Thesis Booklet

Long Read Sequencing of the Varicella-Zoster Virus and Vaccinia Virus Transcriptome

István Prazsák M.Sc.

University of Szeged

Faculty of Medicine

Department of Medical Biology

Doctoral School of Interdisciplinary Medicine

Supervisors: Dóra Tombácz Ph.D and Zsolt Boldogkői Ph.D., Ds.C.

Szeged

2021

Publications directly related to the subject of the thesis

Tombácz, Dóra; Prazsák, István; Maróti, Zoltán; Moldován, Norbert; Csabai, Zsolt; Balázs, Zsolt; Dénes, Béla; Kalmár, Tibor; Snyder, Michael; Boldogkői, Zsolt: Long-read Assay Sheds New Light on the Transcriptome Complexity of a Viral Pathogen and on Virus-Host Interaction– SCIENTIFIC REPORTS 10: 13822 (2020)

IF: 3,998 Q1 (2019)

Boldogkői, Zsolt; Balázs, Zsolt; Moldován, Norbert; Prazsák, István; Tombácz, Dóra: Novel Classes of Replication-associated Transcripts Discovered in Viruses RNA BIOLOGY 16: 2 pp. 166-175., 10 p. (2019)

IF: 5,216 Q1 (2018)

Tombácz, Dóra, István Prazsák, Attila Szűcs, B Dénes, Michael Snyder, and Zsolt Boldogkői. 2018. "Dynamic Transcriptome Profiling Dataset of Vaccinia Virus Obtained from Long-Read Sequencing Techniques." GIGASCIENCE.

IF: 7,267 D1 (2017)

Tombácz, Dóra, István Prazsák, Norbert Moldován, Attila Szűcs, and Zsolt Boldogkői. 2018. "Lytic Transcriptome Dataset of Varicella-Zoster Virus Generated by Long-Read Sequencing." FRONTIERS IN GENETICS 9.

IF: 4,151 Q1 (2017)

Prazsák, István; Moldován, Norbert; Balázs, Zsolt; Tombácz, Dóra; Megyeri, Klára; Szűcs, Attila; Csabai, Zsolt; Boldogkői, Zsolt Long-read sequencing uncovers a complex transcriptome topology in Varicella-zoster virus BMC GENOMICS 19 Paper: 873 (2018)

IF: 3.73 Q1 (2017)

Prazsak, I, D Tombacz, A Szucs, B Denes, M Snyder, and Z Boldogkoi. 2018. "Full Genome Sequence of the Western Reserve Strain of Vaccinia Virus Determined by Third-Generation Sequencing." GENOME ANNOUNCEMENTS 6 (11).

new name: Microbiology Resource Announcements: IF:0 Q3

Abbreviations

asRNA: antisense RNA

cDNA: copy DNA

CTO: close to replication origin

CV-1: African green monkey (*Cercopithecus aethiops*) immortalized kidney cell line

cxRNA: complex RNA

lncRNA: long non-coding RNA

LoRTIA: Long-read RNA-Seq Transcript Isoform Annotator (toolkit)

LRS: long read sequencing

MOI: multiplicity of infection

MRC-5: human (*Homo sapiens*) immortalized lung fibroblast cell line

ncRNA: non-coding RNA

NGS: next generation of sequencing

nroRNA: near to replication origin RNA

NTO: near the transcription origin

ONT: Oxford Nanopore Technologies (Ltd)

ORF: open reading frame

Ori: origin of replication

OriS: origin of replication at US genomic region

p.i.: post infection (time point)

PacBio: Pacific Biosciences (Ltd)

PAS: polyA signal

sncRNA: small non-coding RNA

SRS: short read sequencing

TES: transcription end site

TIS: translation initiation site

TSS: transcription start site

VACV: Vaccinia virus

VACV-WR: Western Reserve (strain of VACV)

VZV: Varicella-zoster virus

Introduction

In order to understand how viruses are capable of conducting their life cycles and manipulating host cells it is important to know how their genes are expressed. Existing sequencing methods have limitations in capturing all the possible isoforms of transcripts produced by viruses. We circumvent those difficulties through the use of long-read sequencing (LRS). Short-read sequencing (SRS) technologies such as Illumina instruments(Goodwin, McPherson, and McCombie 2016), BGI(Jeon et al. 2019), or Ion Torrent sequencers(Rothberg et al. 2011) can produce reads which typically ranges between 75 and 300 bps. This means that complete viral mRNAs cannot be obtained in a single read, impairing the ability to deconvolute overlapping transcript isoforms. LRS technologies routinely generate reads in excess of 10 kb(Pollard et al. 2018). Another limitation of SRS is that RNA cannot be sequenced directly leaving modified nucleotides unknown(Harel et al. 2019). Two technologies currently dominate the LRS area, the Pacific Biosciences' (PacBio) single-molecule real-time (SMRT) sequencing and Oxford Nanopore Technologies' (ONT) nanopore sequencing. Transcriptome annotations are available for the human genome and the major model organisms but oversimplified for viruses. Subsequently, alternative transcript structures remain undetected, and the presence of overlapping transcription units can muddle expression level estimates and impair subsequent interpretations(Depledge, Mohr, and Wilson 2018).

