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

Ph.D. Thesis

István Prazsák M.Sc.

University of Szeged

Faculty of Medicine

Department of Medical Biology

Doctoral School of Interdisciplinary Medicine

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

Szeged

2020

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

Publications indirectly related to the subject of the thesis

István Prazsák, Norbert Moldován, Zsolt Csabai, Ákos Harangozó, Zsolt Balázs, Attila Szűcs, Klára Megyeri, Dóra Tombácz and Zsolt Boldogkői: Hidden complexity of the Varicella-Zoster Virus transcriptome revealed by long-read sequencing Acta Microbiologica et Immunologica Hungarica 2019. 18th International Congress of the Hungarian Society for Microbiology

IF: 1.079 Q3 (2018)

Dora, Tombacz, Prazsak Istvan, Torma Gabor, Szucs Attila, Snyder Michael, and Boldogkoi Zsolt. 2018. “Comprehensive Analysis of Vaccinia Virus Transcriptome by Third-Generation Sequencing.” VIROLOGY & MYCOLOGY 7 (1): 61–61.

IF: 0

Olah, P, D Tombacz, N Poka, Z Csabai, I Prazsak, and Z Boldogkoi. 2015. “Characterization of Pseudorabies Virus Transcriptome by Illumina Sequencing.” BMC MICROBIOLOGY 15. doi:10.1186/s12866-015-0470-0.

IF: 2,581 Q1

Takács, IF, D Tombácz, B Berta, I Prazsák, N Póka, and Z Boldogkoi. 2013. “The ICP22 Protein Selectively Modifies the Transcription of Different Kinetic Classes of Pseudorabies Virus Genes.” BMC MOLECULAR BIOLOGY 14: 1–12.

IF: 2,057 Q2

Publications not related to the subject of the thesis

Olasz, F., Tombácz, D., Torma, G., Csabai, Z., Moldován, N., Dörmő, Á., Prazsák, I., Mészáros, I., Magyar, T., Tamás, V., Zádori, Z., & Boldogkői, Z. Short and Long-Read Sequencing Survey of the Dynamic Transcriptomes of African Swine Fever Virus and the Host Cells. FRONTIERS IN GENETICS: 11, 758. (2020)

IF: 3,789 Q1 (2019)

Gál, János, Gábor Kovács, Richárd Bagyó, Gábor Vári, and István Prazsák. 2016. “A New Loureedia Species on Overgrazed Former Cork Oak Forest in Morocco (Araneae: Eresidae).” ANIMAL WELFARE ETHOLOGY AND HOUSING SYSTEMS 12 (1): 11–19.

IF: 0

Kovacs, G, I Prazsak, J Eichardt, G Vari, and H Gyurkovics. 2015. "A New Ladybird Spider from Hungary (Araneae, Eresidae)." ZOOKEYS (494): 13–30.

IF: 0.938 Q2

Prazsák, István, and Péter Maróy. 2008. "Study of the Genetic Relationships of Oribatid Mites (Acari, Oribatida) Using Nucleotide Sequences." ACTA BIOLOGICA SZEGEDIENSIS 52 (2): 347–348.

IF: 0 Q4

Puskás, I, I Prazsák, A Farsang, and P Maróy. 2008a. "Antropogén Hatásra Módosult Fizikai, Kémiai És Biológiai Tulajdonságok Értékelése Szeged És Környéke Talajaiban." AGROKÉMIA ÉS TALAJTAN 57 (2): 261–280.

IF: 0 Q4

Puskás, I, I Prazsák, A Farsang, and P Maróy 2008b. "Physical, Chemical and Biological Aspects of Human Impacts on Urban Soils of Szeged (SE Hungary)." JOURNAL OF ENVIRONMENTAL GEOGRAPHY 1 (1–2): 11–21.

