Long-read Sequencing Uncovers a Complex Transcriptome Topology in Varicella Zoster Virus

István Prazsák[1†], Norbert Moldován[1†], Dóra Tombácz[1], Klára Megyeri[2], Attila Szűcs[1], Zsolt Csabai[1], Zsolt Boldogkői[1*]

[1]Department of Medical Biology, Faculty of Medicine, University of Szeged, Szeged, 6720, Hungary

[2]Department of Medical Microbiology and Immunobiology, Faculty of Medicine, University of Szeged, Szeged, 6720, Hungary

[†]these two authors contributed equally to this work

[*] To whom correspondence should be addressed

Tel: +36-62-545-595

Fax: +36-62-545-131

E-mail: boldogkoi.zsolt@med.u-szeged.hu

Abstract

Background

Varicella zoster virus (VZV) is a human pathogenic alphaherpesvirus harboring a relatively large DNA molecule. The VZV transcriptome has already been analyzed by microarray and short-read sequencing analyses. However, both approaches have substantial limitations when used for structural characterization of transcript isoforms, even if supplemented with primer extension or other techniques. Among others, they are inefficient in distinguishing between embedded RNA molecules, transcript isoforms, including splice and length variants, as well as between alternative polycistronic transcripts. It has been demonstrated in several studies that long-read sequencing is able to circumvent these problems.

Results

In this work, we report the analysis of VZV lytic transcriptome using the Oxford Nanopore Technologies sequencing platform. These investigations have led to the identification of 114 novel transcripts, including mRNAs, non-coding RNAs, polycistronic RNAs and complex transcripts, as well as 10 novel spliced transcripts and 27 novel transcription start site isoforms and transcription end site isoforms. A novel class of transcripts, the nroRNAs are described in this study. These transcripts are encoded by the genomic region located in close vicinity to the viral replication origin. We also show that the VZV latency transcript (VLT) exhibits a more complex structural variation than formerly believed. Additionally, we have detected RNA editing in a novel non-coding RNA molecule.

Conclusions

Our investigations disclosed a composite transcriptomic architecture of VZV, including the discovery of novel RNA molecules and transcript isoforms, as well as a complex meshwork of transcriptional read-throughs and overlaps. The results represent a substantial advance in the annotation VZV transcriptome and in understanding the molecular biology of the herpesviruses in general.

Keywords: herpesvirus, varicella zoster virus, transcriptome, Oxford Nanopore Technologies, long-read sequencing, RNA editing

Background

Varicella zoster virus (VZV) is one of the nine herpesviruses that infect humans. It is the etiological agent of chickenpox (varicella) caused by primary infection and shingles (zoster) is due to reactivation from latency, which is established in the spinal and trigeminal ganglia [1]. VZV belongs to the *Varicellovirus* genus of the subfamily *Alphaherpesvirinae*, and is closely related to the pseudorabies virus (PRV) a veterinary *Varicellovirus* and to the herpes simplex virus type-1 (HSV-1) belonging to the *Simplexvirus* genus of alphaherpesviruses. The investigation of the VZV pathomechanism is not easy due to its highly cell-associated nature and the lack of animal models, except the SCID mouse model [2, 3]. The VZV virion is ~80-120 nm in diameter and is composed of an icosahedral nucleocapsid surrounded by a tegument layer [4]. The outer virion component is an envelope derived from the host cell membrane with incorporated viral glycoproteins [5]. The VZV genome is composed of a ~125-kbp double-stranded DNA molecule containing at least 70 annotated open reading frames (ORFs). The viral DNA consists of two unique genomic regions (UL and US), and a pair of inverted repeats (IRs) surrounding the US region. Three genes (ORF62, ORF63 and ORF64) located at the IR regions are therefore in duplicate [6, 7]. VZV encodes six genes (ORFs S/L, 1, 2, 13, 32, and 57), which are not present in HSV [8]. It has been long held that VZV does not express the RNA molecule homologues to the HSV latency-associated transcripts (LAT) [9], however, a recent study reported the identification of VZV latency-associated transcript (VLT) overlapping the ORF61 gene. It has also been thought that VZV lacks the α-TIF protein encoded by the vp16 gene that activates the transcription of immediate-early (IE) genes during the initial events of the virus life cycle [10]. VZV ORF10 together with ORF4, 62 and 63 are transcriptional transactivators that are all present in the virus tegument [11, 12]. ORF10 is supposed to substitute the function of vp16 in the transactivation of ORF62 (homologue of HSV ICP4), but unlike VP16, it is unable to control the expression of ORF4, ORF63, and ORF61 [13, 14].

The viral replication is primarily controlled by the regulation of transcription, which is carried out through a sequential activation of viral genes. First, the IE genes are expressed, which is followed by the activation of the early (E) and then the late (L) kinetic classes of genes [15]. The IE protein ORF62 is the major transactivator of the viral genes, which recruits the general transcriptional apparatus of the host cell and thereby controls the expression of other viral genes [16, 17]. The VZV E genes encode the proteins required for the DNA replication, while L genes code for the structural elements of the virus. Polycistronism represents an inbuilt characteristic of the herpesvirus genome [18,19]. However, the herpesvirus multigenic transcripts differ from the bacterial polycistronic transcripts in that the downstream genes on the viral RNA molecules are untranslated. An exception to this rule was found in the Kaposi's sarcoma-associated herpesvirus [19, 20]. The functional significance of this organization of the herpesviral transcripts has not yet been ascertained. Alternative splicing expands the RNA and protein repertoire with respect of the one gene, one RNA/protein situation. In contrast to the beta and gammaherpesviruses, splicing events are rare among the alphaherpesviruses [21, 22]. In VZV only a few genes have been shown to produce spliced mRNAs so far, which include the ORF0 (also referred as ORFS/L) located at the termini of UL region, the UL15 homologue ORF42/45, the glycoprotein M (gM) encoding ORF50, and the newly discovered, multiple spliced VLT [6,9,23,74,75].

RNA editing includes the adenosine deaminase acting on RNA1 (ADAR1) enzyme that catalyzes the C-6 deamination of adenine (A) to inosine (I). The I is recognized as guanine (G) by the reverse transcriptase (RT) enzyme [26], therefore it can be detected with cDNA sequencing methods. In hyper-edited sites the ADAR1 transforms multiple adjacent As on an RNA strand. A G-enriched neighborhood has been described at the RNA editing sites [27] with an upstream uracil (U) exerting a stabilizing effect on the RNA-ADAR complex [28]. Additionally, RNA hyper-editing has been shown to play a crucial role in the viral life cycle by dodging the host's immune system [29] and also in the control of DNA replication [30].

The translation of eukaryotic mRNAs occurs according to the scanning model, where the 40S ribosomal subunits scans the RNA strand in 5' to 3' direction and initiates the translation at the first AUG they encounter [31]. Mammalian mRNAs contain an essential sequence context for translation initiation, known as the Kozak consensus sequence [32]. A purine at -3 and G at +4 position has the strongest binding effect for the translational machinery, while AUGs with a different context tend to be overlooked by the ribosomal subunits (leaky scanning) [33]. Upstream ORFs in the 5'-untranslated regions (UTRs) of the RNAs have been shown to exert a regulatory effect on the protein synthesis through a process called translation reinitiation [31, 34]. In short upstream (u)ORFs translation reinitiation shows a positive correlation with distance between stop codon of uORF and the AUG of downstream ORF [35].

The hybridization-based microarray and the second-generation short-read sequencing (SRS) techniques have revolutionized genome and transcriptome research, including herpesviruses [36–39]. However, both techniques have limitations compared to long-read sequencing (LRS), including Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) platforms. Microarray and SRS techniques perform poorly in the detection of multiple-intron transcripts, transcript length variants, and polycistronic RNA molecules, as well as transcriptional overlaps. Additionally, the LRS sequencing platforms are capable of determining the transcript ends with base-pair precision, without the need of any supplementary method.

Similarly to other LRS techniques, ONT cDNA sequencing is afflicted by sample degradation, resulting in false transcriptional start sites (TSSs). This problem can be mitigated by using cap selection. Another source of non-specificity is the presence of nucleotide sequences complementary to the MinION strand switching oligonucleotide, which can lead to either template switching [40, 41] or false priming [42]. False priming with the oligo(dT) primer can also occur at A-rich regions of the transcripts. These errors contribute to artifactual 3'-ends of the reads.

LRS techniques have already been used in genome and transcriptome studies of viruses [43–46], including herpesviruses [18, 21, 22, 47–49]. These reports have revealed a hidden transcriptome complexity, which especially included the discovery of long RNA molecules and a large variation of transcript isoforms. ONT and PacBio-based studies have also detected a number of embedded transcripts with in-frame truncated (t)ORFs in several herpesviruses, such as PRV [18, 43], HSV-1 [50], and human cytomegalovirus (HCMV) [21]. Additionally, a large extent of transcriptional read-through and overlaps has also been uncovered by these studies [18, 21, 22]. It has been proposed that the transcriptional overlaps may be the result of a transcriptional interference mechanism playing a role in genome-wide regulation of gene expressions [51].

In this work, we used the ONT MinION sequencing technique to investigate the structural aspects of the polyadenylated VZV transcripts. We report numerous novel transcripts, transcript isoforms, and yet unknown splice events. This study also explores a complex meshwork of transcriptional overlaps. Additionally, we report and analyze a hyper-editing event in a VZV transcript.

Methods

Cells and viral infection

Human primary embryonic lung fibroblast cell line (MRC5) was obtained from the American Type Culture Collection (ATCC) and grown in DMEM supplemented with antibiotic/antimycotic solution and 10% fetal bovine serum (FBS) at 37°C in a 5% CO_2 atmosphere. The live attenuated OKA/Merck strain of varicella-zoster virus (VZV) was cultured at 37°C in MRC5 cell line, the cells were harvested by trypsinization, when the monolayers had displayed specific cytopathic changes. For subsequent propagation of the virus, infected cells were used to inoculate MRC5 cultures previously grown to full confluence at a ratio of 1:10 infected to uninfected cells. The cultures were then incubated at 37 °C for 5 days, when the cytopathic effect was near 100%.

RNA purification

Total RNA was isolated immediately after collecting the infected cells, using the Nucleospin RNA Kit (Macherey-Nagel) according to the manufacturer's instruction. Briefly, infected cells were collected by centrifugation and the cell membrane was disrupted using the lysis buffer provided in the kit. Genomic DNA was digested with the RNase-free rDNase solution (supplied with the kit). Samples were eluted in a total volume of 50µl nuclease-free water. To eliminate residual DNA contamination a subsequent treatment with the TURBO DNA-free Kit (Thermo Fisher Scientific) was conducted. The RNA concentration was measured with a Qubit 2.0 Fluorometer, with the Qubit RNA BR Assay Kit (Thermo Fisher Scientific).

Poly(A) selection

Thirty-five µg of total RNA was pipetted in separate DNA LoBind Eppendorf tubes (Merck). The poly(A)$^+$ RNA fraction was isolated from the samples using the Oligotex mRNA Mini Kit (Qiagen). RNA samples were stored at -80°C until use.

Oxford Nanopore MinION cDNA sequencing and barcoding

The cDNA library was prepared using the Ligation Sequencing kit (SQK-LSK108; Oxford Nanopore Technologies) according to the modified 1D strand switching cDNA by ligation protocol. In short, the polyA(+) RNA fraction was reverse transcribed using an oligo(d)T-containing primer [(VN)T20 (Bio Basic, Canada). The RT reaction was carried out as recommended by the 1D protocol, using SuperScript IV enzyme (Life Technologies); a strand-switching oligo [containing three O-methyl-guanine RNA bases (PCR_Sw_mod_3G; Bio Basic, Canada)] was added to the sample. The cDNA was amplified by using KAPA HiFi DNA Polymerase (Kapa Biosystems) and Ligation Sequencing Kit Primer Mix (part of the 1D Kit) following the ONT 1D Kit's manual. End repair was carried out on the samples using NEBNext End repair / dA-tailing Module (New England Biolabs), which was followed by "barcoding": the C11 barcode (ONT PCR Barcoding Kit 96; EXP-PBC096) was ligated to the sample according to the 1D PCR barcoding (96) genomic DNA (SQK-LSK108) protocol, Barcode Adapter ligation step. The barcoded- sample was amplified by PCR using KAPA HiFi DNA Polymerase, as well as the C11 PCR barcode according to the 1D PCR barcoding protocol. The PCR product was end-repaired, then it was followed by adapter ligation using the sequencing adapters supplied in the kit and NEB Blunt/TA Ligase Master Mix (New England Biolabs). The cDNA sample was purified between each step using Agencourt AMPure XP magnetic beads (Beckman Coulter). The concentration of cDNA library was determined using a Qubit 2.0 Fluorometer through use of the Qubit (ds)DNA HS Assay Kit (Thermo Fisher Scientific). Samples were loaded on R9.4 SpotON Flow Cells, while the base calling was performed using Albacore v2.1.10.