Herpesviruses are a family of dsDNA viruses, divided in three subfamilies, i.e., α- (with HSV-1, HSV-2, and VZV), β- (CMV, HHV6, and HHV7) and γ-Herpesvirus (EBV and KSHV), that differ in their genetic content, infection sites and pathogenesis(De Pelsmaeker et al. 2018). Herpesviruses are among the most successful human pathogens; almost all member of the human population is infected by at least one species of Herpesviruses. Among the three main Herpesvirus subfamilies VZV has evolved similar strategies after primary infection to establish latency in dorsal root ganglia(Levin et al. 2003).

Reactivation can occur after a long period to cause painful epidermal lesion along dermatomes. VZV causes chickenpox during primary infection, and establishes life-long latency in ganglia, from where it can reactivate to cause a painful condition, herpes zoster (shingles)(Lungu et al. 1995).

The VZV genome consists of 125 kb of linear, double-stranded DNA (dsDNA) comprising one long (UL) and one short unique (US) region, each flanked by inverted repeats(Davison and Scott 1986). 71 open reading frames (ORFs) have been annotated in VZV, three of which are duplicated in the inverted repeat regions (*orf62, orf63, orf64*). The relative expressivity of many of these genes has been characterized by microarray and SRS in lytic phase(Baird et al. 2014; Cohrs, Hurley, and Gilden 2003). Functional sncRNAs playing role in regulation of replication has been found by SRS technique in VZV(Bisht et al. 2020; Markus et al. 2017). Viral latency and reactivation is of particular interest. It has been widely accepted, that latency associated transcripts regulate gene expression during the switch of the latent and lytic phase of Herpesviruses. Recently, one of the missing puzzle in VZV transcriptomics was found by SRS technique namely that a multiple spliced RNA, antisense to *orf61* in latently infected human trigeminal ganglia is expressed in neurons(Depledge, Ouwendijk, et al. 2018).

Vaccinia virus (VACV) is member of the family Poxviridae, which also includes the Variola virus, the causative agent of smallpox, which was one of the deadliest disease in the history. Active immunization with VACV led to the successful eradication of smallpox worldwide and since then, VACV has been used as a viral vector for the development of recombinant vaccines in gene therapies, cancer immunotherapies, and oncolytic therapies. Unlike Herpesviruses it replicates in the cytoplasm in cytoplasmic viral factories and its genome permits cloning large foreign sequences(Chard et al. 2015; Dénes et al. 2014; Hung et al. 2007).

VACV strain Western Reserve (VACV-WR) has linear, dsDNA genome of 195kbp, it is flanked by inverted terminal repeat (ITR) sequences which form covalently closed hairpin termini. VACV-WR contains 215 ORFs. Early studies on VACV transcriptome were performed with Northern-blot(Baldick and Moss

1993; Wittek and Moss 1982)combined with *in vitro* expression studies(Ahn et al. 1992; Ahn, Jones, and Moss 1990; Amegadzie, Ahn, and Moss 1991; Bajszár et al. 1983; Venkatesan, Gershowitz, and Moss 1982) and resolved expression profiles of single genes(Cooper and Moss 1979), while microarray data were used to analyze the time course changes of gene expression of all VACV ORFs and host genes(Rubins et al. 2008). SOLiD and Illumina deep sequencing platforms were also used to determine the precise 5′ and 3′ ends of early VACV RNAs(Yang et al. 2010a; Yang, Bruno, et al. 2011a). Late RNAs showed pervasive transcription initiation and termination(Yang et al. 2012) and these results, obtained with SRS techniques, suggest extensive read-through, particularly in the late phase of the infection(Yang et al. 2010a). SRS technique has detected distinctive cis-regulatory elements for early, intermediate and late promoters, additionally ribosome profiling identified several translation initiation sites(Yang et al. 2015).

Cap-selected and non-cap-selected sequencing libraries were used in our study to infer transcriptome structure of VZV and VACV. A variety of previously undescribed transcript isoforms, short, embedded transcripts, long transcripts overlapping many ORFs and new splice variants are described here, as well as confirmation of already known transcripts.

Aims of the study

Transcriptome of VZV and VACV – two members of large, DNA viruses – have been analyzed in detail by long read sequencing.

In order to characterize and reannotate the transcripts produced in lytic infection, and provide a new, comprehensive transcriptomic landscape of these viruses the aims of the study were the following:

-base-pair precision determination of 5’ ends of transcripts by cap-selected cDNA library sequencing;

-determination of 3’ ends of transcripts by sequencing oligod(T)-captured transcripts;

-validation and annotation of VZV and VACV RNAs by linking TSS and TES positions together;

-annotation of transcript variants, detection of transcript isoforms.

Nanopore sequencing enables the detection of complex network of intron-exon junctions, therefore new, alternative splicing detection in VZV transcriptome was also goal of the study.

Methods

Propagation of VZV, purification of RNA and sequencing of VZV's cDNA libraries

The live attenuated OKA/Merck strain of varicella zoster virus (VZV) was cultured at 37°C in human primary embryonic lung fibroblast cell line (MRC5), cells were then incubated at 37 °C for 5 days, when the cytopathic effect was near complete. Total RNA was isolated using the Nucleospin RNA Kit (Macherey-Nagel) according to the manufacturer's instructions. The cDNA library was prepared using the Ligation Sequencing kit (ONT SQK-LSK108) according to the modified 1D strand switching cDNA by ligation protocol then the polyA(+) RNA fraction was reverse transcribed (RT) using an oligo(d)T-containing primer.

