionally related genes clustered in the same genomic loci. *Frontiers in Genetics*. 2012;3:122. doi:10.3389/fgene.2012.00122.
8. Tombácz D, Tóth JS, Petrovszki P, Boldogkői Z. Whole-genome analysis of pseudorabies virus gene expression by real-time quantitative RT-PCR assay. BMC Genomics. 2009;10:491. doi: 10.1186/1471-2164-10-491.

9. Virologie. Volume 7, Numéro 5, 319-28, septembre-octobre 2003
10. Travis J Taylor, Mark A. Brockman, Elizabeth E. McNamee, and David M. Knipe Frontiers in Bioscience 7, d752-764, March 1, 2002
11. Klupp BG Hengartner CJ, Mettenleiter TC, Enquist LW: Complete, annotated sequence of the pseudorabies virus genome. J Virol 2004, 78:424–440.
12. Pomeranz LE, Reynolds AE, Hengartner CJ: Molecular biology of pseudorabies virus: impact on neurovirology and veterinary medicine. Microbiol Mol Biol Rev 2005, 69(3):462- 500.
13. Backovic M, DuBois RM, Cockburn JJ, et al. Structure of a core fragment of glycoprotein H from pseudorabies virus in complex with antibody. *Proceedings of the National Academy of Sciences of the United States of America*. 2010;107(52):22635-22640. doi:10.1073/pnas.1011507107.
14. Arii J, Uema M, Morimoto T, et al. Entry of Herpes Simplex Virus 1 and Other Alphaherpesviruses via the Paired Immunoglobulin-Like Type 2 Receptor α . *Journal of Virology*. 2009;83(9):4520-4527. doi:10.1128/JVI.02601-08.
15. Peeters, B., et al., Pseudorabies virus envelope glycoproteins gp50 and gII are essential for virus penetration, but only gII is involved in membrane fusion. J Virol, 1992. 66(2): p. 894-905.
16. Roizman, B., and D. M. Knipe. , Herpes simplex viruses and their replication. 2001: p. 2399–2460.
17. Leelawong M, Guo D, Smith GA. A Physical Link between the Pseudorabies Virus Capsid and the Nuclear Egress Complex. *Journal of Virology*. 2011;85(22):11675-11684. doi:10.1128/JVI.05614-11.
18. Homa F, Huffman J, Toropova K, Lopez H, Makhov A, Conway J. Structure of the pseudorabies virus capsid: comparison with herpes simplex virus type 1 and differential binding of essential minor proteins. *Journal of molecular biology*. 2013;425(18):3415-3428. doi:10.1016/j.jmb.2013.06.034.
19. Ben-Porat, T., and A. S. Kaplan., Molecular biology of pseudorabies virus. 1985:p. 105–173.
20. Fuchs, W., et al., The UL48 tegument protein of pseudorabies virus is critical for intracytoplasmic assembly of infectious virions. Journal of Virology, 2002.

76(13): p. 6729-6742.

21. Tombácz D, Tóth JS, Petrovszki P, Boldogkői Z. Whole-genome analysis of pseudorabies virus gene expression by real-time quantitative RT-PCR assay. BMC Genomics. 2009;10:491.
22. York IA, Roop C, Andrews DW, Riddell SR, Graham FL and Johnson DC: A cytosolic herpes simplex virus protein inhibits antigen presentation to CD8 lymphocytes. Cell 1994, 77:525-535.
23. Gu H. Infected cell protein 0 functional domains and their coordination in herpes simplex virus replication. *World Journal of Virology*. 2016;5(1):1-13. doi:10.5501/wjv.v5.i1.1.
24. Suazo PA, Ibañez FJ, Retamal-Díaz AR, et al. Evasion of Early Antiviral Responses by Herpes Simplex Viruses. *Mediators of Inflammation*. 2015;2015:593757. doi:10.1155/2015/593757.
25. Juillard F, Tan M, Li S, Kaye KM. Kaposi’s Sarcoma Herpesvirus Genome Persistence. *Frontiers in Microbiology*. 2016;7:1149. doi:10.3389/fmicb.2016.01149.
26. Satoshi Taharaguchi, Tsutomu Kobayashi, Saori Yoshino, Etsuro Ono, Analysis of regulatory functions for the region located upstream from the latency-associated transcript (LAT) promoter of pseudorabies virus in cultured cells, Veterinary Microbiology, Volume 85, Issue 3, 22 March 2002, Pages 197-208
27. Taharaguchi S1, Yoshino S, Amagai K, Ono E.The latency-associated transcript promoter of pseudorabies virus directs neuron-specific expression in trigeminal ganglia of transgenic mice. J Gen Virol. 2003 Aug;84(Pt 8):2015-22.
28. Aleman N, Quiroga MI, Lopez-Pena M, Vazquez S, Guerrero FH and Nieto JM: Induction and inhibition of apoptosis by pseudorabies virus in the trigeminal ganglion during acute infection of swine. J Virol 2001, 75(1):469-479
29. Shearwin KE, Callen BP, Egan JB. Transcriptional interference – a crash course. *Trends in genetics : TIG*. 2005;21(6):339-345.

doi:10.1016/j.tig.2005.04.009.

30. Transcriptional Interference in Convergent Promoters as a Means for Tunable Gene Expression Antoni E. Bordoy, Usha S. Varanasi, Colleen M. Courtney, and Anushree Chatterjee ACS Synthetic Biology Article ASAP DOI: 10.1021/acssynbio.5b00223
31. Palmer AC, Egan JB, Shearwin KE. Transcriptional interference by RNA polymerase pausing and dislodgement of transcription factors. *Transcription*. 2011;2(1):9-14. doi:10.4161/trns.2.1.13511.
32. Prescott EM, Proudfoot NJ. Transcriptional collision between convergent genes in budding yeast. *Proceedings of the National Academy of Sciences of the United States of America*. 2002;99(13):8796-8801. doi:10.1073/pnas.132270899.
33. Elías-Arnanz M, Salas M. Bacteriophage phi29 DNA replication arrest caused by codirectional collisions with the transcription machinery. *The EMBO Journal*. 1997;16(18):5775-5783. doi:10.1093/emboj/16.18.5775.
34. Katayama S., Tomaru Y., Kasukawa T., Waki K., Nakanishi M., Nakamura M., Nishida H., Yap C. C., Suzuki M., Kawai J., Suzuki H., Carninci P., Hayashizaki Y., Wells C., Frith M., Ravasi T., Pang K. C., Hallinan J., Mattick J., Hume D. A., Lipovich L., Batalov S., Engström P. G., Mizuno Y., Faghihi M. A., Sandelin A., Chalk A. M., Mottagui-Tabar S., Liang Z., Lenhard B., Wahlestedt C. RIKEN Genome Exploration Research Group; Genome Science Group (Genome Network Project Core Group); FANTOM Consortium. (2005). Antisense transcription in the mammalian transcriptome. *Science* 309 1564–1566
35. Pertea M. The Human Transcriptome: An Unfinished Story . *Genes*. 2012;3(3):344-360. doi:10.3390/genes3030344.
36. Luthra R, Chen H, Roy-Chowdhuri S, Singh RR. Next-Generation Sequencing in Clinical Molecular Diagnostics of Cancer: Advantages and Challenges. Farah CS, Cho WC, eds. Cancers. 2015;7(4):2023-2036. doi:10.3390/cancers7040874.
37. Ma Y, Shi N, Li M, Chen F, Niu H. Applications of Next-generation

Sequencing in Systemic Autoimmune Diseases. Genomics, Proteomics & Bioinformatics. 2015;13(4):242-249. doi:10.1016/j.gpb.2015.09.004.

38. Arrieta M-C, Stiemsma LT, Amenyogbe N, Brown EM, Finlay B. The Intestinal Microbiome in Early Life: Health and Disease. Frontiers in Immunology. 2014;5:427. doi:10.3389/fimmu.2014.00427.
39. Li K-Y, Wang J-L, Wei J-P, et al. Fecal microbiota in pouchitis and ulcerative colitis. *World Journal of Gastroenterology*. 2016;22(40):8929-8939. doi:10.3748/wjg.v22.i40.8929.
40. Hodkinson BP, Grice EA. Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. *Advances in Wound Care*. 2015;4(1):50-58. doi:10.1089/wound.2014.0542.
41. Bentley DR., Balasubramanian S., Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008;456:53–59
42. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics*. 2016;14(5):265-279. doi:10.1016/j.gpb.2016.05.004.
43. Chin CS1, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data.Nat Methods. 2013 Jun;10(6):563-9. doi: 10.1038/nmeth.2474. Epub 2013 May 5.
44. Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. *Journal of genetics and genomics = Yi chuan xue bao*. 2011;38(3):95-109. doi:10.1016/j.jgg.2011.02.003.
45. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46. doi: 10.1038/nrg2626. Review. PMID: 19997069
46. Perea C, De La Hoz JF, Cruz DF, et al. Bioinformatic analysis of genotype by sequencing (GBS) data with NGSEP. *BMC Genomics*. 2016;17(Suppl 5):498. doi:10.1186/s12864-016-2827-7.
47. Song C-X, Clark TA, Lu X-Y, et al. Sensitive and specific single-molecule

sequencing of 5-hydroxymethylcytosine. *Nature methods*. 2011;9(1):75-77. doi:10.1038/nmeth.1779.

48. Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. *Nature biotechnology*. 2013;31(11):1009-1014. doi:10.1038/nbt.2705.
49. D Strahan R, Uppal T, Verma SC. Next-Generation Sequencing in the Understanding of Kaposi's Sarcoma-Associated Herpesvirus (KSHV) Biology. Mak J, Walker P, Gilbert MT, eds. *Viruses*. 2016;8(4):92. doi:10.3390/v8040092.
50. Wille CK, Nawandar DM, Henning AN, et al. 5-hydroxymethylation of the EBV genome regulates the latent to lytic switch. *Proceedings of the National Academy of Sciences of the United States of America*. 2015;112(52):E7257-E7265. doi:10.1073/pnas.1513432112.
51. Sun Z1, Guo Y1, Li M2, Yao Z2. Genotype analysis of varicella-zoster virus isolates from suburban Shanghai Municipal Province, China. J Med Microbiol. 2016 Feb;65(2):123-8. doi: 10.1099/jmm.0.000208. Epub 2015 Dec 9.
52. Wang WD1, Lee GC2, Kim YY1, Lee CH1. A Comparison between Low- and High-Passage Strains of Human Cytomegalovirus. J Microbiol Biotechnol. 2016 Oct 28;26(10):1800-1807. doi: 10.4014/jmb.1604.04045.
53. Oh J, Sanders IF, Chen EZ, et al. Genome Wide Nucleosome Mapping for HSV-1 Shows Nucleosomes Are Deposited at Preferred Positions during Lytic Infection. Nevels M, ed. *PLoS ONE*. 2015;10(2):e0117471. doi:10.1371/journal.pone.0117471.
54. Kwok H, Chiang AKS. From Conventional to Next Generation Sequencing of Epstein-Barr Virus Genomes. Mak J, Walker P, Gilbert MT, eds. *Viruses*. 2016;8(3):60. doi:10.3390/v8030060.
55. Majerciak V, Ni T, Yang W, Meng B, Zhu J, Zheng ZM. A viral genome landscape of RNA polyadenylation from KSHV latent to lytic infection. PLoS Pathog. 2013;9(11):e1003749. doi:10.1371/journal.ppat.1003749.
56. Gatherer D, Seirafian S, Cunningham C, Holton M, Dargan DJ, Baluchova K, et al. High-resolution human cytomegalovirus transcriptome. PNAS.

2011;108(49):19755–60.

57. van Beurden SJ, Gatherer D, Kerrc K, Galbraithd J, Herzykd P, Peetersa BPH, et al. Anguillid herpesvirus 1 transcriptome. J Virol. 2012;86(18):10150–61.
58. Anselmo A, Flori L, Jaffrezic F, Rutigliano T, Cecere M, Cortes-Perez N, et al. Co-expression of host and viral microRNAs in porcine dendritic cells infected by the pseudorabies virus. PLoS One. 2011;6(3), e17374. doi:10.1371/journal.pone.0017374.
59. Wu YQ, Chen DJ, He HB, Chen DS, Chen LL, Chen HC, et al. Pseudorabies virus infected porcine epithelial cell line generates a diverse set of host microRNAs and a special cluster of viral microRNAs. PLoS One. 2012;7(1), e30988. doi:10.1371/journal.pone.0030988.
60. Piedade D, Azevedo-Pereira JM. The Role of microRNAs in the Pathogenesis of Herpesvirus Infection. Ploss A, ed. *Viruses*. 2016;8(6):156. doi:10.3390/v8060156.
61. Nsiah YA1, Rapp F. Role of latency-associated transcript in herpes simplex virus infection. Intervirology. 1991;32(2):101-15.
62. 62 Rossetto CC, Pari GS. PAN's Labyrinth: Molecular Biology of Kaposi's Sarcoma-Associated Herpesvirus (KSHV) PAN RNA, a Multifunctional Long Noncoding RNA. Zheng Z-M, ed. *Viruses*. 2014;6(11):4212-4226. doi:10.3390/v6114212.
63. Juranic Lisnic V, Babic Cac M, Lisnic B, Trsan T, Mefferd A, Das Mukhopadhyay C, et al. Dual analysis of the murine cytomegalovirus and host cell transcriptomes reveal new aspects of the virus-host cell interface. PLoS Pathog. 2013;9(9), e1003611. doi:10.1371/journal.ppat.1003611.
64. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
65. Chin CS1, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013 Jun;10(6):563-9. doi: 10.1038/nmeth.2474. Epub 2013

May 5.

66. Istrail S, Sutton GG, Florea L, et al. Whole-genome shotgun assembly and comparison of human genome assemblies. *Proceedings of the National Academy of Sciences of the United States of America*. 2004;101(7):1916-1921. doi:10.1073/pnas.0307971100.
67. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105–11. doi: 10.1093/bioinformatics/btp120.
68. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9. doi: 10.1038/nmeth.1923.
69. Rutherford K, Parkhill K, Crook J, Horsnell T, Rice P, Rajandream MA, et al. Artemis: sequence visualization and annotation. Bioinformatics. 2000;16(10):944–5. doi: 10.1093/bioinformatics/16.10.944.
70. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6. doi: 10.1038/nbt.1754.
71. Krzywinski, M. et al. Circos: an Information Aesthetic for Comparative Genomics. Genome Res (2009) 19:1639-1645
72. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2008. ISBN 3-900051-07-0
73. Ahmed F, Kumar M, Raghava GP. Prediction of polyadenylation signals in human DNA sequences using nucleotide frequencies. In Silico Biol. 2009;9(3):135–48.
74. Tcherepanov V, Ehlers A, Upton C. Genome Annotation Transfer Utility (GATU): rapid annotation of viral genomes using a closely related reference genome. *BMC Genomics*. 2006;7:150. doi:10.1186/1471-2164-7-150.
75. Baumeister J, Klupp BG, Mettenleiter TC. Pseudorabies virus and equine herpesvirus 1 share a nonessential gene which is absent in other herpesviruses and located adjacent to a highly conserved gene cluster. J

Virol. 1995;69(9):5560–7.

76. Hatfield L, Hearing P. The NFIII/OCT-1 binding site stimulates adenovirus DNA replication in vivo and is functionally redundant with adjacent sequences.*Journal of Virology*. 1993;67(7):3931-3939.
77. Coutinho TJD, Franco GR, Lobo FP. Homology-Independent Metrics for Comparative Genomics. *Computational and Structural Biotechnology Journal*. 2015;13:352-357. doi:10.1016/j.csbj.2015.04.005.
78. Tempel S., Tahi F. A fast ab-initio method for predicting miRNA precursors in genomes. Nucleic Acids Res. 2012;40:e80. doi: 10.1093/nar/gks146.
79. Jiang P., Wu H., Wang W., Ma W., Sun X., Lu Z. MiPred: Classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res. 2007;35:W339–W344. doi: 10.1093/nar/gkm368.
80. Colgan DF, Manley JL. Mechanism and regulation of mRNA polyadenylation. Genes Dev. 1997;11: 2755-2766.
81. Proudfoot NJ. Ending the message: poly(A) signals then and now. *Genes & Development*. 2011;25(17):1770-1782. doi:10.1101/gad.17268411.
82. Rakesh S. Laishram, Poly(A) polymerase (PAP) diversity in gene expression – Star-PAP vs canonical PAP, FEBS Letters, Volume 588, Issue 14, 27 June 2014, Pages 2185-2197, ISSN 0014-5793, http://dx.doi.org/10.1016/j.febslet.2014.05.029.
83. Yamada S, Imada T, Watanabe W, Honda Y, Nakajima-Iijima S, Shimizu Y, et al. Nucleotide sequence and transcriptional mapping of the major capsid protein gene of pseudorabies virus. Virology. 1991;185:56–66. doi: 10.1016/0042-6822(91)90753-X.
84. Klupp B, Kern H, Mettenleiter TC. The virulence-determining genomic Bam HI-fragment 4 of pseudorabies virus contains genes corresponding to the UL15 (partial), UL18, UL19, UL20, and UL21 genes of herpes simplex virus and a putative origin of replication. Virology. 1992;191:900–8. doi: 10.1016/0042-6822(92)90265-Q.
85. Fuchs W, Klupp BG, Granzow H, Leege T, Mettenleiter TC. Characterization

of pseudorabies virus (PRV) cleavage-encapsidation proteins and functional complementation of PRV pUL32 by the homologous protein of herpes simplex virus type 1. J Virol. 2009;83(8):3930–43. doi: 10.1128/JVI.02636-08.