IF: 0 Q4

Table of content

Introduction	1
Techniques for Transcriptome Analysis.....	2
The <i>next generation</i> sequencing technique	3
Short read sequencing techniques	4
Illumina	5
Long read sequencing techniques.....	6
Nanopore sequencing	7
Characteristics of VZV	9
Etymology and history of chickenpox.....	9
Structure and biology of VZV	10
Pathogenesis of VZV.....	11
Latency	12
Reactivation from latency	12
Clinical complications, prevention and therapy	13
Transcriptome research of VZV	13
Characteristics of VACV.....	14
History and origin of Vaccinia virus	15
Etymology of smallpox and vaccinia	16
Vaccines used for the eradication of smallpox.....	16
Genome and gene organization of VACVs	17
Life cycle and replication of VACV	18
Transcriptomics of VACV	18
Aims of the study	20
Methods.....	21
Propagation of VZV and purification of RNA	21
Oxford Nanopore MinION oligo(d)T sequencing and barcoding of VZV's cDNA	21
Cap selection, cDNA synthesis and sequencing.....	21
Targeted sequencing of NTO transcripts.....	22

Data analysis and alignment.....	22
Annotation of transcripts and splice junctions	22
Analysis of hyperediting	23
Cell culture and propagation of VACV; RNA purification of VACV	23
PacBio RSII & Sequel sequencing.....	23
ONT MinION platform – cDNA sequencing of polyA(+) RNAs.....	24
ONT MinION platform – Library preparation from Cap-selected samples	24
Analysis of reads, mapping and annotation of TSS and TES positions and transcripts	24
Nomenclature of transcripts	25
Prediction of <i>cis</i> -regulatory sequences.....	25
Results of VZV analysis.....	26
Structure of VZV transcriptome; number of sequenced reads in VZV transcriptome	26
Non-coding transcripts of VZV transcriptome.....	29
5'- and 3'-UTR isoforms of VZV transcriptome.....	29
Splice sites and splice isoforms in VZV transcriptome.....	30
Near-replication-origin (nro)RNAs – a novel class of transcripts.....	32
Polycistronic and complex transcripts.....	32
Transcriptional overlaps	33
Upstream ORFs	33
RNA editing	34
Results of VACV analysis.....	34
Genome reconstruction of VACV	34
Long-read transcriptome sequencing of VACV	35
Transcription start and end sites (TSSs & TESs)	36
Nomenclature and annotation of novel transcripts	37
Novel mono-, bi- and polycistronic RNAs.....	38
Novel putative protein-coding genes.....	38
Antisense RNAs	40
Intragenic non-coding transcripts	40
Replication-associated transcripts	40

Miscellaneous, complex transcripts.....	41
Cis-regulatory sequence prediction of novel transcripts	41
Transcriptional overlaps	42
Discussion	43
Genome reconstruction.....	43
Transcriptome annotation.....	43
Multigenic transcripts.....	45
Splicing.....	47
RNA editing	48
nroRNAs.....	48
Conclusion.....	50
Acknowledgements	51
Supplementary documents of VZV	52
Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus	52
Supplementary documents of VACV	61
Long-read Assays Shed New Light on the Transcriptome Complexity of a Viral Pathogen.....	61
References	80

Abbreviations

.5 : embedded (transcript isoform)

as- : antisense (transcript)

asRNA: antisense RNA

ASFV: African swine fever virus

bp: base pair

c- : complex (transcript isoform)

cDNA: copy DNA

CTO: close to replication origin

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

cxRNA: complex RNA

dA-tailing: deoxyadenosine tailing

dNTP: deoxynucleoside triphosphate

dRNA: direct RNA

E: early (transcript)

EHV: equine herpesvirus

FBS: Fetal bovine serum

HCMV: human cytomegalovirus

HHV-3: human herpesvirus type 3

HSV: Human herpesvirus

I: intermediate (transcript)

IGV: Integrative Genome Viewer (software)

ITR: inverted terminal repeat

-l : long (transcript isoform)

L: late (transcript)

lncRNA: long non-coding RNA

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

MAPQ: mapping quality (score)

MOI: multiplicity of infection

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

LRS: long read sequencing

mRNA: messenger RNA

ncRNA: non-coding RNA

NGS: next generation of sequencing

nroRNA: near to replication origin RNA

nt(s): nucleotide(s)

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 (timepoint)

PacBio: Pacific Biosciences (Ltd)

