MULTIPLATFORM ANALYSIS OF HERPESVIRUS TRANSCRIPTOMES

Ph.D. THESIS

ZSOLT CSABAI

Department of Medical Biology

Doctoral School of Interdisciplinary Medicine

Faculty of Medicine

University of Szeged

Supervisors: Dr. Dóra Tombácz and prof Dr. Zsolt Boldogkői

Szeged

2017

Publications directly related to the subject of thesis

Csabai Zsolt*, Takács Irma, Michael Snyder, Boldogkői Zsolt, Tombácz Dóra
Evaluation of the impact of ul54 gene-deletion on the global transcription and DNA replication of pseudorabies virus
ARCHIVES OF VIROLOGY pp. 1-16. (2017)
IF: 2,058

Póka Nándor*, **Csabai Zsolt***, Pásti Emese, Tombácz Dóra, Boldogkői Zsolt
Deletion of the us7 and us8 genes of pseudorabies virus exerts a differential effect on the expression of early and late viral genes
VIRUS GENES (2017)
IF: 1,431

Tombácz D*, **Csabai Z***, Oláh P, Havelda Z, Sharon D, Snyder M, Boldogkoi Z
Characterization of novel transcripts in pseudorabies virus
VIRUSES-BASEL 7:(5) pp. 2727-2744. (2015)
IF: 3,042

Balázs Zsolt, Tombácz Dóra, Szűcs Attila, **Csabai Zsolt**, Megyeri Klára, Alexey N Petrov, Michael Snyder, Boldogkői Zsolt
Long-Read Sequencing of Human Cytomegalovirus Transcriptome Reveals RNA Isoforms Carrying Distinct Coding Potentials
SCIENTIFIC REPORTS 7: Paper 10.1038/s41598-017-16262-z. 9 p. (2017)
IF: 4,259

Tombácz Dóra, **Csabai Zsolt**, Szűcs Attila, Balázs Zsolt, Moldován Norbert, Donald Sharon, MIchael Snyder, Boldogkői Zsolt
Long-Read Isoform Sequencing Reveals a Hidden Complexity of the Transcriptional Landscape of Herpes Simplex Virus Type 1
FRONTIERS IN MICROBIOLOGY (2017)
IF: 4,076

Olah P, Tombacz D, Poka N, **Csabai Z**, Prazsak I, Boldogkoi Z
Characterization of pseudorabies virus transcriptome by Illumina sequencing.
BMC MICROBIOLOGY **15:** Paper 130. 9 p. (2015)
IF: 2,581

Tombacz D, Balazs Z, **Csabai Z**, Moldovan N, Szucs A, Sharon D, Snyder M, Boldogkoi Z:
Characterization of the Dynamic Transciptome of a Herpesvirus with Long-read Single Molecule Real Time Sequencing.
SCIENTIFIC REPORTS accepted for publication, 2017. jan. 26.
IF: 4,259

Tombácz D, **Csabai Z**, Oláh P, Balázs Z, Likó I, Zsigmond L, Sharon D, Snyder M, Boldogkői Z
Full-Length Isoform Sequencing Reveals Novel Transcripts and Substantial Transcriptional Overlaps in a Herpesvirus
***PLOS ONE* 11:**(9) Paper e0162868. 29 p. (2016)
IF: 2,806

Publications indirectly related to the subject of thesis

Moldován Norbert, Balázs Zsolt, Tombácz Dóra, **Csabai Zsolt**, Szűcs Attila, Michael Snyder, Boldogkői Zsolt
Multi-platform Analysis Reveals a Complex Transcriptome Architecture of a Circovirus
VIRUS RESEARCH 237: pp. 37-46. (2017)
IF:2,628

Tombácz Dora, Moldovan Norbert, Balazs Zsolt, **Csabai Zsolt**, Michael Snyder, Boldogkoi Zsolt:
Genetic Adaptation of porcine Circovirus Type 1 to Cultured Porcine kidney Cells Revealed be Single-molecule Long-read Sequencing Technology.
GENOME ANNOUNCEMENTS , Accepted, scheduled publication date: 2017. febr 2.
IF: 0

Szűcs Attila, Moldován Norbert, Tombácz Dóra, **Csabai Zsolt**, Michael Snyder, Boldogkői Zsolt
Long-Read Sequencing Reveals a GC Pressure during the Evolution of Porcine Endogenous Retrovirus
GENOME ANNOUNCEMENTS ,5:(40) Paper e01040-17. 2 p. (2017)
IF: 0

Tombacz D, Sharon D, Olah P, **Csabai Z**, Snyder M, Boldogkoi Z
Strain Kaplan of Pseudorabies Virus Genome Sequenced by PacBio Single-Molecule Real-Time Sequencing Technology.
GENOME ANNOUNCEMENTS 2:(4) Paper e00628-14. 3 p. (2014)
IF: 0

Unrelated publications to the subject of thesis

Tombácz Dóra, Maróti Zoltán, Kalmár Tibor, **Csabai Zsolt**, Balázs Zsolt, Takahashi Shinichi, Palkovits Miklós, Snyder Michael, Boldogkői Zsolt
High-Coverage Whole-Exome Sequencing Identifies Candidate Genes for Suicide in Victims with Major Depressive Disorder
SCIENTIFIC REPORTS 7: Paper 7106. 11 p. (2017)
IF: 4,259

Szenasi T, Kenesi E, Nagy A, Molnar A, Balint BL, Zvara A, **Csabai Z**, Deak F, Boros Olah B, Mates L, Nagy L, Puskas LG, Kiss I
Hmgb1 can facilitate activation of the matrilin-1 gene promoter by Sox9 and L-Sox5/Sox6 in early steps of chondrogenesis.
BIOCHIMICA ET BIOPHYSICA ACTA-GENE REGULATORY MECHANISMS 1829:(10) pp. 1075-1091. (2013)
IF: 5,440

Cumulative impact factor: 36,839

List of abbreviations

ASP	antisense promoter
asRNA	antisense RNA
AST	antisense transcript
AZURE	antisense transcript in the IR US overlapping region
CTO	close to replication origin
cDNA	complementary DNA
DMEM	Dulbecco's Modified Eagle Medium
dNTP	Deoxynucleotide triphosphate
dsDNA BR	double stranded DNA broad range assay kit
E	early
EDTA	Ethylenediaminetetraacetic acid
FBS	fetal bovine serum
gE	gylcoprotein E
GFP	green fluorescent protein
gI	gylcoprotein I
HCMV	Human cytomegalovirus
HSV-1	Human herpesvirus 1
IE	immediate-early
IGV	integrative genome viewer
IRL	internal repeat long
IRS	internal repeat short
IsoSeq	Isoform Sequencing
Ka	Kaplan strain
L	late
LAT	latency associate transcript
LLT	long latency transcript
lncRNA	long non coding RNA
MOI	multiplicity of infection
MRC-5	human lung fibroblast cells
ncRNA	non coding RNA

NOIR	non coding RNA in the inverted repeat
ORF	open reading frame
PacBio	Pacific Biosciences
PAS	polyadenylation signal
PA- Seq	polyadenylated sequencing
PBS	phosphate-buffered saline
PK-15	porcine kidney cell line
PRV	pseudorabies virus
PSSM	position specific scoring matrices
PTO	proximal to the origin
RACE	Rapid amplification of cDNA ends
RNA HS	RNA High Sensitivity
rRNA	ribosomal RNA
RT^2-PCR	Reverse transcription linked real-time PCR
SMRT	Single Molecule Real-time
TES	transcription end site
TRL	terminal repeat long
TRS	terminal repeat short
TSS	transcription start site
UL	unique long
US	unique short
UTR	untranslated region
VERO	African green monkey immortalized kidney epithelial cell line
wt	wild type

Tabel of contents

Introduction

Herpesviridae is a large family of double stranded (ds)DNA viruses. Herpresviruses have more than 100 members, 8 of these are human pathogen (herpes simplex virus types 1 and 2, varicella-zoster virus, cytomegalovirus, Epstein-Barr virus, human herpesvirus 6 (variants A and B), human herpesvirus 7, and Kaposi's sarcoma virus or human herpesvirus 8.). Worldwide ~90 % of population have been infected with one of these viruses.

Herpesviruses are nuclear replicating viruses; transcription, genome replication and capsid assembly occur in the host cell nucleus. The intracellular trafficking of these viruses is connected to Golgi transport.

Herpesviruses are divided into three major groups: alphaherpesviruses, betaherpesviruses, gammaherpesviruses (Figure.1).

The life cycle of all herpesviruses in their natural host can be divided into lytic and latent infections. During a lytic infection the virus is replicate and newly synthesized particles are released into the surrounding medium. During a latent infection viral replication is suppressed. Viral latency is a hallmark of all known herpesviruses. [1]

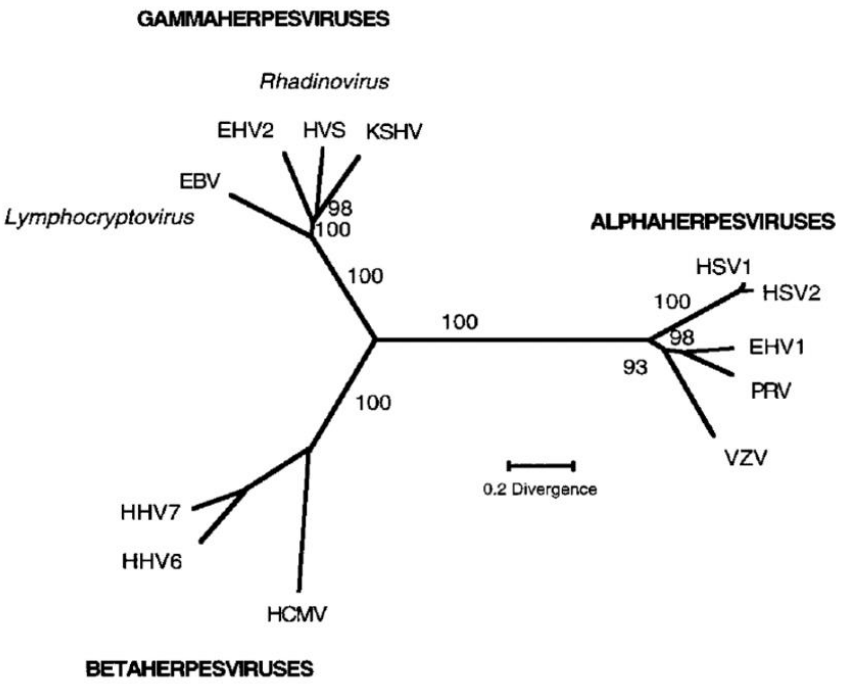


Figure.1 Phylogenetic tree of herpesviruses*. The phylogenetic tree was constructed by comparing the amino acid sequences of the major capsid protein gene* [2].

Herpesvirus genes are expressed in a coordinated temporal cascade and grouped into three kinetic classes, immediate-early (IE), early (E) and late (L) [3]. The IE proteins are required for the transcription of both E and L genes. The E genes typically encode proteins that play a role in DNA replication, while the L genes specify the structural components of the virus.

Pseudorabies virus

The pseudorabies virus (PRV), an alphaherpesvirus with a broad host range, causes fatal encephalitis in a wide variety of animals, with the exception of its natural reservoir, the adult pig. It is a commonly employed model organism in studies of the molecular pathogenesis of herpesviruses [4,5], for labelling neural circuits [6–8] and for the delivery of genetically-encoded fluorescent activity markers to the neurons [9]. The genomes of viruses are very compact, composed mainly of protein-coding genes and short intergenic regions. The PRV genome contains a unique long (UL) and unique short (US) region, the latter bracketed by inverted repeat (IR) sequences (TRL/IRL, TRS/IRS, respectively). PRV DNA (upgraded with our own data) contains 67 protein-coding and 20 RNA genes (KJ717942.1). Similarly as for other herpesviruses, most of the PRV genes are organized into polycistronic transcriptional units, which are typical in prokaryotes, but rare in higher-order organisms [10–12].

Herpesvirus genes are expressed in a temporally ordered cascade and grouped into three kinetic classes. The protein products of immediate-early (IE) genes are required for the transcription of both early (E) genes encoding the synthetic machinery of DNA, and late (L) genes specifying the structural elements of the virus. L genes can be subdivided into leaky late (L1 or E/L) and true late (L2 or L) classes depending upon whether DNA replication is an absolute prerequisite for their expression (this is the case for L2 genes). While the herpes simplex virus (HSV) expresses 5 IE genes, the PRV genome contains only a single one, the immediate-early 180 (*ie180*) gene, which encodes an essential transcriptional transactivator [13]. Kinetic analysis of the PRV transcriptome confronts a serious problem in due to the polycistronic organization of the viral genes. Previous approaches on the analysis of the herpesvirus transcriptome have used microarrays [14], Illumina sequencing, and real-time reverse transcription PCR (RT^2-PCR) analysis [15]. However, the identification of transcript isoforms, including splice and length variants, with these techniques is difficult or impossible.

Kinetic studies of the herpesvirus transcriptome faces a significant challenge due to the overlapping nature of the viral transcripts. The typical architecture of polycistronic units is characterized by varying transcription start sites (TSSs) that are caused by the control of distinct promoters, and shared transcription end sites (TESs). As an example, the following transcripts are produced from a tetracistronic unit: 1-2-3-4, 2-3-4, 3-4 and 4, where '1' represents the most upstream gene, while '4' is the most downstream gene within the given unit.

Alternative splicing expands the information content of genomes by producing multiple messages from a single gene. Investigations aimed at cracking the 'splicing code' have concluded that it is determined by multiple interactions between *cis*- and *trans*-acting factors, but the precise mechanisms are not well understood [16]. The spliced isoforms can have similar or antagonistic functions [17]. The following spliced PRV RNAs have so far been described: US1 [18], UL15 [19]) and LLT [20].

Despite the fact that hardly more than one per cent of the mammalian genome encodes protein sequences, a large proportion of the DNA is transcriptionally active, producing non-coding RNA molecules (ncRNAs; [21]). Debate is ongoing as to whether this pervasive transcription represents mere transcriptional noise for the most part, or whether these transcripts have still unidentified functions [22]. The most abundant and least annotated class of ncRNAs is the long ncRNAs (lncRNAs [23]), which are defined as transcripts exceeding 200 nucleotides. A large proportion of murine and human DNA was recently reported to encode a wide variety of lncRNAs [24]. Many protein-coding genes specify lncRNAs transcribed from the plus strand as templates, which are called antisense lncRNAs. The latency-associated transcript (LAT) of HSV has been described as the first lncRNA of herpesviruses [25]. A spliced 8.4-kb antisense RNA, termed long latency transcript (LLT), is synthesized from the complementary DNA strand of *ie180* and *ep0* genes and is controlled by the LAT promoter of PRV [13].
The original genome sequence of the PRV was a composite generated from six viral strains [26] and determined using the traditional Sanger method. The complete genome of PRV strain Ka and other strains have been sequenced both by Sanger capillary sequencing [27] and by Illumina deep sequencing [28–30]. We used the Pacific Biosciences (PacBio) platform to sequence the wild-type Kaplan (Ka) strain of the PRV genome [31].

The us7 and us8 genes

The *us8* and *us7* genes encode the membrane glycoproteins E (gE) and I (gI), respectively, which form heterodimers [32, 33], and are thus often mentioned as gE/gI proteins. These genes

are well-conserved across alphaherpesviruses, and are involved in the control of cell-to-cell spread in both cultured cells and animals [34-36]. This dimeric protein molecule binds to the Fc receptors at mucosal surfaces [37], and plays a role in sorting virion particles to cell junctions [38]. It has been demonstrated that spreading of the virus along the axons in an anterograde manner requires the presence of both proteins [39, 40]. The gE/gI heterodimer along with the US9 protein has also been shown to be required for the axonal transport of viral capsids and glycoproteins in an anterograde manner [41-44]. Although the replication of viral DNA has been described to be unaffected in *us7*/*us8* mutants, their virulence is attenuated even in cell cultures due to the reduced efficiency of virion formation and direct cell-to-cell spreading [45]. The level of attenuation is severely increased when glycoprotein M (gM) was also deleted [46]. The gE and gI proteins may also have separate functions, seeing that the gE null mutant is less virulent than the gI mutant [47]. However, gE/gI mutants exhibit few, tractable phenotype in cultured cells [47]. Glycoproteins E and I are intensely investigated due to their importance in many aspects of the PRV lifecycle. Revealing the molecular mechanisms of viral spread leads not only to a better understanding of the virus itself, but also paves the way for designing better anti-viral drugs or vaccines [48, 49]. Furthermore, the mutation of *us7*/*us8* genes is important in studies on the mapping of neural circuits because it eliminates the spreading capability of PRV in an anterograde manner, and thereby rendering this virus an ideal retrograde multi-synaptic tracing tool [50], [7],[9].

The ul54 gene

Herpes simplex virus type 1 (HSV-1), the prototype of the *Alphaherpesvirinae* subfamily, has five IE genes (*icp4, icp0, icp27, icp22*, and *icp47*; [51, 52]), while in comparison, PRV has only a single true IE gene, the *ie180* (homologous to the *icp4* of HSV-1; [53]). It has been shown that *ep0* and *ul54* genes of PRV (homologous to *icp0* [20] and the *icp27* [54 of HSV, respectively] are expressed in the E phase of infection. PRV lacks the *icp47* gene and there is no consensus has been reached as to whether the *us1* gene (ICP22 in HSV) is expressed with IE [55] or E [18] kinetics. The least characterized among the above genes in PRV is the *ul54* gene [54, 57, 58].

The *ul54* gene is located on the unique long (UL) region of the PRV genome and it forms a tandem cluster along with 2 other genes (*ul53, ul52*), which produce 3' coterminal transcripts. PRV *ul54* gene is composed of 1,164 nucleotides and encodes a protein of 361 amino acids.