86. Zhang G, Leader DP. The structure of the pseudorabies virus genome at the end of the inverted repeat sequences proximal to the junction with the short unique region. J Gen Virol.1990;71(10):2433–41. doi: 10.1099/0022-1317-71-10-2433.

87. Priola SA, Stevens JG. The 5′ and 3′ limits of transcription in the pseudorabies virus latency associated transcription unit. Virology. 1991;182(2):852–6. doi: 10.1016/0042-6822(91)90628-O.

88. Cheung AK. Cloning of the latency gene and the early protein 0 gene of pseudorabies virus. J Virol. 1991;65(10):5260–71.

89. Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R. Comprehensive splice-site analysis using comparative genomics. Nucl Acids Res. 2006;34(14):3955–67. doi: 10.1093/nar/gkl556.

90. Griffiths A, Coen DM. An unusual internal ribosome entry site in the herpes simplex virus thymidine kinase gene. *Proceedings of the National Academy of Sciences of the United States of America*. 2005;102(27):9667-9672. doi:10.1073/pnas.0504132102.

91. Yu T, Chen F, Ku X, Fan J, Zhu Y, Ma H, Li S, Wu B, He Q. Growth characteristics and complete genomic sequence analysis of a novel pseudorabies virus in China. Virus Genes. 2016 Aug;52(4):474-83. doi: 10.1007/s11262-016-1324-z. Epub 2016 Mar 24.

92. Gu Z, Hou C, Sun H, et al. Emergence of highly virulent pseudorabies virus in southern China. *Canadian Journal of Veterinary Research*. 2015;79(3):221-228.

93. Sozzi E, Moreno A, Lelli D, Cinotti S, Alborali GL, Nigrelli A, Luppi A, Bresaola M, Catella A, Cordioli P.Genomic characterization of pseudorabies virus strains isolated in Italy. Transbound Emerg Dis. 2014 Aug;61(4):334-40. doi: 10.1111/tbed.12038. Epub 2013 Jan 18.a

94. Müller T, Klupp BG, Freuling C, Hoffmann B, Mojcicz M, Capua I, Palfi V,

Toma B, Lutz W, Ruiz-Fon F, Gortárzar C, Hlinak A, Schaarschmidt U, Zimmer K, Conraths FJ, Hahn EC, Mettenleiter TC. Characterization of pseudorabies virus of wild boar origin from Europe. Epidemiol Infect. 2010 Nov;138(11):1590-600. doi: 10.1017/S0950268810000361. Epub 2010 Mar 12.

I.

Oláh *et al. BMC Microbiology* (2015) 15:130
DOI 10.1186/s12866-015-0470-0

BMC Microbiology

RESEARCH ARTICLE Open Access

Characterization of pseudorabies virus transcriptome by Illumina sequencing

Péter Oláh[†], Dóra Tombácz[†], Nándor Póka, Zsolt Csabai, István Prazsák and Zsolt Boldogkői[*]

Abstract

Background: Pseudorabies virus is a widely-studied model organism of the *Herpesviridae* family, with a compact genome arrangement of 72 known coding sequences. In order to obtain an up-to-date genetic map of the virus, a combination of RNA-sequencing approaches were applied, as recent advancements in high-throughput sequencing methods have provided a wealth of information on novel RNA species and transcript isoforms, revealing additional layers of transcriptome complexity in several viral species.

Results: The total RNA content and polyadenylation landscape of pseudorabies virus were characterized for the first time at high coverage by Illumina high-throughput sequencing of cDNA samples collected during the lytic infectious cycle. As anticipated, nearly all of the viral genome was transcribed, with the exception of loci in the large internal and terminal repeats, and several small intergenic repetitive sequences. Our findings included a small novel polyadenylated non-coding RNA near an origin of replication, and the single-base resolution mapping of 3′ UTRs across the viral genome. Alternative polyadenylation sites were found in a number of genes and a novel alternative splice site was characterized in the *ep0* gene, while previously known splicing events were confirmed, yielding no alternative splice isoforms. Additionally, we detected the active polyadenylation of transcripts earlier believed to be transcribed as part of polycistronic RNAs.

Conclusion: To the best of our knowledge, the present work has furnished the highest-resolution transcriptome map of an alphaherpesvirus to date, and reveals further complexities of viral gene expression, with the identification of novel transcript boundaries, alternative splicing of the key transactivator EP0, and a highly abundant, novel non-coding RNA near the lytic replication origin. These advances provide a detailed genetic map of PRV for future research.

Keywords: Alphaherpesvirus, RNA-Seq, Polyadenylation, Gene expression, Viral genomics

Background

Pseudorabies virus (PRV, Suid Herpesvirus 1), also known as Aujeszky's disease virus, a herpesvirus belonging in the subfamily *Alphaherpesvirinae,* infects swine populations and causes economic losses worldwide. PRV is widely used in studies of the molecular pathomechanism of herpesviruses [1], as a tract-tracing tool for mapping neuronal circuitries [2, 3] and for the delivery of genetically encoded fluorescent activity markers [4]. The transcription of herpesviruses is strictly regulated by cascade-like processes. Three temporal classes of viral genes can be distinguished in terms of the time of their activation during the viral life cycle: initially, the immediate-early (IE) genes are expressed, whose protein products are transcription factors. PRV has a single IE-class gene, *ie180*, which is the major regulator of viral gene expression. The early (E) genes typically play roles in the replication of viral DNA, while most of the late (L) genes code for structural elements of the virus. The PRV genome is arranged into two unique protein coding regions, the unique long (UL) and unique short (US) regions, flanked by the internal and terminal repeats (IR and TR). The genome of PRV is large among viruses, but much smaller than those of cellular organisms, and especially the mammalian genome. The whole transcriptome analysis of PRV can therefore be performed by real-time RT-PCR, a technique, which provides an accurate platform for the temporal analysis of transcription in both wild-type [5] and mutant viral strains [6]. However, PCR can target only a small genomic region, and information related to transcript lengths, splicing, alternative initiation and

[*] Correspondence: boldogkoi.zsolt@med.u-szeged.hu
[†]Equal contributors
Department of Medical Biology, Faculty of Medicine, University of Szeged, Szeged, Hungary

termination of transcription, unknown transcripts, etc. is not provided. Furthermore, PCR is inconvenient for the detection of novel transcripts. Coding sequences and their related transcripts have been widely studied in PRV [5, 7], together with the microRNA expression in both the lytic and latent phases of the viral life cycle [8, 9], whereas other sources of non-coding transcription, alternative transcript termination and alternative splicing have not yet been analyzed at a genome-wide level. In order to complement previous RT-PCR based studies, we have carried out high-throughput sequencing of the total RNA and polyA(+) RNA fractions of PRV during lytic infection. Transcriptome-wide profiling has led to the discovery of novel regulatory RNAs and an accurate assessment of their expression in several members of the *Herpesviridae* (human cytomegalovirus: [10], anguillid herpesvirus 1: [11]). These studies have discovered highly abundant long non-coding RNAs (lncRNAs), while in addition, the characterization of the MAT ncRNA in murine cytomegalovirus has shown its role not only as a lncRNA, but also coding for an ORF with potential regulatory functions [12]. Host-pathogen interaction studies have also revealed dramatic changes in expression levels of a range of host regulatory- and non-coding RNAs during lytic infection with varicella zoster virus [13]. Recent findings suggest that, similarly as in eukaryotes, alternative transcript termination might be an important regulatory mechanism in herpesvirus gene expression [14]. Indeed, the assessment of 3′ UTRs in PRV strain Kaplan (Ka) identified three genes, each containing two alternative termination sites, while also indicating individual polyadenylation (PA) sites of genes previously recognized as being exclusively transcribed in polycistronic RNAs and not possessing their own PA sites. The PA sites have also been categorized in terms of relative expression levels by determining the overall frequency of proximal and distal PA-site usage per gene.

Results and discussion

Assessment of the PRV transcriptome by total RNA sequencing and PA-Seq

For the investigation of the lytic PRV transcriptome, porcine kidney (PK-15) epithelial cells were infected with a high dose (10 pfu) of PRV strain Ka. Samples were gathered up to 24 h post-infection (p.i.) in order to capture all RNA species during lytic infection for sequencing library preparation. Both random hexamer-primed and oligo(dT)-primed libraries were prepared in order to assess total RNA and mRNA transcripts separately. In our modified polyadenylation sequencing (PA-Seq) protocol [14], total RNA was reverse-transcribed by using custom designed oligo(T10-VN) anchored primers containing standard Illumina strand-specific adaptor sequences. The two-nucleotide anchor sequence ensures the annealing of primers at exactly the PA site of mRNAs, providing considerably fewer reads that contain redundant adenine homopolymer stretches, with more useful sequence information resulting for the given depth of sequencing. PA peaks were detected by using HOMER [15] in strand-specific mode, with adjustments for viral cDNA peak calling and a cutoff of 50 reads per base position. PA peaks occurred on both strands, mainly in accordance with previously existing ORF annotations, and also long non-coding RNAs, including the latency-associated transcript (LAT) and the long-latency transcript (LLT).

Both the RNA integrity measurements during the sample preparation, and the low signal-to-noise ratio in the 1 kb region surrounding the PA peaks during the analysis indicated high library quality. Sequencing of the total RNA isolates of infected cells yielded a data set of ~ 208 million 100 bp paired-end reads for the random hexamer-primed library, of which 1.3 million reads aligned to the viral genome version KJ717942.1, and the majority of the remaining sequences aligned to the host organism genome *Sus scrofa* 10.2. PA-Seq resulted in ~ 103 million single- end, 50 bp reads, with 10 million reads aligning to the above-mentioned viral reference.

PRV transcriptome profiling

Nearly all of the viral genome was transcribed, with the exception of highly repetitive sequences within the terminal and internal repeats that do not encode any RNA species. Similarly, there was no detectable transcription at intergenic repeat regions, which were earlier predicted to be transcriptional insulators [16]. Significant transcription at these insulator sequences was observed only in two convergently oriented gene pairs, *ul44-ul26* and, to a lesser extent, *ul35-ul36*. Here, the alternative transcript termination indicates that leaky transcription traverses the intergenic repeat boundaries with lengths of 109 bp and 443 bp, respectively. On the other hand, non-transcribed, repetitive regions were markedly present between ORF-1 and *ul54*; *ul46* and *ul27*; *ul40* and *ul41*; and *ul11* and *ul10*. In these boundary regions, no expression was observable. A high percentage of the transcription is committed to producing a newly identified non-coding RNA, CTO ("close to OriL"), located between the *ul21* gene and the oriL, between bases 63673–63958 on the complementary strand of genome KJ717942.1. The CTO (RPKM = 1.6×10^6 in the total RNA library) and US1 (RPKM = 2.32×10^5) encoding the ICP22 homolog *Rsp40* immediate-early regulatory protein are the most abundant transcripts. Although we examined lytic infection, the two latency-associated transcripts (LAT and LLT) were found to be expressed at a low level, and not at sufficient coverage to determine splicing donor and acceptor sites. Transcription of the hypothetical ORF1.2 [17] sequence was also detected, involving 5′ upstream regions, although single-base localization of the transcription start site is complicated by the presence of several repeats in the genomic sequence in the interval 730–960 bp. On the use of PA-Seq, 3′ transcript boundaries can be accurately identified between and within gene clusters (Fig. 1). In convergently oriented clusters, more extensive overlaps include coding regions of the opposite genes, potentially giving rise to transcriptional interference between the interacting partners [18]. An example of such a relation is between *ul30* and *ul31*, with a tail-to-tail overlap of 80 nucleotides. Here, the expression of *ul30* mRNA exceeds that of *ul31*, with considerable antisense expression over the latter gene, possibly due to transcriptional read-through from *ul30.* As anticipated, convergent genes with more than ~45 bp separating their respective PA signals (PAS) do not demonstrate detectable transcriptional overlap, ranging from *ul18-ul15* (45 bp) to *ul46-ul27* (632 bp), while convergent gene pairs in closer proximity exhibit longer 3′ UTR overlapping regions. A short, 3′-overlapping antisense non-coding transcript (termed SANC) was also found adjacent to the PA site of the *ul21* gene, near OriL (64558–64674 on reference genome KJ717942.1), with an expression of RPKM = 1.67×10^3 in the total RNA library, the highest non-coding antisense expression in our samples. The various overlaps between the viral genes are presented in Table 1. These overlaps may affect the expression of adjacent genes. It is hypothesized that these interactions form a regulatory network controlling the transcription cascade of herpesviruses [18].

Splice sites in the PRV transcriptome

For splice site analysis, total RNA reads were aligned to PRV strain Ka genome KJ717942.1. All possible splice donor and acceptor sites were considered, with a lower bound of at least 10 supporting reads. Through the exclusion of low-coverage junction candidates, artifacts possibly occurring due to mispriming or template switching during amplification steps could be neglected (Additional file 1). The initial set of splice acceptor and donor sites contained 97 candidate splice junctions, with 49 sites above the threshold coverage. This set contained several permutations of the *us1* 3′ UTR splice junctions, which were screened for the presence of short anchor regions and high mismatch ratios within these anchors. After screening for anchors of <5 bases, consistent splice junctions were readily identifiable. The remaining, high-coverage splice sites are denoted as follows: (D + 10000^A + 12000), with D denoting the donor site, A the acceptor site, and +/− the DNA top and bottom strand, respectively, along the coordinates of the splice junction. Splice sites have previously been characterized in the protein-coding region of *ul15* [19, 20] (D −76165^A −73285), and in the 3′ UTR of *us1* [21] (D +115592^A +115713; D +115766^A +115921), present in both terminal and internal repeats, while one site in the non-coding RNA LLT [22] (D +97765; A +102403) was expressed at an insufficient level for accurate splice site identification. A low percentage of reads also mapped outside the assigned acceptor and donor sites (Fig. 1). A novel alternative splice site was characterized in *ep0*, the homolog of *Herpes simplex* 1 ICP0 [23], which is also a spliced gene, but expressed in the immediate-early class in HSV-1. The newly characterized *ep0* alternative splicing consists of two potential donor sites at (D −97480) and (D −97528) and the acceptor site at (A −97389) (Fig. 1). While the splice junction formed between the acceptor and proximal donor sites conforms to the rule of GT/AG nucleotides comprising ~99 % of junctions in eukaryotic organisms [24], the junction formed with the distal donor site contains GT/CG bases. Experimental validation of the novel splice site has been carried out by RT-PCR, using two primer sets (Additional file 2) designed approximately 100 bp upstream and downstream of the splice site, followed by polyacrylamide gel electrophoresis. The experiments confirmed the presence of the novel isoform during lytic infection robustly after visualization (Fig. 2).