PAS: polyA signal

PBS: phosphate-buffered saline solution

PCR: polymerase chain reaction

PR: post replicative

PRV: pseudorabies virus

PTO: proximal to origin (of replication)

qPCR: quantitative polymerase chain reaction

rDNase: recombinant DNase I

ROI: reads of insert

RPMI medium: Roswell Park Memorial
Institute medium (for cell cultures)

rRNA: ribosomal RNA

R_t : normalized relative expression ratio of a
given transcript in a given time point of
infection

R_t : relative, normalized expression ratio of
transcripts in a given time point (t)

RT reaction: reverse transcription reaction

RT: reverse transcription

-s : short (transcript isoform)

SCID-hu mouse — an immunodeficient mouse
transplanted with relevant organs of the human
immune system

SMRT cell: Single Molecule, Real-Time
Sequencing (chip)

sncRNA: short non-coding RNA

SOLiD: Sequencing by Oligonucleotide
Ligation and Detection

SRS: short read sequencing

ssDNA: single-stranded DNA

ssRNA: single-stranded RNA

TES: transcription end site

tr- : truncated (transcript)

tRNA: transfer RNA

TSS: transcription start site

UL: unique long (region of the genome)

US: unique short (region of the genome)

UTR: untranslated region (of a mRNA)

VACV: Vaccinia virus

VF: viral factory

VZV: Varicella zoster virus

WR: Western Reserve (strain of VACV)

Introduction

Throughout our history, humans have been battling viruses. Many viral infections had large impacts on human societies; whole populations, perhaps even cultures have been destroyed by infective agents. The first outbreaks of viral epidemics were probably closely related to the increased number of human communities and to agricultural development in the Neolithic period around 10.000 years ago^{1,2}. The continuous interplay between host and viral pathogens has recently been recognized as a key driver of mammalian evolution³. After a pandemic or epidemic, the infected population goes extinct or tries to adapt to the new conditions. New findings suggest that the virome of mammals has influenced the genetic diversity of hosts and an astonishing number of gene product adaptations that have developed since humans' divergence from chimpanzees have been driven by coevolved viruses^{3,4}. It is worth noting – although retroviruses are beyond of the scope of this thesis - that the genomes of animals are saturated with many inactive viral copies (retrotransposons and retroposons). Many studies suggest that retroviruses have had an impact on the development of the mammalian uterus, enabling the fine tuning of placental gene expression^{5,6}. We also know that retroviral elements offered adaptive advantages in the brain development of early human species, among other animals^{7,8}.

Generally, the genome keeps information for replication, transcription and translation of genes. The transcriptome of viruses is the whole bulk of RNAs produced by the genome and it is of particular interest in virology. The transcriptome represents the activity of the whole genome at a certain stage in a viral life cycle or functional state. Transcripts are the products of transcription, including messenger RNA (mRNA) and non-coding RNA (ncRNA). The mRNA, a single-stranded intermediate molecule, is considered to be the molecule that allows the conversion of genetic information from DNA to protein, while ncRNA influences gene expression, protein synthesis, transport, chromosome integrity and other fundamental processes at various levels⁹. Therefore, the study of the transcriptome promotes understanding of the activities of cells, tissues and organisms. Transcriptome sequencing, also known as RNA-Seq, is a recently developed technique that quantifies the transcriptome in given samples and thus promotes transcriptome analysis. Viral RNAs regulate, control and enable the propagation of new viral particles. Therefore, they play a crucial role in the viral life cycle. Nowadays, revolutionary novel technical achievements in the area of sequencing offer completely new insight into the transcriptomes of many organisms, including viruses^{10,11}. To become acquainted with the structure of a viral transcriptome, it is prerequisite to understand the operation of viruses.

The genomes of DNA viruses are structurally either double-stranded or single-stranded¹². Double-stranded DNA (dsDNA) viruses are a diverse group of viruses, including many pathogenic agents. A relatively large genome and plenty of viral genes are characteristic features of the dsDNA viruses. Biological information is organized in a sequential manner and is packaged in the nucleocapsid of viruses. In the host cell, genetic information is arranged on a series of chromosomes in the nucleus. With

the discovery of the structure of DNA and RNA, there was a great desire to decipher this information, which was considered to be the genetic blueprint of living organisms.