Several functions of ICP27 and their homologs have been revealed, however, such as the regulation of transcription [57] and DNA replication [59-62], as well as the shut off of host

protein synthesis [63], and the usage of polyadenylation sites [64], as well as viral growth however. The deletion of *ul54* gene has been shown to result in severe growth defects [65]. A previously published study [62] reported that the *ul54* gene and its protein product are not essential for PRV growth and replication in tissue culture; however, the mutant virus exhibited reduced growth ability. It has also been shown that this multifunctional protein is not essential for host shut-off and that its absence causes aberrant accumulation of late proteins at the early phase of infection in a cell-type dependent manner [62]. Two recently published studies revealed that the ICP27 plays also a role in the nucleo-cytoplasmic trafficking and in nucleolar-targeting [66, 67], respectively. Despite the accumulating data on the function of the *ul54* gene, its precise role in the virus lifecycle remains still poorly understood.

Herpes simplex virus

Herpes simplex virus type 1 (HSV-1) is a human pathogenic alphaherpesvirus from the *Herpesviridae* family. Herpes is a lifelong infection, which often has mild or no symptoms. The most common symptoms of viral infection are cold sores. HSV-1 can cause acute encephalitis in immunocompromised patients. According to WHO's first global estimates, worldwide more than 3.7 billion people under the age of fifty are infected with HSV-1 [68]. The HSV-1 genome is composed of a unique long (UL) and a unique short (US) region, both being bracketed by inverted repeats (IRLs and IRSs, respectively). According to earlier annotations, the HSV-1 DNA contains 89 protein-coding, 10 long non-coding (lnc)RNA genes and several micro RNAs [74].

It has been demonstrated that a large part of the mammalian genome is transcribed, producing a large variety of non-coding (nc)RNA molecule [75]. There is a current debate on whether the genome-wide expression of ncRNAs merely represents transcriptional noise, or whether these molecules might have yet undisclosed functions [76]. The most abundant class of the non-coding transcripts are the lncRNAs [77], which are defined as RNA molecules that exceed 200 nucleotides. A large number of lncRNAs have recently been described in mice and humans [78]. Several genomic regions containing protein-coding genes also encode antisense lncRNAs from the complementary DNA strand.

Various methods have already been used for the analysis of the herpesvirus transcriptome including microarrays [80], Illumina sequencing [81], multi-time-point real-time reverse transcription PCR (qRT-PCR) analysis [82], and PacBio SMRT sequencing [123]. Next-generation sequencing platforms have only been used for analyzing the transcriptional activity along the viral genome [81].

Cytomegalovirus

The Human Cytomegalovirus (HCMV) is a human pathogenic beta-herpesvirus that can cause life-threatening infections in new-born infants and immunocompromised patients. Congenital HCMV infections can lead to severe malformations or even death [83]. The genome of the HCMV is one of the largest in the *Herpesviridae* family, and its coding potential is not fully understood. The number of its protein coding sequences ranged from 164 [84-86] to 220 [87], while a recent study identified 751 individuals, translationally active open reading frames (ORFs) by ribosome profiling [88]. Most of these novel ORFs were short ORFs with potential regulatory functions, or N-terminally truncated versions of already annotated proteins. The HCMV genome contains a Unique Long (UL) region and a Unique Short (US) region, each bracketed by terminal and internal repeats. Laboratory HCMV strains such as the Towne strain, which was used in our experiments, have undergone substantial genetic alterations compared to the wild-type (wt) virus, which severely affected their pathogenicity.

HCMV, similarly to other herpesviruses, has a complex transcriptional architecture; alternative transcription initiation [89], alternative splicing events [86;90], and polycistronic transcripts [91] all increase the coding potential of the viral genome. Splicing in herpesviruses is relatively rare [91], over 100 splice junctions have been described in HCMV [86;88;90] – many of which are alternatively spliced.

Short-read sequencing analysis has demonstrated that the complete HCMV genome is transcriptionally active during lytic infection [86]. It has also revealed numerous splice sites, and confirmed many of the previously detected. The HCMV genome, however, continues to be insufficiently annotated [91]. Generating accurate transcript models using short-read sequencing data is challenging [93], and methods with high specificity, such as Northern-blotting and rapid amplification of cDNA ends (RACE) techniques are too laborious to be used for mapping the transcriptome of complex organisms. On the other hand, long-read sequencing is capable of determining the base composition of full-length transcripts, which enables distinction between transcript isoforms and alternative splicing events, and thus renders it a powerful tool in transcript discovery [94].

To date, among the major shortcomings of long-read sequencing methods are their low throughput and relatively high rate of error [95]. While the latter does not typically pose a challenge in transcriptomic analyses, the low coverage of larger genomes means that the analysis is at a greater disposition for picking up erroneous signals. RNA-degradation and

template switching are both potential sources of errors. 5’-truncated RNAs are frequently seen as a result of RNA-degradation, which hinder the detection of alternative internal TSSs. Template switching can produce false transcript isoforms, such as false splice junctions or chimeric reads in a homology-dependent manner [96].

Aims

- To re-evaluate the currently available knowledge concerning the structures of PRV transcripts by using Illumina HiSeq and PacBio RS II platform (PA-Seq and random primer-based RNA-Seq), which can identify all poly(A)$^+$ RNA molecules generated in cultured porcine kidney (PK-15) cells productively infected with the virus.
- Using PacBio long-read sequencing technology for the characterization of the global lytic transcriptome of HSV-1. Application an amplified isoform sequencing (Iso-Seq) protocol that based on PCR amplification of the cDNAs prior to sequencing.
- Our focus was to identify novel transcripts, transcript isoforms, novel splice junctions, and to determine the coding potential of these transcripts, in HCMV RNA population in human fibroblast cells during lytic infection.
- Generate an *ul54*-KO virus and examine the effects of the mutation on the replication and global transcription of PRV by using quantitative real-time-PCR and reverse transcription (qRT)-PCR platforms.
- Characterisation the dynamic transcriptome of *us7/us8*-deleted PRV in comparison with the wild-type (*wt*) virus, using a multi-time-point quantitative reverse transcriptase-based real-time PCR technique.

-

Material and methods

Viruses, cells and infection

An immortalized PK-15 epithelial cells were used for Pseudorabies virus strain Kaplan. For the propagation of strain KOS of HSV-1, immortalized kidney epithelial cell line (Vero) was used, isolated from African green monkey and for the HCMV strain Towne (ATCC VR-977) was grown in human lung fibroblast cells [MRC-5; American Type Culture Collection (ATCC)] in DMEM supplemented with 10% FBS (Gibco Invitrogen), and 100µl penicillin-streptomycin 10K/10K mixture (Lonza). PK-15 cells were cultivated in Dulbecco's modified Eagle medium supplemented with 5% foetal bovine serum (Gibco Invitrogen) with 80µg gentamycin/ml at 37°C, under 5% CO_2, Vero cells were grown in DMEM (Gibco/Thermo Fisher Scientific),with 10% foetal bovine serum (Gibco Invitrogen) and 100µl penicillin-streptomycin 10K/10K mixture (Lonza)/ml and 5% CO_2 at 37°C supplemented with 5% foetal bovine serum (Invitrogen/Thermo Fisher Scientific) and 80µg of gentamycin per ml (Invitrogen/Thermo Fisher Scientific) at 37°C in an atmosphere of 95% air, 5% CO_2. Cells were infected at a multiplicity of infection (Table 1-infection). Infected cells were incubated until 1 hour at 37°C. Followed by removal of the virus suspension and washing twice with phosphate-buffered saline (PBS). After the addition of new medium to the cells, they were incubated.

Cell line	Virus	Pfu/cell	Infection times (hour)
PK-15	PRV (Kaplan)	10	1,2,,4,6,8,12,18,24
PK-15	PRV (Kaplan)	0,1	1,2,4,6,8
PK-15	PRV Δ*us7*Δ*us8* (Kaplan)	10	1,2,,4,6,8,12,18,24
PK-15	PRV Δul54 (Kaplan)	0,1	1,2,4,6,8
VERO	HSV-1 (KOS)	1	1,2,4,6,8,12
MRC-5	HCMV (Town)	0,05	1,3,6,12,24,72,96,120

Table 1. ***Infection conditions***

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 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 $5 \times 10_6$.

RNA and DNA sample preparation

Total RNAs were isolated from infected cells by using Nucleospin RNA Kit (Macherey-Nagel) as was suggested by the manufacturer. In summary, the cells were collected by centrifugation and lysed by a buffer (kit component). Potential genomic DNA contamination was digested by RNase-free rDNase solution (supplied with the kit). Samples were eluted in nuclease-free water (part of the kit) in a total volume of 60 µl. Next, possible residual DNA contamination was eliminated by using the TURBO DNA-free Kit (Ambion/Thermo Fisher Scientific). RNA concentration was calculated using a Qubit 2.0 Fluorometer instrument (through use of the Qubit RNA BR Assay Kit (Life Technologies/Thermo Fisher Scientific). The RNA samples were stored at - 80 ºC until use.

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.

DNA was isolated from the cells with Qiagen DNeasy Blood & Tissue Kit following the manual's Spin-Column Protocol for Purification of Total DNA from Animal Blood or Cells. The quantity of DNA was measured by Qubit 2.0 Fluorometer, using Qubit dsDNA BR Assay Kit (Life Technologies/Thermo Fisher Scientific). The purified DNA samples were stored at -20 ºC until use.

Reverse transcription real-time PCR

Single-stranded (ss)cDNA production was carried out by using SuperScript III reverse transcriptase (Invitrogen/Thermo Fisher Scientific) and gene-specific primers. Briefly, the reaction mixtures containing RNA, primer, SuperScript III enzyme, buffer, and dNTP mix were incubated at 55 ºC for 1 h. Finally, the reaction was terminated by heating at 70 ºC for 15 min. Samples were diluted 10-fold with nuclease-free water (Ambion/Thermo Fisher Scientific).

The entire viral genome or single-stranded cDNAs were used as templates for the amplification of specific sequences by Rotor-Gene Q real-time PCR cycler (Qiagen) and Absolute QPCR SYBR Green Mix (Thermo Fisher Scientific). To ensure the accuracy, the following controls were used: no-RT, no-primer, no-template, as well as loading control (pig 28S rRNA). Purified the viral DNA was also used to verify the specificity of the primers.

PCR and Real-time RT PCR

The cDNAs were amplified with the Veriti Thermal Cycler (Applied Biosystems), using AccuPrime™ GC-Rich DNA Polymerase (Invitrogen). The running conditions were as follows: 3 min at 95°C, followed by 30 cycles of 92°C for 30 s (denaturation), 60°C for 30 s (annealing), and 72°C for 10 s (extension). Final elongation of 10 min at 72°C was set.

Reverse transcription reactions were carried out by using 70 ng of total RNA as template, Superscript III enzyme (Life Technologies) and anchored oligo(dT) primers.
Real-time PCR reactions were performed in a volume of 20 μl with Absolute QPCR SYBR Green Mix (Thermo Scientific) containing 7 μl of cDNA solution diluted 10-fold, 1.5 μl of forward and 1.5 μl of reverse primers (10 μM each). 28S ribosomal (r)RNA was used as a reference gene in each run. The conditions for the PCR amplification were as follows: 15 min at 95°C for the enzyme activation, followed by 30 cycles of 94°C for 25 s (denaturation), 60°C for 25 s (annealing), and 72°C for 6 s (synthesis). Relative expression ratios (R) were calculated

$$R = \frac{(E_{sample\ \max})^{Ct_{sample\ \max}}}{(E_{sample})^{Ct_{sample}}} : \frac{(E_{ref\ \max})^{Ct_{ref\ \max}}}{(E_{ref})^{Ct_{ref}}},$$

via the following formula: where E is the amplification efficiency, Ct is the threshold cycle number, “sample” refers to the examined viral transcript and “ref” is 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.

Calculation of R_x, $R_{x'}$ and $R_{x'}r$ values

The mean expression value (E^{Ct}) of all examined mRNAs (*total*) in a given sample was used as a normalization factor for the transcripts (*sample*) real-time quantitative PCR data as described earlier [30] to obtain R_x values. Here we used the E^{Ct} values for the calculation of the expression values instead of using the Cts alone.

$$R_x = \frac{E_{sample}^{\ Ct_{sample}}}{\overline{E_{other}^{\ Ct_{other}}}} \qquad R_{x'} = \frac{R_x sample}{\overline{R_x all}} \qquad R_{x'}r = \frac{R_{x'} mu\tan t}{\overline{R_{x'} wt}}$$

Pearson's correlation analysis

Pearson's correlation coefficient (r) was calculated to analyse the change of the gene expression pattern, using the following formula: $r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)S_x S_y}$

The Pearson correlation coefficient (r) is a number between −1 and +1 that measures the linear relationship between two variables (X and Y), which are the R_x values of the same gene in different genetic backgrounds. X and Y are the average values, n is the sample number, and SX and SY are the standard errors of the mean values for X and Y, respectively. A positive r value means a positive association, while a negative value for the correlation indicates an inverse association.

PacBio RS II sequencing

Non-amplified Iso-Seq protocol

cDNA synthesis The Poly(A^+) fractions of total RNAs were quantified through use of the Qubit RNA HS Assay Kit (Life Technologies), followed by conversion to cDNAs with the SuperScript Double-Stranded cDNA Synthesis Kit (Life Technologies; the included first strand enzyme was changed to SuperScript III Reverse Transcriptase). The reverse transcription (RT) reactions were primed with Anchored Oligo$(dT)_{20}$ primers (Life Technologies). The cDNAs obtained were quantified with the Qubit HS dsDNA Assay Kit (Life Technologies).

Library preparation and sequencing 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). SMRTbell templates were bound to polymerases by using the DNA polymerase binding kit XL 1.0 (part #100-150-800) and v2 primers. The polymerase-template complexes were bound to magbeads with the Pacific Biosciences MagBead Binding Kit. The SMRTBell libraries were analysed for length and concentration through use of the Agilent 2100 Bioanalyzer. DNA sequencing was carried out with a Pacific Biosciences RS II sequencer using P5-C3 chemistry. Movie lengths were 180 min.

Amplified Iso-Seq protocol

cDNA synthesis Poly(A)$^+$ RNAs were purified from total RNA samples by using the Oligotex mRNA Mini Kit (Qiagen), and were converted to cDNAs. The cDNA production and the SMRTbell library preparation were carried out via the protocol described by PacBio: Isoform Sequencing (Iso-Seq) using the Clontech SMARTer PCR cDNA Synthesis Kit and No Size Selection (for the analysis of short transcripts) or Manual Agarose-gel Size Selection (analysis of long transcripts). Briefly, the first-strand cDNAs were generated by using the SMARTer PCR cDNA Synthesis Kit (Clontech).

No size selection: Single-stranded cDNAs (sscDNAs) were amplified by PCR (16 cycles, based on the Test Amplification), using the KAPA HiFi Enzyme (Kapa Biosystems). 500 ng of each cDNA sample was used for the SMRTbell template preparation, using the PacBio DNA Template Prep Kit 2.0. Manual Agarose-gel Size Selection: KAPA HiFi Enzyme was used for the PCR reactions. Two different PCR reactions were used to obtain different transcripts. Twelve PCR cycles and 1:45 min extension were set for the amplification of transcripts between 2-3 kb. Fifteen cycles and 3 min extension were used for the longer transcripts.

The random primer-based PacBio sequencing was carried out exactly as above, except that instead of the oligodT primer, adapter-linked GC-rich random primers were used for the RT reaction.

Library preparation and sequencing SMRTbell templates were bound to polymerases by using the DNA/Polymerase Binding Kit P6 (P/N 100-356-300) and v2 primers. The polymerase-template complexes were bound to magbeads with the PacBio MagBead Binding Kit. The samples were analysed on the Agilent 2100 bioanalyzer. Sequencing reactions were performed by using the PacBio RS II sequencer with DNA Sequencing Reagent 4.0 (P/N 100-356-200). Movie lengths were 240 min (one movie was recorded for each SMRT cell). The PacBio Iso-Seq protocol (SMRT Analysis version v2.3.0.[97].) was used for transcriptome data analysis.

Data analysis and visualization

Reads were mapped to the reference genome (KJ717942.1 [98]; X14112, FJ616285) using BLASR and GMAP alignment tools. Visualization and data analysis were carried out in SMRT Analysis v2.3.0. Reads of mapping quality >20 and chimeric reads were discarded. Repeat-

spanning reads were counted as only a single occurrence. The Bio.motifs package of Biopython [99] was used for identification of the potential polymerase II binding sites and polyA signals. The polymerase II binding motifs were obtained from the JASPAR PolII database [100]. The JASPAR count matrices were converted to the position-weight-matrices and position-specific scoring matrices (PSSM) using the PRV-specific background. The PSSMs were used for a motif search. Score thresholds were generated for the PRV background sequence. Generation of the PSSM for the polyA signal was based on literature data [101]. The PolyApred support vector machine-based method was also used for the prediction of polyadenylation signals [102].

Illumina sequencing

For Illumina sequencing, RNAs were isolated from cells in various stages of infection up to 24 h pi, and afterwards mixed for library preparation, in order to obtain a wide spectrum of PRV transcripts (mixed infection kinetics). Strand-specific total RNA libraries were prepared for paired-end 100 nt sequencing by using the Illumina-compatible ScriptSeq v2 RNA-Seq Library Preparation Kit (Epicentre). For PA-Seq, a single-end library was constructed through the use of custom anchored adaptor-primer oligonucleotides with an oligo(VN)T20 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. Transcriptome sequencing was performed on an Illumina HiScanSQ platform. FastQC v0.10.1 was used to check the quality of raw read files. Reads were aligned to the pig genome (assembly: Sscrofa10.2) and subsequently to the PRV KJ717942.1 reference genome, using Tophat v2.09 [103]. The ambiguous reads were eliminated from further analysis. The mapping for PA-Seq analysis was performed by using the Bowtie2 program [104], followed by peak detection with HOMER in strand-specific mode, with adjustments for the peak qualities of oligo(dT) primed libraries. We used in-house scripts for the assignment of peak categories based on the following criteria for the abundant transcripts: the presence of a PAS in the 50-nt region upstream from the poly(A) site and the presence of at least 2 consecutive adenine mismatches in at least 10 independent reads at the poly(A) site. Annotation and visualization were carried out in the Artemis Genome Browser v15.0.0 [105]and IGV v2.2 [106].