Cap selection, cDNA synthesis and sequencing

For the precise determination of TSSs, the ONT's 1D strand switching cDNA by ligation protocol was combined with a 5'-cap specific protocol. The cDNA sample was prepared from total RNA using the TeloPrime Full-Length cDNA Amplification Kit (Lexogen). Reverse transcription (RT) reaction was performed at 46°C for 50min according to the kit's recommendations by using an oligo(d)T-containing primer (5' -> 3': TCTCAGGCGTTTTTTTTTTTTTTTT). The RT product was purified by using spin column-based method (silica columns are from the Lexogen kit). A double-strand specific ligase enzyme (Lexogen Kit) was used to ligate an adapter to the 5'C of the cap of the RNA. The reaction was carried out at 25°C, overnight. The sample was washed by applying the silica-column method. The Second-Strand Mix and the Enzyme Mix (both from the TeloPrime Kit) as well as a Verity Thermal Cycler (Applied Biosystems) were used to produce double-stranded cDNAs according to the Kit's guide. In order to generate sufficient amount of cDNAs for MinION library preparation, samples were amplified by endpoint PCR following the TeloPrime Kit's manual. The generation of the sequencing-ready library from this sample is based on the 1D strand switching cDNA by ligation protocol from ONT. The Ligation Sequencing kit (SQK-LSK108, ONT) and NEBNext End repair / dA-tailing Module (New England Biolabs) was used to repair the cDNA ends. This step was followed by the 1D adapter

ligation, which was carried out according to the 1D protocol, using the NEB Blunt/TA Ligase Master Mix (New England Biolabs). We used barcoding for the better *in silico* identification of the transcripts' ends. We found this method useful because it helped the base-pair precision mapping of TSSs. The library was sequenced on an ONT R9.4 SpotON Flow Cells.

Data analysis and alignment

Reads resulting from ONT sequencing were aligned to the reference genome of VZV (GeneBank accession: NC_001348.1) and the host cell genome (Homo sapiens - GRCh37, BioProject number: PRJNA31257) using GMAP v2017-04-24[53]. In order to annotate the 5' ends, the last 16nt of the MinION 5' strand switching adapter or the last 16nt of the cap selection adapter was aligned in a window of -10 nt and +30 nt from the first mapped position of a read using the Smith-Waterman algorithm, with a match cost of +2, a mismatch cost of -3, a gap opening cost of -3, and a gap extension cost of -2. Read ends with a score below 14 were considered false 5' ends and were discarded. A 5'-end position was considered a TSS if the number of reads starting at this position was significantly higher than at other nucleotides in the region surrounding this start position. For this the Poisson-probability (Poisson[k_0;λ]) of k_0 read starting at a given nucleotide in the -50 nt to +50 nt window from each local maximum was calculated with $\lambda = \frac{\sum_{i=-50}^{50} k_i}{101}$. The 5'-ends of the low-abundant long reads were inspected individually using the Integrative Genome Viewer (IGV) [54]. Poly(A) tails were defined using the same algorithm and parameters, by mapping 15 nt of homopolymer As or Ts to the soft-clipped region on the read's end. Read ends with a score below 14 or with more than 5 As/Ts directly upstream their position were considered artefacts of false priming, and were discarded. Transcription end sites (TESs) were defined using the same criteria for the Poisson-probability as in case of TSSs. Reads passing both criteria and present in more than three copies were considered as "certain transcript isoforms". Those with less than three copies or the longest unique reads with questionable 5' ends were considered "uncertain transcript isoforms". Reads with sequencing adapters or poly(A) tails on their both ends were discarded, except for complex transcripts, which were individually inspected using IGV. Reads with a larger than 10nt difference in their 5' or 3' ends were considered novel length isoforms (L: longer 5' UTR, S: shorter 5' UTR, AT: alternative 3' termination). Short length isoforms harboring a truncated version of the known open reading frame (ORF) were considered novel putative protein coding transcripts, and designated as '.5'. If multiple putative protein coding transcripts were present, then the one with the longest 5' UTR was designated '.5' and its shorter versions were labelled in an ascending order. Transcripts with TESs located within the ORFs of genes (therefore lacking STOP codons) or with TSSs within the coding regions without in-frame ORFs were both considered non-coding transcripts.

Multigenic transcripts containing at least two genes standing in opposite were named complex transcripts. If a TSS was not obvious at these transcripts, we assumed that they start at the closest upstream annotated TSSs. Splice junctions were accepted if the intron boundary consensus sequences (GT and AG) were present in at least ten sequencing reads and if the frequency of introns was more than 1% at the given region.

To assess the homology of protein products possibly translated from the ORFs of novel transcript isoforms we used the online BLASTP suit[55], with an expected threshold of 10.

To evaluate the effect of hyper-editing on the secondary structure of RNAs, we used the RNAstructure Software suite [56] with the following parameters: temperature=310,15 K, Max % Energy Difference=10, Max Number of Structures=20, Window Size=0. To simulate the presence of inosine, we changed the edited adenines to guanine. To assess the secondary structure of the ORF62-AS2-ORF62 dsRNA hybrid we used the first 467 bases of the ORF62 labeled ORF62-5' (from genomic position 109,204 to 108,738) as one of our starting sequences, and the whole sequence of ORF62-AS2

as the other one. Reads were visualized using the Geneious [57] software suite and IGV. GC-, CCAAT- and TATA-boxes were annotated using the Smith-Waterman algorithm for canonical motifs [58].

Results

Analysis of the VZV transcriptome with third-generation sequencing

In this work, the ONT MinION RNA-Seq technique was used for profiling the genome-wide expression of the lytic VZV transcripts (**Figure 1**). Both cap-selected and non-cap-selected protocols were applied for the analyses. Sequencing of the non-cap-selected sample yielded 511,886 reads of which 57,888 mapped to the VZV reference genome (GeneBank Accession: NC_001348.1) with an average mapped read length of 1,349 bp, and an average coverage of 625, while 453,998 reads mapped to the host cell genome (Homo sapiens - GRCh37, BioProject number: PRJNA31257). Sequencing of the cap-selected samples yielded 827,608 reads, of which 509,531 mapped to the viral genome, with an average coverage of 1,200, and 318,077 reads mapping to the host cell. This latter technique performed poorly in VZV, which was indicated by the short average aligned read length (294 bp) (**Figure 2 panel a.**). Intriguingly, we obtained the same poor result with PRV [22] and HSV-1 (our unpublished results), however, the cap-selection technique performed very well in the analysis of the transcriptomes of HCMV, the baculovirus *Autographa californica* multiple nucleopolyhedrosis virus (AcMNPV) and vaccinia virus (our unpublished results). These results together with an analysis of the nucleotide distribution at the vicinity of the 5'-ends in the cap-selected datasets (**Figure 2 panel b.**) demonstrate that the read length of cap-selected RNA-s is not determined by the GC-content (PRV: 73%, HSV-1: 68%, VZV: 46%), but by a yet unknown factor, which is common in alphaherpesviruses.

In total, we detected 1,124 5'-ends and 255 3'-ends in the non-cap-selected dataset, and 1,472 5'-ends and 279 3'-ends in the cap-selected dataset. The number of 5'-ends qualifying as a TSS was 10.86% of the total 5'-end positions, while 32.95% of the total 3'-end positions turned out as TES (Table 1). We excluded 49 5'-ends and 16 3'-ends from the non-cap-selected and 22 5'-ends and 16 3'-ends from the cap-selected datasets because they proved to be the products of false priming.

Table 1 The number of read ends following each step of filtering.

	Non-cap-selected		Cap-selected	
Filter	**5' ends**	**3' ends**	**5' ends**	**3' ends**
None	1,124	255	1,472	279
Peak analysis and Poison probability	151	92	246	116
Template switching/false priming	102	76	180	100

Earlier results with primer extension, S1 nuclease assay and Illumina sequencing determined the transcript ends of some of the RNA molecules of VZV. Using the ONT sequencing platform, we confirmed 18 previously known TSSs and nine TESs. Additionally, we annotated 124 new TSSs and 71 new TESs listed in **Supplementary Table 1**. The sequences upstream of the TSSs and TESs were analyzed *in silico* to detect putative GC-, CCAAT- and TATA-box motifs, and Poly(A) signals (PASs), by motif searching (**Table 2**).

Table 2 Putative promoter sequences and Poly(A) signals (PASs) of the VZV transcripts.

	Search range (nt)	Nr. of motifs	Nr. of transcripts	Mean Distance	SD
GC-box	-150 to 0	41	55	93,28	33
CCAAT-box	-150 to -40	19	27	93,91	23
TATA-box	-35 to -25	26	44	31,67	1,5

PAS	-50 to 0	61	169	16,11	8,86

The search range position 0 is the position of the TSS in GC-, CCAAT- and TATA-boxes and is the position of the TES for PAS. The distance was calculated between the position of the TSS/TES and the last nucleotide of the motif.

In this work, we detected an enrichment of As and Us in an interval of up to -10 nt upstream of the TES, and an abundant GT-rich region downstream in the +10 nt interval (**Figure 3.**). We annotated altogether 246 novel transcripts of which 149 were confirmed by at least three reads, listed in **Supplementary Table 2** and **Supplementary Table 1**. Additionally, we mapped the TSS and TES of 37 previously detected transcripts lacking former mapping of RNA end positions (**Supplementary Table 1**).

Putative mRNAs

Transcripts embedded into a larger RNA molecules are easily detected by the LRS techniques. If such a transcript contains an in-frame ORF shorter [truncated (t)ORF] than that if the host ORF, it can be considered as a putative mRNA potentially coding for an N-terminally abridged polypeptide. Earlier *in silico* approaches have annotated the VZV ORFs [6, 7] one of which (the orf33.5) is a tORF. In this study, we report the identification of 28 embedded transcripts containing tORFs. We could identify promoters for 12 of these transcripts, located at a mean distance of 93.85±32.18 nt (mean ± SD) from their TSSs (**Supplementary Table 3**). Three of these transcripts (ORF13.5, ORF54.5-53 and ORF62.3) contain strong Kozak consensus sequences near their AUGs, while 19 have at least one of the two essential nucleotides present on their -3 or +4 positions. Twenty-five of the possibly protein coding transcripts contain a PAS at a 19.44±8.91 (mean ± SD) distance upstream their TESs (**Supplementary Table 3**).

Determination of the ends of mRNAs with annotated ORFs

In this work, we determined the TSSs and TESs of 25 mRNAs whose ORFs had earlier been described (**Figure 1 Supplementary Table 1).** The TSS and TES of these monocistronic mRNAs were not characterized before by any other means. Twelve of these transcripts have a promoter sequence 51.57±33.31 (mean ± SD) upstream their TSSs, and 19 have a canonical PAS upstream their TESs at a mean distance of 14.96±11.59 (mean ± SD).

Non-coding transcripts

The ncRNAs are transcripts without the ability to encode proteins. They can overlap the coding sequences of the genes in the same orientation, or in the opposite orientation [antisense (as)RNAs], or they can be located at the intergenic regions. The 5'-truncated (TR) transcripts have their own promoters but share the PASs and TESs with the 'host' transcript, while the 3'-truncated (NC) RNAs are controlled by the same promoters as the canonical transcript in which they are embedded but have no in-frame ORFs. Twenty-three of the novel non-coding transcripts are long non-coding (lnc)RNAs exceeding 200 bps in length per definition, while five of them are small non-coding (snc)RNAs with a size below 200 bps.

In this work, we identified 20 embedded ncRNAs, of which 8 were 5'-truncated, while 12 were 3'-truncated transcripts (**Supplementary Table 1**). We also detected one intergenic and seven antisense ncRNAs, all of them being controlled by their own promoters. In total, 17 of the novel ncRNAs contain a canonical promoter sequence 60.4±35.19 nt (mean ± SD) upstream their TSSs, while 21 have a PAS at a 16.35±8.87 (mean ± SD) distance of their TESs. ORF42-43-AS and ORF35-AS overlap multiple mRNAs. ORF42-43-AS stands in antisense orientation with respect of ORF42/ORF45-SP-1 and ORF42/ORF45-SP-2, while in tail-to head orientation with ORF43 and ORF43-44. ORF35-AS stands in antisense orientation with the polycistronic transcript ORF35-34-33 and in tail-to-head orientation with ORF36 and ORF36-S and ORF36-37. (**Figure 1**). The LAT RNA has been described in every member of the alphaherpesvirus subfamily [59, 60], including VZV [9]. We confirmed the existence of four

previously detected lytic isoforms of VLT (VLT_{ly}) of which two are TSS isoforms, one is a splice variant and one is both a TSS isoform and a splice variant (**Figure 4, Supplementary Table 2**).

5'- and 3'-UTR isoforms

The 5'-UTR isoforms (TSS variants) start upstream or downstream of the TSS of the earlier annotated transcripts, and their expression is regulated by their own promoters, while 3'-UTR isoforms (TES variants) contain distinct PAS and their polyadenylation occurs upstream or downstream of the TES of the main transcript. In this report, we detected 19 novel 5'-UTR length variants, 7 being shorter and 12 being longer than the earlier annotated transcript isoforms. We found canonical promoter sequences in 10 of the 5'-UTR isoforms at a distance of 94.93±38.58 (mean ± SD). Additionally, we detected 8 3'-UTR length isoform, all with a canonical PAS 22±7.65 (mean ± SD) upstream their TESs. An intriguing finding is the putative TSS at the genome position +4 belonging to the rare abundance transcript ORF0-1-L-C and ORF0-1-2-L-C at the extreme termini of UL region (TR_L), which suggests the existence of a promoter located on the other terminal repeat (TR_S) of the genome. This putative promoter is supposed to be active in the circular genome.