Sequencing is the method to decode the nucleic acids, the so-called sequence of nucleotides. Cracking the code of life was a long-awaited dream in science and had been seen as an almost impossible task when Holley and colleagues were able to first determine (in 1965) the structure and serial nucleotide composition of a transfer RNA¹³. Two years later the first DNA molecule pieces were determined at the end of a lambda phage genome¹⁴. A couple of years later Walter Fiers and his team sequenced the first gene, a viral coat protein coding gene, and in 1976 the complete nucleotide-sequence of MS2 bacteriophage, the first RNA genome and transcriptome ever to be sequenced was published^{15,16}. Historically, transcriptome sequencing preceded the sequencing of the first DNA genome, which was the genome of phage phiX174¹⁷, by two years.

Techniques for Transcriptome Analysis

Analysis of transcripts at a molecular level was made possible by the discovery of reverse transcriptase (RT) and, later, the development of the polymerase chain reaction (PCR)^{18–21}. Combining these two techniques was a hallmark for transcriptome studies. mRNA is thought to be more susceptible to hydrolysis than DNA, because of the presence of ribose's 2'-hydroxyl group²². Therefore, producing complementary DNA copies from any mRNA molecule by *in vitro* reverse transcription, allows mRNA analysis without the degradation of samples. *In vitro* gene expression studies were time consuming and laborious, in spite of precise techniques for the analysis of single gene products. The technical development of reverse transcription and PCR has increased the quantity and isoforms of viral transcripts, including Herpesviruses or Poxviruses, being analysed^{23–27}. A drawback of these techniques is that they cannot measure the entire transcriptome at a single time point, only parts of it.

Hybridization reactions between single-stranded nucleic acids without enzymes have long been known^{28–30}. Using this phenomenon to develop a hybridization assay on the surface of a miniaturized chip was the birth of transcriptomics. A two dimensional cDNA array, later called a microarray, of known nucleotide sequences, utilizing the principles of Watson-Crick base pairing of labeled samples, has provided an opportunity to monitor viral genes or the gene expression of many cellular genes^{31–33}. This technique was commercialized by the company Affymetrix - cofounded by the Hungarian-rooted scientist Stephen Fodor³⁴. He used solid-phase oligonucleotide synthesis combined with automated photolithography, fluorescent-labeled nucleic acids and massive bioinformatics to implement microarray data analysis in large scale genome studies^{32,33}. Due to their size, it was relatively easy to produce whole genome arrays of all known and predicted open reading frames for a given viral pathogen³⁷. Microarray studies have been predominantly applied in the last three decades to measure and compare gene expression levels, including the global transcriptome changes of host cells during viral infections^{38–41}. Microarrays revolutionized genomics, clinical diagnostics and transcriptomics in

the '90s, although hybridization-based methods suffer from many erroneous factors and 'a priori' assumptions limiting their use in exhaustive, comprehensive whole transcriptomic and epigenomic studies⁴²⁻⁴⁴. Hybridization probes are designed based on the current annotation of a particular genome, so genomic regions that have no probe coverage are invisible. cDNA-derived expressed sequence tag (EST) libraries can circumvent the drawbacks of hybridization based methods, and catch unknown mRNAs as well, but the difficulty of this technique is that each tag of isolated mRNA is transformed into bacteria, then sequenced one by one⁴⁵. Serial Analysis of Gene Expression (or SAGE) is a similar method. It uses a small piece of each mRNA (typically 14-21bp) as a tag, which is reverse transcribed, biotinylated and enzymatically fragmented into short cDNA, then they are isolated on streptavidin beads and ligated together⁴⁶. This library of concatenated sequence tags can be sequenced to annotate transcriptionally active parts of any genome⁴⁷. This technique was used in early host-pathogen transcriptome studies, e.g. numerous low abundance transcripts of Herpesviruses were detected⁴⁸. However, a pitfall of SAGE is that only small pieces of transcripts are represented from an mRNA, therefore a significant number of transcripts fail to map and are missing from the whole transcriptome analysis.