Frequency of alternative polyadenylation correlated to weak and strong PA signals and flanking motifs

Through the use of the highly sensitive PA-Seq method, the 3′-end of the PRV transcripts was identified by the presence of poly(A) tails. The use of anchored oligo(dT) primers resulted in the accurate mapping of polyadenylation sites, while providing a high coverage for quantitative analyses (Fig. 1). The most highly abundant transcripts (CTO, *us1*, *ul31*, and *ul35*) were in accordance with the random hexamer-primed data, while the greater resolution provided by PA-Seq also allowed the identification of transcripts that were of low abundance or difficult to detect by other sequencing methods or RT-PCR, such as the genes of the *ul6-ul9* convergently oriented cluster. The PAS-usage of eukaryotic organisms is thought to be well conserved, with the canonical AAUAAA being the most widely used signal 10–30 nucleotides upstream of the cleavage site [25, 26]. Not surprisingly, the analysis of the PAS motifs indicated that strong polyadenylation peaks correspond to the AAUAAA signal (<90 %), while AUUAAA is the second most widely used element (~10 %) (Fig. 3). Two further signal motifs were the less conserved USE GU-rich element, >30 nucleotides upstream of the cleavage site [25], and DSE, >20 nucleotides downstream of the cleavage site. In humans, it has been shown that when multiple PA sites are used, the 3′ -most tends to use the AAUAAA signal, while the inner signals tend to vary considerably from the consensus [27]. In our PA-Seq samples, only 6 genomic positions containing the canonical AATAAA sequence proved to be unused PA signals, 3 motifs residing inside coding sequences (+9072-9077; −52929-52934; −78490-78495) and one motif located directly

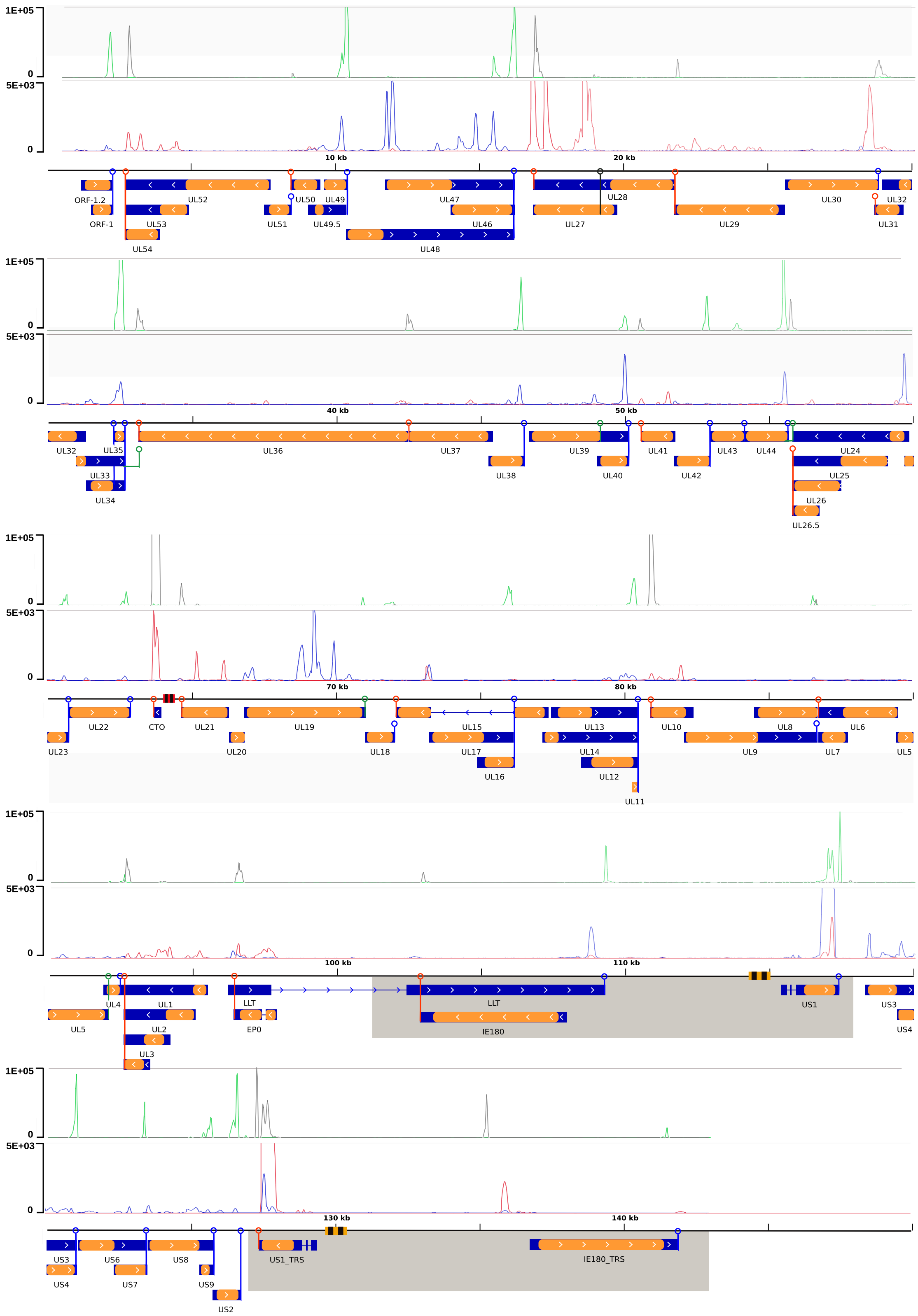


Fig. 1 (See legend on next page.)

(See figure on previous page.)

Fig. 1 Transcriptional map of the PRV genome identified by total RNA sequencing and PA-Seq. Genetic map: orange: coding sequences, blue: transcripts, red striped rectangle: OriL palindrome, yellow striped rectangles: OriS palindromes, grey: internal and terminal repeat regions, blue circles: PA site on + strand, red circles: PA site on −strand, green circles: alternative PA site on + strand, black circles: alternative PA site on −strand. Expression levels (in coverage per base): upper box: PA-Seq expression, green: +strand read coverage, black: −strand read coverage, lower box: totalRNA sequencing, blue: +strand coverage, red: −strand coverage

upstream of *us3* (+118308-118313), and not corresponding to any viral transcript. The remaining signal was located within the large repeat regions, and therefore present in two copies, (−117738-117743) and (+126996-127001). On the other hand, canonical PAS that were previously considered inactive demonstrated pronounced polyadenylation peaks, providing alternative transcript termination sites for genes *ul35, ul44* and *ul22.* In both cases, the usage frequency of the distal PA site was at least an order of magnitude lower than those of the proximal ones.

The PA-Seq method additionally revealed polyadenylation peaks in transcripts encoded by upstream genes of tandem gene clusters. These included the polyadenylation of *ul19.* This transcript has previously been detected in strain Indiana-Funkhauser [28], with the non-canonical PAS ATATAAA; in our PA-Seq samples, we have confirmed the active use of this site in strain Ka. A similar arrangement was found in *ul28*, although no conservative PAS was detectable upstream of the well-defined PA-peak at base position 18960. Though PA peaks within the clusters of the US region were markedly above the background signal and correlated well with the coding sequence boundaries, these signals were several orders of magnitude weaker than the commonly observed polyA peaks, making them difficult to validate. The tandemly oriented *ul4* transcript has been hypothesized to be 3′ coterminal with the *ul5* transcript [16], as the canonical PAS directly downstream of *ul5* is inside the *ul4* ORF. However, PA-Seq peaks were found at the 3′ ends of both genes, showing that the PAS of *ul5* is also active. The most abundant transcript during PRV lytic infection proved to be a previously unknown non-coding RNA of 286 bp, located between genes *ul21* and *ul22*, and named CTO. This long non-coding RNA is characterized by an irregular GC composition, where the third-position GC content increases sharply in all three reading frames in the length of the transcript. Based on sequence similarity search, this arrangement is not present in the close relatives of PRV, such as varicella zoster, herpes simplex and bovine herpesviruses. On the other hand, PRV strains Becker, Bartha and HeN1 show <99 % sequence similarity with strain Kaplan in the CTO genomic region. An alternative PA site was also observed, about 120 nucleotides downstream from the main PAS. An in-depth characterization of the transcript is presented in [29].

Table 1 The organization of alternative splicing, overlapping gene clusters, polycistronic RNAs and alternative polyadenylation events in the PRV genome

Detected splice sites			Alternative polyadenylation			Convergently overlapping gene clusters		Divergently overlapping genes		Tandem gene clusters
Gene	Donor site	Acceptor site	Gene	Alternative polyadenylation signal	Coordinate					
UL15	D −76165	A −73285	UL35	AAUAAA	33133–33138	UL51	UL50	UL52	UL51	UL52-UL54
US1	D +115592	A +115713	UL44	AAUAAA	55768–55773	UL30	UL31, UL32	UL50	UL49.5	UL48-UL46
	D +115766	A +115921	UL22	N/A[a]	63624	UL33, UL34, UL35	UL36	UL29	UL30	UL31-UL32
US1	D −129158	A −129037	UL19	AUAUAAA	71005–710011	UL44	UL26.5, UL26, UL25, UL24	UL32	UL33	UL33-UL35
	D −128984	A −128829	UL28	N/A[a]	18960	UL8, UL9	UL6, UL7	UL37	UL38	UL39-UL40
EP0	D −97480	A −97389	UL5	AAUAAA	92065–92070			UL41	UL42	UL24-UL26.5
	D −97528		CTO	N/A[a]	63538			UL24	UL23	UL17-UL16
								UL21	UL20	UL14-UL11
								UL15	UL14	UL9-UL8
								UL10	UL9	UL7-UL6
								UL6	UL5	UL1-UL3.5
										US3-US4
										US6-US7
										US8-US9

[a] No prediction available for canonical or non-canonical polyA signal using PolyApred server

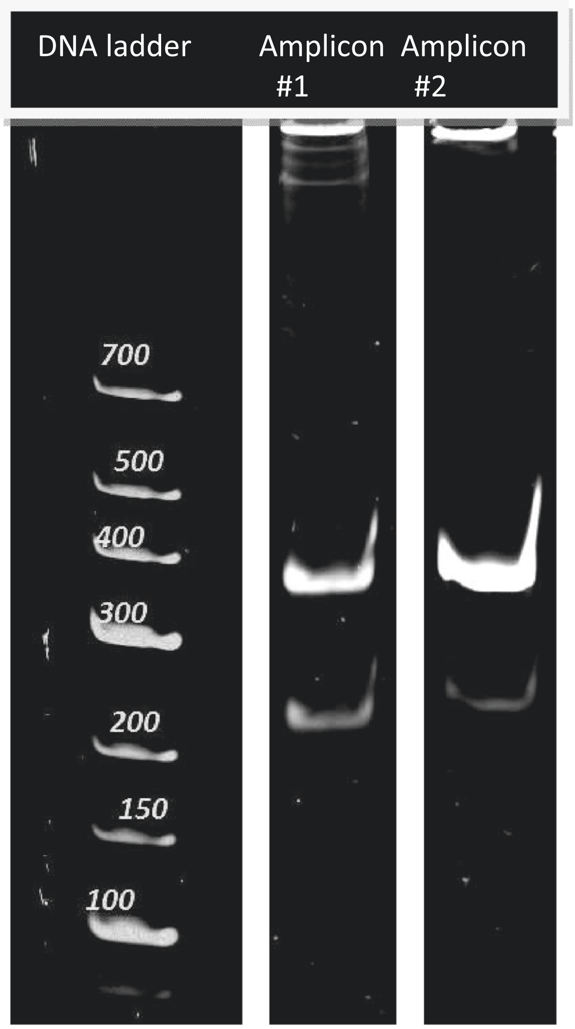


Fig. 2 RT-PCR validation of the novel splice site of the EP0 gene, using two different primer sets. 12 % acrylamide gel electrophoresis, Gel Red staining

Transcription overlaps

We assessed the various transcript overlaps, including parallel (tandem), divergent and convergent overlaps (Fig. 4, Table 1). Most of the PRV genes are organized into tandem gene clusters producing polycistronic RNAs (Table 1). An interesting feature of organization is that all of the upstream genes of the clusters end within the downstream genes. Similarly, the divergent gene pairs overlap in every case. Theoretically, this phenomenon may be explained by the restriction of the viral genome length. However, since these overlaps are not too extensive, they probably provide a regulatory mechanism for transcription. The distant convergent genes are separated from each other by repetitive sequences which were found to be heavily methylated (this latter result will be published elsewhere), indicating a mechanism with a likely function for the prevention of transcriptional collisions. Closely located convergent gene pairs transcriptionally overlap or can overlap (alternative transcriptional termination) themselves (Table 1). In the *ul35* and *ul44* genes, these alternative termination sites traversed intergenic repetitive regions, previously considered to be transcriptional barriers. This finding indicates that low-frequency “leaky” transcription occurs more often than anticipated in PRV. Although the function of these PA sites is unknown, it is noteworthy that a highly similar arrangement was present between convergent gene clusters *ul9-ul8* and *ul7-ul6*, with the difference that a strongly repetitive sequence resembling the above-mentioned intergenic repeats in both length and base content, was found within the comparatively long 3′ UTR of *ul7*.

Conclusion

The single-base resolution map of pseudorabies transcripts revealed that the compact, 143 kbp genome of PRV is transcribed pervasively, with the exception of loci in the large inverted repeats and short intergenic sequences. In addition to previously known splice sites, a novel junction was characterized in the transactivator

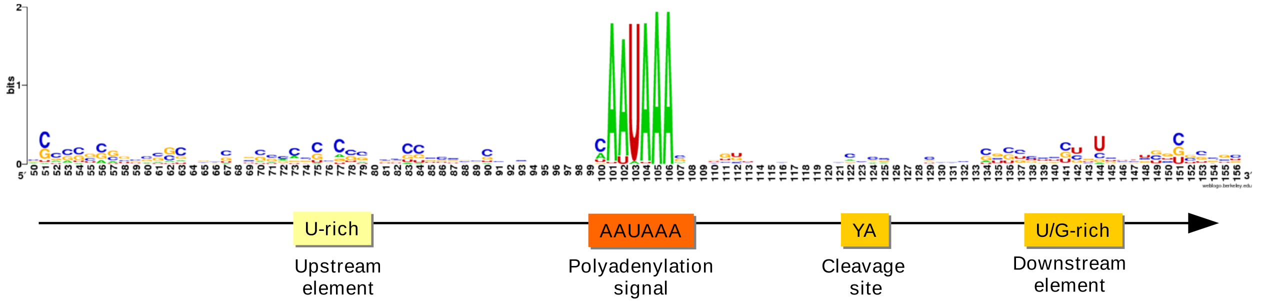


Fig. 3 PA signal consensus of PRV and schematic organization of eukaryotic PA sites. The sequence logo represents nucleotide occurrence frequencies within the +/−50 bp region of PA signals in the viral genome. Colored boxes indicate the functional elements affecting eukaryotic polyadenylation

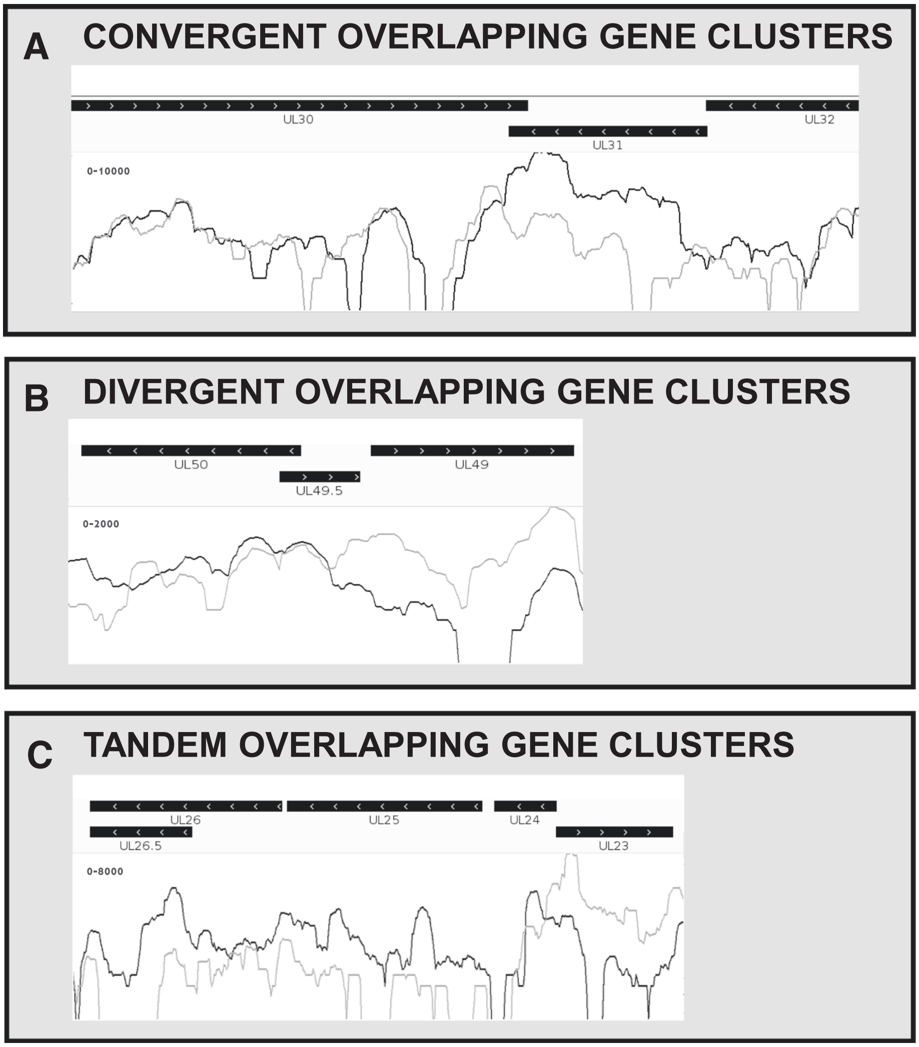


Fig. 4 Convergent (**a**), divergent (**b**) and tandem (**c**) overlaps in the PRV genome, as shown by random-hexamer primed samples. Extensive transcriptional overlaps are frequent throughout the condensed viral genome. Black boxes: coding sequences, white arrows: gene orientation, grey line graph: positive strand expression, black line graph: negative strand expression

ep0, while the splice sites of lytic genes were confirmed at a high depth of coverage. Polyadenylation signal usage was found to be more frequent than previously predicted, with alternative PAS in genes *ul35*, *ul44*, *ul19*, *ul28* and *ul5*. While alternative transcript termination is a major regulatory factor in eukaryotic organisms, to date there is limited data for viruses in this field. The region of the lytic replication origin was also found to express a novel, highly abundant ncRNA, named CTO, along with a short, 3′ overlapping ncRNA of *ul21*, termed SANC. Other pervasively transcribed regions include the ORF1.2 5′ UTR. The described PRV transcript isoforms and non-coding RNAs help guide future research in the possible regulatory mechanisms of alphaherpesviruses.