Northern blot analysis

Northern blotting was performed as described by Ausubel et al. [116] in the following way: 10 μg RNA from PRV-infected PK-15 cells and 10 μg RNA from non-infected cells were separated on 1% formaldehyde agarose gel. The RNA was blotted by capillary blotting to a positively charged nylon membrane (Hybond-N, Amersham). Two non-overlapping probes—both mapping within the UL36.5 transcript—were used for the hybridization (Fig 2). Probes 1 and 2 were amplified (see primers in S1 Table) then probes were radiolabeled with DecaLabel DNA Labelling Kit (Thermo Scientific) using 2 MBq α-^{32}P dCTP. Hybridization was performed at 42°C overnight in 50 ml hybridization solution (0.1% SDS, 5x Denhardt's solution, 30% deionized formamide, 5 mMEDTA, 0.9MNaCl, 50 mMNa_2HPO_4, 100 μg/ml fragmented herring spermDNA, α-32P labelled probe). After hybridization, the filters were washed three times at 65°C with the washing solution (0.2xSSC, 0.1% SDS). Scanning and analysing of the results were done using Typhoon™ FLA 9000 imager and Image Quant 5.0 software.

Generation of the ul54 -deleted virus

The PRVul54-KO mutant virus was constructed as described in the following: as a first step, the BamHI fragment containing the entire ul54 gene was isolated from agarose gel, and then was subcloned to pRL525 cloning plasmid [117]. The resulting recombinant plasmid was used as a template for the PCR amplification of the two arms of the flanking sequences providing homology with the targeted genomic region of the PRV. A unique EcoRI site was inserted in place of a 1,017-bp segment (located within 2901 and 3919 bps) within to ul54 gene with the PCR reaction (Figure 2). This was then followed by the insertion of a green-fluorescent protein (GFP; pEGFP-N1 vector, Clontech) gene expression cassette (Clontech) bracketed by EcoRI sites into the EcoRI site of the targeting sequence. The resulting recombinant plasmid was used as a transfer construct for the generation of the knockout virus. The linearized targeting plasmid was transfected along with the purified wild-type (wt, strain Kaplan) viral DNA into porcine kidney (PK-15) immortalized epithelial cells. The recombinant virus was generated by homologous recombination, followed by isolation on the basis of fluorescence of cells infected by the recombinant virus using an inverted fluorescent microscope (Olympus iX71). The first isolate was plaque-purified, which was repeated until the contaminating wt virus was eliminated.

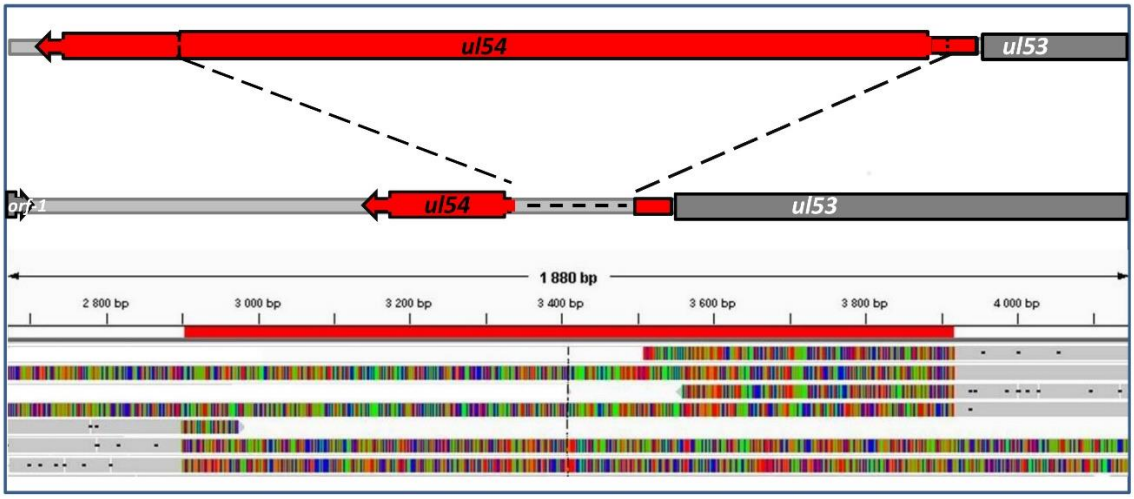


Fig 2. Deletion of the ul54 gene of PRV. *Almost the entire ul54 gene was eliminated by a technique based on homologous recombination.* ***A****: this part of the figure shows the schematic representation of the inserted GFP* expression *cassette (illustrated at the top), as well as the knocked-out region of the PRV genome.* ***B****: Integrative Genomics Viewer (IGV) representation showing the presence of the mutation*

Generation of us7/us8-KO PRV

The BamHI-7 fragment of the viral genome - containing the gE and gI coding region - was isolated and subcloned to pRL525 cloning vector. The 1,855-bp StuI–AgeI DNA fragment was replaced with an EcoRI linker. The GFP reporter gene cassette was modified to contain EcoRI sites at both ends, and this construct was inserted in place of the StuI–AgeI fragment. Removal of the 1,855 bp fragment from the PRV genome resulted in the inactivation of both *us7* and *us8* genes of the virus (Figure 3).

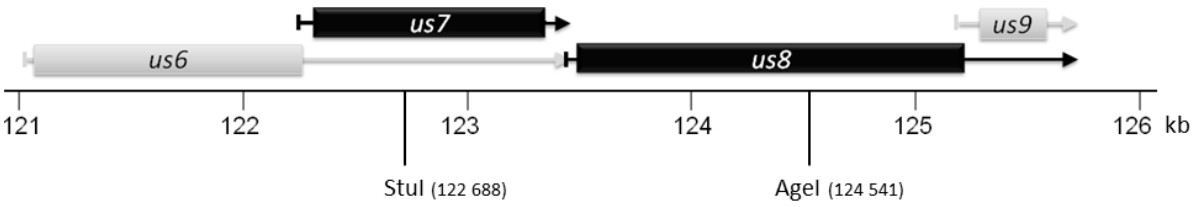


Figure 3. Genetic structure of the mutant virus *The us7/us8-KO virus was generated by the deletion of the 1,855-bp StuI–AgeI fragment from the PRV genome*

Results and Discussion

Pseudorabies Virus

Determination of 5’- and 3’-ends of the PRV transcripts.

The full-length PRV transcripts have mainly been predicted *in silico*. Most of the abundant PRV transcripts have been detected by Northern blot hybridization, and close to a third of them have been analysed by S1 nuclease mapping or primer extension methods [106;107]. However, these techniques can-not determine the 5’-ends of the transcripts with base pair precision. Using PacBio analysis, we determined the exact 5’- and 3’-ends of the RNA molecules. We found that most of the transcripts were initiated (transcription start site; TSS) at 15 to 31 bp downstream from the TATA box, with a mean of 23.5 bp (median: 23; mode: 22) (the GenBank sequence KJ717942.1 has been upgraded with the new data). We additionally determined the complete nucleotide sequences of the 5’-UTRs of 5 transcripts UL6, UL11, UL23, UL33 and UL36 RNAs, which have not been annotated with any method previously. The amplified Iso-Seq method was found to be superior to the non-amplified protocol in the establishment of accurate 5’-ends and the detection of low-abundance transcripts due to the attachment of an oligonucleotide to the upstream region of the first strand of the cDNAs. This latter technique provided complete nucleotide sequences spanning the entire length of the transcripts from the poly(A)-tail to the 5'-end without any sequence loss. PRV genes were found to be controlled by several TATA box sequence variants producing different levels of transcripts.

Genes controlled by alternative promoters.

PacBio sequencing revealed 19 coding and non-coding genes possessing alternative transcription start sites ranging from 2 to 6 TSSs (*Supplement S4 Table Tombácz et al. 2016*). However, TATA-less transcripts have been reported to be common in eukaryotic organisms [109]. We found that the *ul50* gene contains three active TATA sequences, one located within the *ul49.5* gene and the other two within the *ul49* gene. Functional alternative TATA boxes were also identified in the *ul49.5*, *ul32*, *ul44*, *ul21*, *ul5* and *us8* protein-coding genes. Both previously annotated TATA boxes of the *us3* gene were shown to be active in our experiments.

It emerged that the *orf-1* gene is controlled in a more sophisticated manner than previously believed: besides the *orf-1* gene, four longer transcripts were identified upstream of the known TSS, which were named ORF-1M1 and M2 (M: medium) and ORF-1L1 and L2 (L: long). Among these five transcripts, only the ORF-1 and the ORF-1M are controlled by a TATA box. The presence of an ORF suggests that this gene encodes a protein. However, the distribution of the GC content within the ORF exhibited a pattern intermediate between the intergenic and protein-coding sequences, which might indicate that *orf-1* is a pseudogene with accumulated mutations destroying the protein-coding function and therefore the GC preference of the third codon positions as well. For an explanation, the distribution of the high GC content within the three reading frames of the PRV protein-coding genes exhibits a special bias: the G +C bases are primarily accumulated at the silent (generally third) codon positions to an extent close to 100% [110]. This phenomenon provides a unique method for prediction of the coding sequences, since the UTRs and the intergenic sequences do not display such a base distribution pattern. However, the close location of AT-rich packaging and cleavage sequencing can distort the codon usage at the upstream region of the *orf-1* gene. Additionally, we showed that only one of the two predicted TATA boxes is active for the *ul37* and *ul42* genes, at least in the PK-15 cells, since we did not obtain any reads from the distant sequences. Our data also revealed that the *ul9* and the *ul41* genes use different TATA elements from those of predicted *in silico*. The *ul10* gene was shown to produce six different isoforms differing in their TSSs.

Transcripts with alternative poly(A) signals.

It was earlier demonstrated that the genes in many eukaryotic organisms produce alternative transcription end sites (TESs) [111–113]. We also observed this phenomenon in the PRV RNAs. Two kinds of variation were distinguished: transcripts were produced by using alternative PASs (*S2, S3 and S4 Tables Tombácz et al. 2016*) or the same PASs (*S5 Table Tombácz et al. 2016*). Alternative PASs can be separated from each other by one or more genes, or they can be located adjacent to each other without intervening coding genes. Such latter PASs were detected in the *ul44* gene with both Illumina and PacBio sequencing, and in the *ul22* and *ul35* genes by only Illumina sequencing. We also detected PAS variants with the method described by Beaudoing and colleagues [101] in 6 additional PRV genes: *ul27*, *ul35*, *ul44*, *ul22 cto-s* and *us2*. PRV transcripts were found to contain several PAS sequence variants, but only the AATAAA sequence is abundant.

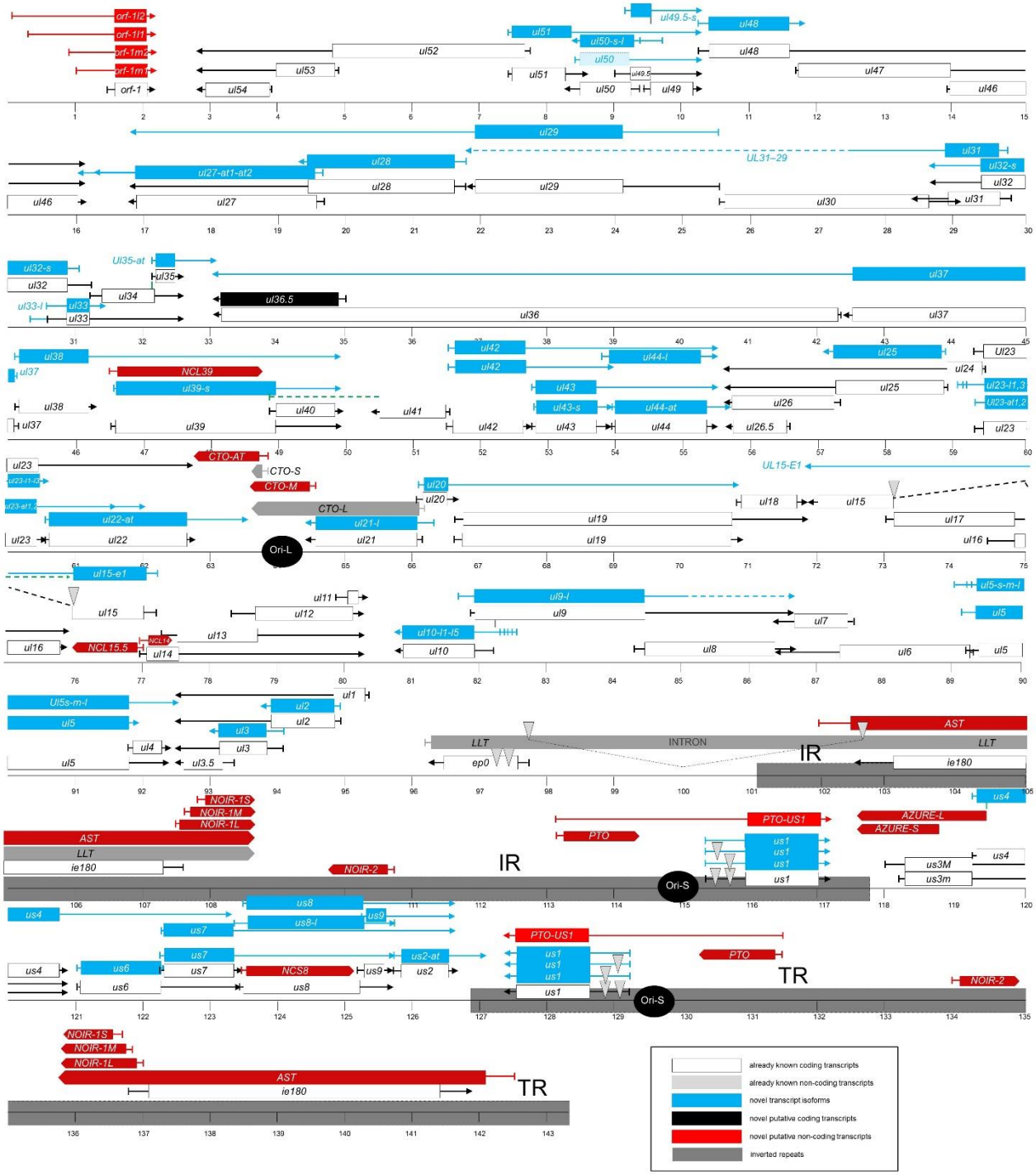


Fig 4. The transcriptome of the pseudorabies virus. *The updated version of the PRV genome is composed of 67 protein-coding genes and 20 putative non-coding genes. The coding transcripts described earlier are depicted as white rectangles with black frames. In polycistronic transcripts, only the most upstream ORFs are illustrated by rectangles, while the downstream genes are represented by arrow lines. The already identified non-coding RNAs are represented as dark-grey arrow-rectangles. The novel putative protein-coding gene (ul36.5) is illustrated as a black rectangle and the novel putative non-coding genes are*

depicted as red arrow lines. The novel mono- and polycistronic transcripts are indicated with blue rectangles and arrow lines. The light-blue rectangle with dashed border lines indicates a transcript containing antisense sequences of the ul50 gene at its most upstream region. The long dark-grey rectangles represent the two IR regions of the viral genome; black circles indicate the three replication origins. Abbreviations within the name of the transcripts: `s': small TSS variant; `m': medium-size TSS isoform; `l': long TSS variant; `at': alternative TES variant.

Novel transcripts.

Our study revealed a novel gene, named *ul36.5*, which is embedded within the larger *ul36* gene. The presence of an in-frame ORF within the *ul36* ORF suggests that this gene encodes a shorter version of UL36 protein. This RNA molecule is generated by alternative transcription initiation, resulting in an 1808 bp-long nested transcript located within the interval 34,934–33,127 nts on the KJ717942.1 genome, and sharing 3'-termi-nals with the UL36 transcript. The *ul36.5* gene might produce a truncated version of the UL36 tegument protein, and is composed of 467 amino acids. A putative TATA box for the *ul36.5* gene is predicted in the interval 34,959–34,964 nts on the PRV genome. We detected a set of nearly uniform ROI coverage along the length of the gene in every sequencing run, indicating that its expression is regulated independently of the *ul36* gene. We found that UL36.5 was an abundant transcript, while the *ul36* gene was expressed at a low level. The expression of UL36.5 transcript was verified by Northern blot analysis.

We have identified 19 novel putative ncRNAs (Fig 4, *S2, S6 Tables Tombácz et al. 2016*). The *ul15* gene is composed of two exons and an intron; the latter includes the *ul16* and *ul17* genes, both oriented oppositely to the *ul15* gene. The first exon of the *ul15* gene (*ul15-e1*) was also demonstrated to be transcribed separately from the second exon, resulting in a poly(A)-transcript, which we have named NCL15.5 (non-coding UL15.5). This transcript is generated by alternative transcription termination and is 11 nts longer than *ul15-e1*. We identified a weak PAS (AATACA) between 76,173 and 76,168 nts of the PRV genome. The 3'-ends of the NCL15.5 and UL16 transcripts are located immediately adjacent to each other, without an intergenic region between them. Although *ncl15.5* encodes the entire amino acid sequence of *ul15-e1*, no stop codon is available in the coding frame. Hence, protein molecules could be produced only by special mechanisms, based, for example, on the use of a distal stop codon in another reading frame through a process such as ribosomal frameshifting, as in the *gag* and *pol* genes of HIV [114]. The amount of NCL15.5 RNA is low relative to that of spliced UL15 RNA (~3.5%).

We recently reported the discovery and characterization of an overlapping ncRNA pair, CTO-S and CTO-L, sharing a common PAS. We identified an additional 3'-end coterminal transcript, termed CTO-M (close to the *ori*L–medium size). The transcription of this 960 nt-long ncRNA is predicted to be initiated from a sequence overlapping the PAS of the *ul21* gene. It was demonstrated earlier that PASs can also function as TATA boxes in herpesviruses [115]. No ORF was detected within this RNA gene. The GC distribution of *cto-m* shows no GC-preference in any reading frames, which suggests that this is a non-coding gene.