On the other hand, the rapid development of next generation sequencing (NGS) techniques seems to provide solutions to the problems of RNA microarrays, because NGS rely on direct sequencing of the entire transcriptome. These cost-effective, new, high-throughput techniques have shown the existence of hitherto unknown transcripts, transcript isoforms, splice events, and new cis-regulatory elements of the mammalian transcriptome^{49,50}. NGS operates with base-pair precision and provides an overall molecular snapshot of the transcriptomic landscape of gene expression even at single cell resolution⁵¹.

The next generation sequencing technique

In order to understand what *next generation* sequencing means, a brief description of *first-generation* sequencing technique is discussed. Determination of protein primary structure inspired many researchers to determine the primary structure of RNA and DNA in the '60s. As Frederick Sanger – who sequenced the first protein – pointed out, “one of the main difficulties in sequencing resides in the separation step: in almost all sequencing work, fractionation has been a crucial factor, and progress in sequencing has often depended on progress in fractionation methods”⁵². A tRNA from yeast was determined by Holley in 1965 using serial enzymatic cleavages followed by several chromatography steps¹³. The whole process took 3 years and more than 100 kg of yeast was used as the starting material to obtain 1g of pure, 77 nucleotide-long tRNA^{53,54}. The first landmark in genome sequencing was the meticulous work of Fiers and colleagues on the MS2 bacteriophage genome published in 1975¹⁵. It is a 4kb long positive single-stranded RNA (+ssRNA) genome, which was small enough for an early sequencing attempt, however working with RNA genomes is challenging due to the instable nature of ssRNA. After creating small pieces of the genome using overlapping enzymatic fragmentation, it took

almost 6 years to puzzle them together with 2-dimensional chromatography⁵⁵. By 1977, Sanger and Coulson had used the ‘plus and minus’ method to sequence the first DNA genome of phiX174 bacteriophage using radioactive labeled P-isotopes^{17,56}. This method allowed a 50-80 base sequence to be resolved in a single experiment, however it was laborious and time-consuming and was suitable for only single-stranded DNA⁵². Inspired by Kornberg’s experiments, Sanger and Coulson improved their sequencing method, which relied on stopping the polynucleotide synthesis with a nucleoside-analogue and they could read almost 500 bp of DNA at once^{52,57,58}. Sanger’s method has become the most commonly used technology to sequence DNA. Many improvements have been made in the intervening years e.g. fluorescently-labeled terminators instead of radioactively-labeled nucleotides made the detection of fragments safer, introduction of capillary gel electrophoresis provided high-confidence base calling (Figure 1A), recombinant new enzymes enabled sequencing up to 1000 bps⁵⁹. This technique was the basis for sequencing the entire human genome and transcriptome and has made the genomic revolution of medicine over the last four decades a possibility^{60,61}. Sanger sequencing is still in use today for sequencing individual fragments of DNA/cDNA. Despite its many advantages, it is not able to achieve low-cost, high-throughput sequencing, partly because detection and synthesis are separated during the sequencing process and it cannot handle millions of bases simultaneously.

Short read sequencing techniques

Following the terminology in the history of sequencing, next-generation sequencing (NGS) refers to second, third and fourth generation sequencing techniques. Nonetheless, other authors and I prefer to distinguish NGS based on the length of reads, namely short- and long-read sequencing techniques, because the latter nomenclature better depicts the essence of NGS⁶². Short-read sequencing (SRS) techniques have dramatically changed the market of whole genome sequencing in the last decade due to their extremely high cost-efficiency. The common principle of these techniques is *sequencing-by-synthesis* (except in the case of SOLiD - *sequencing-by-ligation* - system). The sequencing reaction takes place on the surface of a chip to which many thousands of samples can be attached, thus achieving massive parallel sample handling. SRS-techniques differ from first generation (or *chain termination*) sequencing techniques in that they do not need gel electrophoresis, because signal detection and sequence reading occur simultaneously, not separately, which results in highly efficient throughput.