For the precise determination of TSSs, the abovementioned 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); RT reaction was performed by using the oligo(d)T-containing primer. Adapter was ligated to the 5'C of the cap of the RNA by a double-strand specific ligase enzyme (Lexogen Kit). The Second-Strand Mix and the Enzyme Mix (both from the TeloPrime Kit) was used to produce double-stranded cDNAs according to the kit's guide. Then samples were amplified according by the TeloPrime Kit's manual. The PCR products were end-repaired, and then it was followed by adapter ligation using the sequencing adapters supplied in the kit. The cDNA sample was purified between each step using Agencourt AMPure XP magnetic beads (Beckman Coulter). Samples were loaded on R9.4 SpotON Flow Cells, while the base calling was performed using Albacore v2.1.10 software. Targeted long-read sequencing was also performed to validate the transcripts near to the replication origo of VZV. To evaluate the effect of hyper-editing on the secondary structure of RNAs, we used the 'RNAstructure Software' suite(Reuter and Mathews 2010). To simulate the presence of inosine, we changed the edited adenines to guanine. Reads were visualized using the Geneious (Kearse et al. 2012) software suite and IGV.

Propagation of VACV, RNA purification of VACV

African green monkey kidney fibroblast (CV-1) cells were infected with VACV-WR at 10 MOI/cell and were incubated at 37°C for 1,2,3,4,6,8,12 and 16 hours in 5% CO_2 atmosphere. Total RNA was extracted from the VACV infected cells in various stages of viral infection using the Macherey-Nagel RNA Kit, according to the Kit's manual. Poly(A) RNA fractions were isolated from the total RNA samples by using the Oligotex mRNA Mini Kit (Qiagen) following the protocol. The rRNAs were removed from the samples for the analysis of non-polyA RNAs using the Ribo-Zero Magnetic Kit (Illumina).

PacBio RSII & Sequel sequencing of VACV cDNAs

cDNAs were generated from the polyA(+) RNA fractions by applying the SMARTer PCR cDNA Synthesis Kit (Clontech), according to the PacBio 'Isoform Sequencing (Iso-Seq) using the Clontech SMARTer PCR cDNA Synthesis Kit and No Size Selection' protocol. The samples derived from different infection time points 1, 4, 8, and 12 hours p.i.) were mixed together for the RSII library preparation, however the time points (1, 2, 3, 4, 6, and 8 hours pi) were used individually for the production of libraries for the Sequel sequencing. A cDNA sample was prepared form an rRNA-depleted RNA mixture (from the 1, 4, 8, and 12 hours) with modified random hexamer primers instead of the SMARTer Kit's oligo(d)T-containing oligo. Samples were used for SMRTbell template preparation, using the PacBio DNA Template Prep Kit 1.0. The PacBio's MagBead Kit (v2) was used for the reaction. Sequencing movies were captured using the RSII and Sequel machines. The polyA(+) RNA fraction was used for cDNA sequencing on the ONT's MinION device. RNAs from different infection time points were converted to cDNAs according to the ONT 1D SQK-LSK108 protocol. From the samples an RNA mixture were prepared for sequencing but the time points were also sequenced individually. Libraries were generated by using the above mentioned 1D ligation kit and protocol, according to the kits' manuals, as it is described above.

Data analysis and alignment

Bioinformatical steps were performed in Ubuntu/Linux operational system in python3. Reads resulting from ONT sequencing were aligned to the reference genome of VZV (NC_001348.1) and the host cell genome (Homo sapiens - GRCh37; PRJNA31257) using GMAP v2017-04-24 (Wu and Watanabe 2005). The PacBio ROIs and the ONT raw reads of VACV were aligned to the reference genome of the VACV (LT966077.1) and to the host cell (*Chlorocebus sabaeus*; GCA_000409795.2) using minimap2 aligner. In order to annotate the 5' and 3' ends of reads, an algorithm was developed, which constituted the core algorithm of a new transcript annotation toolkit called LoRTIA (Long-read RNA-Seq Transcript Isoform Annotator). The detailed description of the computation is found in(Prazsák et al. 2018). Possible artefacts of false priming and template switching were excluded from the dataset.

Annotation of transcripts and splice junctions of VZV transcripts

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 labeled 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. The relative abundance of the transcript isoforms was calculated by dividing the number of reads with the total mapped read counts of the dataset. A ± 10nt-variation was allowed at both the TSS and TES of sequencing read for considering them as certain transcript isoforms. To assess the homology of protein products possibly translated from the ORFs of novel transcript isoforms we used the online BLASTP suite (Altschul et al. 1997), with an expected threshold of 10. 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.

Analysis of hyperediting

To evaluate the effect of hyper-editing on the secondary structure of specific VZV RNAs, we used the 'RNAstructure Software' suite(Reuter and Mathews 2010). To simulate the presence of inosine, we changed the edited adenines to guanine. Reads were visualized using the Geneious (Kearse et al. 2012) software suite and IGV.

Results of VZV analysis

We detected in the VZV experiments 614 192, 66 455, and 328 118 mapped reads from the cap-selected, non-cap-selected, and targeted methods respectively. The total number of host and viral reads was 10 338 565. The average read length was the longest in the non-cap-selected library of 1349bps. In total, we detected 1124 5'-ends and 255 3'-ends in the non-cap-selected read population, and 1428 5'-ends and 279 3'-ends in the cap-selected dataset. We analyzed the end positions of reads with LoRTIA transcript annotator tool to obtain TSS and TES positions. The number of 5'-ends qualifying as a putative TSS was 10.86% of the total 5'-end positions, while 32.95% of the total 3'-end positions turned out as TES. Furthermore, we excluded 49 5'-ends and 16 3'-ends from the non-cap-selected and 21 5'-ends and 16 3'-ends from the cap-selected datasets because they proved to be the products of false priming. Using the ONT MinION sequencing platform, we confirmed 18 previously known TSSs and nine TESs. Additionally, we could annotate 124 new putative TSSs and 71 TESs, if we used the common section of cap- and non-cap-selected methods.