We also identified a longer variant of CTO-S, termed CTO-AT (AT: alternative transcription termination; 787 bps), overlapping the UL22-AT, which is an isoform of the UL22 transcript with longer 3'-UTR.

We identified three 3'-coterminal non-coding polyadenylated transcripts, named NOIR1-S, NOIR1-M and NOIR1-L (non-coding RNA in the inverted repeat; short, medium or long variant), located in the IR region of PRV, with sizes of 1293 nts, 1425 nts and 1485 nts, respectively. The *noir1* genes are arranged in different orientations compared to the *ie180* gene. These transcripts are more abundant than the IE180 mRNA. The common PAS of the *noir1* genes is localized in the DNA segment 109278–109283 nts, but we could not identify TATA boxes upstream of these genes, except for the shortest transcript. We also detected longer overlapping 3' co-terminal transcripts, which may be incomplete reads from the LLT driven by the LAT promoter.

We also detected a 902 nt-long non-coding transcript (termed NOIR2) located in the IR region. The *noir2* gene is situated at a distance of 2721 bp downstream of the *ie180* gene and arranged in a parallel orientation with this transactivator gene, and in a convergent orientation with the *noir1* genes. We could not identify either a putative TATA box or typical PAS sequence (only atypical: AATAGA). We obtained relatively few reads with both PacBio and Illumina platforms at each time point from this RNA as compared with an average viral transcript. The lack of ORFs indicates that all four *noir* genes are lncRNAs. No GC-preference is observed at any reading frames of the *noir* genes, which suggests that they are non-coding genes.

Two overlapping ncRNA genes were also identified near the *ori*S sequences, which were named PTO (proximal to the *ori*S; 1098 bps) and PTO-US1 (4087 bps). No complete ROIs were produced from this latter transcript; we can only assume that it is initiated from the nearby *pto* promoter. This transcript overlaps with the *ori*S of the PRV. The PTO-US1 can also be considered to be a US1 transcript with a very long alternative 5'-UTR. No GC-preference is observed at any reading frames, which suggests that *pto* is a non-coding gene.

An additional pair of coterminal ncRNAs encompassing the US-IR boundary region was identified, and termed AZURE (antisense transcripts in the IR-US overlapping region). The *azure-s* gene codes for an 1198 nt-long lncRNA, which partially overlaps the *us3* gene and a 294-nt segment of the IR region; *azure-l* encodes a 2039 nt-long transcript overlapping partially the *us4* gene and fully the *us3* gene. The PA signal of the AZURE transcripts was located between 117,737 and 117,742 nts, but no TATA boxes were detected for the gene encoding this RNA molecule. There is no GC-preference in those sequences of the *azure* gene, which do not overlap with the *us3* gene, which suggests that *azure* is a non-coding gene.
We could detect existence of the antisense transcript (AST) controlled by the antisense promoter (ASP), which was predicted by Cheung [20]. The AST and LLT are coterminal with the NOIR-1 transcripts.

Truncated transcripts.

We detected three polyadenylated transcripts, termed NCL14 (433 bp), NCL39 (2234 bp) and NCS8 (NCS: non-coding US; 1534 bp), whose expressions were pre-maturely terminated, and therefore lacked their stop codons. The potential function of these truncated RNA molecules
The expression of the non-parallel overlapping ncRNAs was verified by traditional and quantitative PCR analyses and the remainder of the transcripts was confirmed by Illumina sequencing and/or random-primer based Iso-Seq sequencing. We did not detect the UL8.5, the LAT transcripts and the spliced version of LLT with any of the sequencing techniques used in this study.

Novel transcript isoforms.

It has been currently thought that the expression of the herpesvirus genes follows the general scheme that most of the downstream genes in a tandem gene cluster can be transcribed either as monocistronic mRNAs or as downstream genes in polycistronic (bi-, tri- or tetracistronic) transcripts, while the upstream genes are expressed only as parts of polycistronic RNA molecules. However, our investigations revealed that many upstream genes of 3'-coterminal gene clusters are also expressed as monocistronic transcripts, and some genes believed to have only their own transcriptional termination also share common 3'-ends with other tandem genes (*Fig 4, S3 Table Tombácz et al. 2016*). We found that, in addition to polycistronic expression,

nine genes were also transcribed individually. Moreover, we identified several previously unrecognized tandem polycistronic units: six bicistronic, seven tricistronic and one tetracistronic transcripts, for which we could determine the 5'-ends (Fig 4, *S2 and S7 Tables Tombácz et al. 2016*). Earlier annotations suggest that the *ul20* gene is expressed exclusively as a monocistronic RNA, whereas we found that the UL20-19 bicistronic RNAs are represented in a much higher pro-portion than the UL20 transcript in infected PK-15 cells. Our study revealed altogether 30 genes that produce novel transcription variants by utilizing one proximate and one or more distant additional PA signals. The newly discovered mono- and polycistronic transcripts are typically expressed at low levels, which explain why they went undetected previously. The question arises as to whether the only function of these transcripts is to contribute to the proteome of the infected cells, or if they also play other additional roles.

Complex transcripts

Complex transcripts contain genes situated in opposite orientations relative to each other (Fig 4). We identified two full-length complex transcripts: UL51-50-49.5–49 and UL50-49.5–49 (*Table 1a Tombácz et al. 2016*). The latter transcript is initiated from the sequence overlapping the PAS of the *ul51* gene. Our investigations revealed a widespread expression of very long polycistronic RNAs belonging to this category, whose upstream sequences could not be determined with the PA-Seq techniques. We illustrated these low-abundance RNA molecules as if they were controlled by the promoters of the closest upstream genes oriented in the right direction and gave *ad hoc* names accordingly. However, it is possible that they are shorter and initiated by still unidentified promoters, or even longer driven by more distant promoters. We identified 15 transcripts with incomplete 5'-ends (*Table 1b Tombácz et. al. 2016).* Five complex transcripts (UL18-15E2-17-16, AZURE-US1-PTO-NOIR2, US2-US1-PTO-NOIR2, NOIR1-NOIR2-PTO-US1and UL31-30-29) were detected only by strand-specific random primer-based PacBio sequencing. With the exception of UL31-30-29, we could not identify their TSSs (their TESs were also undetected; *Table 1c Tombácz et al. 2016*). The presence of the antisense RNA sequences on these transcripts allows the confirmation of their existence by RT^2-PCR analysis and Illumina sequencing, which were carried out in each case (*Table 1 Tombácz et al. 2016*) We also analyzed those parts of the genome for antisense RNA expression which did not produce PacBio reads at all by RT^2-PCR and Illumina sequencing, and found extensive transcription from the complementary regions of practically every protein-coding

and non-coding gene (*Table 1d Tombácz et al. 2016).* We must assume that the entire PRV genome produces very long low-abundance complex transcripts.

Categorization of transcripts with regard to expression profile kinetics

The lytic PRV transcriptome was characterised by the quantification of the polyadenylated RNA molecules produced by the virus during productive infection in cultured cells.

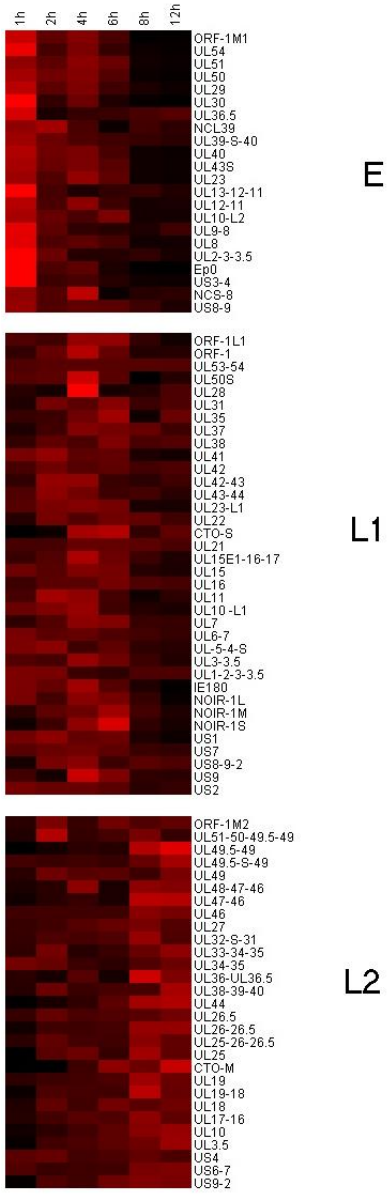


Figure 5. Heat map representation of the transcriptional kinetics measured by RT-qPCR. *The PRV transcripts were clustered according to their F_x values by k-means clustering, using Euclidean distance similarity metrics. Red rectangles indicate high, black rectangles indicate low relative expression values.*

The kinetic categorisation of viral transcripts was based on the principle that the E genes are expressed at a high level in the first stage of the viral life cycle, and produce a relatively low amount of gene product later; meanwhile the L2 genes are expressed only at a low level during the first hours of infection, while becoming more abundant later. The L1 genes are characterised by intermediate kinetic profiles. We used k-means clustering of the F_x values of the transcripts based on Pearson-correlation to create three clusters (Figure 5). Figure 6 shows that the overall E, L1 and L2 gene expressions significantly differ from one another. E genes exhibit high relative expression prior to the onset of DNA replication, which by 12h p. i. then declines considerably. The L2 genes exhibit inverse kinetics compared to those of E genes. The L1 genes behave differently than the E genes at the beginning of infection, while later the expression curve of these became similar. To the contrary, L1 genes display expression dynamics similar to those of L2 genes during the first hours of infection, which becomes different during the second half of infection. The ratio of E gene products in comparison to their maximal values was higher than in the L genes during the first 4h of infection. Independently of the kinetic classes, the amounts of gene products increase between 6 to 8h p. i.; apparently due in part to the multiplication of the viral DNA molecules. The amounts of E gene products typically decrease around 8 and 12h, which was contrary to the increasing quantity of L2 transcripts within this time period.

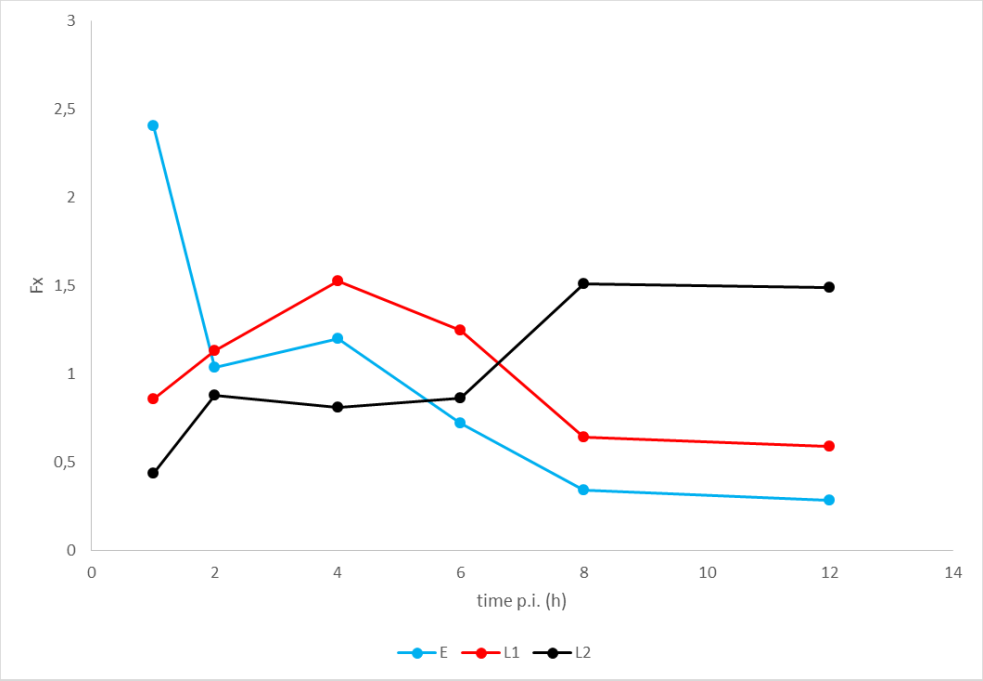


Figure 6. The mean Fx values of the different kinetic classes of PRV transcripts. *The relative expression rate of early genes is high in the early hours of infection, while those of the late genes are higher in later hours. The L1 transcripts exhibit an intermediate expression profile.*

The effect of the *us7/us8* deletion on the expression of PRV genes

We compared the expression levels of 37 PRV genes in *wt* and mutant ($PRV_{us7/us8\text{-}KO}$) backgrounds throughout a 24 h period of viral infection. PK-15 cells were infected with high titre (MOI=10) for both viruses. Comparison of the global transcriptome of the two viruses revealed a relatively higher level of gene expression at the first 6 h in all of the three kinetic classes in the *wt* virus genes ($R_r = R_{gEgI\Delta}/R_{wt} < 1$) (*Table 2. Poka et. al. 2017*). However, this trend reversed in the late phase of infection: with some exceptions, the genes of the mutant virus were expressed in relatively higher levels. For the *us6* and *us9* genes a special reason may account for the irregular behaviour: they are bracketing the deleted region, therefore, their expressions are likely to be affected by the genetic modification itself [e.g. *us6* and *us7* genes share a common poly(A) signal]. These two genes were removed from the further analysis. Comparison of the impact of the mutation on the different kinetic classes of PRV genes revealed that at the early phase of infection the L genes are affected to the greatest extent (lowest average transcript level), while at the later period the expression of IE and E genes are those affected the most (highest average transcript level) **(**Figure 7). The increase of gene expression in the mutant virus is the largest in the *ep0* gene, which produces up to 23.66-fold transcript level compared to the *wt* virus.

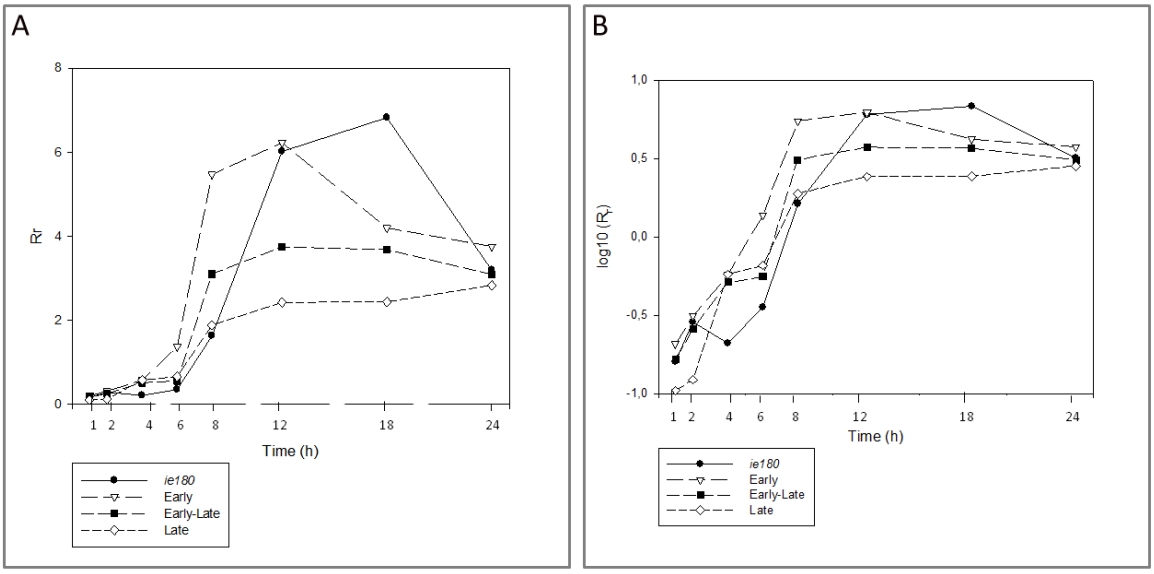


Figure 7. Plots of the average R_r values of the IE, E, E/L and L transcripts versus time. *These plots show that the effect of the us7 and us8 gene products on the other PRV genes increases (decrease in $\overline{R_r}$ values) between 1 and 12h pi, with the higher after 6h pi. It can be seen that the*

stability of the E mRNAs and/or the expression of the E genes are strongly effect affected by the gE and gI proteins, the same was observed in the ie180 gene with a time shift, while this effect is much weaker for the L transcripts and intermediate for the E/L transcripts.

In order to evaluate the effect of the *us7*/*us8* deletion on the expression dynamics of the PRV genes, Pearson's correlation coefficients were calculated for the R values of the average values of the kinetic classes of the genes in the two viruses (Table 4 *Poka et al. 2017*). As a result, we determined that the mutation exerted a moderate global effect on the transcription dynamics of PRV genes: a higher effect on the average E gene expression (r=0.804), an intermediate effect on the average E/L genes (r=0.825) and a less significant effect on the average L genes (r=0.915). However, comparison of the R_{Δ} values of mRNAs of the two PRV revealed that deletion results in an unexpected trend in gene expression dynamics: except in three genes in four time points, the transcript levels continuously increase within the 1 to 12h infection period in the mutant virus, while this tendency cannot be clearly observed especially in the E genes of the *wt* virus (Table 5). The same is true for the 12-18h interval, where the genes of the null mutant behave in a more regular manner (global decrease of transcript levels) than those of the *wt* virus.