Methods

Virus, cells and infection

For the propagation of strain Kaplan of PRV, immortalized PK-15 epithelial cells were applied. PK-15 cells were cultivated in Dulbecco's modified Eagle medium supplemented with 5 % fetal bovine serum (Gibco Invitrogen) with 80 μg gentamycin/ml at 37 °C, under 5 % CO_2. The virus stock used for the experiments was prepared as follows: rapidly-growing semi-confluent PK-15 cells were infected at a multiplicity of infection of 0.1 plaque-forming unit (pfu)/cell and were incubated until a complete cytopathic effect was observed. The infected cells were frozen and thawed three times, followed by low-speed centrifugation (10,000 g) for 20 min. The cell debris was removed, while the supernatant was concentrated and further purified by ultracentrifugation through a 30 % sugar cushion at 24,000 rpm for 1 h, using a Sorvall AH-628 rotor. The number of cells in a culture flask was 5×10^6. A high multiplicity of infection (10 pfu/cell) was used for the infection of PK-15 cells. Infected cells were incubated for 1 h, followed by removal of the virus suspension and washing with phosphate-buffered saline (PBS). After the addition of new medium to the cells, they were incubated for 1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 or 24 h p.i. Mock-infected cells, but otherwise treated in the same way as the infected cells, were used as controls.

Isolation of RNAs

RNA was extracted from samples of each individual time point of infection by using the NucleoSpin RNA II Kit (Macherey-Nagel GmbH and Co. KG), as described previously [5]. Briefly, after the cells had been collected by centrifugation and lysed with buffer containing chaotropic ions, the nucleic acids were docked to a silica column. The DNA was removed with RNase-free DNase solution (supplied with the NucleoSpin RNA II Kit). Finally, the RNAs were eluted from the column in RNase-free water (supplied with the kit). To eliminate the residual DNA contamination, all RNA samples were treated by an additional digestion with Turbo DNase (Ambion Inc.). The concentrations of the RNA samples were measured by spectrophotometric analysis with a BioPhotometer Plus instrument (Eppendorf). RNA samples were stored at −80 °C until further use.

cDNA library preparation

Strand-specific total RNA libraries were prepared for paired-end 100 bp sequencing by using the Illumina compatible ScriptSeq v2 RNA-Seq Library Preparation Kit (Epicenter). For polyA-sequencing, a single-end library was constructed through the use of custom anchored adaptor-primer oligonucleotides with an oligo(VN)T_{20} primer sequence. Anchored primers compensate for the loss in throughput due to the high fraction of reads containing solely adenine bases on the use of conventional oligo(dT) primers.

Illumina sequencing

Transcriptome sequencing was performed on an Illumina HiScanSQ platform at the Genomic Medicine and Bioinformatic Core Facilty of the University of Debrecen. Quality assessment of raw read files was achieved with FastQC v0.10.1. Reads were aligned to the respective host genome (*Sus scrofa*, assembly: Sscrofa10.2) and subsequently to the PRV genome (KJ717942.1), using Tophat v2.09 [30]; ambiguous reads were discarded. For PA-Seq, mapping was carried out with Bowtie v2. [31], followed by peak detection using HOMER in strand-specific mode, with adjustments for the peak qualities of oligo(dT) primed libraries. Peak categories were assigned by using in-house scripts, based on the following criteria: the presence or absence of a PAS in the 50 bp region upstream from the PA site and the presence of at least 2 consecutive adenine mismatches in at least 10 independent reads at the PA site. Annotation and visualization were carried out in the Artemis Genome Browser v15.0.0 [32] and IGV v2.2 [33]. GC bias in the alignments was inspected by using the Bioconductor R package. The prediction of canonical and non-canonical PAS was carried out using PolyApred [34].

RT-qPCR analysis of alternative splicing

For the validation of splicing events, two sets of primers were designed with lengths from 19 to 23 nucleotides, approximately 100 bp upstream and downstream of the splice site, detailed in Additional file 2. Reverse transcription was performed in 5 μl of solution containing 0.02 μg of total RNA, 2 pmol of the gene-specific primer, 0.25 μl of dNTP mix, 1 μl of 5× First-Strand Buffer, 0.25 μl (50 units/μl) of SuperScript III Reverse Transcriptase (Invitrogen) and 1 U of RNAsin (Applied Biosystems Inc.). The mixture was incubated at 55 °C for 60 min. The reaction was stopped at 70 °C for 15 min. No-RT control reactions (RT reactions without Superscript III enzyme) were run to test the potential viral DNA contamination by conventional PCR. RNA samples with no detectable DNA contamination were used for RT-qPCR reactions.

Real-time quantitative PCR experiments were carried out for each sample in triplicate, on a Rotor-Gene 6000 cycler (Corbett Life Science). Reactions were carried out in 20-μl mixtures containing 7 μl of ×10 dilution cDNA, 10 μl of ABsolute qPCR SYBR Green Mix (Thermo Fisher Scientific), 1.5 μl of forward and 1.5 μl of reverse primers (10 μM each). The running conditions were as follows: [1] 15 min at 95 °C, 30 cycles of 94 °C for 25 s (denaturation), 60 °C for 25 s (annealing), and 72 °C for 6 s (extension). Products were visualized on 12 % polyacrylamide gel stained with Gel Red dye, gel images were acquired using a ProteinSimple AlphaImager HV gel documentation system.

Availability of data

Raw data from PA-Seq and RNA-Seq experiments are deposited in the European Nucleotide Archive under accession code PRJEB9526. The PRV genomic sequence used for mapping is available in Genbank, with accession number KJ717942.1.

Additional files

Additional file 1: Splice site analysis using totalRNA sequencing data.

Additional file 2: Primer sequences for the Real-Time RT PCR analysis.

Abbreviations

PRV: Pseudorabies virus; PA-Seq: Polyadenylation sequencing; PAS: Polyadenylation signal; RPKM: Reads per kilobase per million; qRT-PCR: quantitative reverse transcription PCR.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DT and ZC performed the reverse transcription reactions, purified PRV RNA, and propagated PK-15 cells and viruses. ZB coordinated the study, propagated viruses, drafted and corrected the manuscript. OP performed sequencing data analysis and drafted the manuscript. IP propagated PK-15 cells. NP took part in data analysis. All authors have read and approved the final manuscript.

Acknowledgment

Library preparation, sequencing and primary bioinformatic analysis of the samples was performed at the Genomic Medicine and Bioinformatic Core Facility of the University of Debrecen, Medical Faculty.
This research was supported by grant (SH/7/2/8) of the Swiss-Hungarian Cooperation Programme to Zsolt Boldogkői.

Received: 16 March 2015 Accepted: 19 June 2015
Published online: 01 July 2015

References

1. Szpara ML, Kobiler O, Enquist LW. A common neuronal response to alphaherpesvirus infection. J Neuroimmune Pharmacol. 2010;5(3):418–27.
2. Card JP, Enquist LW. Transneuronal circuit analysis with pseudorabies viruses. Curr Protoc Neurosci. 2001;68:1.5.1–1.5.39. doi:10.1002/0471142301.ns0105s68.
3. Boldogkői Z, Sík A, Dénes A, Reichart A, Toldi J, Gerendai I, et al. Novel tracing paradigms-genetically engineered herpesviruses as tools for mapping functional circuits within the CNS: present status and future prospects. Prog Neurobiol. 2004;72(6):417–45.
4. Boldogkői Z, Bálint K, Awatramani GB, Balya D, Busskamp V, Viney TJ, et al. Genetically timed, activity sensor and rainbow transsynaptic viral tools. Nat Methods. 2009;6:127–30.
5. Tombácz D, Tóth JS, Petrovszki P, Boldogkői Z. Whole-genome analysis of pseudorabies virus gene expression by real-time quantitative RT-PCR assay. BMC Genomics. 2009;10:491.
6. Tombácz D, Tóth JS, Boldogkoi Z. Effects of deletion of the early protein 0 gene of pseudorabies virus on the overall viral gene expression. Gene. 2012;493(2):235–42.
7. Szpara ML, Tafuri YR, Parsons L, Shamim SR, Verstrepen KJ. A Wide extent of inter-strain diversity in virulent and vaccine strains of alphaherpesviruses. PLoS Pathog. 2011;7(10), e1002282. doi:10.1371/journal.ppat.1002282.
8. Anselmo A, Flori L, Jaffrezic F, Rutigliano T, Cecere M, Cortes-Perez N, et al. Co-expression of host and viral microRNAs in porcine dendritic cells infected by the pseudorabies virus. PLoS One. 2011;6(3), e17374. doi:10.1371/journal.pone.0017374.
9. Wu YQ, Chen DJ, He HB, Chen DS, Chen LL, Chen HC, et al. Pseudorabies virus infected porcine epithelial cell line generates a diverse set of host microRNAs and a special cluster of viral microRNAs. PLoS One. 2012;7(1), e30988. doi:10.1371/journal.pone.0030988.
10. Gatherer D, Seirafian S, Cunningham C, Holton M, Dargan DJ, Baluchova K, et al. High-resolution human cytomegalovirus transcriptome. PNAS. 2011;108(49):19755–60.
11. van Beurden SJ, Gatherer D, Kerrc K, Galbraithd J, Herzykd P, Peetersa BPH, et al. Anguillid herpesvirus 1 transcriptome. J Virol. 2012;86(18):10150–61.
12. Juranic Lisnic V, Babic Cac M, Lisnic B, Trsan T, Mefferd A, Das Mukhopadhyay C, et al. Dual analysis of the murine cytomegalovirus and host cell transcriptomes reveal new aspects of the virus-host cell interface. PLoS Pathog. 2013;9(9), e1003611. doi:10.1371/journal.ppat.1003611.
13. Jones M, Dry IR, Frampton D, Singh M, Kanda RK, Yee MB, et al. RNA-seq analysis of host and viral gene expression highlights interaction between varicella zoster virus and keratinocyte differentiation. PLoS Pathog. 2014;10(1):e1003896. doi:10.1371/journal.ppat.1003896.
14. Majerciak V, Ni T, Yang W, Meng B, Zhu J, Zheng ZM. A viral genome landscape of RNA polyadenylation from KSHV latent to lytic infection. PLoS Pathog. 2013;9(11):e1003749. doi:10.1371/journal.ppat.1003749.
15. Heinz S, Benner C, Spann N, Bertolino E. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38(4):576–89.
16. Klupp BG, Hengartner CJ, Mettenleiter TC, Enquist LW. Complete, annotated sequence of the pseudorabies virus genome. J Virol. 2004;78(1):424–40.
17. Baumeister J, Klupp BG, Mettenleiter TC. Pseudorabies virus and equine herpesvirus 1 share a nonessential gene which is absent in other herpesviruses and located adjacent to a highly conserved gene cluster. J Virol. 1995;69(9):5560–7.
18. Boldogkői Z. Transcriptional interference networks coordinate the expression of functionally related genes clustered in the same genomic loci. Front Genet. 2012;5(3):122.
19. Klupp B, Kern H, Mettenleiter TC. The virulence-determining genomic Bam HI-fragment 4 of pseudorabies virus contains genes corresponding to the UL15 (partial), UL18, UL19, UL20, and UL21 genes of herpes simplex virus and a putative origin of replication. Virology. 1992;191:900–8.
20. Fuchs W, Klupp BG, Granzow H, Leege T, Mettenleiter TC. Characterization of pseudorabies virus (PRV) cleavage-encapsidation proteins and functional complementation of PRV pUL32 by the homologous protein of herpes simplex virus type 1. J Virol. 2009;83(8):3930–43.
21. Zhang G, Leader DP. The structure of the pseudorabies virus genome at the end of the inverted repeat sequences proximal to the junction with the short unique region. J Gen Virol. 1990;71(10):2433–41.
22. Priola SA, Stevens JG. The 5′ and 3′ limits of transcription in the pseudorabies virus latency associated transcription unit. Virology. 1991;182(2):852–6.
23. Cheung AK. Cloning of the latency gene and the early protein 0 gene of pseudorabies virus. J Virol. 1991;65(10):5260–71.
24. Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R. Comprehensive splice-site analysis using comparative genomics. Nucl Acids Res. 2006;34(14):3955–67.
25. Colgan DF, Manley JL. Mechanism and regulation of mRNA polyadenylation. Genes Dev. 1997;11:2755–66.
26. Neilson JR, Sandberg R. Heterogeneity in mammalian RNA 3′ end formation. Exp Cell Res. 2010;316(8):1357–64.
27. Beaudoing E, Freier S, Wyatt JR, Claverie J-M, Gautheret D. Patterns of variant polyadenylation signal usage in human genes. Genome Res. 2000;10(7):1001–10.
28. Yamada S, Imada T, Watanabe W, Honda Y, Nakajima-Iijima S, Shimizu Y, et al. Nucleotide sequence and transcriptional mapping of the major capsid protein gene of pseudorabies virus. Virology. 1991;185:56–66.
29. Tombácz D, Csabai Z, Oláh P, Havelda Z, Sharon D, Snyder M, et al. Characterization of novel transcripts in pseudorabies virus. Viruses. 2015;7(5):2727–44.
30. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105–11.
31. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
32. Rutherford K, Parkhill K, Crook J, Horsnell T, Rice P, Rajandream MA, et al. Artemis: sequence visualization and annotation. Bioinformatics. 2000;16(10):944–5.
33. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6.
34. Ahmed F, Kumar M, Raghava GP. Prediction of polyadenylation signals in human DNA sequences using nucleotide frequencies. In Silico Biol. 2009;9(3):135–48.

II.

Viruses **2015**, *7*, 2727-2744; doi:10.3390/v7052727

OPEN ACCESS

viruses

ISSN 1999-4915

www.mdpi.com/journal/viruses

Article

Characterization of Novel Transcripts in Pseudorabies Virus

Dóra Tombácz [1,†], Zsolt Csabai [1,†], Péter Oláh [1], Zoltán Havelda [2], Donald Sharon [3], Michael Snyder [3] and Zsolt Boldogkői [1,*]

[1] Department of Medical Biology, Faculty of Medicine, University of Szeged, Somogyi B. u. 4., Szeged H-6720, Hungary; E-Mails: tombacz.dora@med.u-szeged.hu (D.T.); csabai.zsolt@med.u-szeged.hu (Z.C.); olah.peter@med.u-szeged.hu (P.O.)

[2] Agricultural Biotechnology Center, Institute for Plant Biotechnology, Plant Developmental Biology Group, Szent-Györgyi A. u. 4, Gödöllő H-2100, Hungary; E-Mail: havelda@abc.hu

[3] Department of Genetics, School of Medicine, Stanford University, 300 Pasteur Dr., Stanford, CA 94305-5120, USA; E-Mails: dsharon@stanford.edu (D.S.); mpsnyder@stanford.edu (M.S.)

[†] These two authors contributed equally to this work.

[*] Author to whom correspondence should be addressed; E-Mail: boldogkoi.zsolt@med.u-szeged.hu; Tel.: +36-62-545595.

Academic Editor: Eric O. Freed

Received: 15 April 2015 / Accepted: 18 May 2015 / Published: 22 May 2015

Abstract: In this study we identified two 3′-coterminal RNA molecules in the pseudorabies virus. The highly abundant short transcript (CTO-S) proved to be encoded between the *ul21* and *ul22* genes in close vicinity of the replication origin (OriL) of the virus. The less abundant long RNA molecule (CTO-L) is a transcriptional readthrough product of the *ul21* gene and overlaps OriL. These polyadenylated RNAs were characterized by ascertaining their nucleotide sequences with the Illumina HiScanSQ and Pacific Biosciences Real-Time (PacBio RSII) sequencing platforms and by analyzing their transcription kinetics through use of multi-time-point Real-Time RT-PCR and the PacBio RSII system. It emerged that transcription of the CTOs is fully dependent on the viral transactivator protein IE180 and CTO-S is not a microRNA precursor. We propose an interaction between the transcription and replication machineries at this genomic location, which might play an important role in the regulation of DNA synthesis.