New coding and non-coding transcripts and splice sites

Transcripts embedded into larger RNA molecules are easily detected by the LRS techniques. Our analysis reported the identification of 25 embedded viral transcripts containing5'-truncated ORFs. Eighteen of these *possibly protein-coding* VZV transcripts contain a canonical PAS. We found twenty-three novel non-coding transcripts, which are belonging to 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. We also detected one intergenic and seven antisense ncRNAs. Until now five genes had reported spliced variants. We identified 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. Furthermore, two novel splice variants encoded by the ORF42/45 gene were detected. Furthermore, we found ten (!) novel splice junctions of ORF63 splice variants coinciding the recently described VLTly splice variants. Using *in silico* methods, we have detected 44 potential (u)ORFs on the 5'-UTRs of 81 VZV transcripts.

nroRNAs, complex transcripts, and RNA editing

Our research group reported that Herpesviruses express transcripts located near the replication origins (nroRNA) (Boldogkői et al. 2019; Tombácz et al. 2015). We identified nine nroRNAs starting in the proximity of VZV OriS. The NTO1v1, NTO1v3, and NTO1v5 are the spliced long TSS variants of ORF63 producing a long fusion transcript. nroRNAs are present in other member of alpha Herpesviruses (*Figure 1.*). We identified 33 hitherto unknown, long polycistronic RNAs in VZV transcriptome, including four complex transcripts. Furthermore, 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 $\mathrm{VLT}_{\mathrm{ly}}$ isoforms.

One of the replication origo associated transcript, the NTO3 shows a very high frequency of A to G substitution, which is interestingly not present in the overlapping reads of this region. We found that 58% of all substitutions are A-

>G 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) making 22.07% of all As of the transcript edited. When forming an intramolecular *in silico* 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. This suggests a hyper-editing event in NTO3.

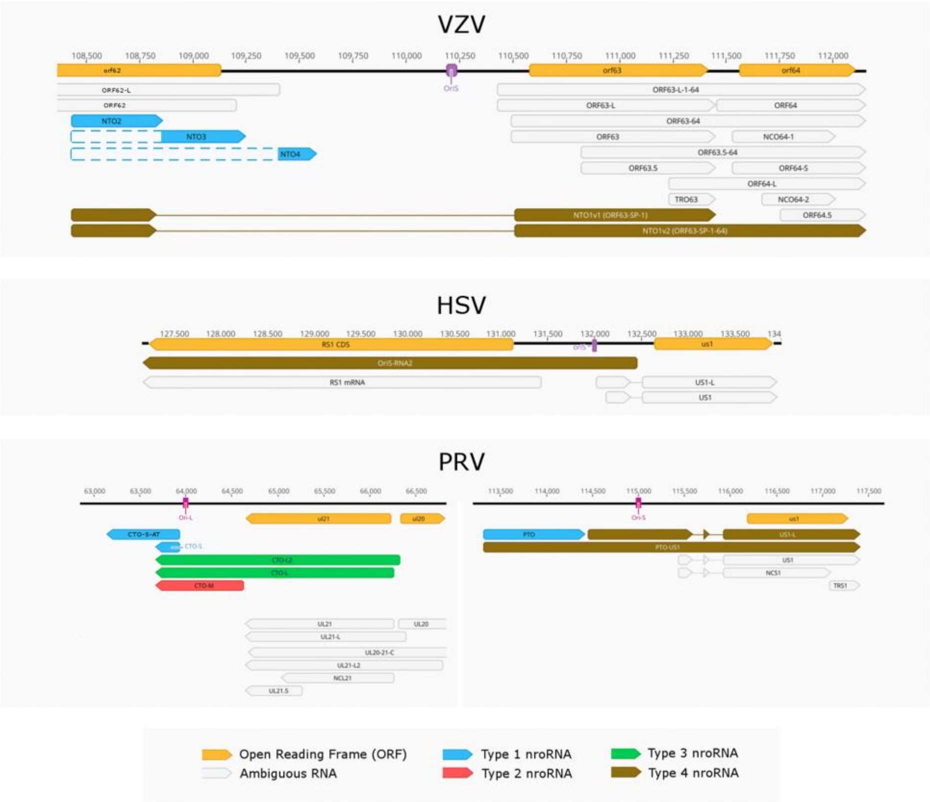

Figure 1. Types of near-replication-origin (nro)RNAs of three alphaherpesviruses. The nroRNAs can be classified in four distinct types according to their position to the replication origin and coding capacity. Type 1 nroRNAs are non-coding RNAs that do not overlap the Ori. Type 2 nroRNAs are ncRNAs that overlap the Ori. Type 3 nroRNAs are mRNAs with alternative TESs and long 3'UTRs and overlapping the Ori, while type 4 nroRNAs are mRNAs with alternative TSSs and long 5'UTRs overlapping the Ori

Results of VACV analysis

Due to the high read number and coverage, we reconstructed from transcriptome data the genome of VACV as following: viral reads were mapped to the reference genome of Vaccinia Virus Western Reserve (WR) strain (GenBank accession no. NC_006998). The mapped reads - sequenced by both PacBio and ONT techniques – have an average read count per each nucleotide 289 and 43 in the case of PacBio and ONT respectively. Every nucleotide of VACV genome was covered with reads especially in late time points of viral life cycle, e.g. 8h p.i. around 80%, and at 12h p.i. the entire VACV genome was transcriptionally active. We used the PacBio cDNA data for the reconstruction of VACV genome sequence. The length of the reconstructed genome sequence of the WR strain of VACV was composed of 194,888 base pairs. The average GC-content was 33.3%. WR strain differs in 163 point mutations from the reference genome, mainly in the F region.