The effect of the ul54 gene deletion on the expression of PRV genes

We also investigated the effect of the mutation on the viral gene expressions. PK-15 cells were infected with either the wt or $PRV_{ul54\text{-}KO}$ virus, using low MOI (0.1 pfu/cell) for the infection. The expressions of PRV genes were monitored within an 8 h period of time. The reason for using a short infection period was to exclude that the mature viruses released from the infected cells initiate a new cycle of replication in the non-infected cells in the later periods. We obtained that the mutation exerts a drastic effect on the PRV transcriptome. Compared to the wt, the $PRV_{ul54\text{-}KO}$ exhibits aberrant expression of ie180 gene and several E, E/L and L genes. The effect of the mutation on individual PRV genes was examined by using the R_r values, which was calculated as the ratio of the R values of mutant and wt virus at each time point. We also calculated the impact of mutation on the average E, E/L and L transcripts (Figure 8, *Table 4 Csabai et al. 2017*) It can be seen that the E genes on average are negatively affected by the mutation at the first 1 h post infection (p.i.), while the expression levels become the same in the two viral backgrounds by 2 h p.i. This is in contrast with both the E/L and L genes, which are over-expressed in the mutant virus within this period. This latter result is consistent with the observation made by Schwartz and colleagues, who have shown the accumulation of late viral

product at the early period of infection [123]. An overall decline of transcript levels was observed at 4 h p.i. which may be related with the differential effect of the initiation of DNA synthesis on the two genotypes. The genes belonging in different kinetic classes behave dissimilarly by 6 h p.i.: there is a significant fall in the rate of expression of L genes, while the E genes appear to become unaffected by the mutation. Finally, all kinetic classes of PRV genes become considerably suppressed by 8 h p.i. Note that gene expressions are significantly lower in the mutant than in the wt background at 4 h p.i. in all PRV genes, except the latency-associated transcript (LAT), which exhibits a 7.7 fold increase of expression in $PRV_{ul54\text{-}KO}$ (*Table 3 Csabai et al. 2017*). By far the highest elevation in gene expression is detected in *ul53* gene (8.25 fold) in the mutant genome at 6 h p.i., and furthermore, this gene is the only gene that is expressed at a higher level at 8 h p.i. in the $PRV_{ul54\text{-}KO}$ than in the wt background. This phenomenon may be explained by the fact that *ul53* and the deleted ul54 genes are adjacent to each other on the viral genome, and *ul54* might exert a cis- or trans-acting suppressive effect on *ul53*, which is non-existent in the mutant virus. Intriguingly, the *ul52* gene, which also produces co-terminal transcripts with the ul54 gene, is transcribed at a lower level ($R_r = 0.39$) in the mutant virus at this stage of infection.

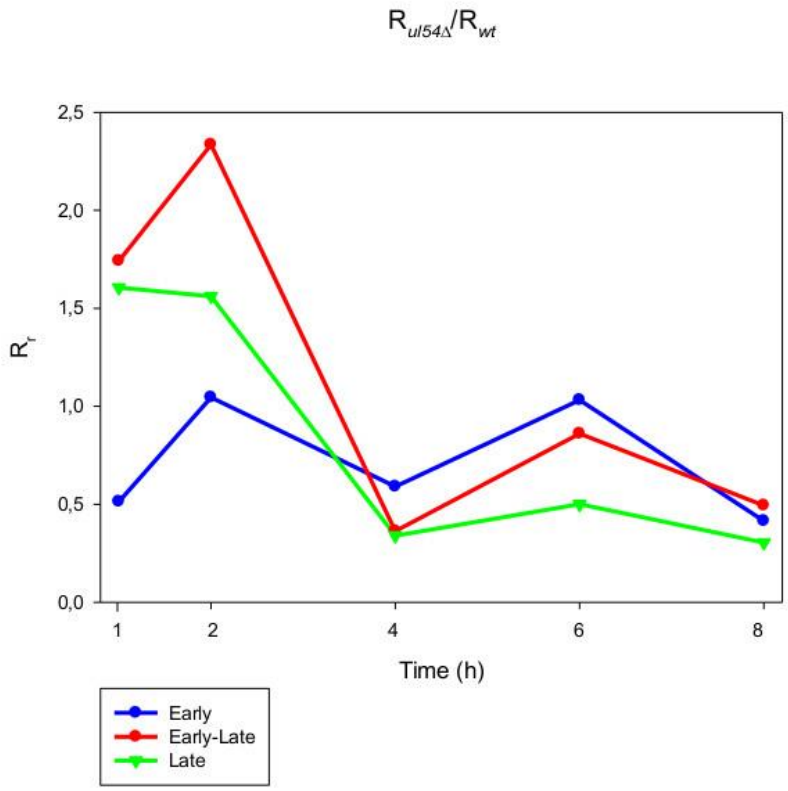


Fig 8 The impact of the ul54 mutation on the expression of PRV genes. *This plot shows the average R_r values of the three kinetic classes of PRV genes. The late genes are up-regulated at the early stage, while they are down-regulated at the late stage of infection in the mutant*

background. The early genes are down-regulated at 1h and 8h of infection in the ul54-KO virus. Black-filled circles with a straight line indicate the measured average R_r values of the E genes; White-filled triangles with a dashed line represent the E/L genes, while the values of L genes are labeled by black-filled squares with a dense dashed line.

Abrogation of ul54 gene function also affected the DNA synthesis of the mutant virus. The onset of replication exhibits in a more than 2-h delay in $PRV_{ul54\text{-}KO}$ and a low copy number of DNA is produced compared to the wt virus (Figure 9). The expression of the cto gene encoding a non-coding transcript is also delayed by two hours and expressed in a very low amount in the mutant background. The CTO is supposed to interact with the DNA replication; the correlation between the delays of the two processes may therefore not be a coincidence. The time slip in the initiation of viral replication has a great impact on the gene expression, especially for the L genes from 4 h p.i., whose expressions are dependent on the replication. At a later stage of the viral life cycle a global repression of gene expression in the mutant background was observed. This phenomenon is explained by the low copy number of the mutant viral DNA.

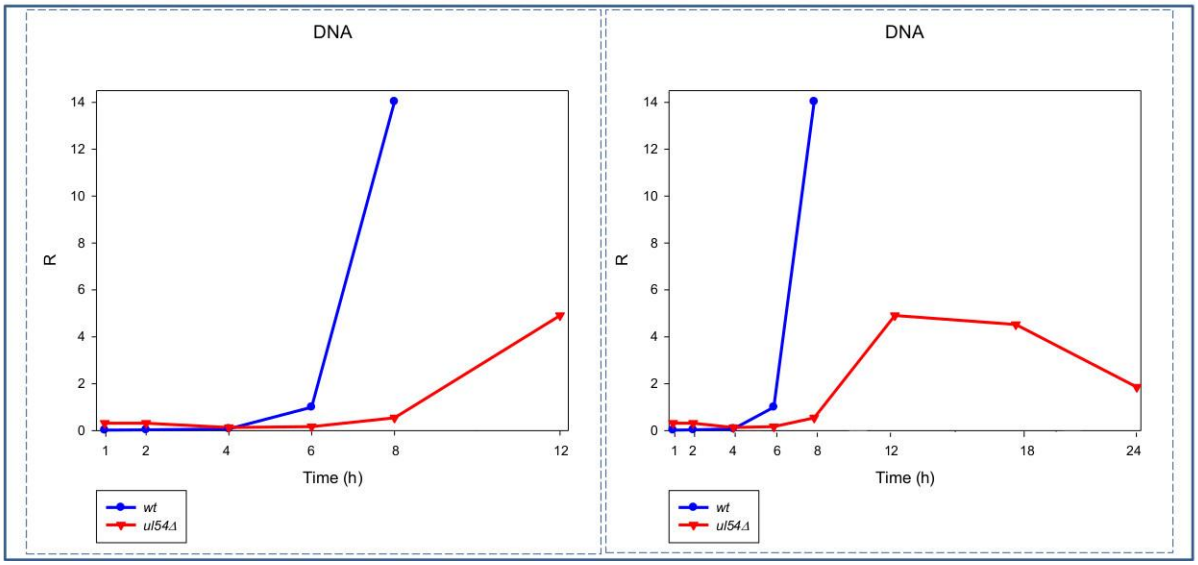


Fig 9 Replication of the PRV DNA in the wt and mutant viruses. *The abrogation of ul54 gene leads to a significant decreasing effect on the DNA replication. These plots show the dynamics of the DNA replication during the first 12 h pi (a) or between 1-24 h p.i. (b) in the mutant virus, as well as between the 1-8h in the wt PRV.*

Herpes simplex virus-1 (HSV-1)

Novel putative protein-coding genes

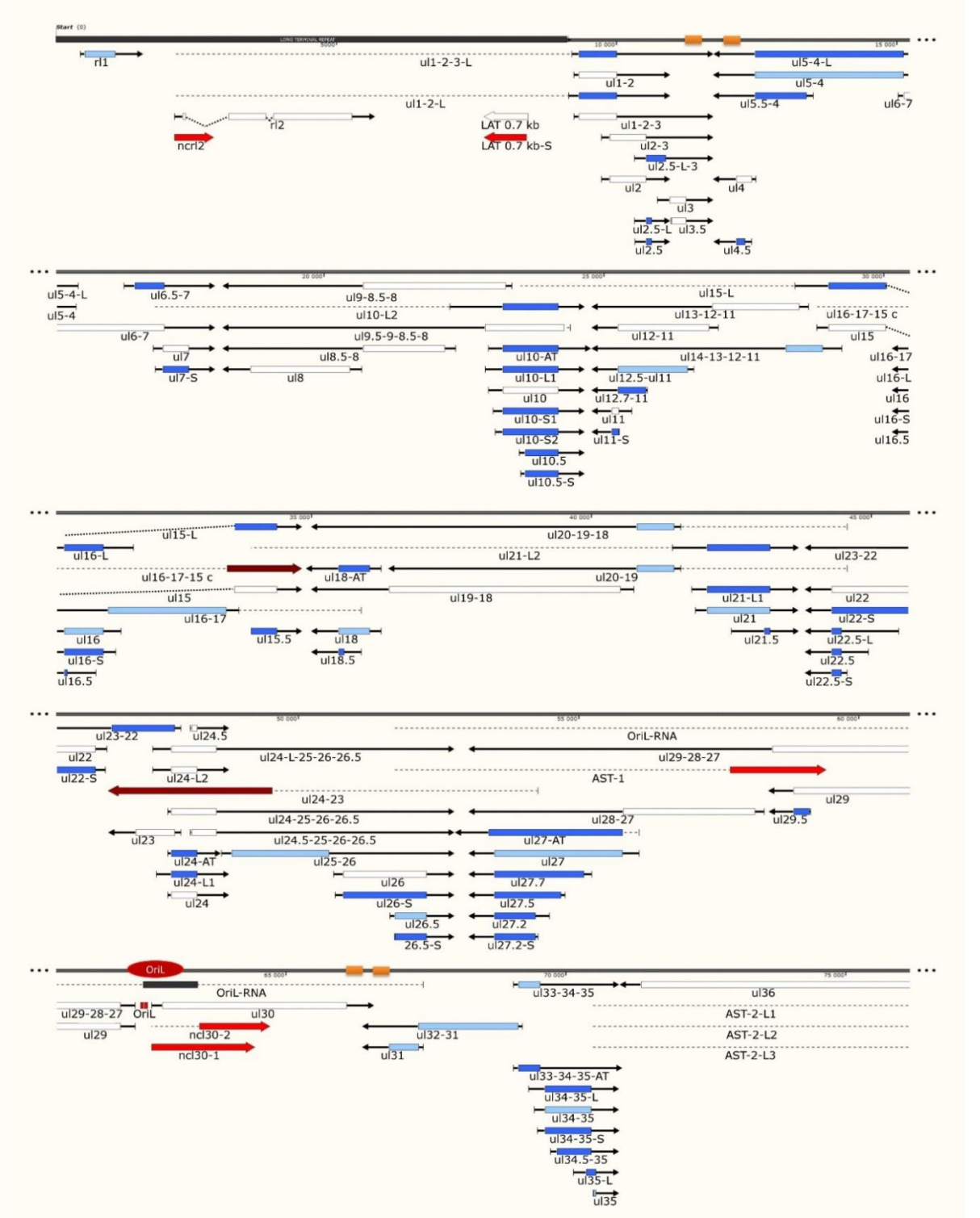
Our investigations revealed 34 novel putative protein-coding genes (Figure 10, *Table 1, Table S4 Tombácz et al. 2017*), which is a fairly large number in a well-characterized viral genome. Each new gene is embedded into already annotated protein-coding genes. The mRNAs from these 5' truncated genes are generated by alternative transcription initiation from an intragenic promoter within the larger host gene. The first in-frame AUG triplets are supposed to correspond to the first amino acids of the putative protein molecules. With the exception of seven genes (*ul5.5; ul15.5; ul21.5; ul24.5; ul27.2; ul41.5;* and *ul53.5*), the remaining novel genes were expressed at markedly lower levels than the host genes.

Novel non-coding transcripts

We identified ten novel ncRNAs, including a LAT variant, four antisense (as) and four 3' truncated transcripts, as well as a putative RNA molecule overlapping the replication origin (Ori-L) of the virus (Figure 10**,** *Table 2 Tombácz et al. 2017*). The novel 0.7 kb LAT-S transcript is a TSS isoform of the 0.7-kb LAT [118]. We detected polyadenylated asRNAs (termed antisense transcripts, AST) transcribed from the complementary DNA strands. The existence of ASTs was verified by strand-specific PCR. We also detected lncRNAs, which were produced from protein-coding genes, but their transcriptions were prematurely terminated, and as a result lack stop codons. The potential function of these 3' truncated RNA molecules remains obscure. The homologue of NCS8 transcript has been described in PRV. Furthermore, we identified a putative Ori-L-overlapping ncRNA with uncertain orientation, TSS and TES. We cannot exclude the possibility that it is not an individual RNA molecule but is rather the upstream region of a putative longer TSS isoform of the UL30 transcript. We did not detect HSV ncRNAs homologous to the PRV CTO-S transcript located in the close vicinity of the Ori-L of this virus. The probable reason for this is that the Ori-L is situated at different genomic loci in the two viruses. We could however detect the Ori-S-overlapping RNA described by Voss and Roizman [119].

Determination of the 5’ and 3’ termini of the HSV transcripts

We precisely mapped the 5’ and 3’ ends of the HSV transcripts (*Table S3 Tombácz et al. 2017*) and re-annotated those that had not been well established earlier. Our amplified Iso-Seq technique is capable of identifying the full-length transcripts without any loss at the terminal sequences.



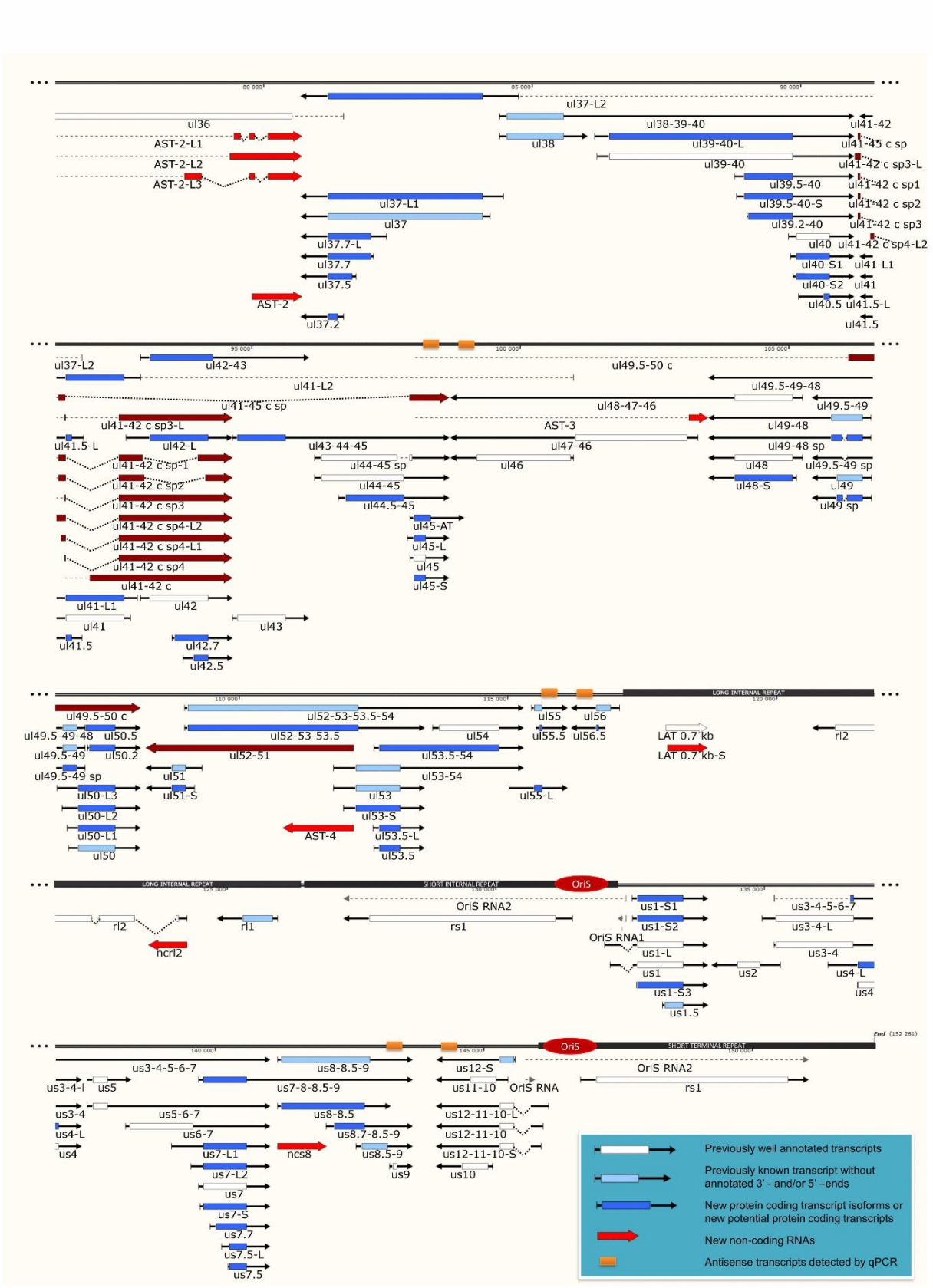


Figure 10. The transcriptome of the herpes simplex virus *The current version of the HSV-1 genome is composed of 115 protein-coding genes and 19 putative non-coding RNAs (continued from the previous page). The coding transcripts identified earlier are depicted as arrow-rectangles with white*

boxes (indicating the ORFs), the previously known transcripts (without annotated TSS and TES positions) are illustrated as light blue arrow-rectangles. The novel potential protein coding genes are labeled by dark blue rectangular arrows. The already identified lncRNAs are represented as dark-grey arrow lines, while the novel putative lncRNAs are depicted as red arrow lines. The novel polycistronic transcripts are indicated with dark blue rectangular arrows. The dark-red rectangular arrows show complex transcripts. The long black boxes represent the repeat regions of the HSV genome; red ovals represent the three replication origins. Abbreviations within the name of the transcripts: 'S': short TSS variant; 'L': long TSS variant; 'AT': alternative TES variant.