Keywords: non-coding RNA; pseudorabies virus; herpesvirus; RNA sequencing; DNA replication

1. Introduction

Pseudorabies virus (PRV), an alphaherpesvirus related to the human pathogens herpes simplex virus (HSV) and varicella-zoster virus, infects a wide range of mammalian species, including experimental rodents and pigs, the reservoir of the virus. PRV is commonly used in investigations of the molecular pathogenesis of herpesviruses [1,2], for the mapping of neural circuits [3–5] and for the delivery of genetically encoded fluorescence activity markers to the central nervous system [6] and cardiomyocytes [7]. During the past few years, a large variety of non-coding RNAs (ncRNAs) have been revealed in both cellular organisms and viruses. Micro (mi)RNAs (the best known ncRNAs) typically act to decrease the target mRNA level [8]. These transcripts are generated through the processing of long precursor RNA molecules. MicroRNAs have been detected in α-herpesviruses (HSV: [9], and PRV: [10,11]), betaherpesviruses (human cytomegalovirus (HCVM): [12]), and gammaherpesviruses (Epstein-Barr virus (EBV): [13]). These transcripts have been shown to play various roles, including the switch between the latent and lytic phases, evasion of host immune surveillance and apoptosis inhibition [14]. Long ncRNAs (lncRNAs) are the most abundant group of ncRNAs [15]. Numerous protein-encoding genes have been shown to specify antisense (as)-lncRNAs transcribed from the complementary DNA strands as templates. Large proportions of the mouse and human genomes have recently been reported to express lncRNAs [16,17]. The functions of these transcripts are still largely unknown. Many lncRNAs are involved in the regulation of transcription, such as *XIST* [18] and *HOTAIR* [19], or post-transcriptional regulation [20], or have structural roles [17]. Studies of multiple model systems have revealed that lncRNAs can function as modular scaffolds, forming extensive networks between chromatin regulators and various ribonucleoproteins [21]. Several polyadenylated lncRNAs have recently been demonstrated to be highly abundant in herpesviruses, including RNA2.7 in HCMV, accounting for nearly half of the total gene expression in RNA-Seq studies [22], and the widely-studied PAN RNA in Kaposi's sarcoma-associated herpesvirus [23], which has diverse roles during the viral life cycle [24]. The HSV latency-associated transcript (LAT) was the first identified as-lncRNA molecule [25] in alphaherpesviruses. A spliced 8.4-kb RNA, termed the long latency transcript (LLT), is generated from the complementary DNA strand of *ie180* and *ep0* genes under the control of the LAT promoter of PRV [26]. The expression of as-lncRNAs has also been detected in some other HSV genes [27–29]. Moreover, several antisense long non-coding transcripts have been discovered in HCMV [30] and EBV [31].

2. Materials and Methods

2.1. Cells and Viruses

An immortalized porcine kidney epithelial cell line (PK-15) was used for the propagation of PRV. The cells were cultivated in Dulbecco's modified Eagle medium supplemented with 5% fetal bovine serum (Gibco Invitrogen, Carlsbad, CA, USA) and 80 μg gentamycin/mL at 37 °C under 5% CO_2. The virus stock used for the kinetic analyses was prepared as follows: rapidly-growing semi-confluent PK-15 cells were infected at a multiplicity of infection (MOI) of 0.1 pfu/cell, and then incubated at 37 ℃ under 5% CO_2 until a complete cytopathic effect was observed. The infected cells were next frozen and thawed three times, followed by centrifugation at 10,000 × *g* for 15 min. The titer of the virus stock was determined by using the same cell type. For the transcription kinetic experiments, cells were infected at either a low (0.1 pfu/cell) or a high MOI (10 pfu/cell), and then incubated for 1 h. This was followed by removal of the virus suspension and washing with phosphate-buffered saline. Infected cells were incubated for various periods of time following the addition of new medium to the cells.

For Illumina DNA sequencing we mixed infected cells, which were incubated for 1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 or 24 h. For PacBio analysis, infected cells were incubated 1, 2, 4, 6, 8 or 12 h p.i. For Real-Time RT-PCR, infected PK-15 cells were incubated for 1, 2, 4, 6, 8, 12 or 24 h. Mock-infected cells, which were otherwise treated in the same way as the infected cells, were used as controls.

2.2. Generation of Recombinant Viruses

The generation of *ep0* and *vhs* gene-deleted viruses was described elsewhere (*vhs*-KO: [32], *ep0*-KO: [33]). Briefly, the desired viral genes were deleted by targeted mutagenesis using homologous recombination. Following subcloning of the target region of PRV, a *lacZ* gene expression-cassette was inserted in place of the genes to be deleted in both mutants. Mutant viruses were selected on the basis of the blue plaque phenotype.

2.3. RNA Isolation for RNA-Seq and Real-Time RT-PCR

Total RNA was purified by using the Nucleospin RNA kit (Macherey-Nagel), following the kit protocol. Cells were collected by low-speed centrifugation, lysed in a buffer containing the chaotropic ions needed for the inactivation of RNases and providing the conditions for the binding of nucleic acids to a silica membrane. Contaminating DNA was removed with RNase-free rDNase solution (included in the kit). The isolated total RNA was treated by means of the TURBO DNA-free™ Kit (Life Technologies) to remove potential residual DNA contamination. RNA concentration was determined by Qubit 2.0, and RNA integrity was assessed by using an Agilent 2100 Bioanalyzer. Samples were stored at −80 °C.

2.4. Illumina HiScanSQ cDNA Sequencing

Preparation of cDNA libraries—strand-specific total RNA libraries were prepared for sequencing through use of the Illumina ScriptSeq v2 RNA-Seq Library Preparation Kit (Epicentre, Madison, WY USA) for random hexamer primed amplification and the sequencing of 2 × 100 bp fragments. For PA-Seq, a single-end library was constructed by using custom-anchored adaptor-primer oligonucleotides with an oligo(VN)T_{20} primer sequence. Anchored primers compensate for the loss in throughput due to the high fraction of reads containing solely adenine bases when conventional oligo(dT) primers are used.

Transcriptome sequencing was performed on an Illumina HiScanSQ platform, generating ~200 million paired-end reads of 100 bp length and ~105 million 50 bp single-end reads. The quality assessment of the raw read files was achieved with FastQC v0.10.1. Reads were aligned to the respective host genome (*Sus scrofa*, assembly: Sscrofa10.2) and subsequently to the PRV genome (KJ717942.1) by using Tophat v2.09. [34]; ambiguous reads were discarded. For PA-Seq, mapping was carried out with Bowtie v2. [35], and polyA peaks were detected through the use of in-house scripts, based on the criteria of the presence of a PA signal in the 50 bp region upstream from the PA site and the presence of at least two consecutive adenine mismatches in at least 10 independent reads at the PA site. Annotation and visualization were carried out with the Artemis Genome Browser v15.0.0 [36]. Any GC bias of the alignments was inspected with the Bioconductor R package.

2.5. PacBio RS II cDNA Sequencing

2.5.1. PolyA RNA Purification

Polyadenylated RNAs were isolated from the total RNA samples by using the Oligotex mRNA Mini Kit (Qiagen, Venlo, The Netherlands) according to the kit instructions for the Oligotex mRNA Spin-Column Protocol.

2.5.2. cDNA Synthesis

The PolyA RNA samples were quantified with the Qubit RNA HS Assay Kit (Life Technologies, Carlsbad, CA, USA) and converted to cDNAs with the SuperScript Double-Stranded cDNA Synthesis Kit (Life Technologies). RT reactions were primed with an Anchored Oligo(dT)$_{20}$ primer (Life Technologies). The cDNAs were quantified with the Qubit HS dsDNA Assay Kit (Life Technologies) and quality was assessed with the Agilent 2100 bioanalyzer.

2.5.3. Library Preparation, Sequencing and Data Collection

SMRTbell libraries were generated by using the PacBio DNA Template Prep Kit 2.0 and the Pacific Biosciences template preparation and sequencing protocol for Very Low (10 ng) Input 2 kb libraries with carrier DNA (pBR322, Thermo Scientific, Waltham, MA, USA). SMRTbell templates were bound to polymerases by using the DNA polymerase binding kit XL 1.0 (part #100-150-800) and v2 primers.

Polymerase-template complexes were bound to magbeads with the Pacific Biosciences MagBead Binding Kit, and sequencing was carried out on the Pacific Biosciences RSII sequencer with C3 sequencing reagents. Movie lengths were 180 min (one movie was recorded for each SMRT Cell). Subread filtering and alignment were carried out in SMRT Pipe v2.2.0. Visualization and data analysis were performed in SMRT Analysis v2.2.0.

2.6. Normalization of PacBio Data with Mitochondrial Transcripts

The read counts of viral transcripts at each time-point were normalized to mitochondrial read counts, aligned to the *Sus scrofa* 10.2 MT chromosome sequence. The following mitochondrial genes were used for the normalization: ATP6; ATP8; CYTB; ND1; ND2; ND3; ND4; ND4L; ND5; ND6; COX1; COX2 and COX3. While the degradation of cytoplasmic mRNAs during alphaherpesvirus infection has been previously shown [37,38], no such evidence is known for mtRNAs. Although recent studies have shown the steady decrease of mtDNA levels in Vero cells expressing the UL12.5 gene of

HSV-1 [39]. We chose the mtRNAs as reference RNAs because the UL12.5 gene is absent from the PRV genome.

2.7. Reverse Transcription

RT reactions were carried out with 70 ng of total RNA with the use of Superscript III enzyme (Life Technologies) and gene-specific primers or oligo(dT) primers.

2.8. Real-Time PCR

Real-Time PCR reactions were performed in a volume of 20 µL with Absolute QPCR SYBR Green Mix (Thermo Scientific) containing 7 µL of 10-fold diluted cDNA, 1.5 µL of forward and 1.5 µL of reverse primers (10 µM each; Table 1A). 28S ribosomal (r)RNA was used as a reference gene in each run. The PCR amplification conditions were as follows: 15 min at 95 ℃ for the enzyme activation, followed by 30 cycles of 94 ℃ for 25 s (denaturation), 60 ℃ for 25 s (annealing), and 72 ℃ for 6 s (extension).

Table 1. Primer sequences for the Real-Time RT PCR analysis.

	Name	**Sequence (5′-3′)**	**Genomic Position**
A	CTO-S fw	GACGATCCGGCGGTCCCA	63858–63875
	CTO-S rev	GCGCCACAACCCGGAGC	63915–63931
	CTO-L fw	GTG TCG CGG ACA GAG AAT GG	64604–64623
	CTO-L rev	GGC CCA GTA CCT GTT TCA GC	64708–64727
B	T7-CTO-out fw	TAATACGACTCACTATAGGGAGAGGTCTCTAAGGGGGAACCAG	63605–63626
	SP6-CTO-out rev	ATTTAGGTGACACTATAGAAGNGCCGAAAAATTCGCACATACC	63989–64008

(underline: T7 and SP6 promoter sequences, respectively).

Relative expression ratios (R) were calculated via the following formula:

$$R = \frac{(E_{\text{sample·max}})^{Ct_{\text{sample·max}}}}{(E_{\text{sample}})^{Ct_{\text{sample}}}} : \frac{(E_{\text{ref·max}})^{Ct_{\text{ref·max}}}}{(E_{\text{ref}})^{Ct_{\text{ref}}}} \quad (1)$$

where E is the amplification efficiency, Ct is the threshold cycle number, “sample” refers to the examined PRV transcript and “ref” refers to the 28S rRNA (internal control). The cDNAs were normalized to 28S cDNAs by using the Comparative Quantitation module of the Rotor-Gene Q software (Version 2.3.1, Qiagen), which automatically calculates the efficiency of the reaction. Thresholds were also set by the software.

2.9. Treatment of Cells with CHX

The requirement of *de novo* protein synthesis for CTO production was tested by cycloheximide (CHX) analysis. Cells were incubated in the presence or absence of 100 µg/mL CHX (Sigma-Aldrich, St. Louis, MO, USA) for 1 h prior to virus infection. Mock-infected cells otherwise treated in the same way as infected cells were used as controls.

2.10. Northern Blot Analysis

Traditional Northern blot assay Total RNA was isolated from PK-15 cells through use of TRIzol reagent (Life Technologies) according to the manufacturer's instructions. Samples were denatured in loading buffer for 5 min at 65 °C. Extracted RNA samples (10 ug) were fractionated in formaldehyde/1.2% agarose gel, transferred to a Nytran N membrane (Schleicher & Schuell BioScience, Dassel, Germany) by a capillary method and fixed by ultraviolet cross-linking. The membrane was probed by using the random primed PCR product or the total viral DNA with the DecaLabel DNA Labeling Kit (Fermentas, Vilnius, Lithuania). PCR reactions were carried out with AccuPrime GC-Rich DNA Polymerase (Life Technologies) according to the manufacturer's recommendations (primer sequences Table 1B). The oligonucleotide probe was labeled with [α-^{32}P]CTP. Filter prehybridization was carried out in 50% formamide, 0.5% SDS, 5× SSPE, 5× Denhardt's solution and 20 μg/mL sheared, denatured salmon sperm. The probe was heated for 1 min at 95 °C. Overnight hybridization was carried out at 68 °C. Finally, the hybridization membranes were washed in 2 × SSC 0.1% SDS at 68 °C for ones 10 min, 0.5 × SSC, 0.1% SDS at 68 °C for 10 min, 0.1 × SSC 0.1% SDS at 68 °C for 10 min.

Micro RNA Northern blot analysis two different PCR probes were used. Forward primers were linked with the T7 promoter sequence and reverse primers were linked with the SP6 promoter sequence (Table 1B). Samples (10 μg) were fractionated on denaturing 12% polyacrylamide gels containing 8 M urea, transferred to a Nytran N membrane (Schleicher & Schuell, Germany) by a capillary method and fixed by ultraviolet cross-linking. Prehybridization was carried out in 50% formamide, 0.5% SDS, 5 × SSPE, 5 × Denhardt's solution and 20 μg/mL sheared, denatured salmon sperm DNA. Overnight hybridizations were performed in the same solution at 37 °C. An [α-^{32}P]UTP-labeled RNA probe was used for the hybridization. Membranes were washed twice for 10 min with a solution containing 2 × SSC, 0.1% SDS.

3. Results

3.1. Identification and Structural Characterization of Novel lncRNAs in PRV

The PRV transcriptome was analyzed by means of the Illumina HiScanSQ and Pacific Biosciences (PacBio) RSII sequencing systems. Random hexamer-primed reverse transcription (RT) was used for Illumina sequencing, and oligo(dT)-primed (PA-Seq) RT for both platforms. With these techniques, we detected two novel 3′-coterminal transcripts located between the *ul21* and *ul22* genes, close to the OriL, termed CTOs. The length of the short intergenic lncRNA (CTO-S) is 286 base pairs (bp) and is mapped to bp-s 63673-63958 of the PRV reference genome KJ717942.1 (Figure 1). The attachment of adapter sequences to the Illumina RT primers allowed the analysis of transcription from both DNA strands separately. These investigations revealed that only one of the two DNA strands exhibits transcriptional activity at this genomic region. The long (CTO-L) transcript overlaps OriL, and maps to nucleotides (nt) 63673–66287 (2615 bp). CTO-L originates from the promoter of the *ul21* gene and is produced by the continuation of the RNA polymerase molecule across the transcription termination sequences. CTO-L contains the entire *ul21* gene sequence and is therefore a sense lncRNA. The promoter of the CTO-S transcript was identified in nucleotides 63952–63958 by the Tfsearch algorithm with 96.8% confidence. An Oct1 transcription factor binding site was also discovered at 98.3% confidence in the TransFac database [40].

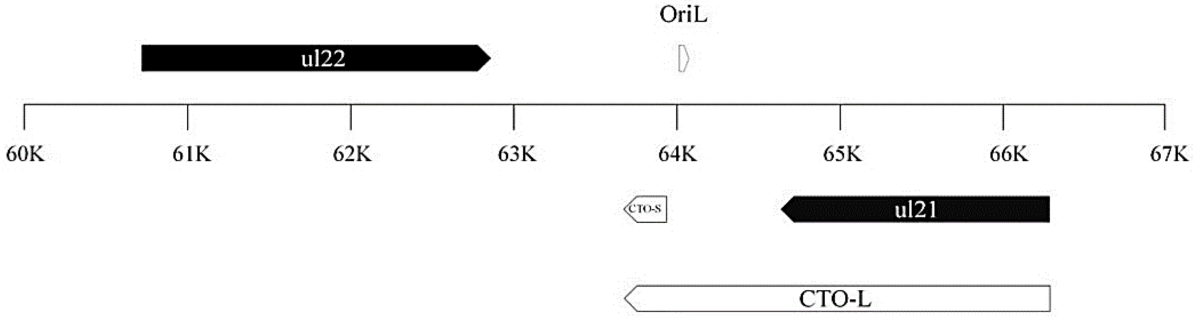


Figure 1. Location of *cto* genes on the PRV genome. Both *cto* transcripts (CTOs) are polyadenylated RNAs with a common 3′ termination. CTO-L is generated by the continuation of transcription after the termination signals of the *ul21* gene. OriL is the replication origin of in the UL region of viral DNA mapped between the *ul21* and *ul22* genes.