To investigate the transcriptomes of VACV we performed polyadenylation sequencing techniques using PacBio isoform sequencing template preparation protocol (Iso-Seq method) for RSII and Sequel platforms, and using the ONT MinION device to sequence cDNAs. We isolated RNA from different time points of infection (at 1, 2 ,3 ,4 ,6 ,8 ,12 and 16 hours), therefore 25 independent sequencing library preparations were used for VACV transcriptome sequencing.

In total, more than 1 115 000 reads of inserts were generated using the PacBio platform. The nanopore sequencing methods yielded about 535 000 viral sequencing reads altogether. We obtained various mean read lengths among the sequenced libraries (1129±126.7 SD in PacBio SMRT and 619±148.2 SD in the ONT Minion samples). The ONT MinION 1D cDNA and the 8h PacBio Sequel libraries could produce the longest aligned reads (>6000bp).

TSSs and TESs

We annotated 1073 TSSs of VACV transcripts, 987 of these have not been previously detected. Eighteen percent of TSSs matched with base-pair precision to those described by Yang and colleagues(Yang et al. 2010b, 2015; Yang, Bruno, et al. 2011b; Yang, Reynolds, et al. 2011). This value increased to 70% when we allowed a ± 10-nt, and to 93% once we allowed a ± 30-nt precision interval for the location of TSSs. Altogether 898 TSSs have been detected in the 'regular' regions of VACV transcriptome. Our analysis revealed that VACV genes express transcripts with multiple shared TSS and TES positions. We found 8191 unique putative viral transcripts using the in house developed LoRTIA pipeline. From the 218 annotated ORFs 175 were expressed alone on separate RNA molecules, representing 1480 transcripts altogether. For the remaining ORFs we could not detect full-length transcripts; however, these regions contain several transcriptional reads lacking a precise TSS and/or TES position, therefore transcript annotation was almost insurmountable (O1L and A11R region). Therefore, we named these regions 'chaotic' and for the others we used the term 'regular' transcriptional region of the viral genome.

New coding and non-coding transcripts

At the 'regular' genomic regions we detected 135 monocistronic coding transcripts of previously annotated ORFs using the LoRTIA software suit, and 12 additional mRNAs did not meet the LoRTIA criteria due to their low abundance. In our study, altogether 292 different kinds of polycistronic transcript isoforms were detected: 43 bicistronic, 137 tricistronic, 92 tetracistronic, 15 pentacistronic, and 5 hexacistronic RNA molecules. We identified novel genes in new intergenic positions, embedded genes (5'-truncted in-frame ORFs within larger canonical ORFs), and short upstream ORFs (uORFs, preceding the main ORFs) *(Figure2.)*. Two genes at novel genomic positions were detected by annotating two novel putative mRNAs that are encoded in intergenic genomic locations: B25.5R/C19.5L in the repeat region and C9.5L in the vicinity of C9L gene. Our study identified 49 novel putative embedded genes with 5'-truncated in-frame ORFs which were shorter than the *in silico* annotated canonical ORFs. With few exceptions, we detected transcriptional activity from both DNA strands along the entire VACV genome. Expression level of mRNAs and antisense RNAs (asRNA) varied between genomic locations. In most cases, the level of asRNA expression was relative low. VACV genes stand in tandem orientation with each other, a few number of convergently and divergently positioned gene pairs are present in VACV genome, leading to few cxRNAs compared with those in Herpesviruses(Tombácz et al. 2016). In VACV, the LoRTIA annotated 30 cxRNA molecules localized within nine genomic regions.

Overlaps

Former studies identified several hundreds of TSSs and PASs, but uncertainty of 3'-ends of RNAs remained due to the high complexity of gene expression and overall read-throughs of upstream genes especially at late time points of infection. Using multiplatform analyses, we revealed an extremely complex meshwork of transcriptional overlaps, which were expressed by almost all viral genes. Altogether 154 tandem, 32 divergent, and 32 convergent gene pairs were found in the VACV genome, with 21 overlapping transcripts between tandem genes, 13 between convergent gene pairs, and 5 between divergent gene pairs. Moreover, closely-spaced tandem genes can form partial and full parallel overlaps.