We found that most of the TSSs were located between 28 to 33nts downstream from the TATA box. Altogether, we were able to identify 46 novel TSSs and 6 TESs in already described transcripts, as well as 16 TSSs in the novel transcripts.

Transcription start site isoforms

PacBio cDNA sequencing uncovered 42 protein-coding and non-coding transcripts with alternative TSSs (Figure 10, *Table 3 Tombácz et al. 2017*). We could detect putative TATA boxes for six TSS variants; however, TATA-less genes have been reported to be common in eukaryotic organisms [109]. Similarly to PRV, we have detected multiple TSSs for the HSV UL10 transcripts which indicate that the complex regulation of this gene is conserved among herpesviruses. We found that the *ul22* and the *us7* gene contain two active TATA sequences, and seven genes (*ul2*; *ul6.5*; *ul12*; *ul22*; *ul51*; *us1*; *us7*) have at least two potential TATA boxes. We identified two transcripts that overlap the Ori-Ss of HSV (Figure 10) at their 5'-UTRs; both are long TSS isoforms of *us1* and *us12* genes. In contrast to the PRV *us1* gene that is located in the IRS, in the HSV only the promoter and a short 5'-UTR of this gene map to the repeat region. The Ori-S-overlapping US1 transcript is homologous to the PRV PTO-US1. Our analysis has revealed extensive minor variations in ~ 68% of the transcripts. These RNA molecules are controlled by the same promoter and vary in length from one to ten nts at their 5'-ends (*Table 3, Tombácz et al. 2017*).

Transcription end site isoforms

It has earlier been shown that many eukaryotic genes produce TES isoforms using alternative PA signal [111-113]. We detected six HSV transcripts, which each had two TES isoforms (Figure 10). Except for three transcripts, we found that the rest of the abundant RNA molecules containing the same PA signal exhibited considerable polymorphism at their 3'-ends (*Table 4 Tombácz et al. 2017).*

Splice isoforms

In this report, we detected 13 novel spliced transcripts (Figure 10, *Table 5 Tombácz et al. 2017).* The most intricate splicing pattern was found in the *ul41-42* and *ul41-45* complex transcripts. We identified two splice variants of the AST-2 antisense transcript. We also detected a new splice isoform of the *ul49* gene, which was found to be expressed in a much higher level than the non-spliced variant. The UL49 and UL49-48 transcripts are the only ones in which the splicing occurs within an ORF, which is translated (in other cases the splicing takes place in either non-coding transcripts or in the downstream genes in polycistronic transcripts, which are non-translated). The splicing within the *ul49* ORF results an in-frame deletion.

Novel polycistronic transcripts

According to the current concept, the HSV genome is organized in such a way that the downstream genes in a tandem gene cluster are transcribed either as monocistronic transcripts or as downstream genes of polycistronic (bi-, tri-, tetra-, or pentacistronic) RNA molecules, whereas the upstream genes are expressed exclusively as parts of polycistronic transcripts. Our earlier investigations revealed that several upstream genes of 3'-coterminal gene clusters of the PRV are also transcribed as monocistronic RNA molecules. However, in this study, except for the transcripts expressed from the embedded genes, we only detected novel polycistronic molecules (*Table 6 Tombácz et al. 2017*) that include transcripts terminated upstream of the co-terminal PA sites. Most of the newly discovered polycistronic RNA molecules are expressed at low levels, which explain why they had previously gone undetected.

Complex transcripts

Complex transcripts are defined as containing at least two genes with opposite orientations. We identified ten full-length complex transcripts in HSV, seven of which had been generated by alternative splicing of the RNA molecule encoded by the genes within the *ul41-45* region (Figure 10). Earlier we had detected the homologs of the UL41/42 and the UL52/51 complex transcripts of HSV-1 also in PRV. Our investigations revealed a widespread expression of very long complex transcripts in PRV whose upstream sequences could not be determined even with the long-read PA-Seq technique. We detected more full-length but fewer partial complex transcripts in HSV than in PRV; the reason for the latter may be that in our previous report we also used random primer-based sequencing that allowed for the capture of more distal upstream

sequences than the PA-Seq technique alone. The RNA molecules with partial sequences were illustrated as if they were controlled by the promoters of the closest upstream genes and were given *ad hoc* names accordingly. However, it is possible that they are shorter and initiated by yet unidentified promoters, or even longer and driven by more distal cis-regulatory sequences.

Cytomegalovirus

Transcription End Sites

The polyadenylated fraction of HCMV RNAs was used to determine the TESs. Oligo(dT) primers are considered specific for the detection of poly(A) tails, and therefore TESs; however, long stretches of (A)s, which can also bind the oligo(dT) primers (albeit with a lower affinity), and therefore produce nonspecific 3'-truncated transcripts. Beside the oligo(dT) primers binding to internal stretches of (A)s, template switching may also be capable of producing 3'-truncated transcripts in a homology-dependent manner. The sequence similarity between stretches of (A)s and the poly(A) tail make these genomic positions predicted sites for template switching. Indeed, we have found 290 potentially spurious TESs, where the genomic positions contained at least three (A)s **(***Dataset S1 Balázs et al. 2017***)**. After discarding these positions, we were able to identify 116 TESs **(***Supplementary Dataset S2 Balázs et al. 2017***).** We were further able to observe that the exact positions of the polyadenylation sites of the transcripts varied by a few nucleotides. In our annotation, the nucleotide where the most reads terminated was designated as the TES. This spread showed a slight dependence from the type of the poly(A) signal **(**Figure 11**)**, and could not be observed in the case of the false poly(A) sites, where reads ended uniformly at the same nucleotide. Unlike real TESs (70%), these positions were rarely (14%) preceded by the most common canonical poly(A) signal (AATAAA). We justify the cut-off values used by the our analysis in fact that the discarded TESs were rarely preceded by the canonical poly(A) signal, while the accepted TESs were preceded by the canonical poly(A) signal as frequently as the host TESs were [101].

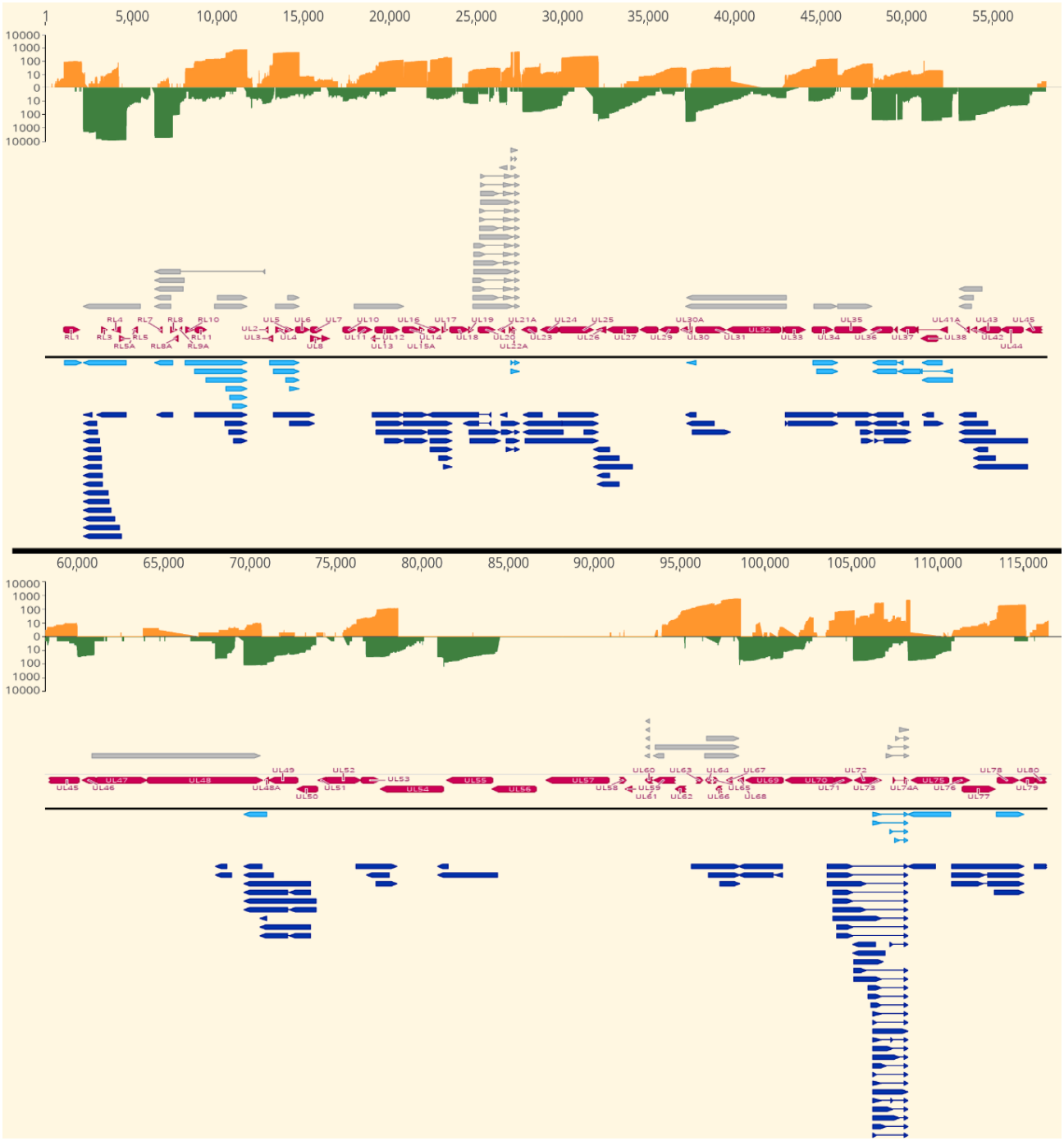
Transcription Start Sites

The applied PacBio isoform sequencing (IsoSeq) protocol allows for the precise identification of TSSs, however, 5'-degradation of the transcripts continues to be a critical issue. Assuming

that sequencing reads start significantly more often at real TSSs than at other genomic positions, we looked for genomic positions where more reads started than it was to be expected in its 101-nt-long region (from 50 nt downstream to 50 nt upstream of the position). The Poisson probability distribution was used to assess significance, similarly to the approach used by Amman et al. [121]; the 101-nt window was chosen, in order to avoid the bias originating from the differential expression of genomic regions. 789 genomic positions were identified as local maxima (±10nt), where at least two reads started at the exact same nucleotide, 248 **(***Dataset S3 Balázs et al 2017***)** of which were accepted as TSSs ($p<6.337*10^{-5}$, Bonferroni $0.05/789=6.337*10^{-5}$). Approximately half (121 out of 248) of the accepted TSSs are verified by reads in both libraries. The reason for this low verification rate is the low read count of the random library. Our analysis also confirmed many TSSs which were published in previous studies using 5'-RACE (as listed with references in *Dataset S4 Balázs et al. 2017*). Even though these studies often investigated HCMV strains other than Towne, the identified TSSs in this study and other studies often matched with base-pair precision.

Novel splice junctions

80 splice junctions were identified when using the criteria of the presence of at least two independent reads and the occurrence of GT dinucleotide adjacent to the donor site, and AG adjacent to the acceptor site. 21 of these splice junctions have not been detected previously, but most of these novel splice junctions share a common donor or acceptor site with previously described splice junctions. Seven other frequent deletions in the reads that did not adhere to this GT-AG rule contained repetitive sequences of varying lengths (3-6 nucleotides) around the deleted segment, and therefore had probably arisen through the course of template switching **(***Dataset S5 Balázs et al. 2017***)**. It is important to note that repetitive elements of similar lengths are found in 33 of the 80 accepted splice junctions as well, but 28 of these splice junctions have either been described previously or they share either donor or acceptor sites with previously described splice junctions. Forty-four (55%) of the detected splice junctions were confirmed by reads in both the poly(A) selected and in the random libraries. A similar portion (10 out of 21 [48%]) of the newly described splice junctions were detected in both libraries.


1
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
45,000
50,000
55,000
60,000
65,000
70,000
75,000
80,000
85,000
90,000
95,000
100,000
105,000
110,000
115,000

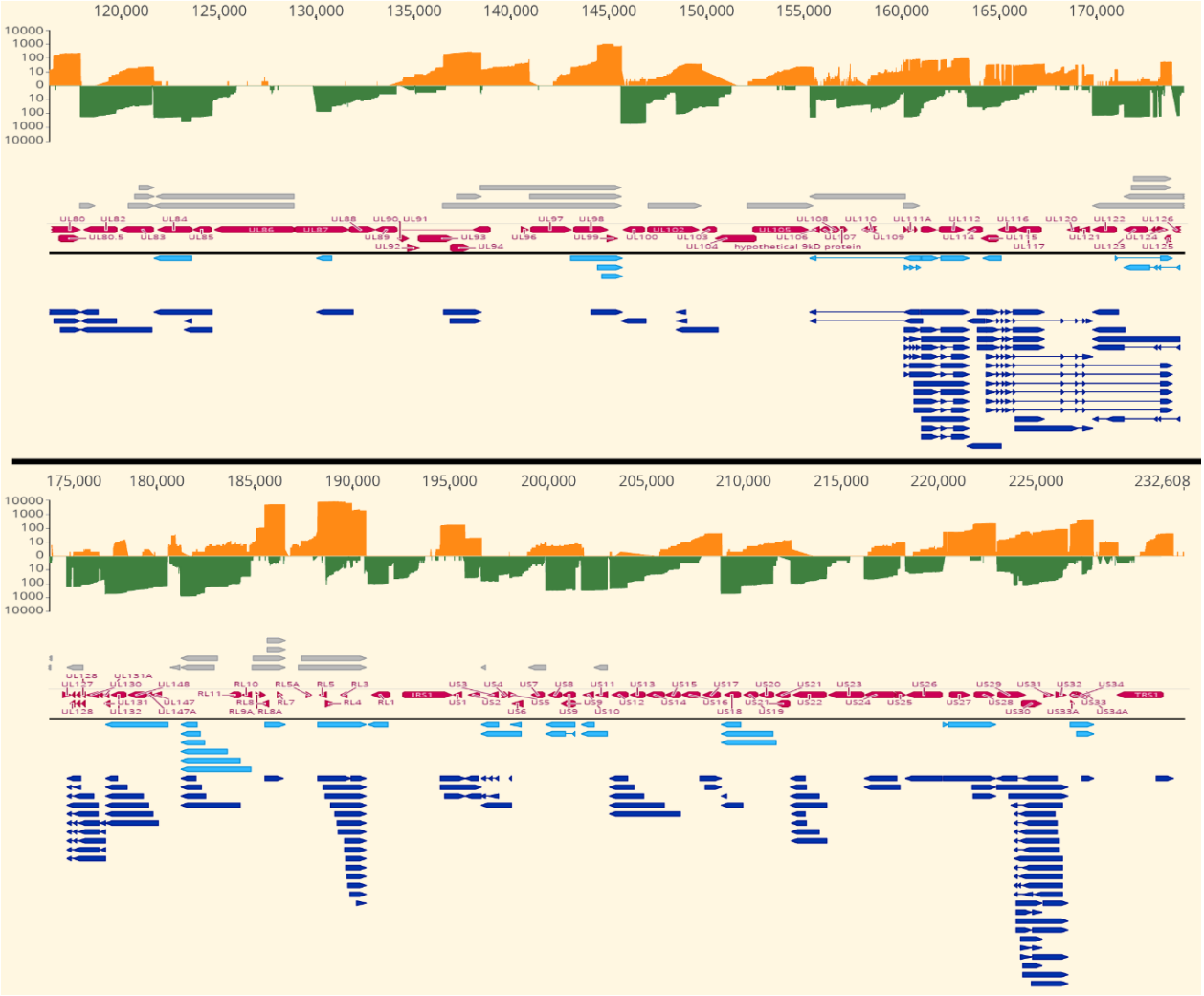


Figure 11. ***Transcriptional landscape of the HCMV***. *(continued from the previous page) The figure presents the coverage and the annotated transcripts on the genome in four blocks. The coverage histogram (above) is drawn on a logarithmic scale, where the orange bars represent coverage on the plus strand and dark green bars represent coverage on the minus strand. The coverage values of oriented reads from the random and poly(A)$^+$ libraries are summed. The annotated transcripts are located below the coverage chart. The previously described transcripts that were not detected by our RNA-sequencing are grey, while the previously described transcripts that our experiments have confirmed have been depicted in light blue, and novel transcript isoforms have been labelled with dark blue. Between the transcripts, canonical ORFs are indicated with purple.*

Transcript annotation

Altogether, 354 HCMV transcripts and transcript isoforms were identified **(**Figure 11**)**; these include already known and novel transcripts, as well as previously unannotated polycistronic transcript variants, 5'-UTR and 3'-UTR isoforms and splice variants. The numbers of reads assigned to each transcript (read starts in the ±10-nt-bin around TSS of the transcript and ends at ±10-nt-bin around the TES of the transcript) is shown in *Dataset S4 Balázs et al. 2017*. Gatherer and colleagues [86] have reported a very high expression rate of the RL4 gene, which was confirmed by our analysis; we obtained an approximately 45-fold average coverage in this genomic location. This has not been captured in the read counts assigned to the transcripts, because the RL4 region contains multiple false poly(A) sites, and reads ending in these false poly(A) sites were not assigned to any transcript. Most of the 291 newly described transcript variants are length variants of already known transcripts, containing different TSSs, TESs, splice junctions or a combination of these. In our analysis, we have also confirmed 63 previously annotated transcripts; the lack of confirmation for other transcript variants does not imply that they are not transcribed in the Towne strain. It could possibly mean that they are expressed at low levels or at other post infection (p.i.) times (our samples contained an equal amount of RNA templates from the different p.i. time points, since viral transcription is more active at late times, it means that late viral transcripts are likely to be overrepresented). Sequencing read length also poses a limitation to transcript identification. The library preparation methods used in this study prefer cDNA sizes between 1 and 2 kbp. This means that very short or very long transcripts could not be detected by this analysis.