Splice sites and splice isoforms

Reverse transcription can produce false introns between repetitive sequences of the template RNA due to the phenomenon of template switching. In order to exclude these artefacts, we removed sequencing reads with low abundance (≤ 1%) and with a repeat of more than six nucleotides next to one of the splice junctions. From the initial set of 10,064 unique splice acceptor and donor site candidates 24 matched our criteria resulting in a total of 16 splice isoforms being above the 1% intron depth level of acceptance. We detected two novel splice variants encoded by the ORF42/45 gene. Furthermore, we detected twelve novel splice sites and confirmed the existence of nine previously described spliced transcripts, all with a consensus GT at the splice donor site and AG at the splice acceptor site.

ORF63, homologue to the HSV-1 and the PRV us1 gene, is one of the main regulators of VZV transcription. In this work, we discovered ten novel splice variants of ORF63. Similarly to the HSV-1 US1 mRNA, NTO1v1 and the NTO1v3 harbors one intron in its 5' UTR, while the NTO1v4 is spliced twice just as the PRV US1 mRNA [47, 50] (**Figure 5**). Ten splice junction of ORF63 splice variants coincide with those of the VLT_{ly} splice variants [9], thus they were labeled as VLT-ORF63-C (**Figure 6**). One of the splice donor sites, present in both NTO1v1 and NTO1v2 is GC, which is different from the canonical splice donor sequence. We detected all of the three ORF50 splice isoforms previously described, but ORF50C occurs in relatively low abundance, below our acceptance threshold. We detected another splice variant in low abundance in the ORF50 cluster, and named it ORF50D (Table 3, Figure 7).

Table 3 Splice junction sites of the VZV transcriptome.

Intron Start	Intron End	Strand	Intron Length	Donor Site	Acceptor Site	Acceptor no	Intron depth	Transcript name	References
585	714	+	129	GT	AG		45,62%	ORF0-1-C-SP	[61]
5004	4155	-	849	GT	AG	□	2,90%	ORF5-4-SP	
16712	16886	+	174	GT	AG	□	5,04%	ORF12-13-SP	
43826	43505	-	321	GT	AG	□	4,11%	ORF24-SP	
43950	43863	-	87	GT	AG	□	1,40%	ORF24-SP-2	
61790	63564	+	1774	GT	AG	□	2,40%	-	
77870	78148	+	278	GT	AG	□	11,29%	ORF42-43-AS	
78938	78039	-	899	GT	AG	□	38,88%	ORF42/ORF45-SP-2	
81537	78039	-	3498	GT	AG		100,00%	ORF42/45-SP-1	[6]

81537	79053	-	2484	GT	AG	?	12,72%	ORF42/ORF45-SP-2	
87759	86782	-	977	GT	AG		4,15%	ORF50B	[23]
87759	86982	-	777	GT	AG		2,33%	ORF50A	[23]
87436	86782	-	654	GT	AG		0,62%	ORF50C	[23]
86982	87436	-	454	GT	AG		0,33%	ORF50D	
10172[illegible]	102420	+	692	GT	AG		87,50%	VLT_{ly}	[9]
10248[illegible]	102853	+	369	GT	AG		100,00%	VLT_{ly}	[9]
10298[illegible]	103827	+	844	GT	AG		2,42%	VLT_{ly}	[9]
10392[illegible]	104293	+	369	GT	AG		3,99%	VLT_{ly}	[9]
10443[illegible]	104768	+	335	GT	AG		5,22%	VLT_{ly}	[9]
10482[illegible]	110509	+	5685	GT	AG	?	2,18%	ORF63-SP2-C	
10883[illegible]	110509	+	1679	*GC*	AG	?	11,12%	NTO1v1 (ORF63-SP 1)	
11492[illegible]	115050	+	130	GT	AG	?	1,04%	ORF67-SP	
12106[illegible]	119388	-	1679	GC	AG	?	10,65%	NTO1v1 (ORF63-SP 1)	

The splice junctions listed in this table passed our criteria of intron detection, with the exception of ORF50C, which is listed as a confirmation, and ORF50D, which is a novel combination of previously known splice sites. Other rare splice variants below the 1% spliced read limit are listed in Supplementary table 2. The TSS and TES belonging to the transcript with splice sites in row 7 could not be determined with certainty.

In four splice isoforms splicing effects the length of the ORFs, all producing N-terminally truncated hypothetical isoforms of previously detected proteins (Table 4).

Table 4 The proteins and hypothetical proteins produced by spliced transcripts.

Transcript	Host protein length (aa)	Spliced protein length (aa)	Intron position	Splice donor	Splice acceptor	Previously detected
ORF5-4-SP	ORF4: 452	452	5' UTR	4141	2783	
ORF12-13-SP	ORF12: 661	233	in frame, 5'truncated	16214	17088	[62]
ORF24-SP	ORF24: 269	137	in frame	44021	43286	[62]
ORF42/45-SP-2	ORF42/45: 747	380	in frame, 5'truncated	82593	78969	
ORF50D	ORF50: 435	-	-	-	-	
NTO1v1 (ORF63-SP-1)	ORF63: 278	278	5' UTR	110581	111417	
ORF63-SP-2-C	ORF63: 278	278	5' UTR	110581	111417	
ORF67-SP	354	311	in frame, 5'truncated	114496	115559	[62]

Near-replication-origin (nro)RNAs – a novel class of transcripts

In our earlier work, we reported [63] that PRV expresses transcripts located near the replication origins (Oris): the CTO family (including CTO-S, CTO-S-AT, CTO-M and CTO-L) at the OriL and the PTOs (PTO and PTO-US1) [18] at the OriS. Expression of nroRNAs has also been described in HSV-1 (OriS-RNS: [64]) and HCMV (OriLyt: [65]; RNA4.9: [66]). The VZV genome contains exclusively OriS, and lacks the replication origin at the UL segment. We identified 9 nroRNAs starting in the proximity of VZV OriS. The NTO1v1, NTO1v3 and NTO1v5 are the spliced long TSS variants of ORF63 while the

NTO1v2, NTO1v4 and NTO1v6 are the spliced long TSS variant of ORF64. NTO2 starts at the same TSS as NTO1v1 but terminates 30 bp downstream of its splice donor site. The NTO3 and NTO4 are positioned downstream of NTO2. A similar transcriptional arrangement can be observed in PRV, where the TSS of the PTO-US1 (positionally similar to the NTO1v1) is located at the same genetic locus as PTO (positionally similar to the NTO2, NTO3 and NTO4, but not homologous) (**Figure 5 panels a and c**). Based on this result, we can distinguish four types of nroRNAs: (1) ncRNAs that do not overlap the Ori (such as CTO-S, CTO-S-AT, PTO, as well as NTO2, NTO3 and NTO4); (2) ncRNAs that do overlap the Ori (such as CTO-M), (3) mRNA isoforms with very long alternative TES (such as CTO-L and CTO-L2); and (4) mRNA isoforms with very long TSS variant [such as PTO-US1, US1-L (PRV), OriS-RNA2 (HSV-1), and the now discovered NTO1 isoforms] (**Figure 7**).

Polycistronic and complex transcripts

A major issue of SRS, as wells as microarray and quantitative PCR approaches is that they have severe limitations in distinguishing between mono- and polycistronic transcripts. In contrast, LRS sequencing is suitable for making this distinction, and it is particularly superior in the detection of low abundance multi-genic transcripts. In this work, we identified 33 novel multigenic RNAs, including 29 polycistronic and 4 complex transcripts. Complex transcripts are multigenic RNAs that contain one or more genes in opposite orientations. Antisense sequences on the complex transcripts are unable to encode proteins. We also detected ten complex transcripts in low abundance in the region of VLT, seven of which are co-terminal with ORF63 and three with ORF64. These transcripts overlap with several oppositely oriented coding sequences and are spliced in a similar manner as the VLT_{ly} isoforms (**Figure 6, Supplementary Table 2**). *In silico* analysis detected an in-frame ORF incorporating the coding sequence of ORF63 (it has an upstream AUG). This results in VLT_{ly}-ORF63-C1, VLT_{ly}-ORF63-C4, VLT_{ly}-ORF63-C5, VLT_{ly}-ORF63-C6 and VLT_{ly}-ORF63-64-C1 encoding hypothetical proteins whose N-terminal is longer with 88 amino acids (aa) than the one encoded by the orf63 gene, while the VLT_{ly}-ORF63-C2; VLT_{ly}-ORF63-C3 and the VLT_{ly}-ORF63-64-C2 encoding hypothetical proteins with 179 aa longer than those coded by orf63 (**Figure 6**), The N-terminal overhangs of the putative proteins show no homology with any other known proteins in online databases.

Transcriptional overlaps

Transcripts can overlap each other in parallel (tandem or head-to-tail), convergent (tail-to-tail) or divergent (head-to-head) manner (**Figure 8**). RNAs identified and annotated with a certain TSS and TES in this work form a total of 486 overlaps, of which 16 are head-to-head, 454 are head-to-tail and 16 are tail-to-tail (**Supplementary Table 4**). The overlaps can be full or partial. Full overlaps can be formed between the RNA molecules encoded by polycistronic transcription units, between embedded and host mRNA molecules, between mRNAs and 3' as well as 5' truncated transcripts, between the TSS and TES isoforms, etc. Partial overlaps can be formed between every transcript type. An overlap is 'hard' if two genes can only produce overlapping transcripts, or 'soft' if both overlapping and non-overlapping transcripts are formed. The soft overlap can be the result of alternative promoter usage (TSS isoforms) or transcriptional readthrough (TES isoforms). An example for the latter case is the ORF17 and ORF18, which produce non-overlapping transcripts, but the TES isoform of ORF18 (ORF18-AT), with its additional 89 bps in the 3'UTR, overlaps with ORF17 in a tail-to-tail manner (**Figure** ### panel c.).

Upstream ORFs

Using *in silico* methods, we detected 44 potential (u)ORFs on the 5'-UTRs of 81 VZV transcripts **(Figure 1)**. Five of the longer TSS variants contain uORFs, while their shorter isoform does not. We have previously described this phenomenon in HCMV [49] and PRV [22]. The average size of an uORF was 54.9 ± 44.64 nt (mean ± SD, median=35 nt), while the average distance of the uORF stop codon from the protein coding ORF's AUG was 174.84 ± 143.58 nt (mean ± SD, median=110 nt). This space between the two uORF and the canonical ORF is enough for a potential reinitiation event. We identified a Kozak consensus sequences in five uORFs.

RNA editing

The sequencing reads of NTO3 transcript show a very high frequency of A to G substitution, which is not present in the overlapping reads of other transcripts. We found that 58% of all substitutions are A->G (**Figure** 9 panel b) in the reads of NTO3, which is significantly higher than the 12.98% in the overlapping transcripts in the same region ($p<0.0001$, Fisher's exact test) (**Figure 9 panel a and c**), making 22.07% of all As of the transcript edited. This suggest a hyper-editing event in NTO3. Using the MEME software suite [67] we could detect a slight enrichment of the GU dinucleotide preceding the editing site resulting in a GUAG motif (the editing site is underlined) (**Figure 8 panel d**). Using the RNAstructure software suite [56], we predicted the secondary structure with the lowest free energy of the NTO3 ssRNA (**Supplementary Figure 1 a and b**) and of the NTO3-ORF62-5' dsRNA hybrid, using both the unedited and edited forms of the asRNA (**Supplementary Figure 1 c and d**). When forming an intramolecular secondary structure, the unedited form has a higher free energy state than the edited form (-143.2 kcal/mol compared to -169.4 kcal/mol), which suggests that hyper-editing confers thermodynamic stability to the secondary structure of the RNA. Twelve of the 17 Is may aid in the formation of stem structures, while in the unedited RNA only four of the As in the same position have a complementing nucleotide. Contrarily, the secondary structure of the dsRNA formed by the edited NTO3 and the ORF62 is slightly less stable than that formed by the unedited form (-818.7 kcal/mol compared to -822.9 kcal/mol), but allows the formation of identical secondary structures.

Discussion

Until now, the VZV transcriptome have been analyzed by Northern blot, primer extension, microarray and Illumina sequencing [17, 68–75]. These techniques have generated useful data, but they have limitations to provide a comprehensive list of the VZV transcripts. In this work, we used the ONT MinION LRS technique for the investigation of the poly(A)+ fraction of the VZV transcriptome. Our results identified altogether 151 novel transcripts, including novel mRNA molecules, monocistronic transcripts, transcript isoforms, as well as multigenic transcripts. Novel splice sites and splice variants were also detected. In this work, we discovered 5 sncRNAs and 23 lncRNAs and annotated their TSSs and TESs with base pair precision.