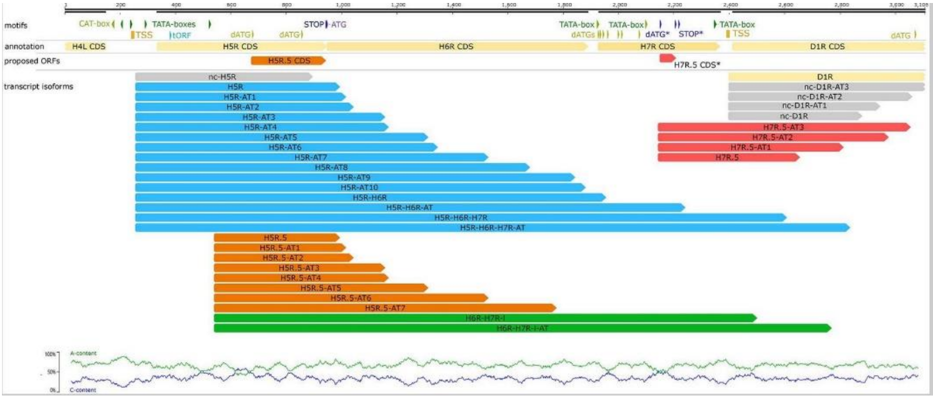

Figure 2. Schematic representation of two examples of the identified putative embedded genes. We detected a number of 5'-truncated transcripts containing short in-frame ORFs. This figure shows transcripts from H5R.5 and H7R.5 ORF regions.

Discussion and conclusion

We have designed experiments to obtain a detailed transcriptional landscape of RNA isoforms of VZV and VACV using ONT MinION sequencing technique combined with PacBio in the latter case. Cap-selected (for an accurate determination of TSSs) and non-cap-selected protocols were applied on the isolated polyA-selected samples. According to our results, both the VZV and the VACV genome is pervasively transcribed and almost all genomic regions are transcriptionally active in the late phase of infection.

Long read sequencing (LRS) technologies (PacBio and ONT) enable researchers to characterize transcriptional complexity including various transcription start sites (TSSs) and transcription end sites (TESs), splicing isoforms, polycistronic RNAs, and polyadenylation sites (PAS), long RNAs encompassing mRNAs, ncRNAs and potential new open reading frames (ORFs). As previously reported, despite their high accuracy, the conventional next generation sequencing (NGS) approaches failed in recovering full-length transcripts and complex transcript structures. LRS has the potential to accurately reflect the transcriptional complexity if proper precautions are taken from quality filtering of reads. The transcripts detected in our VZV experiments were partly isoforms of already known transcripts or previously undescribed transcripts of known genes, including non-coding transcripts and new splice variants. In the VACV experiments, the majority of the described transcripts and almost all of the isoforms were newly annotated. Our experiments revealed a high diversity of transcripts in both viruses definitely.

Although, native RNA sequencing currently suffers from some problems (e.g. lacking full length transcripts, higher error rates), it offers a new horizon in the future of transcriptomics, as a possible gold standard method to assess modifications, quantity and structural variations of transcripts, as it has been demonstrated recently in human and viral samples(Braspenning et al. 2020; Soneson et al. 2019). Our full-length cDNA sequencing approach has opened a new window in the field of Herpesvirus transcriptomics. At the moment only LRS provides a genome wide survey of rare and difficult to detect transcription events with base pair precision.

References

Ahn, B. Y., E. V. Jones, and B. Moss. 1990. "Identification of the Vaccinia Virus Gene Encoding an 18-Kilodalton Subunit of RNA Polymerase and Demonstration of a 5' Poly(A) Leader on Its Early Transcript." *Journal of Virology*.

Ahn, B. Y., J. Rosel, N. B. Cole, and B. Moss. 1992. "Identification and Expression of Rpo19, a Vaccinia Virus Gene Encoding a 19-Kilodalton DNA-Dependent RNA Polymerase Subunit." *Journal of Virology*.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25(17):3389–3402.

Amegadzie, B. Y., B. Y. Ahn, and B. Moss. 1991. "Identification, Sequence, and Expression of the Gene Encoding a M(r) 35,000 Subunit of the Vaccinia Virus DNA-Dependent RNA Polymerase." *Journal of Biological Chemistry*.

Baird, N. L., J. L. Bowlin, R. J. Cohrs, D. Gilden, and K. L. Jones. 2014. "Comparison of Varicella-Zoster Virus RNA Sequences in Human Neurons and Fibroblasts." *Journal of Virology* 88(10):5877–80.

Bajszár, G., R. Wittek, J. P. Weir, and B. Moss. 1983. "Vaccinia Virus Thymidine Kinase and Neighboring Genes: MRNAs and Polypeptides of Wild-Type Virus and Putative Nonsense Mutants." *Journal of Virology*.

Baldick, C. J. and B. Moss. 1993. "Characterization and Temporal Regulation of MRNAs Encoded by Vaccinia Virus Intermediate-Stage Genes." *Journal of Virology*.

Bisht, Punam, Biswajit Das, Paul R. Kinchington, and Ronald S. Goldstein. 2020. "Varicella-Zoster Virus (VZV) Small Noncoding RNAs Antisense to the VZV Latency-Encoded Transcript VLT Enhance Viral Replication." *Journal of Virology*.

Boldogkői, Zsolt, Zsolt Balázs, Norbert Moldován, István Prazsák, and Dóra Tombácz. 2019. "Novel Classes of Replication-Associated Transcripts Discovered in Viruses." *RNA Biology*.

Braspenning, Shirley E., Tomohiko Sadaoka, Judith Breuer, Georges M. G. M. Verjans, Werner J. D. Ouwendijk, and Daniel P. Depledge. 2020. "Decoding the

Architecture of the Varicella-Zoster Virus Transcriptome." *MBio*.