3.2. Transcriptional Analysis of CTO

3.2.1. Illumina RNA-Seq Analysis

We combined transcripts isolated from consecutive time-points of viral infection for Illumina sequencing. CTO-S proved to exhibit a very high expression, with an RPKM (reads per kilobase per million) value of 1.6×10^6 in the random-hexamer primed library, and 45.9% of the total read count in PA-Seq, making this transcript by far the most abundant viral RNA molecule. CTO-L produced only 0.13% of the total reads in the random hexamer-primed library (RPKM = 5×10^{-4}). However, pA-Seq produces more informative data than random hexamer-primed sequencing: in the former case the read numbers are in strict correlation only with the transcript abundance, whereas in the latter case they correlate with the transcript lengths too.

3.2.2. PacBio RNA-Seq Analysis

For the analysis of the transcription kinetics of the CTO length variants, we applied the PacBio RS II system, which is capable of generating significantly longer read lengths than those of second-generation technologies, such as Illumina. The CTO expression was analyzed at 1, 2, 4, 6, 8 or 12 h by using high [10 plaque forming unit (pfu)/cell] infection conditions. Due to template quantity we used the very low input protocol for the template preparation and sequencing, which is not optimal for the detection of small (<700 nt) transcripts, and we therefore observed a very strong bias against CTO-S. Due to the low sensitivity of this technique for small DNA fragments, the real proportion of the two transcripts cannot be precisely ascertained through its use. However, the data obtained could be used to compare the transcription kinetics in the two transcripts (Figure 2). The viral RNA reads were normalized with the pig mitochondrial RNAs, which are thought to resist degradation by the RNase activity of viral proteins. No reads were obtained for either of the CTO transcripts in the first hour of infection. A low amount of CTO-L was detected 2 h post-infection (p.i.). The CTO-S transcript appeared in only the 4 h p.i. samples (Figure 2A). The logarithmic plots demonstrate a slight increase in the dynamics of transcriptions between 4 and 6 h p.i. in both transcripts (Figure 2B), followed by an elevated expression rate, especially in CTO-L, which increased very steeply after 6 h p.i.. However, analysis of the transcriptional activity normalized to the copy number of PRV DNA (determined by Real-Time RT-PCR) demonstrated that the expression from individual DNA molecules was highest at 8 h p.i. for both transcripts (Figure 2C).

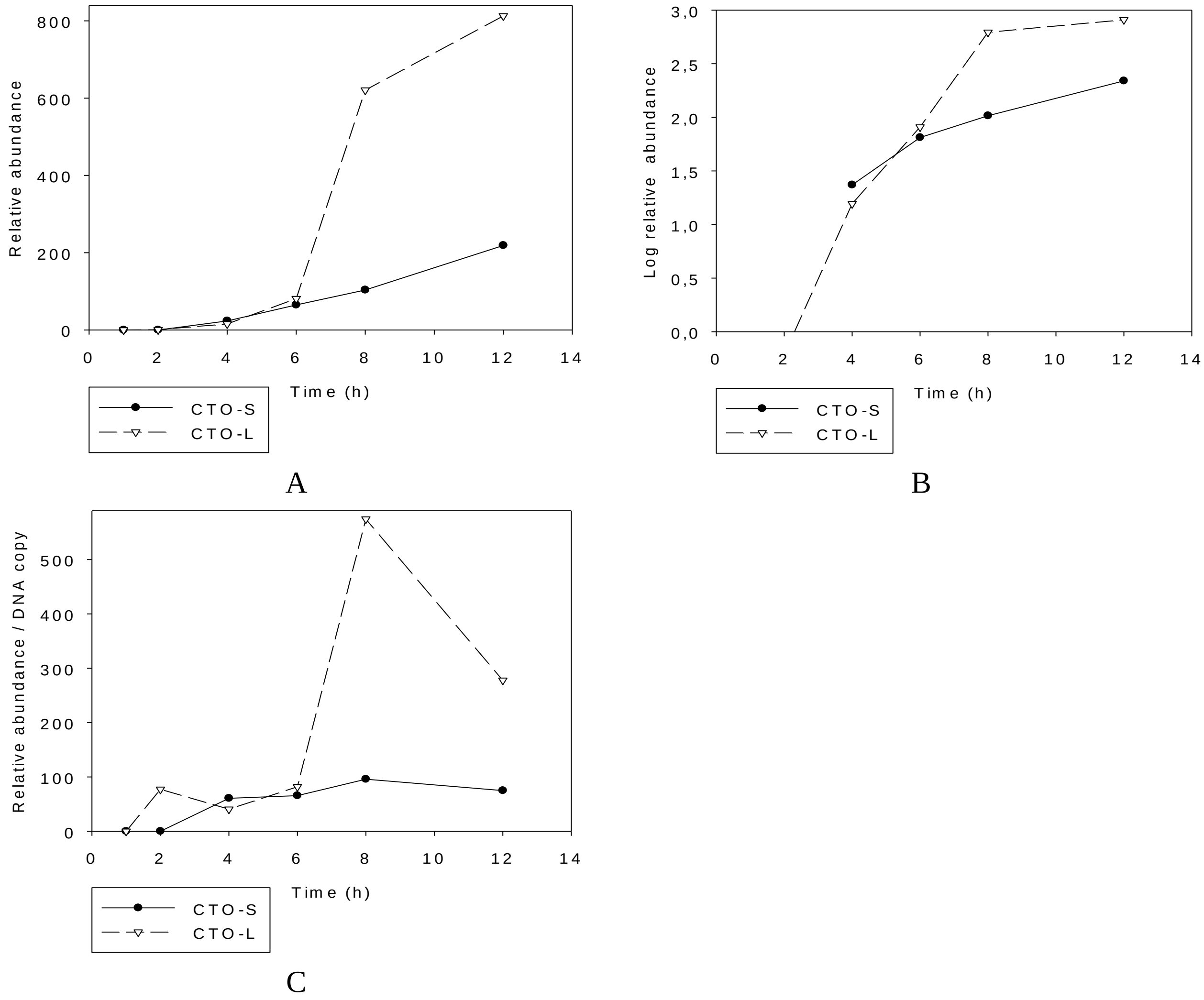


Figure 2. Transcription kinetics of CTO transcripts. The relative abundances of transcripts are depicted on a linear (**A**) and a logarithmic (**B**) scale. All RNA reads obtained by PacBio sequencing were normalized with mitochondrial RNAs. The transcriptional activity of the CTOs was also analyzed by normalizing the data with the relative amount of viral DNAs (**C**).

We examined whether the efficiency of transcriptional readthrough varied in time by comparing the amounts of CTO-L and ul21 mRNA (Figure 3A,B). The data revealed that the ratio of CTO-L to the ul21 transcript increased continuously in time. An examination as to whether this was simply due to a higher transcription rate of individual *ul21* genes did not indicate an ycorrelation between the readthrough efficiency and the transcriptional activity of this gene when the transcript reads were normalized with the DNA copy number (Figure 3C). This suggests that the efficiency of the recognition of transcriptional termination sequences might be regulated by a specialized mechanism.

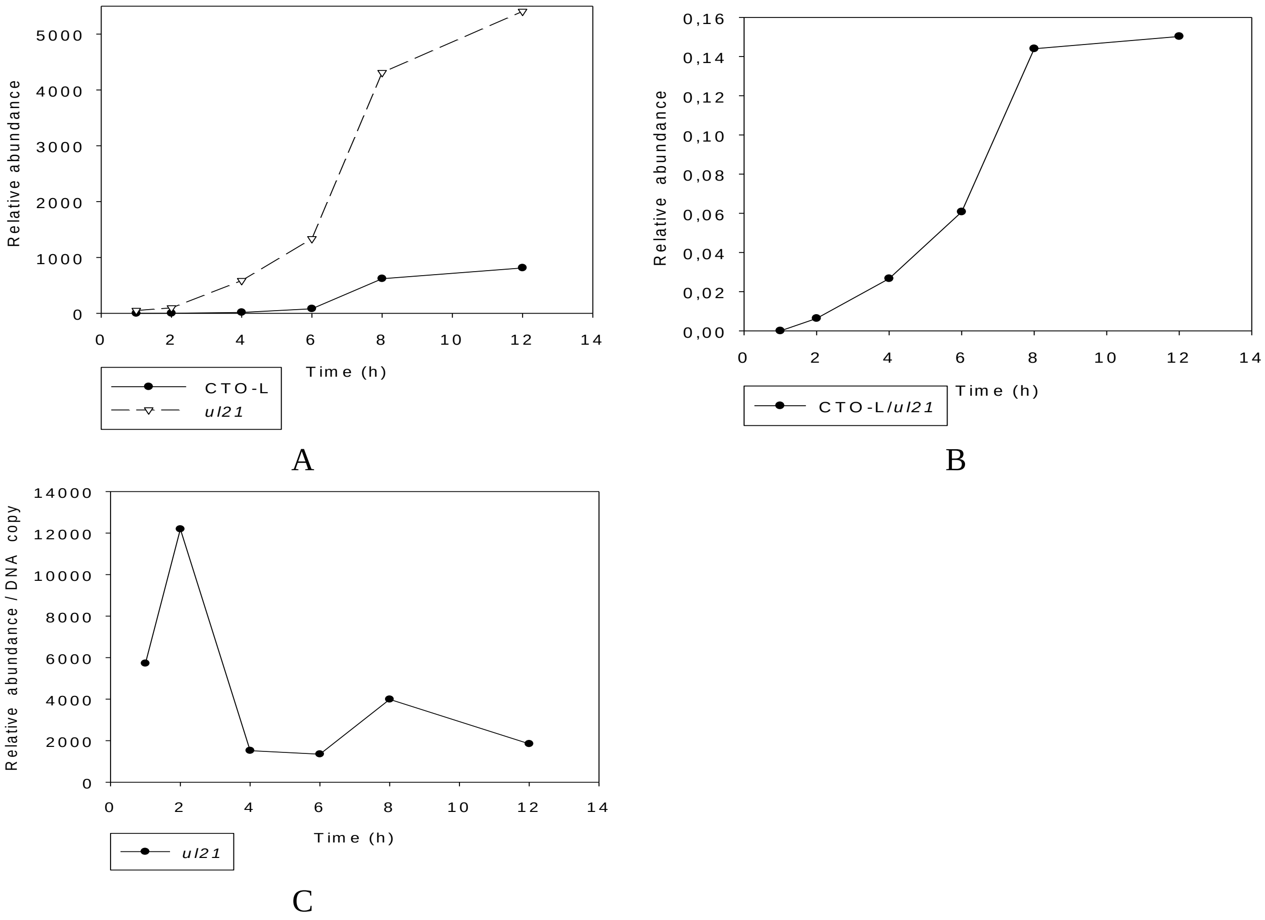


Figure 3. Comparison of the amounts of ul21 and CTO-L transcripts. CTO-L is a readthrough product of the *ul21* gene. (**A**) shows the transcriptional kinetics of the two transcripts, while (**B**) shows the change in readthrough efficiency with time. All RNA reads obtained by PacBio sequencing were normalized with mitochondrial RNAs. (**C**) shows the transcriptional activity normalized to the viral genome.

3.2.3. Multi-Time-Point Real-Time RT-PCR Analysis of CTOs in Wild-Type (wt) and Mutant Backgrounds

Wild-type PRV

Strand-specific priming-based RT was used for the kinetic assay of the abundant CTO-S transcript in both low-titer (0.1 pfu/cell) and high-titer (10 pfu/cell) infection. The method used for the calculation of relative expression ratios (R) was as described earlier [41]. Real-Time RT-PCR analyses confirmed the PacBio and Illumina RNA-Seq results, showing that practically no transcription occurred in the first 2 h of the viral life cycle in the genomic region encoding CTO-S (Figure 3A). In the high-titer infection, CTO-S reached very high levels by 4 h p.i. (Figure 4A), which means that the expression of this transcript is initiated sometime between 2 and 4 h p.i. In the low-titer experiment, however, CTO-S was expressed at a very low level at 4 h, but reached a high level by 6 h p.i. (Figure 4B). Thus, there is a shift in the expression kinetics of CTO-S transcripts in low-pfu as compared with high-pfu experiments. The CTO-L expression was examined by using strand-specific primers for the reverse transcriptions at 1, 2, 4, 6, 8, 12, 18 and 24 h at high-titer infection (Figure 4C) and 1, 2, 4, 6 and

8 h at low-titer infection (Figure 4D). There was no significant expression until 4 h p.i., which confirmed the PacBio sequencing data.

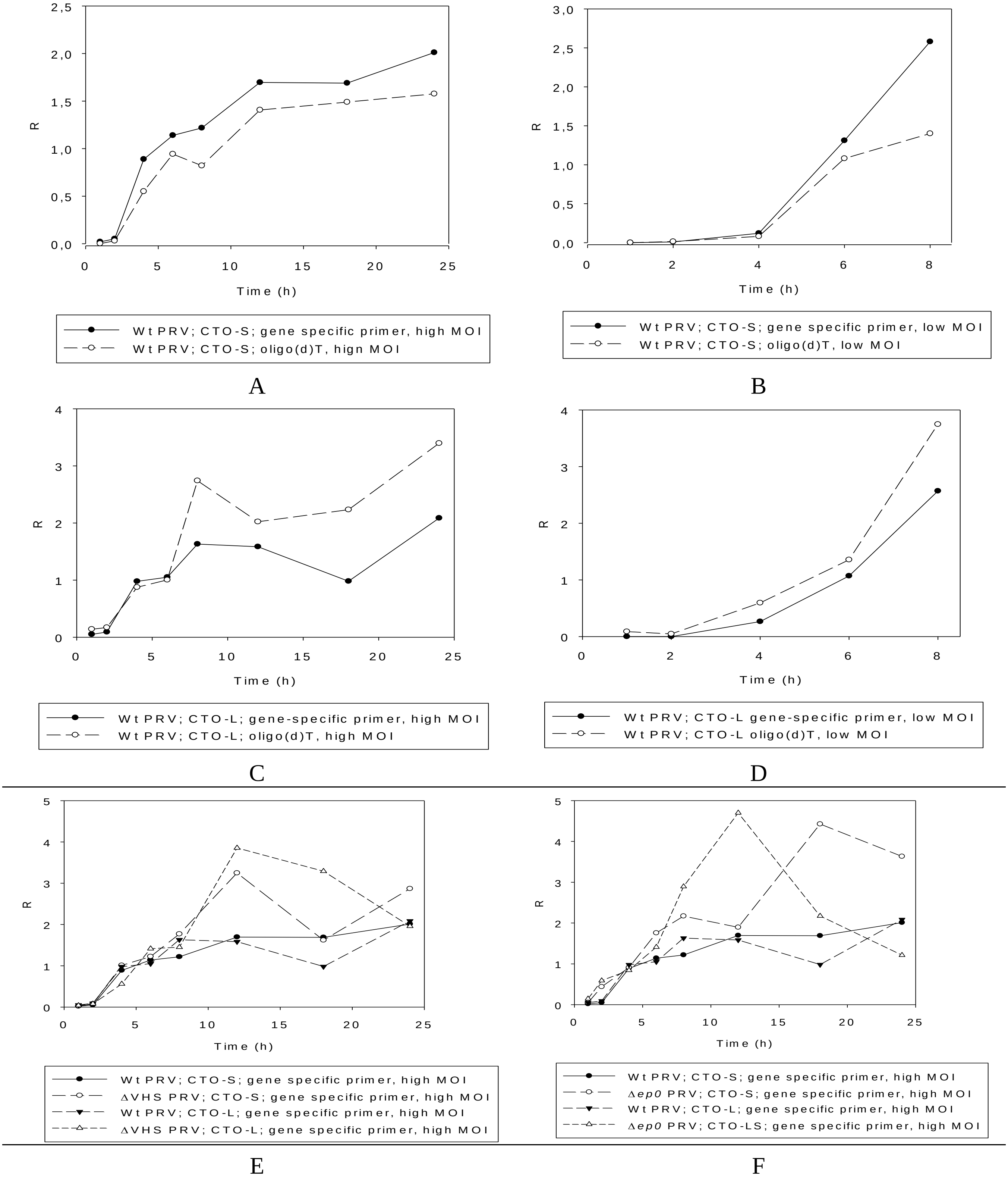


Figure 4. The change in relative expression ratio (R) of the CTO transcripts with time, determined by Real-Time RTR-PCR. (**A**). CTO-S: High-titer (10 pfu/cell) infection. The cDNAs were generated by reverse transcription of CTO-S transcripts through the use of

gene-specific or oligo(dT) primers. (**B**). CTO-S: Low-titer (0.1 pfu/cell) infection. The cDNAs were generated by reverse transcription of CTO-S transcripts through the use of gene-specific or oligo(dT) primers. (**C**). CTO-L: High-titer (10 pfu/cell) infection. The cDNAs were generated by reverse transcription of CTO-S transcripts through the use of gene-specific or oligo(dT) primers. (**D**). CTO-L: Low-titer (0.1 pfu/cell) infection. The cDNAs were generated by reverse transcription of CTO-S transcripts through the use of gene-specific or oligo(dT) primers. (**E**). Expression of CTO-S in *vhs-KO* background following high-titer infection. (**F**). Expression of CTO-S in *ep0-KO* background following high-titer infection.