The enrichment of As and Us upstream of PASs in mammalian systems is well-established [76–78]. These homopolymer A stretches may also cause false priming and template switching events during reverse transcription, which results in false TESs. In our work, we excluded these artefacts by the use of our analysis pipeline, and demonstrated the verity of the annotated TESs by the presence of the canonical GU-rich region in the +10 interval downstream of TESs [79].

Similar to other herpesviruses [21, 50], this study also revealed a complex meshwork of transcriptional read-throughs and overlaps. It has been earlier known that the polycistronic transcription units include co-terminal multigenic RNA molecules, which represent a large extent of parallel overlaps along the entire viral genome. According to our current knowledge, the downstream genes on these long transcripts are untranslated. We can raise the question as to whether the untranslated downstream sequences have any function or if they are mere random read-through products representing transcriptional noise. We can address the same question for the convergently and divergently overlapping RNA segments, and also for the alternative transcriptional overlaps. We have put forward a hypothesis that explains the potential role of this phenomenon which is based on transcriptional interaction between the RNA polymerase molecules at the overlapping regions. Since essentially every gene produces overlapping transcripts and therefore the pairwise interactions can spread along the genome thereby forming a transcriptional interference network (TIN), which alongside the promoter-transcription factor system determines the global gene expression pattern of the viral DNA [51]. It has been previously shown that orf63 has no trans-regulatory effect on the expression of orf62 [80], however deletion of orf63 increases the expression of orf62 [81], suggesting that there is a link between

the regulation of the two genes. A possible explanation could be the transcriptional interference caused by the head-to-head overlaps between ORF62 and NTO1v1 and NTO1v2.

Polycistronic (especially bicistronic) transcripts are also common in eukaryotic organisms, but their generation is explained by trans-splicing and not by transcriptional read-through [82]. However, unprocessed transcripts are difficult to detect because of their short existence, therefore, it cannot be excluded that these transcripts are the result of a transcription read-through mechanism. The predominant occurrence of adjacent genes in the chimeric transcripts suggests that transcriptional readthrough followed by cis-splicing may be the case, that is, the existence of a large extent of transcriptional read-throughs may be not restricted to the herpesviruses but they may represent a general phenomenon. Furthermore, the antisense RNAs produced from their own promoters or by transcriptional read-throughs may hybridize with their sense counterparts thereby initiating RNA interference [83]. Very long complex transcripts form a distinct category among multigenic transcripts because of their oppositely oriented ORFs and increased size. Similarly to other herpesviruses [21, 50] they are present in very low abundance in VZV, however their existence is ambiguous, as they can be formed during reverse transcription by template switching [84]. It has been previously thought that the OriS of VZV lacks any overlapping transcripts [6]. Additionally, Davison and colleagues (1986) hypothesis the existence of non-coding RNA in the intergenic region between orf62 and orf63. We detected several transcripts overlapping OriS, and a small non-coding RNA (the NCO3) between the before-mentioned two genes. This region of the viral transcriptome is structurally similar to the PRV transcriptome. In the vicinity of OriS, we identified two additional novel non-coding transcripts (NTO2 and NTO4) and two 5' elongated and spliced version of ORF63. These nroRNAs are supposed to play a role in the regulation of the viral replication [85] through the interplay between the transcriptional and replication apparatuses [51]. Specifically, the role of the transcription of these RNA molecules might be to interfere with the replication machinery, in order to force the replication fork to an unidirectional progression [63].

Additionally, we detected a hyper-editing process in the novel NTO3 transcript. This phenomenon was previously observed in other members of the herpesvirus family [86–88], but to our knowledge, this is the first observation of A to I hyper-editing in alphaherpesviruses. This process plays a crucial role in the cell's innate immunity [89], while it can be hijacked by some viruses to evade inactivation [29]. Additionally, in hepatitis delta virus hyper-editing is indispensable for replication [30]. A to I editing decreases the affinity of the antisense transcript to the sense RNA, destabilizing their interaction, which may affect the binding of dsRNA enzymes like RNase III homologues [90]. Our *in-silico* analysis shows that despite elevating the level of free energy of the sense-antisense RNA hybrid, hyper-editing does not result in a change of the secondary structure, but in the formation of I·U wobble pairs which are significantly more resistant to possible Dicer cleavage [91], inhibiting RNA interference. It is also possible that I·U pairs play a role in the cleavage and degradation of ORF62 by a process mediated by the Tudor staphylococcal nuclease (Tudor-SN) [92, 93]. Further investigation is needed to elucidate the significance of hyper-editing on the NTO3. An evaluation of gene expression at different time points could shed light on the presence of the asRNA and on the amount of editing in different stages of the viral life cycle. Additionally, miRNA assays of the viral transcriptome in different time points of the infection could prove the effect of the RNAi on the dsRNA formed by the sense-antisense pair.

Splice events are thought to be rare in alphaherpesviruses, however, they appear to be underestimated in the light of LRS techniques. In this work, we enriched the list of spliced transcripts of the lytic phase of the viral infection, and we identified the novel combinations of splice sites in ORF42/45.

At least six VZV-transcripts have been shown to be expressed in latency [94, 95], however recent target enrichment SRS data suggests that VLT and ORF63 are the only two expressed transcripts, maintaining the dormant phase of VZV. In this work Depledge and coworkers showed that the VLT is spliced differently in the latent and lytic phase of the viral lifecycle, and identified several length and splice isoforms during productive infection [9]. We confirmed the existence of four introns of this transcript in the longer, lytic forms of VLT. Additionally, we showed a splice isoform of ORF63 and ORF64

possessing the same introns as the VLT. This suggests that occasionally VLT, ORF63 and ORF64 can form a single transcriptional unit. We found that a novel splice isoform of ORF63 produces a long transcript, which has non-canonical splice donor site (GC, instead of GU). Another rare but clearly visible splice variant of ORF63-L has the same, non-canonical splice donor site. The reason for this could be simply the higher GC content of these regions. Another explanation may be that the splice isoforms of ORF63 are recognized by host cell spliceosomes differentially, in order to perform varying expression pattern during the life cycle of virus. Similar alternative splicing has been described in VLT, assuming distinct functions for it in lytic and latent phase.

The uORFs are supposed to play a regulatory role in the translation of eukaryotic mRNAs. Our results suggests, that at least some of the uORFs could play a role in the production of N-terminally truncated proteins, while most of them have a large-enough space between their stop codons and the protein coding ORF's AUG for a reinitiation event, thus resulting in unaltered protein translation [96]. Further studies implying ribosome profiling and mRNA and protein expression studies are needed to determine the precise function of these uORFs. Nevertheless, the use of alternative promoters for producing TSS variants with or without uORFs may have a role in providing a differential control of translation at distinct stage of the viral lifecycle [21].

Conclusions

This study substantially redefines the VZV transcriptome by identifying a large number of novel RNA molecules and transcript isoforms, as well as revealing a complex pattern of transcriptional overlaps. The extensive transcriptional overlaps may indicate an interaction between the transcription machineries. Additionally, the discovery of nroRNAs suggests an interference between the replication and transcription apparatuses. Besides the significant advance in transcriptome annotation, these data may also help in controlling this virus.

List of abbreviations

ADAR1: adenosine deaminase acting on RNA type 1

AS: antisense

AT: alternative termination

HSV-1: herpes simplex virus type 1

HCMV: human cytomegalovirus

lncRNA: long non-coding RNA

LRS: long-read sequencing

ncRNA: non-coding RNA

nroRNA: near-replication-origin RNA

NTO: near to (replication) origin

ONT: Oxford Nanopore Technologies

PAS: Poly(A) signal

PRV: pseudorabies virus

RT: reverse transcriptase

sncRNA: short non-coding RNA

SRS: short-read sequencing

TES: transcription end site

TR: truncated

TSS: transcription start site

uORF: upstream open reading frame

UTR: untranslated region

VLT: VZV latency transcript

VLT_{ly}: lytic form of VLT

VZV: Varicella Zoster Virus

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The sequencing data and the transcriptome assembly have been uploaded to the European Nucleotide Archive under the project accession number PRJEB25401.

Competing interests

The authors declare that they have no competing interests

Funding

This work was supported by the Swiss-Hungarian Cooperation Programme [SH/7/2/8] to ZBo. The work was also supported by the Bolyai János Scholarship of the Hungarian Academy of Sciences to DT.

Author Contributions

IP analyzed the data, participated in RNA-isolation and sequence alignment and drafted and wrote the manuscript. NM participated in data analysis, carried out the statistical analysis, prepared the figures and revised the manuscript. DT carried out the ONT MinION cDNA sequencing, participated in the design of the study, and took part in drafting the manuscript. AS participated in the sequence alignment and carried out the *in silico* analysis. ZC prepared the RNA, DNA and cDNA samples, carried out the PCR analysis and participated in the MinION sequencing. KM carried out the virus infection and propagated the cells. ZBo conceived, designed and coordinated the study and wrote the manuscript. All authors have read and approved the final version of the manuscript.

References

1. Kennedy PGE. Varicella-zoster virus latency in human ganglia. Rev Med Virol. 2002;12:327–34. doi:10.1002/rmv.362.

2. Cohrs RJ, Gilden DH, Mahalingam R. Varicella zoster virus latency, neurological disease and experimental models: an update. Front Biosci. 2004;9:751–62.

3. Arvin AM. Investigations of the pathogenesis of Varicella zoster virus infection in the SCIDhu mouse model. Herpes. 2006;13:75–80.

4. Zerboni L, Sen N, Oliver SL, Arvin AM. Molecular mechanisms of varicella zoster virus pathogenesis. Nat Rev Micro. 2014;12:197–210.

5. Maresova L, Pasieka TJ, Homan E, Gerday E, Grose C. Incorporation of Three Endocytosed Varicella-Zoster Virus Glycoproteins, gE, gH, and gB, into the Virion Envelope. J Virol. 2005;79:997–1007.

6. Davison AJ, Scott JE. The complete DNA sequence of varicella-zoster virus. J Gen Virol. 1986;67:1759–816.

7. Tillieux SL, Halsey WS, Thomas ES, Voycik JJ, Sathe GM, Vassilev V. Complete DNA sequences of two oka strain varicella-zoster virus genomes. J Virol. 2008;82:11023–44.

8. Kinchington PR, Leger AJS, Guedon J-MG, Hendricks RL. Herpes simplex virus and varicella zoster virus, the house guests who never leave. Herpesviridae. 2012;3:5.

9. Depledge DP, Ouwendijk WJD, Sadaoka T, Braspenning SE, Mori Y, Cohrs RJ, et al. A spliced latency-associated VZV transcript maps antisense to the viral transactivator gene 61. Nat Commun. 2018;9:1167. doi:10.1038/s41467-018-03569-2.

10. Perera LP, Mosca JD, Sadeghi-Zadeh M, Ruyechan WT, Hay J. The Varicella-Zoster virus immediate early protein, IE62, can positively regulate its cognate promoter. Virology. 1992;191:346–54.

11. Kinchington PR, Hougland JK, Arvin AM, Ruyechan WT, Hay J. The varicella-zoster virus immediate-early protein IE62 is a major component of virus particles. J Virol. 1992;66:359–66.

12. Kinchington PR, Bookey D, Turse SE. The transcriptional regulatory proteins encoded by varicella-zoster virus open reading frames (ORFs) 4 and 63, but not ORF 61, are associated with purified virus particles. J Virol. 1995;69:4274–82.

13. Moriuchi H, Moriuchi M, Straus SE, Cohen JI. Varicella-zoster virus open reading frame 10 protein, the herpes simplex virus VP16 homolog, transactivates herpesvirus immediate-early gene promoters. J Virol. 1993;67:2739–46.

14. Che X, Zerboni L, Sommer MH, Arvin AM. Varicella-zoster virus open reading frame 10 is a virulence determinant in skin cells but not in T cells in vivo. J Virol. 2006;80:3238–48.

15. Reichelt M, Brady J, Arvin AM. The Replication Cycle of Varicella-Zoster Virus: Analysis of the Kinetics of Viral Protein Expression, Genome Synthesis, and Virion Assembly at the Single-Cell Level. J Virol. 2009;83:3904–18.

16. Khalil MI, Che X, Sung P, Sommer MH, Hay J, Arvin AM. Mutational analysis of varicella-zoster virus (VZV) immediate early protein (IE62) subdomains and their importance in viral replication. Virology. 2016;492:82–91.

17. Cohrs RJ, Hurley MP, Gilden DH. Array analysis of viral gene transcription during lytic infection of cells in tissue culture with Varicella-Zoster virus. J Virol. 2003;77:11718–32. doi:10.1128/JVI.77.21.11718-11732.2003.

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

19. Kronstad LM, Brulois KF, Jung JU, Glaunsinger BA. Dual short upstream open reading frames control translation of a herpesviral polycistronic mRNA. PLoS Pathog. 2013;9:e1003156. doi:10.1371/journal.ppat.1003156.

20. Talbot SJ, Weiss RA, Kellam P, Boshoff C. Transcriptional analysis of human herpesvirus-8 open reading frames 71, 72, 73, K14, and 74 in a primary effusion lymphoma cell line. Virology. 1999;257:84–94.

21. Balázs Z, Tombácz D, Szűcs A, Csabai Z, Megyeri K, Petrov AN, et al. Long-Read Sequencing of Human Cytomegalovirus Transcriptome Reveals RNA Isoforms Carrying Distinct Coding Potentials.