Chard, Louisa S., Eleni Maniati, Pengju Wang, Zhongxian Zhang, Dongling Gao, Jiwei Wang, Fengyu Cao, Jahangir Ahmed, Margueritte El Khouri, Jonathan Hughes, Shengdian Wang, Xiaozhu Li, Bela Denes, Istvan Fodor, Thorsten Hagemann, Nicholas R. Lemoine, and Yaohe Wang. 2015. "A Vaccinia Virus Armed with Interleukin-10 Is a Promising Therapeutic Agent for Treatment of Murine Pancreatic Cancer." *Clinical Cancer Research*.

Cohrs, Randall J., Michael P. Hurley, and Donald H. Gilden. 2003. "Array Analysis of Viral Gene Transcription during Lytic Infection of Cells in Tissue Culture with Varicella-Zoster Virus." *Journal of Virology* 77(21):11718–32.

Cooper, Jonathan A. and Bernard Moss. 1979. "In Vitro Translation of Immediate Early, Early, and Late Classes of RNA from Vaccinia Virus-Infected Cells." *Virology*.

Davison, A. J. and J. E. Scott. 1986. "The Complete DNA Sequence of Varicella-Zoster Virus." *Journal of General Virology* 67(9):1759–1816.

Dénes, Béla, Nadja Fodor, Andre Obenaus, and István Fodor. 2014. "Engineering Oncolytic Vaccinia Viruses for Non-Invasive Optical Imaging of Tumors." *The Open Biotechnology Journal*.

Depledge, Daniel P., Ian Mohr, and Angus C. Wilson. 2018. "Going the Distance: Optimizing RNA-Seq Strategies for Transcriptomic Analysis of Complex Viral Genomes." *Journal of Virology*.

Depledge, Daniel P., Werner J. D. Ouwendijk, Tomohiko Sadaoka, Shirley E. Braspenning, Yasuko Mori, Randall J. Cohrs, Georges M. G. M. Verjans, and Judith Breuer. 2018. "A Spliced Latency-Associated VZV Transcript Maps Antisense to the Viral Transactivator Gene 61." *Nature Communications* 9(1):1167.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. "Coming of Age: Ten Years of next-Generation Sequencing Technologies." *Nature Reviews. Genetics* 17(6):333–51.

Harel, Noam, Moran Meir, Uri Gophna, and Adi Stern. 2019. "Direct Sequencing of RNA with MinION Nanopore: Detecting Mutations Based on Associations." *Nucleic Acids Research*.

Hung, C. F., Y. C. Tsai, L. He, G. Coukos, I. Fodor, L. Qin, H. Levitsky, and T. C. Wu. 2007. "Vaccinia Virus Preferentially Infects and Controls Human and

Murine Ovarian Tumors in Mice.” *Gene Therapy*.

Jeon, Sol A., Jong Lyul Park, Jong Hwan Kim, Jeong Hwan Kim, Yong Sung Kim, Jin Cheon Kim, and Seon Young Kim. 2019. “Comparison of the MGISEQ-2000 and Illumina Hiseq 4000 Sequencing Platforms for RNA Sequencing.” *Genomics and Informatics*.

Kearse, Matthew, Richard Moir, Amy Wilson, Steven Stones-Havas, Matthew Cheung, Shane Sturrock, Simon Buxton, Alex Cooper, Sidney Markowitz, Chris Duran, Tobias Thierer, Bruce Ashton, Peter Meintjes, and Alexei Drummond. 2012. “Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data.” *Bioinformatics (Oxford, England)* 28(12):1647–49.

Levin, Myron J., Guang-Yun Cai, Michael D. Manchak, and Lewis I. Pizer. 2003. “Varicella-Zoster Virus DNA in Cells Isolated from Human Trigeminal Ganglia.” *Journal of Virology*.

Lungu, Octavian, Paula W. Annunziato, Anne Gershon, Susan M. Staugaitis, Deborah Josefson, Philip Larussa, and Saul J. Silverstein. 1995. “Reactivated and Latent Varicella-Zoster Virus in Human Dorsal Root Ganglia.” *Proceedings of the National Academy of Sciences of the United States of America*.

Markus, Amos, Linoy Golani, Nishant Kumar Ojha, Tatiana Borodiansky-Shteinberg, Paul R. Kinchington, and Ronald S. Goldstein. 2017. “Varicella-Zoster Virus Expresses Multiple Small Noncoding RNAs.” *Journal of Virology* 91(24).

De Pelsmaeker, Steffi, Nicolas Romero, Massimo Vitale, and Herman W. Favoreel. 2018. “Herpesvirus Evasion of Natural Killer Cells.” *Journal of Virology*.

Pollard, Martin O., Deepti Gurdasani, Alexander J. Mentzer, Tarryn Porter, and Manjinder S. Sandhu. 2018. “Long Reads: Their Purpose and Place.” *Human Molecular Genetics*.

Prazsák, István, Norbert Moldován, Zsolt Balázs, Dóra Tombácz, Klára Megyeri, Attila Szűcs, Zsolt Csabai, and Zsolt Boldogkői. 2018. “Long-Read Sequencing Uncovers a Complex Transcriptome Topology in Varicella Zoster Virus.” *BMC Genomics* 19(1):873.

Reuter, Jessica S. and David H. Mathews. 2010. “RNAstructure: Software for RNA Secondary Structure Prediction and Analysis.” *BMC Bioinformatics* 11(1):129.