Novel transcripts

We have identified nine transcripts and their isoforms in the genomic regions where no transcripts have been described to date, and no canonical ORF had been annotated before (**Figure 11**). Eight of these transcripts are antisense to canonical ORFs (UL20, UL36, UL38, UL54, UL115, US1, US17 and US30). Kaye et al. [122] have reported a transcriptional activity antisense to the UL16 mRNA, and have indicated a shorter transcript, partially overlapping with the UL20-AS1 therefore what they described, might have been a shorter variant of this transcript, but no overlapping transcript to the shorter variant, UL20AS2, has been described. The isoforms of UL20AS, UL54AS, US1AS and US17-AS all contain one or more short (<100AA), but translationally active ORFs. The isoforms of the UL36AS UL38AS UL115AS,

and the US30AS, transcripts contain only short ORFs and these ORFs have not shown significant ribosome coverage in [88]. Our analysis also discovered a novel transcript in the short repeat region (designated RS2), which is partially antisense to the RS1. Although ribosome profiling experiments detected no transcriptionally active ORFs in this transcript, it does contain two short ORFs, one well conserved, overlapping with the RS1 ORF, and one less conserved, that is not antisense to the RS1 ORF (**Figure 12**). The transcripts which are antisense to known coding genes, are highly conserved, however, the transcript RS2, which is only partially antisense to the RS1 gene, is significantly less conserved ($p<10^{-5}$, analysis of variance).

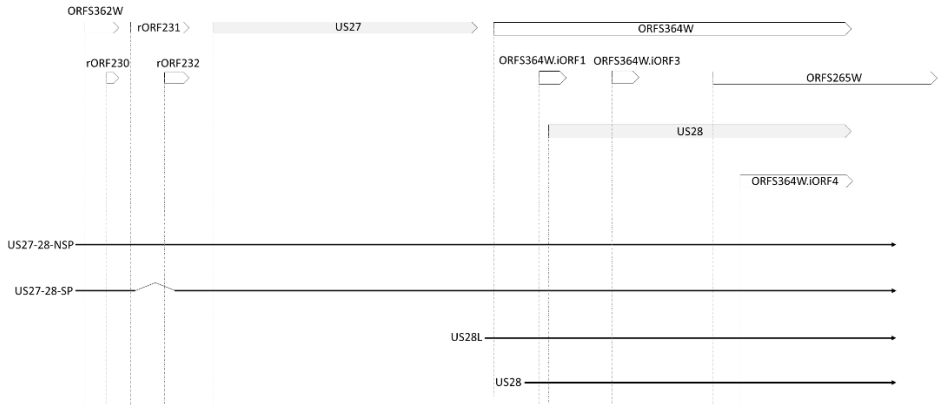


Figure 12. Transcript isoforms contain different ORFs. The figure shows an example of the differential peptide coding capacity (above) of transcript isoforms (below). Canonical ORFs are represented as arrows with a grey background, the other translationally active ORFs are represented as empty arrows and named as published by Stern-Ginossar et al.[88] Dotted vertical lines mark the translational start sites of the ORFs. The transcript isoforms of these two genes can be differentially translated due to polycistronism (US27-28 or only US28), alternative splicing (the splicing in US27 leads to the excision of 75 nucleotides and does not cause frameshift) or alternative transcription initiation (leading to a truncated protein in the cases of ORFS364W and US28).

Conclusion:

Our results revealed the feasibility of the deep sequencing of full-length RNA molecules from the transcriptome of a herpesvirus both at a single-molecule level and in amplified samples. Our investigations essentially redefine the transcriptome of the PRV. We demonstrated that herpesviruses exhibit considerably more genetic complexity than predicted from *in silico* ORF-based genome annotations and gel-based assays. Our investigations uncovered that essentially the entire PRV genome is transcriptionally active, including both DNA strands of the coding and intergenic sequences. Identification of a pervasive genome-wide overlapping pattern of PRV transcripts and of *ori*-overlapping RNA molecules raise the possibility for the potential existence of a genome-wide network exerting joint control on gene expression and replication.

Our investigations revealed an intricate meshwork of transcriptional read-throughs leading to overlapping RNA molecules. It turned out that herpesvirus genes are transcribed in more combinations than it had been previously thought. The number of asRNAs and the complex transcripts of herpesviruses are likely to be underestimated, because most of them may have been undetected due to their non-polyadenylated nature or because they are too long to be identifiable with even a long-read platform.

We demonstrated the utility of long-read sequencing for the investigation of the dynamic transcriptome of a herpesvirus. We have established that this technique can also be applied in the study of processes exhibiting a definite, well-controlled time-course of transcription, such as during viral replication, embryogenesis, tissue regeneration. We have characterised the kinetic properties of several novel PRV transcripts.

We described the generation of a mutant PRV strain with a deletion at the *ul54* locus and the transcriptional characterization of this virus in cultured cells using a real-time RT-PCR technique. We also analyzed the dynamics of viral DNA synthesis and correlated the obtained data on replication with the transcription patterns of the viral genes. We obtained that the abrogation of *ul54* function leads to a differential effect on the various kinetic classes of the PRV genes. This effect may be direct at the early phase of gene expression, but later this

mutation likely exerts its influence on global gene expression at least partly through the DNA replication, which is impeded compared to the *wt* virus.

We investigated the role of gE/gI protein, which may be unrelated with spreading by the analysis of the impact on the mutation on global transcriptome. Our results reveal that the deletion of the *us7* and *us8* genes of PRV leads to significant overall reduction of gene expressions in the first six hours p.i. in every kinetic class of genes without bias toward any of them. However, later (8-24h pi) the genes are upregulated in the *mutant* virus compared to the *wt* virus. This faciliatory effect was much higher on the E and E/L genes compared to the L genes, which is indicated by the decrease of the relative contribution of L gene products to the global viral transcript in the null mutant.

Acknowledgement

First of all I would like to thank for my supervisors Dr. Dóra Tombácz and prof. Zsolt Boldogkői for their support and guidance throughout the Ph.D. program.

I would like to thank my colleagues and staff of the Department of Medical Biology for their assistance, especially for Marianna Ábrahám and Csilla Magyarné-Papdi.

I am grateful for the effective teamwork for our research group members: Dr. Zsolt Balázs, Moldovan Norbert, Dr. Szűcs Attila and former members Dr Péter Oláh, Dr. Nándor Póka.

Last but not at least I give my special thanks for my lovely fiancée Helga and for my family.

References

[1] Carter, J. B., & Saunders, V. A. (2013). *Virology : principles and applications*. Wiley.

[2] Ulrich Koszinowski, P., & Raveendra Pothineni, V. (2010). Proteome-wide production of monoclonal antibodies and study of intracellular localisation for Varicella-zoster virus (VZV). Retrieved from https://edoc.ub.unimuenchen.de/12411/1/Pothineni_Venkata_R.pdf

[3] Harkness, J. M., Kader, M., & DeLuca, N. A. (2014). Transcription of the Herpes Simplex Virus 1 Genome during Productive and Quiescent Infection of Neuronal and Nonneuronal Cells. *Journal of Virology*, *88*(12), 6847–6861. http://doi.org/10.1128/JVI.00516-14

[4] Pomeranz, L. E., Reynolds, A. E., & Hengartner, C. J. (2005). Molecular Biology of Pseudorabies Virus: Impact on Neurovirology and Veterinary Medicine. *Microbiology and Molecular Biology Reviews*, *69*(3), 462–500. http://doi.org/10.1128/MMBR.69.3.462-500.2005

[5] Szpara, M. L., Kobiler, O., & Enquist, L. W. (2010). A Common Neuronal Response to Alphaherpesvirus Infection. *Journal of Neuroimmune Pharmacology*, *5*(3), 418–427. http://doi.org/10.1007/s11481-010-9212-0

[6] Strack, A. M. (1994). Pseudorabies virus as a transneuronal tract tracing tool: specificity and applications to the sympathetic nervous system. *Gene Therapy*, *1 Suppl 1*, S11-4. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8542383

[7] Card, J. P., & Enquist, L. W. (2014). Transneuronal Circuit Analysis with Pseudorabies Viruses. In *Current Protocols in Neuroscience* (Vol. 68, p. 1.5.1-1.5.39). Hoboken, NJ, USA: John Wiley & Sons, Inc. http://doi.org/10.1002/0471142301.ns0105s68

[8] Boldogköi, Z., Sík, A., Dénes, A., Reichart, A., Toldi, J., Gerendai, I., … Palkovits, M. (2004). Novel tracing paradigms--genetically engineered herpesviruses as tools for mapping functional circuits within the CNS: present status and future prospects. *Progress in Neurobiology*, *72*(6), 417–45. http://doi.org/10.1016/j.pneurobio.2004.03.010

[9] Boldogkoi, Z., Balint, K., Awatramani, G. B., Balya, D., Busskamp, V., Viney, T. J., … Roska, B. (2009). Genetically timed, activity-sensor and rainbow transsynaptic viral tools. *Nature Methods*, *6*(2), 127–30. http://doi.org/10.1038/nmeth.1292

[10] Morris, D. R., & Geballe, A. P. (2000). Upstream open reading frames as regulators of mRNA translation. *Molecular and Cellular Biology*, *20*(23), 8635–42. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11073965

[11] Mainguy, G., Koster, J., Woltering, J., Jansen, H., & Durston, A. (2007). Extensive polycistronism and antisense transcription in the mammalian Hox clusters. *PloS One*, *2*(4), e356. http://doi.org/10.1371/journal.pone.0000356

[12] Dossena, S., Nofziger, C., Bernardinelli, E., Soyal, S., Patsch, W., & Paulmichl, M. (2013). Use of the operon structure of the C. elegans genome as a tool to identify functionally related proteins. *Cellular Physiology and Biochemistry : International Journal of Experimental Cellular Physiology, Biochemistry, and Pharmacology*, *32*(7), 41–56. http://doi.org/10.1159/000356623

[13] Li, M.-L., Cui, W., Zhao, Z.-Y., Mo, C.-C., Wang, J.-L., Chen, Y.-L., & Cai, M.-S. (2014). Molecular cloning and characterization of pseudorabies virus EP0 gene. *Indian Journal of Biochemistry & Biophysics*, *51*(2), 100–14. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24980013

[14] Stingley, S. W., Ramirez, J. J., Aguilar, S. A., Simmen, K., Sandri-Goldin, R. M., Ghazal, P., & Wagner, E. K. (2000). Global analysis of herpes simplex virus type 1 transcription using an oligonucleotide-based DNA microarray. *Journal of Virology*, *74*(21), 9916–27. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11024119

[15] Tombácz, D., Tóth, J. S., Petrovszki, P., & Boldogkoi, Z. (2009). Whole-genome analysis of pseudorabies virus gene expression by real-time quantitative RT-PCR assay. *BMC Genomics*, *10*(1), 491. http://doi.org/10.1186/1471-2164-10-491

[16] Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., … Frey, B. J. (2010). Deciphering the splicing code. *Nature*, *465*(7294), 53–59. http://doi.org/10.1038/nature09000

[17] Boise, L. H., González-García, M., Postema, C. E., Ding, L., Lindsten, T., Turka, L. A., … Thompson, C. B. (1993). bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. *Cell*, *74*(4), 597–608. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8358789

[18] Zhang, G., & Leader, D. P. (1990). The structure of the pseudorabies virus genome at the end of the inverted repeat sequences proximal to the junction with the short unique region. *The Journal of General Virology*, *71 (Pt 10)*(10), 2433–41. http://doi.org/10.1099/0022-1317-71-10-2433

[19] Klupp, B. G., Kern, H., & Mettenleiter, T. C. (1992). The virulence-determining genomic BamHI 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*, *191*(2), 900–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1333128

[20] Cheung, A. K. (1991). Cloning of the latency gene and the early protein 0 gene of pseudorabies virus. *Journal of Virology*, *65*(10), 5260–71. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1654441

[21] Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., … Snyder, M. (2004). Global Identification of Human Transcribed Sequences with Genome Tiling Arrays. *Science*, *306*(5705), 2242–2246. http://doi.org/10.1126/science.1103388

[22] Mattick, J. S. (2004). Opinion: RNA regulation: a new genetics? *Nature Reviews Genetics*, *5*(4), 316–323. http://doi.org/10.1038/nrg1321

[23] Mattick, J. S., & Makunin, I. V. (2006). Non-coding RNA. *Human Molecular Genetics*, *15*(suppl_1), R17–R29. http://doi.org/10.1093/hmg/ddl046

[24] Wilusz, J. E., Sunwoo, H., & Spector, D. L. (2009). Long noncoding RNAs: functional surprises from the RNA world. *Genes & Development*, *23*(13), 1494–1504. http://doi.org/10.1101/gad.1800909

[25] Stroop, W. G., Rock, D. L., & Fraser, N. W. (1984). 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. *Laboratory Investigation; a Journal of Technical Methods and Pathology*, *51*(1), 27–38. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6330452

[26]Klupp, B. G., Hengartner, C. J., Mettenleiter, T. C., & Enquist, L. W. (2004). Complete, annotated sequence of the pseudorabies virus genome. *Journal of Virology*, *78*(1), 424–40. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14671123

[27] Grimm, K. S., Klupp, B. G., Granzow, H., Müller, F. M., Fuchs, W., & Mettenleiter, T. C. (2012). Analysis of viral and cellular factors influencing herpesvirus-induced nuclear envelope breakdown. *Journal of Virology*, *86*(12), 6512–21. http://doi.org/10.1128/JVI.00068-12

[28] Szpara, M. L., Tafuri, Y. R., Parsons, L., Shamim, S. R., Verstrepen, K. J., Legendre, M., & Enquist, L. W. (2011). A wide extent of inter-strain diversity in virulent and vaccine strains of alphaherpesviruses. *PLoS Pathogens*, *7*(10), e1002282. http://doi.org/10.1371/journal.ppat.1002282

[29] Yu, T., Chen, F., Ku, X., Zhu, Y., Ma, H., Li, S., & He, Q. (2016). Complete Genome Sequence of Novel Pseudorabies Virus Strain HNB Isolated in China. *Genome Announcements*, *4*(1), e01641-15. http://doi.org/10.1128/genomeA.01641-15

[30] Papageorgiou, K. V, Suárez, N. M., Wilkie, G. S., Filioussis, G., Papaioannou, N., Nauwynck, H. J., … Kritas, S. K. (2016). Genome Sequences of Two Pseudorabies Virus Strains Isolated in Greece. *Genome Announcements*, *4*(1), e01624-15. http://doi.org/10.1128/genomeA.01624-15

[31] 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 Announcements*, *2*(4), e00628-14-e00628-14. http://doi.org/10.1128/genomeA.00628-14

[32] Tirabassi, R. S., Townley, R. A., Eldridge, M. G., & Enquist, L. W. (1997). Characterization of pseudorabies virus mutants expressing carboxy-terminal truncations of gE: evidence for envelope incorporation, virulence, and neurotropism domains. *Journal of Virology*, *71*(9), 6455–64. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9261363

[33] Tirabassi, R. S., & Enquist, L. W. (1998). Role of envelope protein gE endocytosis in the pseudorabies virus life cycle. *Journal of Virology*, *72*(6), 4571–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9573220

[34] Dingwell, K. S., Brunetti, C. R., Hendricks, R. L., Tang, Q., Tang, M., Rainbow, A. J., & Johnson, D. C. (1994). Herpes simplex virus glycoproteins E and I facilitate cell-to-cell spread in vivo and across junctions of cultured cells. *Journal of Virology*, *68*(2), 834–45. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8289387

[35] Maidji, E., Tugizov, S., Jones, T., Zheng, Z., & Pereira, L. (1996). Accessory human cytomegalovirus glycoprotein US9 in the unique short component of the viral genome promotes cell-to-cell transmission of virus in polarized epithelial cells. *Journal of Virology*, *70*(12), 8402–10. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8970961

[36] Mulder, W., Pol, J., Kimman, T., Kok, G., Priem, J., & Peeters, B. (1996). Glycoprotein D-negative pseudorabies virus can spread transneuronally via direct neuron-to-neuron transmission in its natural host, the pig, but not after additional inactivation of gE or gI. *Journal of Virology*, *70*(4), 2191–200. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8642642

[37] Knapp, A. C., Husak, P. J., & Enquist, L. W. (1997). The gE and gI homologs from two alphaherpesviruses have conserved and divergent neuroinvasive properties. *Journal of Virology*, *71*(8), 5820–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9223471

[38] Johnson, D. C., Webb, M., Wisner, T. W., & Brunetti, C. (2001). Herpes simplex virus gE/gI sorts nascent virions to epithelial cell junctions, promoting virus spread. *Journal of Virology*, *75*(2), 821–33. http://doi.org/10.1128/JVI.75.2.821-833.2001

[39] L. W. Enquist, Semin. Virol. 5, 221–231 (1994)

[40] Chang, T. H., & Enquist, L. W. (2005). Neuron-to-Cell Spread of Pseudorabies Virus in a Compartmented Neuronal Culture System. *Journal of Virology*, *79*(17), 10875–10889. http://doi.org/10.1128/JVI.79.17.10875-10889.2005

[41] Babic N, Mettenleiter TC, Ugolini G, Flamand A, Coulon P. Propagation of Pseudorabies Virus in the Nervous System of the Mouse after Intranasal Inoculation. *Virology*. 1994;204(2):616-625. doi:10.1006/viro.1994.1576.