Sci Rep. 2017;7:15989. doi:10.1038/s41598-017-16262-z.

22. Moldován N, Tombácz D, Szűcs A, Csabai Z, Snyder M, Boldogkői Z. Multi-Platform Sequencing Approach Reveals a Novel Transcriptome Profile in Pseudorabies Virus. Front Microbiol. 2018;8:2708. doi:10.3389/fmicb.2017.02708.

23. Sadaoka T, Yanagi T, Yamanishi K, Mori Y. Characterization of the varicella-zoster virus ORF50 gene, which encodes glycoprotein M. J Virol. 2010;84:3488–502.

24. Kemble GW, Annunziato P, Lungu O, Winter RE, Cha T-A, Silverstein SJ, et al. Open Reading Frame S/L of Varicella-Zoster Virus Encodes a Cytoplasmic Protein Expressed in Infected Cells. J Virol. 2000;74:11311–21.

25. Visalli RJ, Nicolosi DM, Irven KL, Goshorn B, Khan T, Visalli MA. The Varicella-zoster virus DNA encapsidation genes: Identification and characterization of the putative terminase subunits. Virus Res. 2007;129:200–11.

26. Walkley CR, Li JB. Rewriting the transcriptome: adenosine-to-inosine RNA editing by ADARs. Genome Biol. 2017;18:205. doi:10.1186/s13059-017-1347-3.

27. Eggington JM, Greene T, Bass BL. Predicting sites of ADAR editing in double-stranded RNA. Nat Commun. 2011;2:319. doi:10.1038/ncomms1324.

28. Matthews MM, Thomas JM, Zheng Y, Tran K, Phelps KJ, Scott AI, et al. Structures of human ADAR2 bound to dsRNA reveal base-flipping mechanism and basis for site selectivity. Nat Struct Mol Biol. 2016;23:426–33. doi:10.1038/nsmb.3203.

29. Zahn RC, Schelp I, Utermöhlen O, von Laer D. A-to-G hypermutation in the genome of lymphocytic choriomeningitis virus. J Virol. 2007;81:457–64. doi:10.1128/JVI.00067-06.

30. Wong SK, Lazinski DW. Replicating hepatitis delta virus RNA is edited in the nucleus by the small form of ADAR1. Proc Natl Acad Sci. 2002;99:15118–23. doi:10.1073/pnas.232416799.

31. Kozak M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene. 2005;361:13–37. doi:10.1016/J.GENE.2005.06.037.

32. Kozak M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell. 1986;44:283–92. http://www.ncbi.nlm.nih.gov/pubmed/3943125. Accessed 16 Aug 2018.

33. Pisarev A V, Kolupaeva VG, Pisareva VP, Merrick WC, Hellen CUT, Pestova T V. Specific functional interactions of nucleotides at key -3 and +4 positions flanking the initiation codon with components of the mammalian 48S translation initiation complex. Genes Dev. 2006;20:624–36. doi:10.1101/gad.1397906.

34. Luukkonen BG, Tan W, Schwartz S. Efficiency of reinitiation of translation on human immunodeficiency virus type 1 mRNAs is determined by the length of the upstream open reading frame and by intercistronic distance. J Virol. 1995;69:4086–94. http://www.ncbi.nlm.nih.gov/pubmed/7769666. Accessed 24 Jul 2018.

35. Kozak M. Constraints on reinitiation of translation in mammals. Nucleic Acids Res. 2001;29:5226–32. doi:10.1093/nar/29.24.5226.

36. Chambers J, Angulo A, Amaratunga D, Guo H, Jiang Y, Wan JS, et al. DNA microarrays of the complex human cytomegalovirus genome: profiling kinetic class with drug sensitivity of viral gene expression. J Virol. 1999;73:5757–66. http://www.ncbi.nlm.nih.gov/pubmed/10364327. Accessed 10 Apr 2018.

37. Ebrahimi B, Dutia BM, Roberts KL, Garcia-Ramirez JJ, Dickinson P, Stewart JP, et al. Transcriptome profile of murine gammaherpesvirus-68 lytic infection. J Gen Virol. 2003;84:99–109. doi:10.1099/vir.0.18639-0.

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

39. Baird NL, Bowlin JL, Cohrs RJ, Gilden D, Jones KL. Comparison of Varicella-Zoster Virus RNA Sequences in Human Neurons and Fibroblasts. J Virol. 2014;88:5877–80. doi:10.1128/JVI.00476-14.

40. Luo GX, Taylor J. Template switching by reverse transcriptase during DNA synthesis. J Virol. 1990;64:4321–8. http://www.ncbi.nlm.nih.gov/pubmed/1696639. Accessed 15 Aug 2017.

41. Cocquet J, Chong A, Zhang G, Veitia RA. Reverse transcriptase template switching and false alternative transcripts. Academic Press; 2006. http://www.sciencedirect.com/science/article/pii/S0888754305003770. Accessed 18 Aug 2017.

42. Kuo RI, Tseng E, Eory L, Paton IR, Archibald AL, Burt DW. Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. BMC Genomics. 2017;18:323. doi:10.1186/s12864-017-3691-9.

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

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

45. Szűcs A, Moldován N, Tombácz D, Csabai Z, Snyder M, Boldogkői Z. Long-Read Sequencing Reveals a GC Pressure during the Evolution of Porcine Endogenous Retrovirus. Genome Announc. 2017;5:e01040-17. doi:10.1128/genomeA.01040-17.

46. Moldován N, Tombácz D, Szűcs A, Csabai Z, Balázs Z, Kis E, et al. Third-generation Sequencing Reveals Extensive Polycistronism and Transcriptional Overlapping in a Baculovirus. Sci Rep. 2018;8:8604. doi:10.1038/s41598-018-26955-8.

47. Tombacz D, Sharon D, Olah P, Csabai Z, Snyder M, Boldogk i Z. Strain Kaplan of Pseudorabies Virus Genome Sequenced by PacBio Single-Molecule Real-Time Sequencing Technology. Genome Announc. 2014;2:e00628-14-e00628-14. doi:10.1128/genomeA.00628-14.

48. 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:e145–e145. doi:10.1093/nar/gkw629.

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

50. Tombácz D, Csabai Z, Szűcs A, Balázs Z, Moldován N, Sharon D, et al. Long-Read Isoform Sequencing Reveals a Hidden Complexity of the Transcriptional Landscape of Herpes Simplex Virus Type 1. Front Microbiol. 2017;8:1079. doi:10.3389/fmicb.2017.01079.

51. Boldogköi Z. Transcriptional interference networks coordinate the expression of functionally related genes clustered in the same genomic loci. Front Genet. 2012;3:122. doi:10.3389/fgene.2012.00122.

52. Mestdagh P, Van Vlierberghe P, De Weer A, Muth D, Westermann F, Speleman F, et al. A novel and universal method for microRNA RT-qPCR data normalization. Genome Biol. 2009;10:R64. doi:10.1186/gb-2009-10-6-r64.

53. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21:1859–75. doi:10.1093/bioinformatics/bti310.

54. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-

performance genomics data visualization and exploration. Brief Bioinform. 2013;14:178–92. doi:10.1093/bib/bbs017.

55. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402. http://www.ncbi.nlm.nih.gov/pubmed/9254694. Accessed 25 Apr 2018.

56. Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics. 2010;11:129. doi:10.1186/1471-2105-11-129.

57. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28:1647–9. doi:10.1093/bioinformatics/bts199.

58. Narang V, Sung W-K, Mittal A. Computational modeling of oligonucleotide positional densities for human promoter prediction. Artif Intell Med. 2005;35:107–19. doi:10.1016/J.ARTMED.2005.02.005.

59. Rock DL, Nesburn AB, Ghiasi H, Ong J, Lewis TL, Lokensgard JR, et al. Detection of latency-related viral RNAs in trigeminal ganglia of rabbits latently infected with herpes simplex virus type 1. J Virol. 1987;61:3820–6. http://www.ncbi.nlm.nih.gov/pubmed/2824816. Accessed 2 May 2018.

60. Jin L, Scherba G. Expression of the pseudorabies virus latency-associated transcript gene during productive infection of cultured cells. J Virol. 1999;73:9781–8. http://www.ncbi.nlm.nih.gov/pubmed/10559288. Accessed 2 May 2018.

61. Kemble GW, Annunziato P, Lungu O, Winter RE, Cha TA, Silverstein SJ, et al. Open reading frame S/L of varicella-zoster virus encodes a cytoplasmic protein expressed in infected cells. J Virol. 2000;74:11311–21. http://www.ncbi.nlm.nih.gov/pubmed/11070031. Accessed 14 Mar 2018.

62. Lenac Rovis T, Bailer SM, Pothineni VR, Ouwendijk WJD, Simic H, Babic M, et al. Comprehensive Analysis of Varicella-Zoster Virus Proteins Using a New Monoclonal Antibody Collection. J Virol. 2013;87:6943–54. doi:10.1128/JVI.00407-13.

63. Tombácz D, Csabai Z, Oláh P, Havelda Z, Sharon D, Snyder M, et al. Characterization of novel transcripts in pseudorabies virus. Viruses. 2015;7:2727–44. doi:10.3390/v7052727.

64. Voss JH, Roizman B. 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. 1988;85:8454–8. doi:10.1073/PNAS.85.22.8454.

65. Huang L, Zhu Y, Anders DG. The variable 3' ends of a human cytomegalovirus oriLyt transcript (SRT) overlap an essential, conserved replicator element. J Virol. 1996;70:5272–81. http://www.ncbi.nlm.nih.gov/pubmed/8764037. Accessed 21 Aug 2018.

66. Kotenko S V., Saccani S, Izotova LS, Mirochnitchenko O V., Pestka S, Baluchova K, et al. Human cytomegalovirus harbors its own unique IL-10 homolog (cmvIL-10). Proc Natl Acad Sci. 2000;97:1695–700. doi:10.1073/pnas.97.4.1695.

67. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37 Web Server:W202–8. doi:10.1093/nar/gkp335.

68. Ostrove JM, Reinhold W, Fan CM, Zorn S, Hay J, Straus SE. Transcription mapping of the varicella-zoster virus genome. J Virol. 1985;56:600–6. http://www.ncbi.nlm.nih.gov/pubmed/2997479. Accessed 14 Mar 2018.

69. Kinchington PR, Vergnes JP, Defechereux P, Piette J, Turse SE. Transcriptional mapping of the varicella-zoster virus regulatory genes encoding open reading frames 4 and 63. J Virol. 1994;68:3570–81. http://www.ncbi.nlm.nih.gov/pubmed/8189496. Accessed 22 Mar 2018.

70. Kato T, Kitamura K, Hayakawa Y, Takahashi M, Kojima A, Sato S, et al. Transcription mapping of

glycoprotein I (gpI) and gpIV of varicella-zoster virus and immunological analysis of the gpI produced in cells infected with the recombinant vaccinia virus. Microbiol Immunol. 1989;33:299–312. http://www.ncbi.nlm.nih.gov/pubmed/2549343. Accessed 11 Apr 2018.

71. Meier JL, Straus SE. Varicella-zoster virus DNA polymerase and major DNA-binding protein genes have overlapping divergent promoters. J Virol. 1993;67:7573–81. http://www.ncbi.nlm.nih.gov/pubmed/8230477. Accessed 11 Oct 2016.

72. Kennedy PGE, Grinfeld E, Craigon M, Vierlinger K, Roy D, Forster T, et al. Transcriptomal analysis of varicella-zoster virus infection using long oligonucleotide-based microarrays. J Gen Virol. 2005;86 Pt 10:2673–84. doi:10.1099/vir.0.80946-0.

73. Grinfeld E, Ross A, Forster T, Ghazal P, Kennedy PGE. Genome-wide reduction in transcriptomal profiles of varicella-zoster virus vaccine strains compared with parental Oka strain using long oligonucleotide microarrays. Virus Genes. 2009;38:19–29. doi:10.1007/s11262-008-0304-3.

74. Cohrs RJ, Lee KS, Beach A, Sanford B, Baird NL, Como C, et al. Targeted Genome Sequencing Reveals Varicella-Zoster Virus Open Reading Frame 12 Deletion. J Virol. 2017;91. doi:10.1128/JVI.01141-17.

75. Markus A, Golani L, Ojha NK, Borodiansky-Shteinberg T, Kinchington PR, Goldstein RS. Varicella-Zoster Virus Expresses Multiple Small Noncoding RNAs. J Virol. 2017;91:e01710-17. doi:10.1128/JVI.01710-17.

76. Legendre M, Gautheret D. Sequence determinants in human polyadenylation site selection. BMC Genomics. 2003;4:7. doi:10.1186/1471-2164-4-7.

77. Tian B, Hu J, Zhang H, Lutz CS. A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res. 2005;33:201–12. doi:10.1093/nar/gki158.

78. Tian B, Graber JH. Signals for pre-mRNA cleavage and polyadenylation. Wiley Interdiscip Rev RNA. 2012;3:385–96. doi:10.1002/wrna.116.