Rothberg, Jonathan M., Wolfgang Hinz, Todd M. Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H. Leamon, Kim Johnson, Mark J. Milgrew, Matthew

Edwards, Jeremy Hoon, Jan F. Simons, David Marran, Jason W. Myers, John F. Davidson, Annika Branting, John R. Nobile, Bernard P. Puc, David Light, Travis A. Clark, Martin Huber, Jeffrey T. Branciforte, Isaac B. Stoner, Simon E. Cawley, Michael Lyons, Yutao Fu, Nils Homer, Marina Sedova, Xin Miao, Brian Reed, Jeffrey Sabina, Erika Feierstein, Michelle Schorn, Mohammad Alanjary, Eileen Dimalanta, Devin Dressman, Rachel Kasinskas, Tanya Sokolsky, Jacqueline A. Fidanza, Eugeni Namsaraev, Kevin J. McKernan, Alan Williams, G. Thomas Roth, and James Bustillo. 2011. "An Integrated Semiconductor Device Enabling Non-Optical Genome Sequencing." *Nature*.

Rubins, Kathleen H., Lisa E. Hensley, George W. Bell, Chunlin Wang, Elliot J. Lefkowitz, Patrick O. Brown, and David A. Relman. 2008. "Comparative Analysis of Viral Gene Expression Programs during Poxvirus Infection: A Transcriptional Map of the Vaccinia and Monkeypox Genomes." *PLoS ONE*.

Soneson, Charlotte, Yao Yao, Anna Bratus-Neuenschwander, Andrea Patrignani, Mark D. Robinson, and Shobbir Hussain. 2019. "A Comprehensive Examination of Nanopore Native RNA Sequencing for Characterization of Complex Transcriptomes." *Nature Communications*.

Tombácz, Dóra, Zsolt Csabai, Péter Oláh, Zsolt Balázs, István Likó, Laura Zsigmond, Donald Sharon, Michael Snyder, and Zsolt Boldogkoi. 2016. "Full-Length Isoform Sequencing Reveals Novel Transcripts and Substantial Transcriptional Overlaps in a Herpesvirus." *PLoS ONE* 11(9).

Tombácz, Dóra, Zsolt Csabai, Péter Oláh, Zoltán Havelda, Donald Sharon, Michael Snyder, and Zsolt Boldogkői. 2015. "Characterization of Novel Transcripts in Pseudorabies Virus." *Viruses* 7(5):2727–44.

Venkatesan, Sundararajan, Alan Gershowitz, and Bernard Moss. 1982. "Complete Nucleotide Sequences of Two Adjacent Early Vaccinia Virus Genes Located Within the Inverted Terminal Repetition." *Journal of Virology*.

Wittek, R. and B. Moss. 1982. "Colinearity of RNAs with the Vaccinia Virus Genome: Anomalies with Two Complementary Early and Late RNAs Result from a Small Deletion or Rearrangement within the Inverted Terminal Repetition." *Journal of Virology*.

Wu, Thomas D. and Colin K. Watanabe. 2005. "GMAP: A Genomic Mapping and Alignment Program for MRNA and EST Sequences." *Bioinformatics (Oxford, England)* 21(9):1859–75.

Yang, Z., D. P. Bruno, C. A. Martens, S. F. Porcella, and B. Moss. 2011a. "Genome-Wide Analysis of the 5' and 3' Ends of Vaccinia Virus Early MRNAs Delineates

Regulatory Sequences of Annotated and Anomalous Transcripts.” *Journal of Virology*.

Yang, Z., D. P. Bruno, C. A. Martens, S. F. Porcella, and B. Moss. 2011b. “Genome-Wide Analysis of the 5’ and 3’ Ends of Vaccinia Virus Early MRNAs Delineates Regulatory Sequences of Annotated and Anomalous Transcripts.” *Journal of Virology*.

Yang, Z., S. E. Reynolds, C. A. Martens, D. P. Bruno, S. F. Porcella, and B. Moss. 2011. “Expression Profiling of the Intermediate and Late Stages of Poxvirus Replication.” *Journal of Virology*.

Yang, Zhilong, Daniel P. Bruno, Craig A. Martens, Stephen F. Porcella, and Bernard Moss. 2010a. “Simultaneous High-Resolution Analysis of Vaccinia Virus and Host Cell Transcriptomes by Deep RNA Sequencing.” *Proceedings of the National Academy of Sciences of the United States of America*.

Yang, Zhilong, Daniel P. Bruno, Craig A. Martens, Stephen F. Porcella, and Bernard Moss. 2010b. “Simultaneous High-Resolution Analysis of Vaccinia Virus and Host Cell Transcriptomes by Deep RNA Sequencing.” *Proceedings of the National Academy of Sciences of the United States of America*.

Yang, Zhilong, Shuai Cao, Craig A. Martens, Stephen F. Porcella, Zhi Xie, Ming Ma, Ben Shen, and Bernard Moss. 2015. “Deciphering Poxvirus Gene Expression by RNA Sequencing and Ribosome Profiling.” *Journal of Virology*.

Yang, Zhilong, Craig A. Martens, Daniel P. Bruno, Stephen F. Porcella, and Bernard Moss. 2012. “Pervasive Initiation and 3′-End Formation of Poxvirus Postreplicative RNAs.” *Journal of Biological Chemistry*.

This work was supported by ÚNKP-20-4 -SZTE 140 New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund.

Nemzeti Kutatási, Fejlesztési
És Innovációs Hivatal

Új Nemzeti
Kiválósági Program