The leading NGS-technique has been pyrosequencing. Its principles have been known since 1993, briefly: during the synthesis of the complement strand, pyrophosphate (PPi) is released and transformed into ATP. The ATP then activates the enzyme luciferase (Figure 1C), which produces a light signal indicating specific base incorporation⁶³. The technique utilizes the advantages of emulsion PCR, all reactions take place on beads in picoliter volumes. Pyrosequencing was commercialized by J. Rothberg’s company, 454, in 2003 and made it possible to sequence parts of the Neanderthal genome for the first time⁶⁴. Although pyrosequencing can read only 300-400 bp, the technique was well implemented in several genotyping assays and genome methylation analyses of Herpesviruses^{65–68} and

Vaccinia viruses⁶⁹. Marketing of the pyrosequencing platform was discontinued in 2013 when the technology became noncompetitive against other SRS platforms.

Illumina

Researchers struggled with the automated Sanger method to sequence the entire human genome, there was a need for a fast, high-throughput, but cheap sequencing technique. Partway through the Human Genome Project a new sequencing idea - invented by two Cambridge scientists, Balasubramanian and Klenerman - brought the Solexa company into being⁷⁰. Their technique uses adapter molecules instead of emulsion PCR. The adapters are passed over a lawn of complementary oligonucleotides bound to a solid phase flow cell. PCR produces a population of clusters from neighboring templates⁵⁹. The company's machine resequenced the PhiX174 genome in 2005. It was a landmark in the history of NGS platforms, therefore Solexa was purchased by Illumina, which has invested in the development of personalized genome sequencers⁷¹. Soon after, Illumina announced that they had succeeded at the \$1000 genome challenge in 2014, which meant a remarkably drastic fall in the cost of whole genome sequencing compared to the 13-year-long Human Genome Project with its \$1 billion bill⁷². The GenBank currently contains more than 36 petabytes (PB) of data and is expected to grow to 43 PB by 2023⁷³. The vast majority of the data have been derived from Illumina sequencing. The popularity of this method can be explained by its low-cost efficacy coupled with extreme sequence accuracy and powerful marketing. Sanger sequencing determines the sequence of any DNA by measuring the mass of the growing strand using chain terminating modified ddNTP-s. One limitation of this method is that it cannot read more than 800 bp long chains, it can only handle a limited number of parallel sequences and evaluation of the results is time consuming. Contrarily, Illumina produces only short reads in a massively parallel manner, so it is able to sequence whole genomes in a defined time period. The principle of this method is sequencing-by-synthesis by reversible chain termination^{74,75}. The DNA or cDNA population needs to be fragmented into 35-75 bp long portions prior to library preparation, then adapters are ligated to bind them to the surface of the so-called flow-cell, where bridge amplification is done. During sequencing the DNA polymerase incorporates one of the four fluorescently labeled modified nucleotides (the reversible terminators, which are protected at their 3' ends) then an image is taken to catch the light signals of the growing strand⁷⁴. The protection group is cleaved to allow incorporation of the next base; thereafter unincorporated nucleotides are washed away (*Figure 1B*). The flow cell is imaged in each cycle resulting in a quasi-error-free base-by-base determination of the complement strand. Although Illumina platforms provide diverse solutions for different, large scale applications, including genomics, metagenomics, whole genome transcriptomics, small RNA transcriptomics, methylation arrays, ribosome profiling, clinical diagnostics, they generate millions of reads and they are not ideal technologies in some ways: amplification bias, high GC-content, homopolymers, difficult sample preparation, long run time, expensive instruments and the generated short reads all hamper their future distribution⁷⁶.

Figure 1. Illustration of the main sequencing techniques. A) Sanger sequencing, B) Illumina sequencing, C) Pyrosequencing D) Ion semiconductor sequencing by Ion Torrent, E) Single-molecule real-time sequencing by Pacific Biosystems, F) Oxford Nanopore. From ResearchGate, approved by the author⁷⁷

One basic question in the technology of sequencing-by-synthesis is how to recognize nucleotide incorporation. There are alternative options to light signal detection in sequencing. The Ion Torrent instrument uses millions of high-sensitivity semiconductors, which are practically ultrasensitive pH meters (Figure 1D), that detect protons (H^+) released when new bases are incorporated into the growing strand during synthesis⁷⁸. This technique allows the reading of 200-600 bp long sequences, filling the gap in read length between SRS and LRS techniques^{74,79}.