79. Hu J, Lutz CS, Wilusz J, Tian B. Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. RNA. 2005;11:1485–93. doi:10.1261/rna.2107305.

80. Kost RG, Kupinsky H, Straus SE. Varicella-Zoster Virus Gene 63: Transcript Mapping and Regulatory Activity. Virology. 1995;209:218–24. doi:10.1006/viro.1995.1246.

81. Hoover SE, Cohrs RJ, Rangel ZG, Gilden DH, Munson P, Cohen JI. Downregulation of varicella-zoster virus (VZV) immediate-early ORF62 transcription by VZV ORF63 correlates with virus replication in vitro and with latency. J Virol. 2006;80:3459–68. doi:10.1128/JVI.80.7.3459-3468.2006.

82. He Y, Yuan C, Chen L, Lei M, Zellmer L, Huang H, et al. Transcriptional-Readthrough RNAs Reflect the Phenomenon of “A Gene Contains Gene(s)” or “Gene(s) within a Gene” in the Human Genome, and Thus Are Not Chimeric RNAs. Genes (Basel). 2018;9:40. doi:10.3390/genes9010040.

83. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, et al. Antisense Transcription in the Mammalian Transcriptome. Science (80-). 2005;309:1564–6. doi:10.1126/science.1112009.

84. Yuan C, Liu Y, Yang M, Liao DJ. New methods as alternative or corrective measures for the pitfalls and artifacts of reverse transcription and polymerase chain reactions (RT-PCR) in cloning chimeric or antisense-accompanied RNA. RNA Biol. 2013;10:957–67. doi:10.4161/rna.24570.

85. Huvet M, Nicolay S, Touchon M, Audit B, d’Aubenton-Carafa Y, Arneodo A, et al. Human gene organization driven by the coordination of replication and transcription. Genome Res. 2007;17:1278–85. doi:10.1101/gr.6533407.

86. Figueroa T, Boumart I, Coupeau D, Rasschaert D. Hyperediting by ADAR1 of a new herpesvirus lncRNA during the lytic phase of the oncogenic Marek’s disease virus. J Gen Virol. 2016;97:2973–88.

doi:10.1099/jgv.0.000606.

87. Gandy SZ, Linnstaedt SD, Muralidhar S, Cashman KA, Rosenthal LJ, Casey JL. RNA editing of the human herpesvirus 8 kaposin transcript eliminates its transforming activity and is induced during lytic replication. J Virol. 2007;81:13544–51. doi:10.1128/JVI.01521-07.

88. Iizasa H, Wulff B-E, Alla NR, Maragkakis M, Megraw M, Hatzigeorgiou A, et al. Editing of Epstein-Barr Virus-encoded BART6 MicroRNAs Controls Their Dicer Targeting and Consequently Affects Viral Latency. J Biol Chem. 2010;285:33358–70. doi:10.1074/jbc.M110.138362.

89. Mannion NM, Greenwood SM, Young R, Cox S, Brindle J, Read D, et al. The RNA-Editing Enzyme ADAR1 Controls Innate Immune Responses to RNA. Cell Rep. 2014;9:1482–94. doi:10.1016/j.celrep.2014.10.041.

90. Nishikura K. Editor meets silencer: crosstalk between RNA editing and RNA interference. Nat Rev Mol Cell Biol. 2006;7:919–31. doi:10.1038/nrm2061.

91. Scadden AD, Smith CW. RNAi is antagonized by A-->I hyper-editing. EMBO Rep. 2001;2:1107–11. doi:10.1093/embo-reports/kve244.

92. Scadden ADJ, Smith CW. Specific cleavage of hyper-edited dsRNAs. EMBO J. 2001;20:4243–52. doi:10.1093/emboj/20.15.4243.

93. Scadden ADJ. The RISC subunit Tudor-SN binds to hyper-edited double-stranded RNA and promotes its cleavage. Nat Struct Mol Biol. 2005;12:489–96. doi:10.1038/nsmb936.

94. Cohen JI, Cox E, Pesnicak L, Srinivas S, Krogmann T. The varicella-zoster virus open reading frame 63 latency-associated protein is critical for establishment of latency. J Virol. 2004;78:11833–40.

95. Nagel MA, Choe A, Traktinskiy I, Cordery-Cotter R, Gilden D, Cohrs RJ. Varicella-Zoster Virus Transcriptome in Latently Infected Human Ganglia. J Virol. 2011;85:2276–87.

96. Poyry TAA, Kaminski A, Jackson RJ. What determines whether mammalian ribosomes resume scanning after translation of a short upstream open reading frame? Genes Dev. 2004;18:62–75. doi:10.1101/gad.276504.

Figure Legends

Figure 1. The VZV transcriptome. Color code: light orange arrow-rectangles: open reading frames (ORFs); black arrow-rectangles: upstream ORFs; dark orange arrow-rectangles: truncated (t)ORFs; gray arrow-rectangles: previously annotated transcripts; blue arrow-rectangles: novel mRNAs; red arrow-rectangles: novel ncRNAs; green arrow-rectangles: novel complex transcripts.

Figure 2. The read length distributions

a. The frequency of the reads of the cap-selected and non-cap-selected sequencing binned by 200 nucleotides shows that the cap-selected reads are skewed towards shorter read lengths.

b. The frequency of each nucleotide in the vicinity of the 5’ ends of the cap-selected reads shows random distribution, suggesting that the short read lengths are not caused by the presence of a specific nucleotide on the RNA but other, yet unknown reason.

Figure 3. The frequency of nucleotides in the vicinity of transcriptional end sites (TES). The weblogo shows an enrichment of A and U bases upstream, while an enrichment of G and U bases downstream of the TES. This pattern is akin with the sequence surroundings of mammalian TESs.

Figure 4. The genomic region of VLT. Color code: light orange arrow-rectangles: open reading frames (ORFs); dark orange arrow-rectangles: truncated (t)ORFs; black arrow-rectangles: upstream ORFs; gray arrow-rectangles: previously annotated transcripts; blue arrow-rectangles: novel mRNAs; red arrow-rectangles: ncRNAs; Reads of the VLT_{ly} transcript isoforms were present in very low abundance, thus their 5’ ends are feathered resembling uncertain TSSs.

Figure 5. Structurally similar regions of three alphaherpesvirus transcripts neighboring OriS. Color code: light orange arrow-rectangles: open reading frames (ORFs); blue arrow-rectangles: mRNAs; red arrow-rectangles: ncRNAs. The ORF63 of VZV, similarly to the HSV-1 US1 has a two-exon-baring splice isoform, while akin with PRV's US1 the three-exon-baring splice isoforms. Both VZV and PRV express non-overlapping non-coding RNAs in the proximity of OriS.

Figure 6. Splice isoforms of ORF63 and ORF64 with a similar splicing pattern as the VLT_{ly}. Color code: light orange arrow-rectangles: open reading frames (ORFs); dark orange arrow-rectangles: spliced ORFs; green arrow-rectangles: complex transcripts; red arrow-rectangles: ncRNAs. AUGs present in the upstream exons of the ORF63 and ORF64 splice isoforms encompasses two spliced ORFs, the vlt-orf63-1 translated to a hypothetical protein product 88 aa longer, while the vlt-orf63-2 translated to a hypothetical protein product 179 aa longer than orf63.

Figure 7. Types of near-replication-origin (nro)RNAs of three alphaherpesviruses. Color code: light orange arrow-rectangles: open reading frames (ORFs gray arrow-rectangles: ambiguous RNAs; blue arrow-rectangles: ncRNAs not overlapping the Ori (type 1); red arrow-rectangles: ncRNA overlapping the Ori (type 2); green arrow-rectangles: mRNAs overlapping the Ori and with long 3' UTRs (type 3); brown arrow-rectangles: mRNAs overlapping the Ori and with long 5' UTRs (type 4).

Figure 8. Types of overlaps between VZV transcripts.

a. Parallel overlaps of the ORF5-4-3 cluster.

b. Divergent overlaps of the ORF8 and ORF9-L and ORF9A-9 transcripts.

c. Convergent overlaps of the ORF17 and ORF18-AT transcripts.

Figure 9. A to I hyper-editing of NTO3.

a. Reads of NTO2 and the overlapping transcripts mapping to the VZV genome visualized with IGV. The orange dots represent G mismatches, indicating editing events.

b. Substitution matrix of the NTO2 reads (n=703 substitutions / 7749 nt, $p<0.0001$, Fisher's exact test).

c. The position and frequency of A->G substitutions on the genomic sequence corresponding to the NTO2 transcript, showing both NTO2 and the overlapping transcripts. Substitutions with high frequencies indicate A to I editing events, while those with low frequency are sequencing errors.

d. The motif surrounding the editing sites. The motif was found using the MEME software suit with an E-value of 0.58 and a log likelihood ratio of 46. The edited adenosine is marked with a *.

Supplementary Figure 1 The secondary structure of NTO3 (a. and b.) as well as the hybrid formed by the ORF62-5' fragment and the NTO3 (c and d).

a. The secondary structure of the unedited ORF62-AS2 with a free energy of -143.2 kcal/mol. The adenines in the editing sites are colored in green.

b. The secondary structure of the edited NTO3 with a free energy of -169.4 kcal/mol. The inosines in the editing sites are colored in orange.

c. The secondary structure of the sense-antisense hybrid composed of the first 467 bases of ORF62 labeled ORF62-5' (gray) and the full sequence of NTO3 (blue), the latter is its unedited form. The free energy of the structure formed by the two molecules is -822.9 kcal/mol.

d. The secondary structure of the sense-antisense hybrid composed of the first 467 bases of ORF62 labeled ORF62-5' (gray) and the full sequence of NTO3 (blue), the later in its edited form. The free energy of the structure formed by the two molecules is -818.7 kcal/mol. The position of I·U base pairs is marked with black arrows.

Additional File Legend

The additional file is in .xlsx format, containing eight Supplementary Tables on separate pages and a References page.

Supplementary Table 1. The list of transcripts with TSS and TES that passed our criteria for validation.

Supplementary Table 2. The list of transcripts with 5' and 3' ends that did not pass our criteria for validation.

Supplementary Table 3. Promoter motifs and poly(A) signals.

Supplementary Table 4. Overlaps between the transcripts listed in Supplementary Table 1.

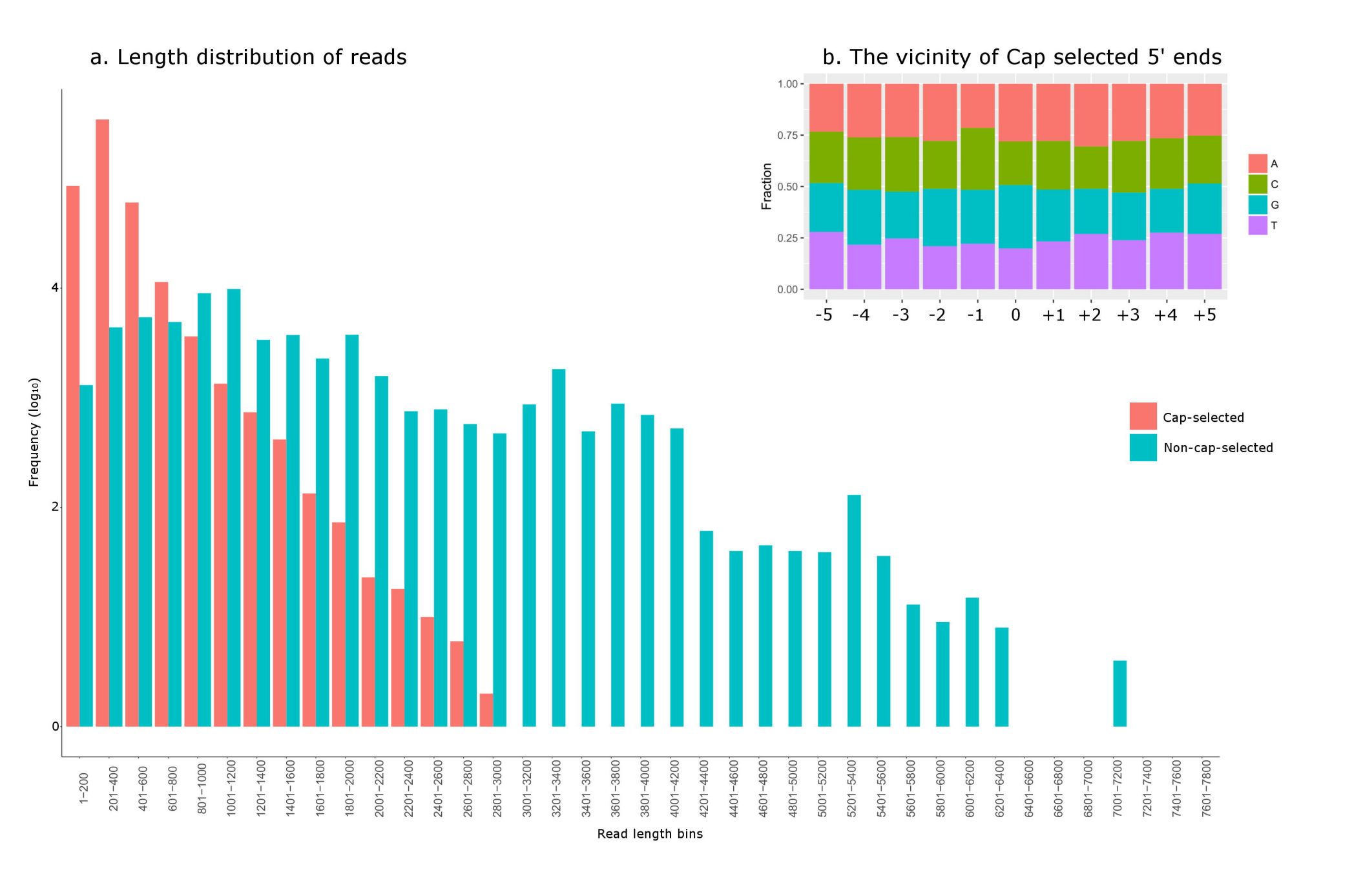
Supplementary Table 5. The full list of 5' ends in the non-cap-selected datasets generated by our pipeline. The p-values were calculated using the Poisson distribution. Positions excluded because of missing TESs are marked by *, while those excluded because of false priming or strand switching are marked with ** in the Notes column. The number of 5' ends in the ±50 vicinity of a given position are shown in the Vicinity column.