[42] Kritas, S. K., Pensaert, M. B., & Mettenleiter, T. C. (1994). Invasion and spread of single glycoprotein deleted mutants of Aujeszky's disease virus (ADV) in the trigeminal nervous pathway of pigs after intranasal inoculation. *Veterinary Microbiology*, *40*(3–4), 323–334. http://doi.org/10.1016/0378-1135(94)90120-1

[43] Kratchmarov, R., Kramer, T., Greco, T. M., Taylor, M. P., Ch'ng, T. H., Cristea, I. M., & Enquist, L. W. (2013). Glycoproteins gE and gI Are Required for Efficient KIF1A-Dependent Anterograde Axonal Transport of Alphaherpesvirus Particles in Neurons. *Journal of Virology*, *87*(17), 9431–9440. http://doi.org/10.1128/JVI.01317-13

[44] Howard, P. W., Wright, C. C., Howard, T., & Johnson, D. C. (2014). Herpes simplex virus gE/gI extracellular domains promote axonal transport and spread from neurons to epithelial cells. *Journal of Virology*, *88*(19), 11178–86. http://doi.org/10.1128/JVI.01627-14

[45] Snyder, A., Polcicova, K., & Johnson, D. C. (2008). Herpes Simplex Virus gE/gI and US9 Proteins Promote Transport of both Capsids and Virion Glycoproteins in Neuronal Axons. *Journal of Virology*, *82*(21), 10613–10624. http://doi.org/10.1128/JVI.01241-08

[46] Brack, A. R., Klupp, B. G., Granzow, H., Tirabassi, R., Enquist, L. W., & Mettenleiter, T. C. (2000). Role of the cytoplasmic tail of pseudorabies virus glycoprotein E in virion formation. *Journal of Virology*, *74*(9), 4004–16. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10756012

[47] Jacobs, L. (1994). Glycoprotein E of pseudorabies virus and homologous proteins in otheralphaherpesvirinae. *Archives of Virology*, *137*(3–4), 209–228. http://doi.org/10.1007/BF01309470

[48] Gu, Z., Dong, J., Wang, J., Hou, C., Sun, H., Yang, W., … Jiang, P. (2015). A novel inactivated gE/gI deleted pseudorabies virus (PRV) vaccine completely protects pigs from an emerged variant PRV challenge. *Virus Research*, *195*, 57–63. http://doi.org/10.1016/j.virusres.2014.09.003

[49] Wu, C.-Y., Liao, C.-M., Chi, J.-N., Chien, M.-S., & Huang, C. (2016). Growth properties and vaccine efficacy of recombinant pseudorabies virus defective in glycoprotein E and thymidine kinase genes. *Journal of Biotechnology*, *229*, 58–64. http://doi.org/10.1016/j.jbiotec.2016.05.009

[50] Ekstrand MI, Enquist LW, Pomeranz LE. The alpha-herpesviruses: molecular pathfinders in nervous system circuits. *Trends Mol Med*. 2008;14(3):134-140. doi:10.1016/j.molmed.2007.12.008.

[51] Anderson KP, Costa RH, Holland LE, Wagner EK. Characterization of herpes simplex virus type 1 RNA present in the absence of de novo protein synthesis. *J Virol*. 1980;34(1):9-27. http://www.ncbi.nlm.nih.gov/pubmed/6246265. Accessed December 4, 2017.

[52] Mackem S, Roizman B. Regulation of herpesvirus macromolecular synthesis: transcription-initiation sites and domains of alpha genes. *Proc Natl Acad Sci U S A*. 1980;77(12):7122-7126. http://www.ncbi.nlm.nih.gov/pubmed/6261240. Accessed December 4, 2017.

[53] Ihara S, Feldman L, Watanabe S, Ben-Porat T. Characterization of the immediate-early functions of pseudorabies virus. *Virology*. 1983;131(2):437-454. doi:10.1016/0042-6822(83)90510-X.

[54] Huang C, Wu C-Y. Characterization and expression of the pseudorabies virus early gene UL54. *J Virol Methods*. 2004;119(2):129-136. doi:10.1016/j.jviromet.2004.03.013.

[55] Ehrlich C, Fuchs W, Mettenleiter TC, Klupp BG. Characterization of the replication origin (OriS) and adjoining parts of the inverted repeat sequences of the pseudorabies virus genome. *J Gen Virol*. 2000;81(6):1539-1543. doi:10.1099/0022-1317-81-6-1539.

[56] 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):2433-2441. doi:10.1099/0022-1317-71-10-2433.

[57] 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):5560-5567. http://www.ncbi.nlm.nih.gov/pubmed/7637001. Accessed December 4, 2017.

[58] Huang Y-J, Chien M-S, Wu C-Y, Huang C. Mapping of functional regions conferring nuclear localization and RNA-binding activity of pseudorabies virus early protein UL54. *J Virol Methods*. 2005;130(1-2):102-107. doi:10.1016/j.jviromet.2005.06.011.

[59] Sacks WR, Greene CC, Aschman DP, Schaffer PA. Herpes simplex virus type 1 ICP27 is an essential regulatory protein. *J Virol*. 1985;55(3):796-805. http://www.ncbi.nlm.nih.gov/pubmed/2991596. Accessed December 4, 2017.

[60] Gruffat H, Batisse J, Pich D, et al. Epstein-Barr virus mRNA export factor EB2 is essential for production of infectious virus. *J Virol*. 2002;76(19):9635-9644. http://www.ncbi.nlm.nih.gov/pubmed/12208942. Accessed December 4, 2017.

[61] Sato B, Sommer M, Ito H, Arvin AM. Requirement of varicella-zoster virus immediate-early 4 protein for viral replication. *J Virol*. 2003;77(22):12369-12372. http://www.ncbi.nlm.nih.gov/pubmed/14581575. Accessed December 4, 2017.

[62] Schwartz JA, Brittle EE, Reynolds AE, Enquist LW, Silverstein SJ. UL54-Null Pseudorabies Virus Is Attenuated in Mice but Productively Infects Cells in Culture. *J Virol*. 2006;80(2):769-784. doi:10.1128/JVI.80.2.769-784.2006.

[63] Hardwicke MA, Sandri-Goldin RM. The herpes simplex virus regulatory protein ICP27 contributes to the decrease in cellular mRNA levels during infection. *J Virol*. 1994;68(8):4797-4810. http://www.ncbi.nlm.nih.gov/pubmed/8035480. Accessed December 4, 2017.

[64] McGregor F, Phelan A, Dunlop J, Clements JB. Regulation of herpes simplex virus poly (A) site usage and the action of immediate-early protein IE63 in the early-late switch. *J Virol*. 1996;70(3):1931-1940. http://www.ncbi.nlm.nih.gov/pubmed/8627719. Accessed December 4, 2017.

[65] Hayashi ML, Blankenship C, Shenk T. Human cytomegalovirus UL69 protein is required for efficient accumulation of infected cells in the G1 phase of the cell cycle. *Proc Natl Acad Sci U S A*. 2000;97(6):2692-2696. doi:10.1073/pnas.050587597.

[66] Li M, Wang S, Cai M, Guo H, Zheng C. Characterization of molecular determinants for nucleocytoplasmic shuttling of PRV UL54. *Virology*. 2011;417(2):385-393. doi:10.1016/j.virol.2011.06.004.

[67] Li M, Wang S, Cai M, Zheng C. Identification of nuclear and nucleolar localization signals of pseudorabies virus (PRV) early protein UL54 reveals that its nuclear targeting is required for efficient production of PRV. *J Virol*. 2011;85(19):10239-10251. doi:10.1128/JVI.05223-11.

[68] Looker KJ, Magaret AS, May MT, et al. Global and Regional Estimates of Prevalent and Incident Herpes Simplex Virus Type 1 Infections in 2012. DeLuca NA, ed. *PLoS One*. 2015;10(10):e0140765. doi:10.1371/journal.pone.0140765.

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

[70] Lim, F., and Filip (2013). HSV-1 as a Model for Emerging Gene Delivery Vehicles. *ISRN Virol.* 2013, 1–12. doi:10.5402/2013/397243.

[71] Macdonald, S. J., Mostafa, H. H., Morrison, L. A., and Davido, D. J. (2012). Genome sequence of herpes simplex virus 1 strain KOS. *J. Virol.* 86, 6371–2. doi:10.1128/JVI.00646-12.

[72] McGeoch, D. J., Dalrymple, M. A., Davison, A. J., Dolan, A., Frame, M. C., McNab, D., et al. (1988). The Complete DNA Sequence of the Long Unique Region in the Genome of Herpes Simplex Virus Type 1. *J. Gen. Virol.* 69, 1531–1574. doi:10.1099/0022-1317-69-7-1531.

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

[74] Du, T., Han, Z., Zhou, G., Zhou, G., and Roizman, B. (2015). Patterns of accumulation of miRNAs encoded by herpes simplex virus during productive infection, latency, and on reactivation. *Proc. Natl. Acad. Sci. U. S. A.* 112, E49-55. doi:10.1073/pnas.1422657112.

[75] Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., et al. (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science* 306, 2242–6. doi:10.1126/science.1103388.

[76] Mattick, J. S. (2004). RNA regulation: a new genetics? *Nat. Rev. Genet.* 5, 316–23. doi:10.1038/nrg1321.

[77] Mattick, J. S., and Makunin, I. V (2006). Non-coding RNA. *Hum. Mol. Genet.* 15 Spec No, R17-29. doi:10.1093/hmg/ddl046.

[78] Wilusz, J. E., Sunwoo, H., and Spector, D. L. (2009). Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.* 23, 1494–504. doi:10.1101/gad.1800909.

[79] Stroop, W. G., Rock, D. L., and Fraser, N. W. (1984). 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. Invest.* 51, 27–38. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6330452 [Accessed February 18, 2016].

[80] Stingley, S. W., Ramirez, J. J., Aguilar, S. A., Simmen, K., Sandri-Goldin, R. M., Ghazal, P., et al. (2000). Global analysis of herpes simplex virus type 1 transcription using an oligonucleotide-based DNA microarray. *J. Virol.* 74, 9916–27. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11024119 [Accessed February 4, 2016].

[81] Harkness, J. M., Kader, M., and DeLuca, N. A. (2014). Transcription of the herpes simplex virus 1 genome during productive and quiescent infection of neuronal and nonneuronal cells. *J. Virol.* 88, 6847–61. doi:10.1128/JVI.00516-14.

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

[83] Rubin, R. H. Impact of Cytomegalovirus Infection on Organ Transplant Recipients. *Clin. Infect. Dis.* **12,** S754–S766 (1990)

[84]Davison, A. J. *et al.* The human cytomegalovirus genome revisited: comparison with the chimpanzee cytomegalovirus genome. *J. Gen. Virol.* **84,** 17–28 (2003).

[85] Dolan, A. *et al.* Genetic content of wild-type human cytomegalovirus. *J. Gen. Virol.* **85,** 1301–1312 (2004).

[86] Gatherer, D. *et al.* High-resolution human cytomegalovirus transcriptome. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 19755–60 (2011).

[87] Murphy, E., Rigoutsos, I., Shibuya, T. & Shenk, T. E. Reevaluation of human cytomegalovirus coding potential. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 13585–90 (2003).

[88] Stern-Ginossar, N. *et al.* Decoding human cytomegalovirus. *Science* **338,** 1088–93 (2012).

[89] Isomura, H. *et al.* Noncanonical TATA sequence in the UL44 late promoter of human cytomegalovirus is required for the accumulation of late viral transcripts. *J. Virol.* **82,** 1638–46 (2008).

[90] Rawlinson, W. D. & Barrell, B. G. Spliced transcripts of human cytomegalovirus. *J. Virol.* **67,** 5502–13 (1993).

[91] Ma, Y. *et al.* Human CMV transcripts: an overview. *Future Microbiol.* **7,** 577–593 (2012).

[92] Sandri-Goldin, R. M. Viral regulation of mRNA export. *J. Virol.* **78,** 4389–96 (2004)

[93] Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10,** 1177–84 (2013).

[94] Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics. Proteomics Bioinformatics* **13,** 278–89 (2015).

[95] Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17,** 333–51 (2016).

[96] Cocquet, J., Chong, A., Zhang, G. & Veitia, R. A. *Reverse transcriptase template switching and false alternative transcripts*. *Genomics* **88,** (2006).

[97] Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*. 2012;13(1):238. doi:10.1186/1471-2105-13-238.

[98] 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 Announc*. 2014;2(4):e00628-14-e00628-14. doi:10.1128/genomeA.00628-14.

[99] Cock PJA, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25(11):1422-1423. doi:10.1093/bioinformatics/btp163.

[100] Bryne JC, Valen E, Tang M-HE, et al. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res*. 2008;36(Database issue):D102-6. doi:10.1093/nar/gkm955.

[101] Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. Patterns of variant polyadenylation signal usage in human genes. *Genome Res*. 2000;10(7):1001-1010. http://www.ncbi.nlm.nih.gov/pubmed/10899149. Accessed December 4, 2017.

[102] Ahmed F, Kumar M, Raghava GPS. Prediction of polyadenylation signals in human DNA sequences using nucleotide frequencies. *In Silico Biol*. 2009;9(3):135-148. http://www.ncbi.nlm.nih.gov/pubmed/19795571. Accessed December 4, 2017.

[103] Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25(9):1105-1111. doi:10.1093/bioinformatics/btp120.

[104] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357-359. doi:10.1038/nmeth.1923.

[105] Rutherford K, Parkhill J, Crook J, et al. Artemis: sequence visualization and annotation. *Bioinformatics*. 2000;16(10):944-945. http://www.ncbi.nlm.nih.gov/pubmed/11120685. Accessed December 4, 2017.

[106] Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24-26. doi:10.1038/nbt.1754.

[107] Braun A, Kaliman A, Boldogköi Z, Aszódi A, Fodor I. Sequence and expression analyses of the UL37 and UL38 genes of Aujeszky's disease virus. *Acta Vet Hung*. 2000;48(1):125-136. doi:10.1556/AVet.48.2000.1.14.

[108] De Wind N, Peeters BP, Zuderveld A, Gielkens AL, Berns AJ, Kimman TG. Mutagenesis and characterization of a 41-kilobase-pair region of the pseudorabies virus genome: transcription map, search for virulence genes, and comparison with homologs of herpes simplex virus type 1. *Virology*. 1994;200(2):784-790. http://www.ncbi.nlm.nih.gov/pubmed/8178460. Accessed December 4, 2017.

[109] Yang C, Bolotin E, Jiang T, Sladek FM, Martinez E. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene*. 2007;389(1):52-65. doi:10.1016/j.gene.2006.09.029.

[110] Boldogköi Z, Murvai J, Fodor I. G and C accumulation at silent positions of codons produces additional ORFs. *Trends Genet*. 1995;11(4):125-126. http://www.ncbi.nlm.nih.gov/pubmed/7732585. Accessed December 4, 2017.

[111] Edwalds-Gilbert G, Veraldi KL, Milcarek C. Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Res*. 1997;25(13):2547-2561. http://www.ncbi.nlm.nih.gov/pubmed/9185563. Accessed December 4, 2017.

[112] Shen Y, Ji G, Haas BJ, et al. Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation. *Nucleic Acids Res*. 2008;36(9):3150-3161. doi:10.1093/nar/gkn158.

[113] Proudfoot NJ. Ending the message: poly(A) signals then and now. *Genes Dev*. 2011;25(17):1770-1782. doi:10.1101/gad.17268411.

[114] Kobayashi Y, Zhuang J, Peltz S, Dougherty J. Identification of a cellular factor that modulates HIV-1 programmed ribosomal frameshifting. *J Biol Chem*. 2010;285(26):19776-19784. doi:10.1074/jbc.M109.085621.

[115] Paran N, Ori A, Haviv I, Shaul Y. A composite polyadenylation signal with TATA box function. *Mol Cell Biol*. 2000;20(3):834-841. http://www.ncbi.nlm.nih.gov/pubmed/10629040. Accessed December 4, 2017.

[116] Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, et al., editors. Current Protocols in Molecular Biology [Internet]. 4th ed. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 1999.

[117] Elhai J, Wolk CP. A versatile class of positive-selection vectors based on the nonviability of palindrome-containing plasmids that allows cloning into long polylinkers. Gene. 1988;68(1):119-138. http://www.ncbi.nlm.nih.gov/pubmed/2851487. Accessed December 5, 2017.

[118] Zhu, J., Kang, W., Marquart, M. E., Hill, J. M., Zheng, X., Block, T. M., et al. (1999). Identification of a Novel 0.7-kb Polyadenylated Transcript in the LAT Promoter Region of HSV-1 That Is Strain Specific and May Contribute to Virulence. Virology 265, 296–307. doi:10.1006/viro.1999.0057.

[119] Voss, J. H., and Roizman, B. (1988). Properties of two 5'-coterminal RNAs transcribed part way and across the S component origin of DNA synthesis of the herpes simplex virus 1 genome. Proc. Natl. Acad. Sci. U. S. A. 85, 8454–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2847162 [Accessed December 18, 2016].

[120] Edwalds-Gilbert, G., Veraldi, K. L., and Milcarek, C. (1997). Alternative poly(A) site selection in complex transcription units: means to an end? Nucleic Acids Res. 25, 2547–61. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9185563 [Accessed February 4, 2016].

[121] Elhai J, Wolk CP. A versatile class of positive-selection vectors based on the nonviability of palindrome-containing plasmids that allows cloning into long polylinkers. Gene. 1988;68(1):119-138. http://www.ncbi.nlm.nih.gov/pubmed/2851487. Accessed December 5, 2017.

[122] Kaye J, Browne H, Stoffel M, Minson T. The UL16 gene of human cytomegalovirus encodes a glycoprotein that is dispensable for growth in vitro. J Virol. 1992;66(11):6609-6615. http://www.ncbi.nlm.nih.gov/pubmed/1328682. Accessed December 5, 2017.

[123] Schwartz JA, Brittle EE, Reynolds AE, Enquist LW, Silverstein SJ. UL54-null pseudorabies virus is attenuated in mice but productively infects cells in culture. J Virol. 2006;80(2):769-784. doi:10.1128/JVI.80.2.769-784.2006.

[124] O'Grady T, Wang X, Höner zu Bentrup K, Baddoo M, Concha M, Flemington EK. Global transcript structure resolution of high gene density genomes through multi-platform data integration. *Nucleic Acids Res*. 2016;44(18):e145-e145. doi:10.1093/nar/gkw629.