Long read sequencing techniques

Although long read sequencing (LRS) platforms have significant technical differences, they routinely generate reads longer than 1000bp and they can even sequence ultra-long reads of 100 000bp, while

SRS techniques have a maximum of 600 bases⁸⁰. The LRS technology emerged in 2010, when Pacific Biosciences (PacBio) commercialized the single-molecule real-time (SMRT®) technology⁸¹ with the first publicly available, long read sequencing instrument. This technique is also based on the *sequencing-by-synthesis* principle, whereby the polymerase enzyme is affixed at the bottom of nano-sized, so-called zero-mode waveguide (ZMW) wells and the DNA is moving. In many thousands of ZMWs fluorescently-tagged nucleotides are incorporated into the newly synthesized strands of DNA (. *Figure 1E*). The reaction happens at picoliter volumes and a superfine optical detection system enables recording of a real time movie of light pulses corresponding to each base incorporated during synthesis⁸². Although accuracy of raw reads generated by PacBio is much lower (~15%) than that of Illumina sequencing, it can be increased. PacBio uses circular templates by ligating hairpin adaptors to both ends of target DNA molecules. As the polymerase copies the circularized template, the DNA is sequenced several times producing a long, concatemer read⁸³. Consensus reads show high accuracy and fidelity, which is comparable with the 99% accuracy value of Illumina sequencing⁸⁴. Our research group was among the first to successfully use the PacBio platform in the sequencing of viral genomes and transcriptomes, including Herpesviruses⁸⁵⁻⁸⁸ and Vaccinia virus⁸⁹.

Nanopore sequencing

The principle of nucleic acid sequencing through nanopores depends on sensitive transmembrane protein biosensors⁹⁰. The nanopore allows charged molecules to move across an insulating layer with the help of the electric potential (. *Figure 1F*). While molecules traverse the pore, the characteristic changes in current are measured and this signal is decoded as the sequence of nucleic acids^{79,91,92}. The technique can be categorized based on the choice of nanopore. Most current efforts implement nanopores derived from bacterial proteins embedded in a lipid membrane or synthetic polymer. The other group is the fully synthetic, solid state nanopores⁹³. Compared to biological nanopores, solid-state nanopores produce higher signal noise during sequencing, although they are physically more stable⁹². The idea of nanopore sequencing was conceived by David Deamer in 1989^{94,95}. Around the same time, Church and Hagan proposed the idea that, driven by electrophoresis, RNA or ssDNA molecules can traverse through a protein pore embedded in a lipid bilayer and the blockade of ionic current can be measured by high precision patch clamp amplifiers⁹⁶. Their idea was experimentally proven in 1996 when Deamer and Kasianowicz published that polynucleic acids can pass through a barrel-like protein called alpha-Hemolysin⁹⁷. Later it was reported that purine bases could be definitively distinguished from pyrimidine bases by their specific signal characteristics, thus achieving a form of sequencing that was totally different from the *sequencing by synthesis* principle⁹⁸. In 2005 Hagan Bayley founded a startup company in Oxford. Ten years later, Oxford Nanopore Technologies (ONT) announced the early access of nanopore sequencer's beta version for the scientific community to test their MinION device and soon received the first result via Twitter⁹⁰. Subsequent developments in nanopore methodology made it possible to commercialize nanopore sequencers. Besides the microelectronic and computational

developments, sequencing chemistry and protein pores had also been improved. The alpha-Hemolysine pore protein was replaced by the genetically engineered MspA protein with which sequencers could read kilobase pair (kbp) long, single polynucleotide chains with relatively high accuracy⁹⁵. Recently, MspA has been replaced with the new generation of genetically designed CsgG pore protein (bacterial amyloid-secreting transmembrane channel) which has greater sensitivity in signal-noise detection^{99,100}. For accurate signal detection, it is crucial to control how nucleic acid chains pass through the nanopore. Therefore several motor proteins have been developed from T7 polymerase, Phi-29 polymerase and, the recently used, Helicase308^{96,100,101}. The motor protein interacts with the nanopore in order ratchet the DNA/RNA chain, base by base, into the hole slowing the sequenced molecule and thereby increasing processivity of the signal recording (*Figure 2*).