Mutant PRVs

Two mutant viruses were used to analyze the CTO-S and CTO-L transcription kinetics in order to evaluate the potential effects of mutations on the expression kinetics of this transcript. The levels of the two transcripts were higher than that of the wt virus in the *vhs-KO* virus (Figure 4E), which is not surprising since the virion host shut-off (VHS) protein plays a role in the destabilization of RNA molecules [42]. We earlier reported similar transcription kinetics for the rest of the PRV genes [32]. The expression kinetics of the CTOs in the *ep0-KO* (ep0: early protein 0) background, however, exhibits an atypical pattern since the transcript levels of other late genes of the wt virus are generally higher than those of the *ep0*-null mutant virus, which is not the case for these lncRNAs (Figure 4F). Additionally, in contrast to the wt virus, the level of CTO-S and CTO-L is relatively high at 2 h p.i. in this mutant virus. Thus, EP0 appears to exert a down-regulatory effect on the transcription of CTOs throughout the whole life cycle of the virus.

3.3. CTO Expression is Controlled by the IE180 Transactivator of PRV

The transcription of PRV genes is controlled by the IE180 transactivator protein. Cycloheximide (CHX), an inhibitor of protein synthesis in eukaryotic cells, completely blocked gene expression, except in the *ie180* gene itself and two as-lncRNA-encoding genes, LAT and LLT. The repression of CTO expression in the presence of CHX indicates that the transcription of these molecules is fully dependent on IE180 (Figure 5). Earlier we had shown that *ul21* gene expression was completely repressed by CHX treatment [41], which—due to their sharing a common promoter—also resulted in the silencing of CTO-L transcription.

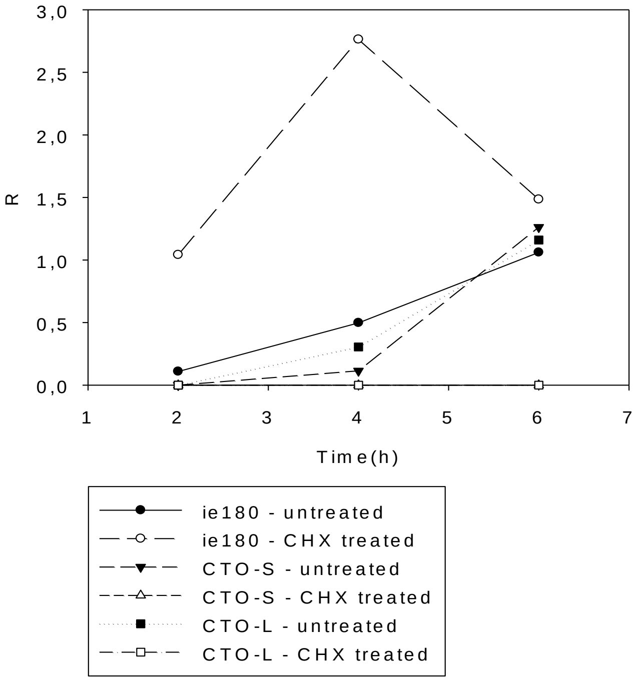


Figure 5. Analysis of transcription following CHX treatment of infected cells. The *ie180* gene is not repressed by cycloheximide (CHX), but CTO-S is totally blocked by this protein synthesis inhibitor, the reason for this being that CTO-S requires the IE180 protein for its expression.

3.4. CTO Expression in the Presence of an Inhibitor of DNA Replication

We additionally investigated the effect of phosphonoacetic acid (PAA), an inhibitor of DNA synthesis, on the transcription kinetics of CTO-S. The method of calculation for the evaluation of the repressive effect of PAA on the transcription of the individual genes was published earlier [41]: $R_{i\text{-PAA}} = R_{6h\text{-PAA}}/R_{6h\text{-UT}}$. In the present study, the average $R_{i\text{-PAA}}$ values were found to be 0.717 for early genes and 0.113 for late genes. The value of $R_{i\text{-PAA}} = 0.184$ for CTO-S and $R_{i\text{-PAA}} = 0.361$ for CTO-L confirmed the result of our kinetic analyses in non-treated samples: these transcripts are expressed with late kinetics.

3.5. Northern Blot and in Silico *Analyses Revealed that CTO-S is not a miRNA Precursor*

Our investigation of whether CTO-S might be a miRNA precursor by using microRNA Northern blot analysis, however, did not detect any transcript with miRNA size in this genomic region (data not shown). Conversely, CTO-S RNA was detected by using traditional Northern blot analysis (Figure S1). Due to the very low copy number, we could not detect CTO-L by Northern blot analysis, however, the existence of this transcript was verified by four independent techniques (PacBio PA-seq, two Illumina RNA-seq methods and Real-time RT-PCR). Sequence analysis of CTO by using the pre-microRNA hairpin prediction tools miRNAFold [43] and miPred [44] yielded negative results in each case. Moreover, previous studies of the miRNA expression in PRV in both porcine dendritic [10] and epithelial [11] cell lines failed to detect miRNAs from the genomic region of CTO.

4. Discussion

In this study we report the identification and characterization of two lncRNAs of pseudorabies virus. The two transcripts share a common poly(A) signal. CTO-S, a short intergenic lncRNA molecule is found to be very abundant, but is not expressed in the first 2 h of viral infection. We have demonstrated that the expression of CTO-S is controlled by the virally-encoded IE180 transactivator. CTO-L is expressed at a relatively low level, produced from the promoter of the *ul21* gene through occasional transcriptional readthrough events across the transcription termination signal of this gene. The levels of CTO-S transcripts in the mutant viruses are higher at every time point than in the wt PRV, indicating a role of these gene products in the stability and/or the regulation of these molecules at the level of transcription. The vicinity of CTO-S to OriL and the overlap of CTO-L with OriL suggest a role of this genomic region in the regulation of DNA replication, which may be based on the interference between the transcriptional and replication machineries, as suggested by Huvet *et al*. [45]. Others did not verify the Huvet model, at least in human cells [46]. Despite this, interference between the two apparatuses may be an existing mechanism; the RNA polymerase molecules transcribing CTO-L might clash with DNA polymerase, thereby preventing the progress of replication in one of the two directions (Figure 6). It has been hypothesized, but never proved, that the synthesis of alphaherpesvirus DNAs starts with δ-type replication, which is followed by a switch to sigma-type replication generating concatemers [47]. The transcription of CTO-S may facilitate replication in another way, through separation of the two DNA strands, thereby helping the progression of DNA polymerase in one direction. In this scenario, the lack of CTO expression in the first few hours of the viral life cycle allows bidirectional δ-type replication; later, the process of CTO transcription itself makes the replication unidirectional through the two mechanisms proposed above. If there is no δ-type replication, and the viral DNA synthesis starts with CTO transcription itself, this makes the replication unidirectional through the two mechanisms proposed above. If there is no δ-type replication, and the viral DNA synthesis starts with concatemer formation immediately, the above putative mechanism might also contribute to the unidirectionality of DNA synthesis. Overall, extensive transcriptional activity near oriL potentially exerts an effect on the DNA replication by determining the orientation of the DNA synthesis and perhaps contributing to the switch from bidirectional to unidirectional replication.

The polyadenylation of CTO-S indicates that this transcript may have additional function(s) in the life cycle of the virus. These transcripts do not have an essential role in the viral replication since two strains (TJ [48] and ZJ01 [49]) contain deletions at this genomic region.

In this putative mechanism, the transcripts are merely by-products. It has been shown in a variety of organisms that a similar mechanism based on the clash between two RNA polymerase molecules in the overlapping region may have a regulatory effect on the transcription through transcriptional interference [50–52].

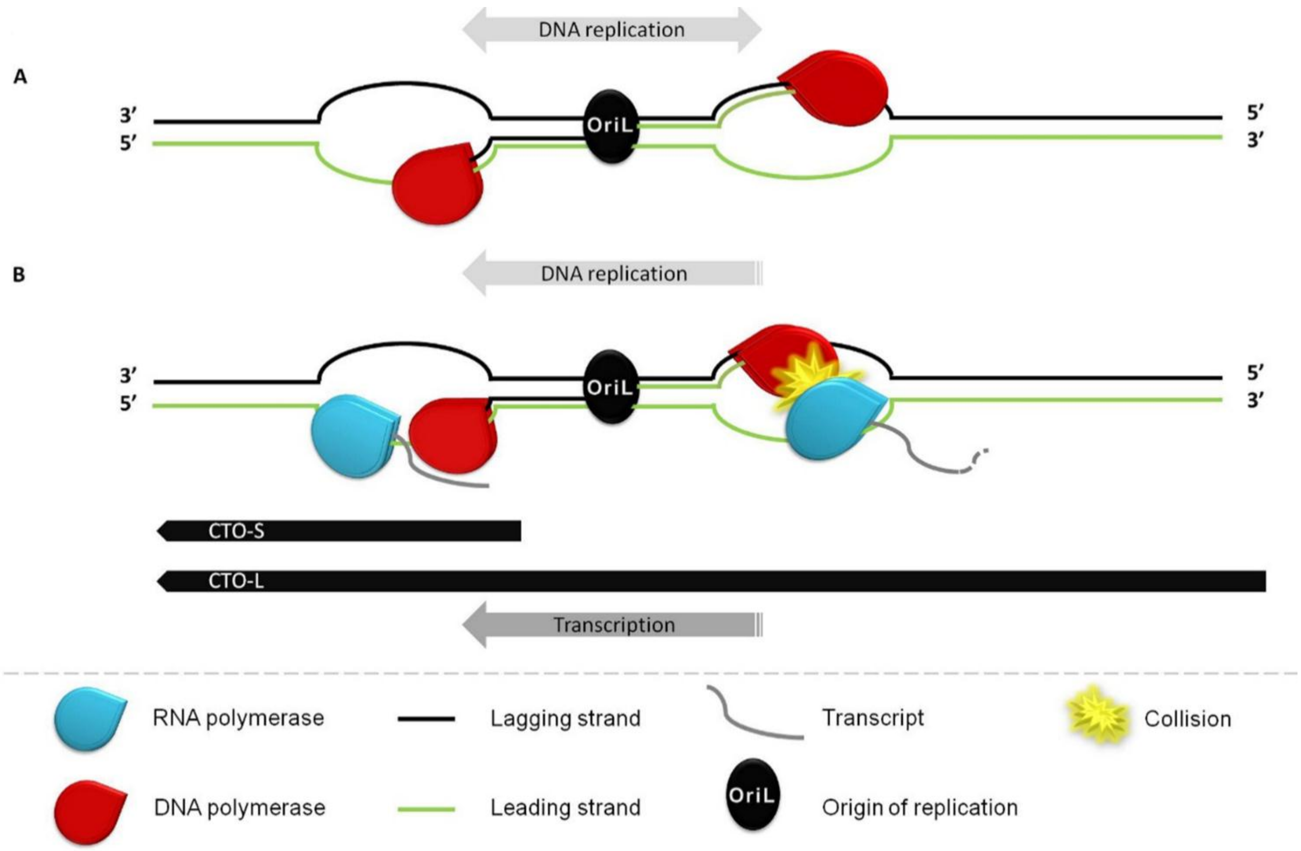


Figure 6. A proposed model for the interactions between the replication and transcription machineries. (**A**) There is no CTO expression in the early stage of infection, and this allows the bidirectional synthesis of DNA (δ type replication). (**B**) Later, the transcription machineries of the CTOs facilitate unidirectional DNA replication through two mechanisms: (1) the DNA polymerase collides with the RNA polymerase synthesizing CTO-L, thereby halting the progression and/or preventing the assembly of the replication machinery (right to OriL); and (2) RNA polymerase transcribing the CTO-S (or CTO-L) facilitates the progression of DNA polymerase in one way through unwinding of the two DNA strands (left to OriL). For the sake of simplicity, the DNA synthesis from the lagging strand is not depicted.

Supplementary Materials

Supplementary materials can be found at http://www.mdpi.com/1999-4915/7/5/2727/s1.

Acknowledgments

This research was supported by the Swiss-Hungarian Cooperation Programme grant SH/7/2/8 to Zsolt Boldogkői, and by CEGS NIH grant 1P50HG007735-01 to Michael Snyder.

Author Contributions

Dóra Tombácz devised and performed Real-Time PCR and PacBio experiments and data analysis, prepared figures and contributed to writing the manuscript. Zsolt Csabai devised and performed

Real-Time PCR and Northern-blot experiments. Péter Oláh performed Illumina analysis and contributed to writing the manuscript. Donald Sharon devised and performed PacBio experiments. Zoltán Havelda coordinated the Northern-blot work. Zsolt Boldogkői generated the recombinant viruses, designed the research plan, organized the study and write the final version of the manuscript. All authors discussed experiments and analysis and collaborated on the final version.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

1. 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.
2. Szpara, M.L.; Kobiler, O.; Enquist, L.W. A common neuronal response to alphaherpesvirus infection. *J. Neuroimmune Pharmacol.* **2010**, *5*, 418–427.
3. Strack, A.M. Pseudorabies virus as a transneuronal tract tracing tool: Specificity and applications to the sympathetic nervous system. *Gene Ther.* **1994**, *1*, S11–S14.
4. Card, J.P.; Enquist, L.W. Transneuronal circuit analysis with pseudorabies viruses. *Curr. Protoc. Neurosci.* **2001**, *1*, doi:10.1002/0471142301.ns0105s09
5. Boldogkői, Z.; Sík, A.; Dénes, A.; 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.
6. Boldogkői, Z.; Bálint, 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.
7. Prorok, J.; Kovács, P.P.; Kristóf, A.A.; Nagy, N.; Tombácz, D.; Tóth, J.S.; Ördög, 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.
8. Janowski, B.A.; Kaihatsu, K.; Huffman, K.E.; Schwartz, J.C.; Ram, R.; Hardy, D.; Mendelson, C.R.; Corey, D.R. Inhibiting transcription of chromosomal DNA with antigene peptide nucleic acids. *Nat. Chem. Biol.* **2005**, *1*, 210–215.
9. Umbach, J.L.; Kramer, M.F.; Jurak, I.; Karnowski, H.W.; Coen, D.M.; Cullen, B.R. MicroRNAs expressed by herpes simplex virus 1 during latent infection regulate viral mRNAs. *Nature* **2008**, *454*, 780–783.
10. Anselmo, A.; Flori, L.; Jaffrezic, F.; Rutigliano, T.; Cecere, M.; Cortes-Perez, N.; Lefèvre, F.; Rogel-Gaillard, C.; Giuffra, E. Co-expression of host and viral microRNAs in porcine dendritic cells infected by the pseudorabies virus. *PLoS ONE* **2011**, *6*, e17374.
11. Wu, Y.Q.; Chen, D.J.; He, H.B.; Chen, D.S.; Chen, L.L.; Chen, H.C.; Liu, Z.F. Pseudorabies virus infected porcine epithelial cell line generates a diverse set of host microRNAs and a special cluster of viral microRNAs. *PLoS ONE* **2012**, *7*, e30988.