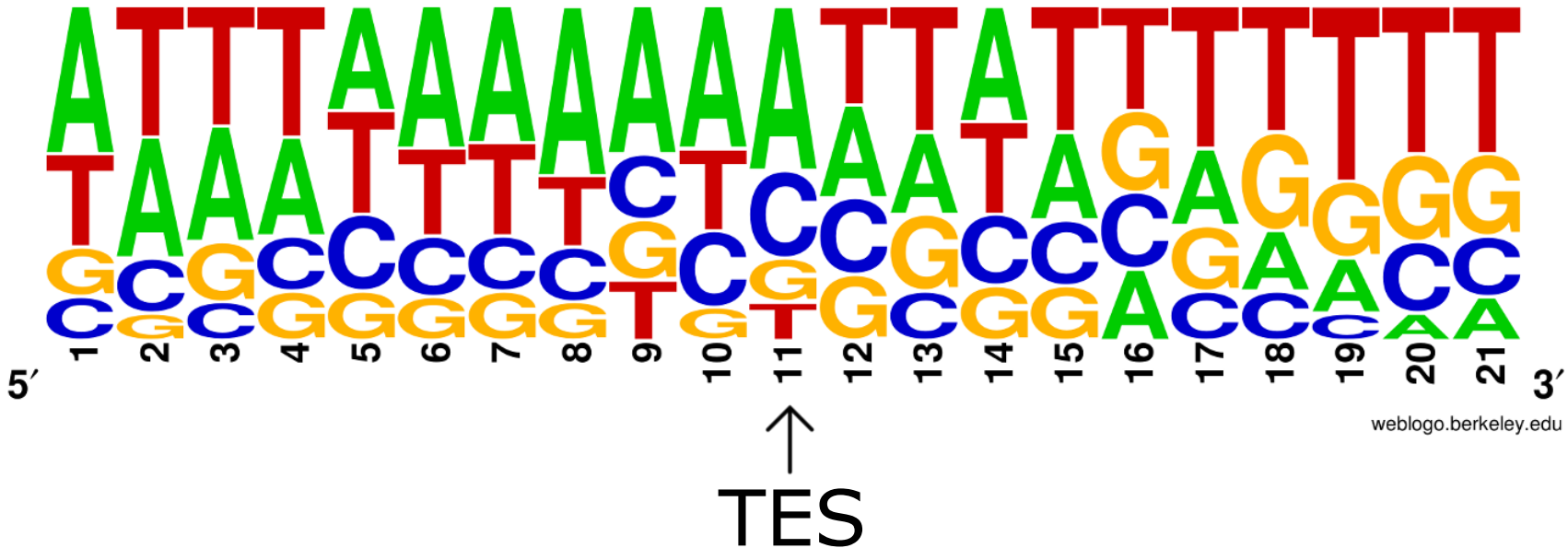
Supplementary Table 6. The full list of 5' ends in the cap-selected datasets generated by our pipeline. The p-values were calculated using the Poisson distribution. Positions excluded because of missing TESs are marked by *, while those excluded because of false priming or strand switching are marked with ** in the Notes column. The number of 5' ends in the ±50 vicinity of a given position are shown in the Vicinity column.

Supplementary Table 7. The full list of 3' ends in the non-cap-selected datasets generated by our pipeline. The p-values were calculated using the Poisson distribution. Positions excluded because of missing TSSs are marked by *, positions excluded because of false priming or strand switching are marked with **. TES with more than 3 but less than 5 adenines in the vicinity of the TES with *y and they were accepted. The number of 3' ends in the ±50 vicinity of a given position are shown in the Vicinity column.

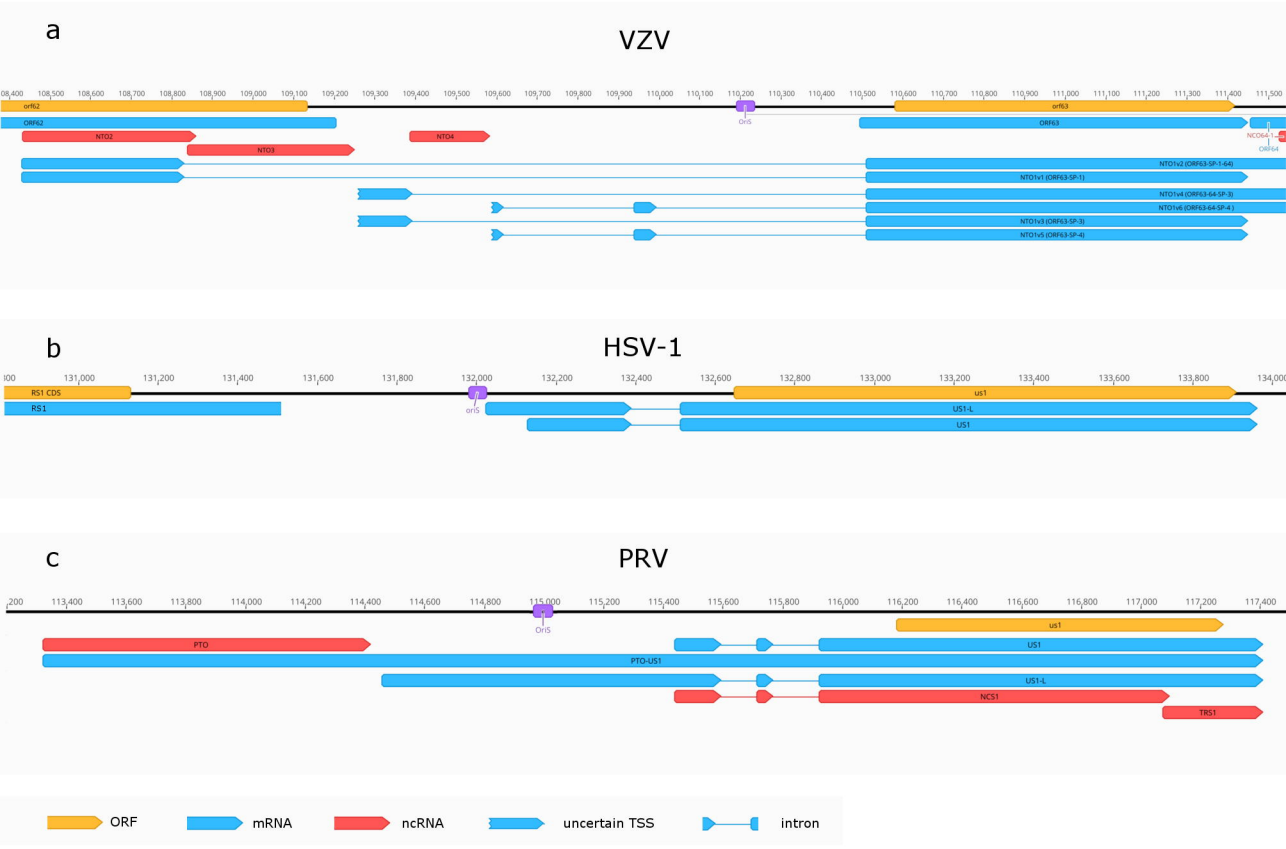
Supplementary Table 8. The full list of 3' ends in the cap-selected datasets generated by our pipeline. The p-values were calculated using the Poisson distribution. Positions excluded because of missing TSSs are marked by *, positions excluded because of false priming or strand switching are marked with **. TES with more than 3 but less than 5 adenines in the vicinity of the TES with *y and they were accepted. The number of 3' ends in the ±50 vicinity of a given position are shown in the Vicinity column.

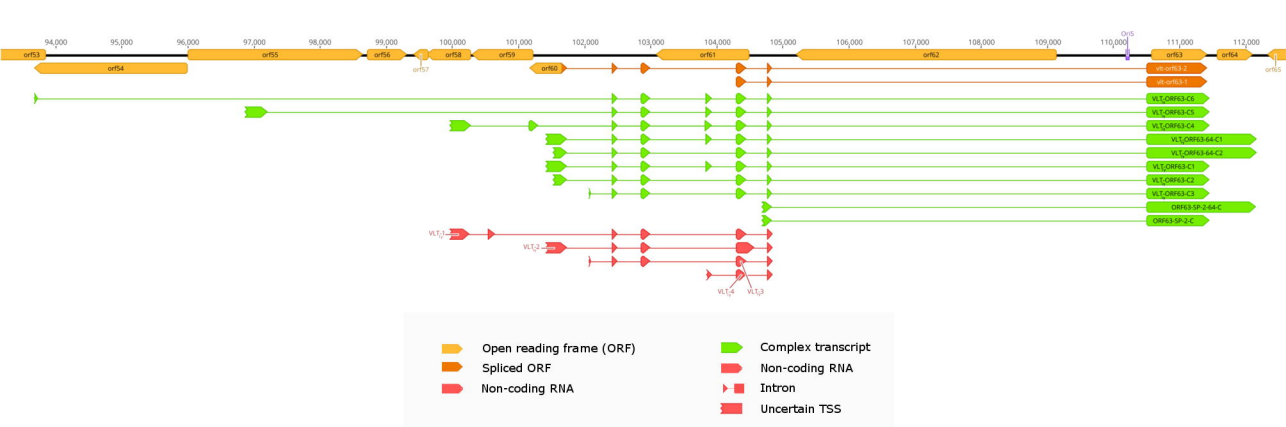

Open reading frame (ORF)
Truncated ORF (tORF)
Upstream ORF (uORF)
Previously annotated transcript
mRNA
Non-coding RNA
Complex transcript
Intron

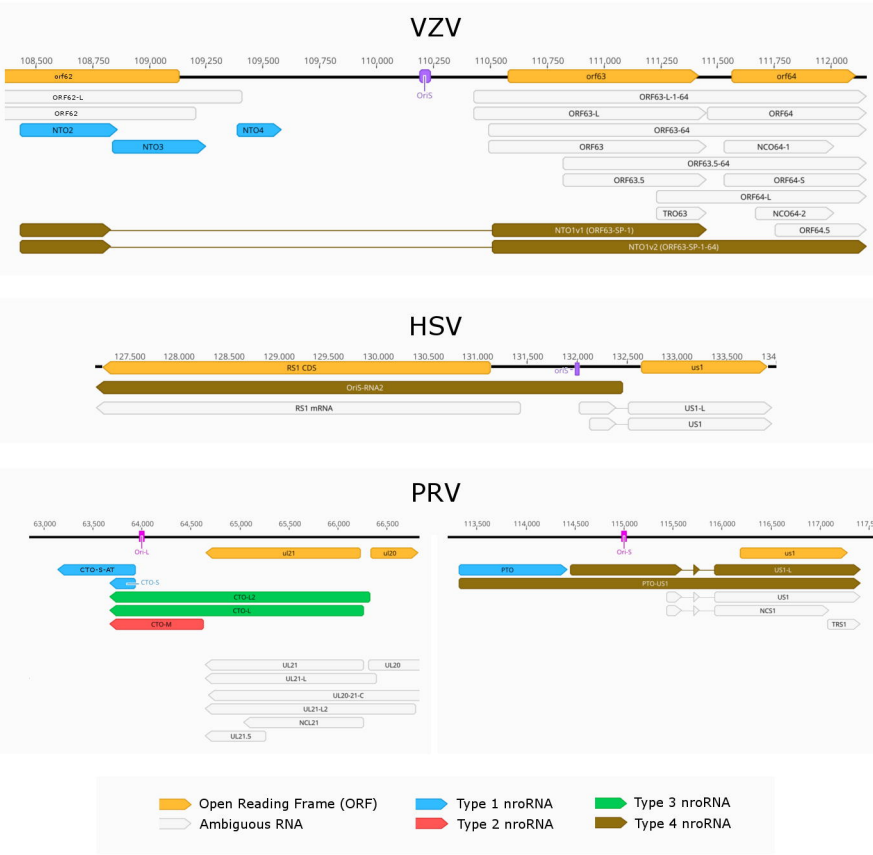

a. Length distribution of reads
b. The vicinity of Cap selected 5' ends
Frequency (log10)
Read length bins
Cap-selected
Non-cap-selected
Fraction
A
C
G
T
1–200
201–400
401–600
601–800
801–1000
1001–1200
1201–1400
1401–1600
1601–1800
1801–2000
2001–2200
2201–2400
2401–2600
2601–2800
2801–3000
3001–3200
3201–3400
3401–3600
3601–3800
3801–4000
4001–4200
4201–4400
4401–4600
4601–4800
4801–5000
5001–5200
5201–5400
5401–5600
5601–5800
5801–6000
6001–6200
6201–6400
6401–6600
6601–6800
6801–7000
7001–7200
7201–7400
7401–7600
7601–7800
-5
-4
-3
-2
-1
0
+1
+2
+3
+4
+5