12. Grey, F.; Antoniewicz, A.; Allen, E.; Saugstad, J.; McShea, A.; Carrington, J.C.; Nelson, J. Identification and characterization of human cytomegalovirus-encoded microRNAs. *J. Virol.* **2005**, *79*, 12095–12099.
13. Pfeffer, S.; Zavolan, M.; Gr ässer, F.A.; Chien, M.; Russo, J.J.; Ju, J.; John, B.; Enright, A.J.; Marks, D.; Sander, C.; Tuschl, T. Identification of virus-encoded microRNAs. *Science* **2004**, *304*, 734–736.
14. Li, K.; Ramchandran, R. Natural antisense transcript: A concomitant engagement with protein-coding transcript. *Oncotarget* **2010**, *1*, 447–452.
15. Mattick, J.S; Makunin, I.V. Non-coding RNA. *Hum. Mol. Genet*. **2006**, *15*, R17–R29.
16. Carninci, P.; Kasukawa, T.; Katayama, S.; Gough, J.; Frith, M.C.; Maeda, N.; Oyama, R.; Ravasi, T.; Lenhard, B.; Wells, C.; *et al*. The transcriptional landscape of the mammalian genome. *Science* **2005**, *309*, 1559–1563.
17. Wilusz, J.E.; Sunwoo, H.; and Spector, D.L. Long noncoding RNAs: Functional surprises from the RNA world. *Genes Dev.* **2009**, *23*, 1494–1504.
18. Zhao, J.; Sun, B.K.; Erwin, J.A.; Song, J.J.; Lee, J.T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **2008**, *322*, 750–756.
19. Tsai, M.C.; Manor, O.; Wan, Y.; Mosammaparast, N.; Wang, J.K.; Lan, F.; Shi, Y.; Segal, E.; Chang, H.Y. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **2010**, *329*, 689–693.
20. Tripathi, V.; Ellis, J.D.; Shen, Z.; Song, D.Y.; Pan, Q.; Watt, A.T.; Freier, S.M.; Bennett, C.F.; Sharma, A.; Bubulya, P.A.; *et al*. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol. Cell.* **2010**, 39, 925–938.
21. Rinn, J.L.; Chang, H.Y. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **2012**, *81*, 145–166.
22. Gatherer, D.; Seirafian, S.; Cunningham, C.; Holton, M.; Dargan, D.J.; Baluchova, K.; Hector, R.D.; Galbraith, J.; Herzyk, P.; Wilkinson, G.W.; *et al*. High-resolution human cytomegalovirus transcriptome. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 19755–19760.
23. Sun, R.; Lin, S.F.; Gradoville, L.; Miller, G. Polyadenylylated nuclear RNA encoded by Kaposi sarcoma-associated herpesvirus. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 11883–11888.
24. Rossetto, C.C.; Tarrant-Elorza, M.; Verma, S.; Purushothaman, P.; Pari, G.S. Regulation of viral and cellular gene expression by Kaposi’s sarcoma associated herpesvirus polyadenylated nuclear RNA. *J. Virol.* **2013**, *87*, 5540–5553.
25. Stroop, W.G.; Rock, D.L.; Fraser, N.W. Localization of herpes simplex virus in the trigeminal and olfactory systems of the mouse central nervous system during acute and latent infections by *in situ* hybridization. *Lab. Investig.* **1984**, *51*, 27–38.
26. Cheung, A.K. Detection of pseudorabies virus transcripts in trigeminal ganglia of latently infected swine. *J. Virol.* **1989**, *63*, 2908–2913.
27. Ward, P.L.; Barker, D.E.; Roizman, B. A novel herpes simplex virus 1 gene, *UL43.5*, maps antisense to the *UL43* gene and encodes a protein which colocalizes in nuclear structures with capsid proteins. *J. Virol.* **1996**, *70*, 2684–2690.

28. Chang, Y.E.; Menotti, L.; Filatov, F.; Campadelli-Fiume, G.; Roizman, B. *UL27.5* is a novel γ2 gene antisense to the herpes simplex virus 1 gene encoding glycoprotein B. *J. Virol.* **1998**, *72*, 6056–6064.
29. Jovasevic, V.; Roizman, B. The novel HSV-1 US5-1 RNA is transcribed off a domain encoding US5, US4, US3, US2 and α22. *Virol. J.* **2010**, *7*, 103.
30. Zhang, G.; Raghavan, B.; Kotur, M.; Cheatham, J.; Sedmak, D.; Cook, C.; Waldman, J.; Trgovcich, J. Antisense transcription in the human cytomegalovirus transcriptome. *J. Virol.* **2007**, *81*, 11267–11281.
31. Iwakiri, D.; Takada, K. Role of EBERs in the pathogenesis of EBV infection. *Adv. Cancer Res.* **2010**, *107*, 119–136.
32. Tombácz, D.; Tóth, J.S.; Boldogkői, Z. Deletion of the virion host shut: Off gene of pseudorabies virus results in selective upregulation of the expression of early viral genes in the late stage of infection. *Genomics* **2011**, *98*, 15–25.
33. Tombácz, D.; Tóth, J.S.; Boldogkoi, Z. Effects of deletion of the early protein 0 gene of pseudorabies virus on the overall viral gene expression. *Gene* **2012**, *493*, 235–242.
34. Trapnell, C.; Pachter, L.; Salzberg, S.L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **2009**, *25*, 1105–1111.
35. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359.
36. Rutherford, K.; Parkhill, K.; Crook, J.; Horsnell, T.; Rice, P.; Rajandream, M.A.; Barrell, B. Artemis: Sequence visualization and annotation. *Bioinformatics* **2000**, *16*, 944–945.
37. Kwong, A.D.; Frenkel, N. The herpes simplex virus virion host shutoff function. *J. Virol.* **1989**, *63*, 4834–4839.
38. Lin, H.W.; Chang, Y.Y.; Wong, M.L.; Lin, J.W.; Chang, T.J. Functional analysis of virion host shutoff protein of pseudorabies virus. *Virology* **2004**, *324*, 412–418.
39. Saffran, H.A.; Pare, J.M.; Corcoran, J.A.; Weller, S.K.; Smiley, J.R. Herpes simplex virus eliminates host mitochondrial DNA. *EMBO Rep.* **2007**, *8*, 188–193.
40. Matys, V.; Kel-Margoulis, O.V.; Fricke, E.; Liebich, I.; Land, S.; Barre-Dirrie, A.; Reuter, I.; Chekmenev, D.; Krull, M.; Hornischer, K.; *et al*. TRANSFAC and its module TRANSCompel: Transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **2006**, *34*, D108–D110.
41. Tombácz, D.; Tóth, J.S.; Petrovszki, P.; Boldogkői, Z. Whole-genome analysis of pseudorabies virus gene expression by real-time quantitative RT-PCR assay. *BMC Genomics* **2009**, *10*, doi:10.1186/1471-2164-10-491.
42. Taddeo, B.; Roizman, B. The virion host shutoff protein (UL41) of herpes simplex virus 1 is an endoribonuclease with a substrate specificity similar to that of RNase A. *J. Virol.* **2006**, *80*, 9341–9345.
43. Tempel, S.; Tahi, F. A fast ab-initio method for predicting miRNA precursors in genomes. *Nucleic Acids Res*. **2012**, *40*, e80.
44. Jiang, P.; Wu, H.; Wang, W.; Ma, W.; Sun, X.; Lu, Z. MiPred: Classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res*. **2007**, *35*, W339–W344.

45. Huvet, M.; Nicolay, S.; Touchon, M.; Audit, B.; d'Aubenton Carafa, Y.; Arneodo, A.; Thermes, C. Human gene organization driven by the coordination of replication and transcription. *Genome Res.* **2007**, *17*, 1278–1285.
46. Necsulea, A.; Guillet, C.; Cadoret, J.C.; Prioleau, M.N.; Duret, L. The relationship between DNA replication and human genome organization. *Mol. Biol. Evol.* **2009**, *26*, 729–741.
47. Ward, S.A.; Weller, S.K. HSV-1 DNA replication. In *Alpharpesviruses*; Caister Academic Press: Norfolk, UK, 2011; pp. 89–112.
48. Gu, Z.; Dong, J.; Wang, J.; Hou, C.; Sun, H.; Yang, W.; Bai, J.; Jiang, P. A novel inactivated gE/gI deleted pseudorabies virus (PRV) vaccine completely protects pigs from an emerged variant PRV challenge. *Virus Res.* **2015**, *195*, 57–63.
49. Luo, Y.; Li, N.; Cong, X.; Wang, C.H.; Du, M.; Li, L.; Zhao, B.; Yuan, J.; Liu, D.D.; Li, S.; *et al*. Pathogenicity and genomic characterization of a pseudorabies virus variant isolated from Bartha-K61-vaccinated swine population in China. *Vet. Microbiol.* **2014**, *174*, 107–115.
50. Osato, N.; Suzuki, Y.; Ikeo, K.; Gojobori, T. Transcriptional interferences in cis natural antisense transcripts of humans and mice. *Genetics* **2007**, *176*, 1299–1306.
51. Gullerova, M.; Proudfoot, N.J. Convergent transcription induces transcriptional gene silencing in fission yeast and mammalian cells. *Nat. Struct. Mol. Biol.* **2012**, *19*, 1193–1201.
52. Boldogkői, Z. Transcriptional interference networks coordinate the expression of functionally related genes clustered in the same genomic loci. *Front. Genet.* **2012**, *3*, doi:10.3389/fgene.2012. 00122. eCollection 2012.

III.

genomeA
Journals.ASM.org

Strain Kaplan of Pseudorabies Virus Genome Sequenced by PacBio Single-Molecule Real-Time Sequencing Technology

Dóra Tombácz,[a] Donald Sharon,[b] Péter Oláh,[a] Zsolt Csabai,[a] Michael Snyder,[b] Zsolt Boldogkői[a]

Department of Medical Biology, Faculty of Medicine, University of Szeged, Szeged, Hungary[a]; Department of Genetics, School of Medicine, Stanford University, Stanford, California, USA[b]

Pseudorabies virus (PRV) is a neurotropic herpesvirus that causes Aujeszky's disease in pigs. PRV strains are widely used as transsynaptic tracers for mapping neural circuits. We present here the complete and fully annotated genome sequence of strain Kaplan of PRV, determined by Pacific Biosciences RSII long-read sequencing technology.

Received 3 June 2014 **Accepted** 26 June 2014 **Published** 17 July 2014

Citation Tombácz D, Sharon D, Oláh P, Csabai Z, Snyder M, Boldogkői Z. 2014. Strain Kaplan of pseudorabies virus genome sequenced by PacBio single-molecule real-time sequencing technology. Genome Announc. 2(4):e00628-14. doi:10.1128/genomeA.00628-14.

Address correspondence to Zsolt Boldogkői, boldogkoi.zsolt@med.u-szeged.hu.

Pseudorabies virus (PRV), also known as Aujeszky's disease virus or suid herpesvirus 1, a member of the *Alphaherpesvirinae* subfamily, causes significant abortion and morbidity in pigs, the natural host of the virus (1). PRV is a useful model organism for studies of the pathogenesis of herpesviruses. The genetically modified strains are powerful tracers for mapping neuronal circuits (2–6), are tools in gene and cancer therapy (7), and serve as viral vectors for gene delivery into mammalian neurons (3, 4) and cardiomyocytes (8); PRVs have also been employed as live vaccines against Aujeszky's disease (9–11). Further, attenuated vaccine strains of PRV are valuable models for novel vaccine development against varicella-zoster virus (VZV) and herpes simplex virus 1 and 2 (HSV-1 and HSV-2, respectively) (12).

The currently available genome sequences of PRV contain several discrepancies, mainly in intergenic repetitive regions (GenBank accession no. JF797218.1), and the totally annotated version of genome sequence is a composite of six different PRV strains (GenBank accession no. NC_006151.1). We have sequenced the PRV Kaplan genome with Pacific Biosciences single-molecule long-read sequencing technology (Pacific Biosciences, Menlo Park, CA, USA) in order to upgrade the draft sequences, reconstruct the GC-rich and repetitive regions of the genome, and extract epigenetic information. The availability of the completely annotated genome and the single-base resolution methylation map of strain Kaplan will aid in understanding the control of viral gene expression at different levels. Investigations of the PRV genome and gene functions are expected to result in the development of effective vaccines and direct practical applications in gene, cancer, and antiviral therapies.

Sequencing of purified virion DNA was carried out on the Pacific Biosciences RSII sequencer. SMRTbell template libraries were prepared from the DNA, as previously described (13, 14), using standard protocols for 6-kb and 20-kb library preparation. Sequencing was performed in five single-molecule real-time (SMRT) cells with P5 DNA polymerase and C3 chemistry (P5-C3) yielding a total of 78,111 reads and an extremely high coverage (1,200×) throughout the genome.

The sequencing reads were processed and mapped to the respective reference sequences with the BLASR mapper (https://github.com/PacificBiosciences/blasr) and the Pacific Biosciences SMRT Analysis pipeline (https://github.com/PacificBiosciences/SMRT-Analysis/wiki /SMRT-Pipe-Reference-Guide-v2.0) using the standard mapping protocol.

The protein-coding genes were predicted by GATU (15). Manual annotation was used to identify other genomic features. Annotation of a previously unknown noncoding RNA (named Close to OriL [CTO]), a newly discovered splice site of the early protein 0 gene, and new isoforms of 11 protein-coding genes are based on RNAseq data (our unpublished data). MicroRNA (miRNA) annotation was based on the precursor miRNAs found in strains NIA-3 and Ea.

The complete genome of strain Kaplan of PRV is characterized as a double-stranded linear DNA composed of 143,423 bp, with an average G+C content of 73.59%. PRV contains 70 protein-coding genes (11 genes have different isoforms), two latency-associated transcripts, and a long noncoding RNA, and its genome predicts 16 miRNAs.

Nucleotide sequence accession number. The complete genome of strain Kaplan of pseudorabies virus was assigned DDBJ/EMBL/GenBank accession no. KJ717942.

ACKNOWLEDGMENTS

This project was supported by the Swiss-Hungarian Cooperation Programme grant (SH/7/2/8) to Z.B. This research was also supported by the European Social Fund in the framework of TÁMOP 4.2.4. A/2-11-1-2012-0001 "National Excellence Program" to D.T.

REFERENCES

1. **Aujeszky A.** 1902. A contagious disease, not readily distinguishable from rabies, with unknown origin. Veterinarius. **12:**387–396. (In Hungarian.)
2. **Card JP, Kobiler O, Ludmir EB, Desai V, Sved AF, Enquist LW.** 2011. A dual infection pseudorabies virus conditional reporter approach to identify projections to collateralized neurons in complex neural circuits. PLoS One **6:**e21141. http://dx.doi.org/10.1371/journal.pone.0021141.
3. **Granstedt AE, Kuhn B, Wang SS, Enquist LW.** 2010. Calcium imaging

of neuronal circuits in vivo using a circuit-tracing pseudorabies virus. Cold Spring Harb. Protoc. 2010:pdb.prot5410. http://dx.doi.org/10.1101/pdb.prot5410.

4. **Boldogkői Z, Balint K, Awatramani GB, Balya D, Busskamp V, Viney TJ, Lagali PS, Duebel J, Pásti E, Tombácz D, Tóth JS, Takács IF, Scherf BG, Roska B.** 2009. Genetically timed, activity-sensor and rainbow transsynaptic viral tools. Nat. Methods **6**:127–130. http://dx.doi.org/10.1038/nmeth.1292.
5. **Song CK, Enquist LW, Bartness TJ.** 2005. New developments in tracing neural circuits with herpesviruses. Virus Res. **111**:235–249. http://dx.doi.org/10.1016/j.virusres.2005.04.012.
6. **Yang M, Card JP, Tirabassi RS, Miselis RR, Enquist LW.** 1999. Retrograde, transneuronal spread of pseudorabies virus in defined neuronal circuitry of the rat brain is facilitated by gE mutations that reduce virulence. J. Virol. **73**:4350–4359.
7. **Boldogkői Z, Nógrádi A.** 2003. Gene and cancer therapy—pseudorabies virus: a novel research and therapeutic tool? Curr. Gene Ther. **3**:155–182. http://dx.doi.org/10.2174/1566523034578393.
8. **Prorok J, Kovács PP, Kristóf AA, Nagy N, Tombácz D, Tóth JS, Ördög B, Jost N, Virág L, Papp JG, Varró A, Tóth A, Boldogkői Z.** 2009. Herpesvirus-mediated delivery of a genetically encoded fluorescent Ca(2+) sensor to canine cardiomyocytes. J. Biomed. Biotechnol. 2009: 361795. http://dx.doi.org/10.1155/2009/361795.
9. **Klingbeil K, Lange E, Teifke JP, Mettenleiter TC, Fuchs W.** 2014. Immunization of pigs with an attenuated pseudorabies virus recombinant expressing the hemagglutinin of pandemic swine origin H1N1 influenza A virus. J. Gen. Virol. **95**:948–959. http://dx.doi.org/10.1099/vir.0.059253-0.
10. **Maresch C, Lange E, Teifke JP, Fuchs W, Klupp B, Müller T, Mettenleiter TC, Vahlenkamp TW.** 2012. Oral immunization of wild boar and domestic pigs with attenuated live vaccine protects against pseudorabies virus infection. Vet. Microbiol. **161**:20–25. http://dx.doi.org/10.1016/j.vetmic.2012.07.002.
11. **Zhu L, Yi Y, Xu Z, Cheng L, Tang S, Guo W.** 2011. Growth, physicochemical properties, and morphogenesis of Chinese wild-type PRV Fa and its gene-deleted mutant strain PRV SA215. Virol. J. **8**:272. http://dx.doi.org/10.1186/1743-422X-8-272.
12. **Szpara ML, Tafuri YR, Parsons L, Shamim SR, Verstrepen KJ, Legendre M, Enquist LW.** 2011. A wide extent of inter-strain diversity in virulent and vaccine strains of alphaherpesviruses. PLoS Pathog. **7**:e1002282. http://dx.doi.org/10.1371/journal.ppat.1002282.
13. **Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW.** 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res. **38**:e159. http://dx.doi.org/10.1093/nar/gkp817.
14. **Clark TA, Murray IA, Morgan RD, Kislyuk AO, Spittle KE, Boitano M, Fomenkov A, Roberts RJ, Korlach J.** 2012. Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. Nucleic Acids Res. **40**:e29. http://dx.doi.org/10.1093/nar/gkr1146.
15. **Tcherepanov V, Ehlers A, Upton C.** 2006. Genome annotation transfer utility (GATU): rapid annotation of viral genomes using a closely related reference genome. BMC Genomics **7**:150. http://dx.doi.org/10.1186/1471-2164-7-150.