5′
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
3′
weblogo.berkeley.edu
TES

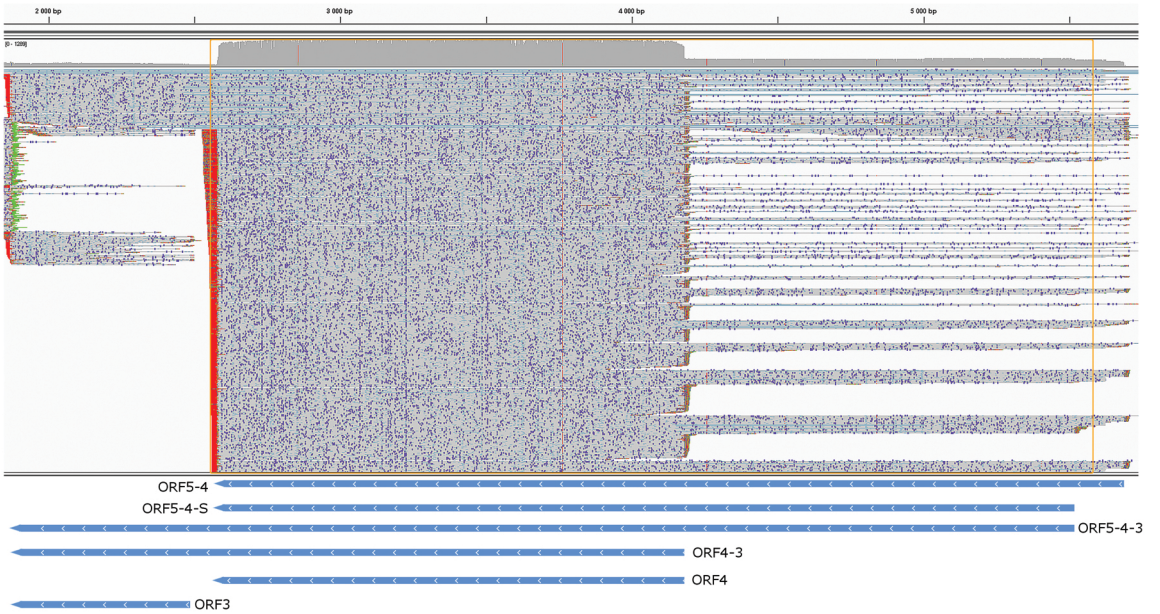

99,500
99,750
100,000
100,250
100,500
100,750
101,000
101,250
101,500
101,750
102,000
102,250
102,500
102,750
103,000
103,250
103,500
103,750
104,000
104,250
104,500
104,750
orf56
orf57
orf58
orf59
orf60
orf61
orf61.5
ORF61-60-59-58-57
ORF60-59-58-57
ORF61.5
ORF57
ORF60-59-AT
ORF61
ORF59-58-57
TRO61
ORF58-57
NCO58
VLT-1
VLT-2
VLT-4
VLT-3
Open reading frame (ORF)
Truncated ORF (tORF)
Upstream ORF (uORF)
mRNA
Non-coding RNA
Intron
Uncertain TSS


a
VZV
orf62
orf63
ORF62
ORF63
NCO64-1
ORF64
NTO2
NTO3
NTO4
NTO1v2 (ORF63-SP-1-64)
NTO1v1 (ORF63-SP-1)
NTO1v4 (ORF63-64-SP-3)
NTO1v6 (ORF63-64-SP-4)
NTO1v3 (ORF63-SP-3)
NTO1v5 (ORF63-SP-4)
oriS
b
HSV-1
RS1 CDS
RS1
us1
US1-L
US1
oriS
c
PRV
us1
PTO
US1
PTO-US1
US1-L
NCS1
TRS1
oriS
ORF
mRNA
ncRNA
uncertain TSS
intron

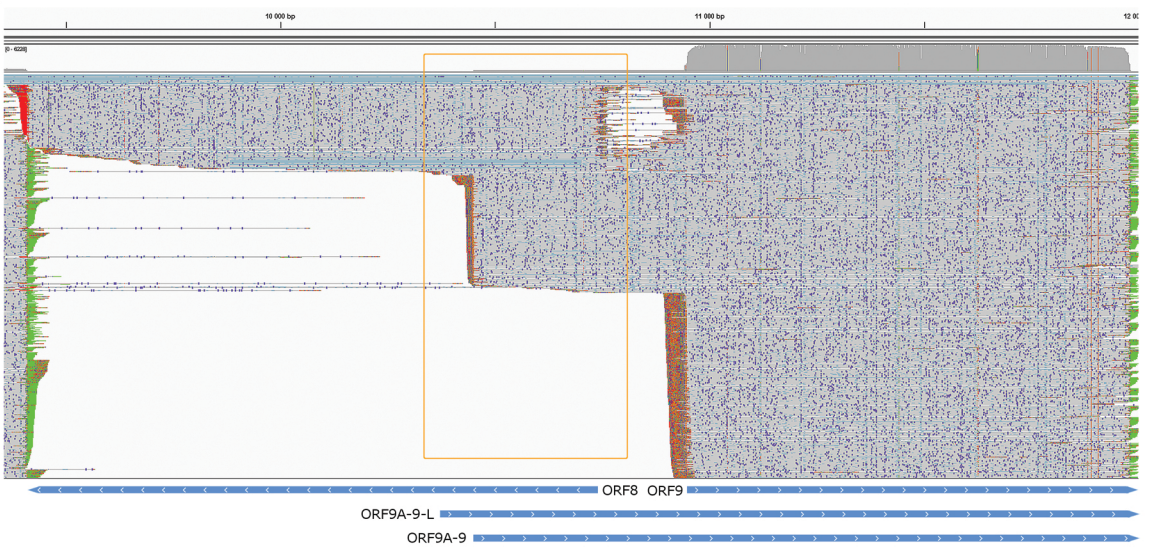

94,000
95,000
96,000
97,000
98,000
99,000
100,000
101,000
102,000
103,000
104,000
105,000
106,000
107,000
108,000
109,000
110,000
111,000
112,000
Ori5
orf53
orf54
orf55
orf56
orf57
orf58
orf59
orf60
orf61
orf62
orf63
orf64
orf65
vlt-orf63-2
vlt-orf63-1
VLT_ORF63-C6
VLT_ORF63-C5
VLT_ORF63-C4
VLT_ORF63-64-C1
VLT_ORF63-64-C2
VLT_ORF63-C1
VLT_ORF63-C2
VLT_ORF63-C3
ORF63-SP-2-64-C
ORF63-SP-2-C
VLT_1
VLT_2
VLT_4
VLT_3
Open reading frame (ORF)
Spliced ORF
Non-coding RNA
Complex transcript
Non-coding RNA
Intron
Uncertain TSS


VZV
108,500
108,750
109,000
109,250
109,500
109,750
110,000
110,250
110,500
110,750
111,000
111,250
111,500
111,750
112,000
orf62
OriS
orf63
orf64
ORF62-L
ORF62
NTO2
NTO3
NTO4
ORF63-L-1-64
ORF63-L
ORF64
ORF63-64
ORF63
NCO64-1
ORF63.5-64
ORF63.5
ORF64-S
ORF64-L
TRO63
NCO64-2
NTO1v1 (ORF63-SP-1)
ORF64.5
NTO1v2 (ORF63-SP-1-64)
HSV
127.500
128.000
128.500
129.000
129.500
130.000
130.500
131.000
131,500
132.000
132,500
133.000
133.500
134
RS1 CDS
OriS
us1
OriS-RNA2
RS1 mRNA
US1-L
US1
PRV
63,000
63,500
64,000
64,500
65,000
65,500
66,000
66,500
OriL
uI21
uI20
CTO-S-AT
CTO-S
CTO-L2
CTO-L
CTO-M
UL21
UL20
UL21-L
UL20-21-C
UL21-L2
NCL21
UL21.5
113,500
114,000
114,500
115,000
115,500
116,000
116,500
117,000
117,500
OriS
us1
PTO
US1-L
PTO-US1
US1
NCS1
TRS1
Open Reading Frame (ORF)
Ambiguous RNA
Type 1 nroRNA
Type 2 nroRNA
Type 3 nroRNA
Type 4 nroRNA

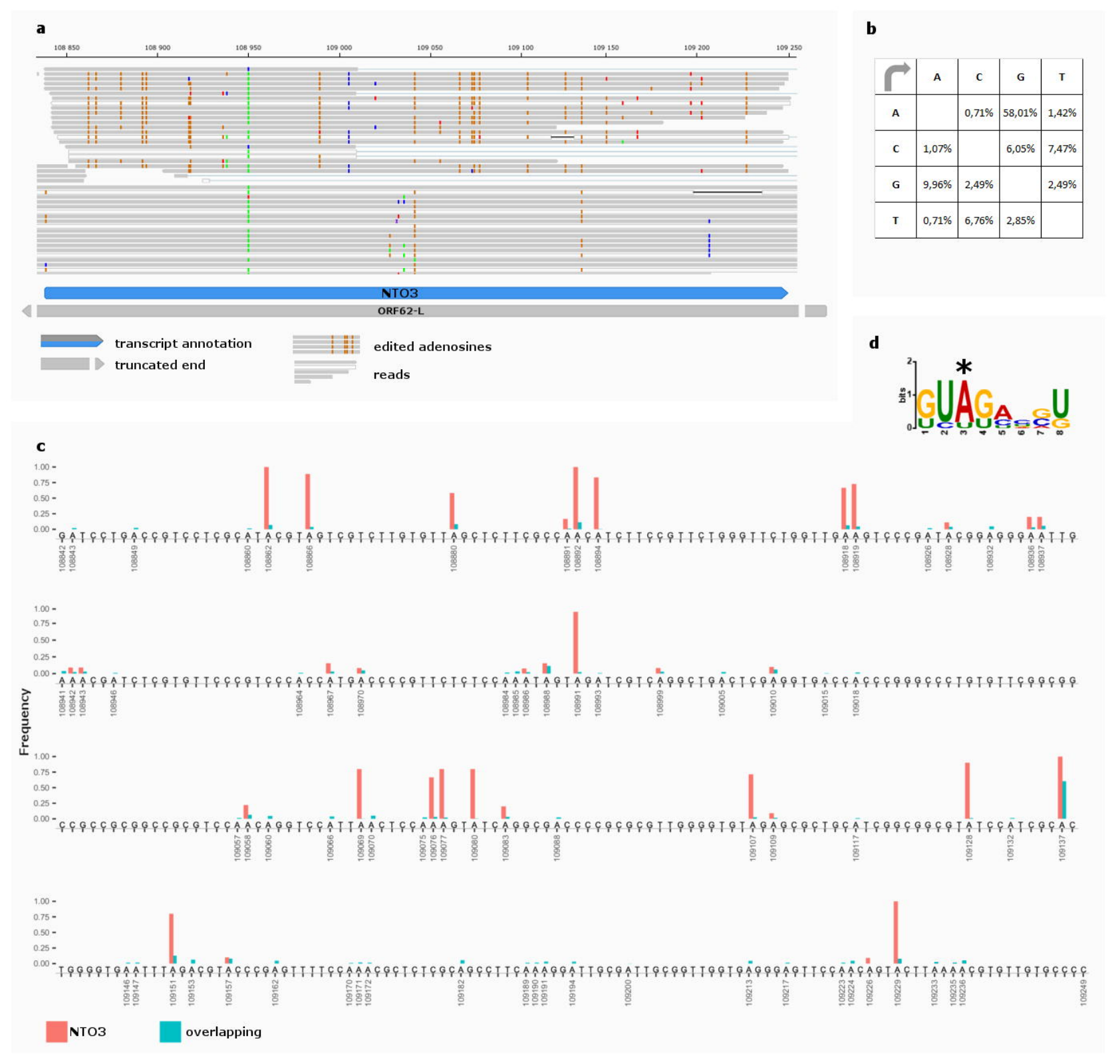
a



b



c


a
108 850
108 900
108 950
109 000
109 050
109 100
109 150
109 200
109 250
NTO3
ORF62-L
transcript annotation
truncated end
edited adenosines
reads
b
A
C
G
T
A
0,71%
58,01%
1,42%
C
1,07%
6,05%
7,47%
G
9,96%
2,49%
2,49%
T
0,71%
6,76%
2,85%
d
bits
c
Frequency
NTO3
overlapping