Figure 2. From left to right, motor protein (green) binds dsDNA -or cDNA- anchored on the nanopore. Motor protein uses ATP to unwind the double helix, it „feeds” the single strand, base-by-base, into the nanopore, meanwhile the electric potential between the two sides of membrane pulls the ssDNA through the pore. Changes in electric current are detected as signals of the nucleotides in the chain. The motor protein dissociates and the pore is ready to receive another new molecule. In: Carlos et al. 2017¹⁰²

The massive evolution of nanopore sequencing techniques has made it possible to read ultra-long, complex sequences, like repeat regions of human chromosomes¹⁰³. Long read sequencing is known for its relatively high error rate, however the early reported 40-60% inaccuracy of raw reads is rapidly approaching nearly 95%-99% reading precision nowadays, depending on the flow cell's chemistry and base calling algorithm^{95,104,105}. Despite scientific skepticism¹⁰⁶, ONT nanopore sequencing has become widely accepted in genomics and transcriptomics as a cost-effective, high-throughput, single molecule, long-read sequencing technique. An outstanding advantage of ONT MinION sequencer is that it can sequence DNA in real time at single base resolution without the need for expensive optical instruments, or amplification (*Figure 2*). Additionally, it is pocket sized and is able to read DNA modifications. We can utilize the benefits of ONT sequencing particularly in transcriptomics. Full-length reads compensate for the relatively high error rates, therefore robust detection of splice isoforms, length isoforms, alternative start and stop sites even in overlapping, complex transcriptomic environments are available. Because the direction of reads mirrors their natural position, directionality of transcripts can be determined without extra elaboration. Furthermore, base modifications can easily be identified from

thorough analyses of raw signal files. A disadvantage of ONT sequencer is the above-mentioned relatively high error rate, which means that without high coverage the alignment of reads is sometimes not precise enough for variant calling. On the other hand, this phenomenon can be overlooked in transcriptomics studies, where, not the precise sequence of RNAs, but their length and position are the most important information in transcript discovery. We can claim that nanopore sequencing is the most powerful tool in our arsenal in structural transcriptomics studies. To the best of our knowledge, our research group reported the first application of this technology in the field of viral transcriptome research using different members of Herpesviruses^{11,85,107–110}.

In my thesis, I try to elucidate how sequencing viral RNA in lytic infections reveals new insights into viral transcriptomes and the complexity of the viral transcriptomic landscape. In the following chapters I focus on two members of large, double-stranded DNA viruses – namely Varicella-zoster and the Vaccinia virus - and the analysis of their detailed transcriptome is presented. I created and analyzed various transcriptome libraries to obtain genome wide information about transcript isoforms of the above mentioned viruses.

Characteristics of VZV

Etymology and history of chickenpox

The origin of the name chickenpox (or chicken pox) is unclear, but according to the most likely explanation, it distinguishes the milder infection from the more serious smallpox (variola) and syphilis (also known as the great pox). Pox is a variant spelling of pocks, the plural form of the Old English pock (=pustule from Latin) meaning a small blister or pimple on the skin containing pus¹¹¹. The first known and recorded use of the word chickenpox was in 1694¹¹². One possible explanation about the history of the word chickenpox is found in Aronson's work: in Shakespeare's time a chequeen (later named chicken) was a gold coin, which had been devalued. Perhaps the disease was inappreciable compared to smallpox or great pox, like the value of the chequeen, which was either used as cheap decoration on a dress (spangle) or in a hazard game as coins (called chicken stake)¹¹³. According to other explanations the name likely comes from chickens' association with weakness and wimpiness¹¹¹. The term Herpes (meaning to creep in Greek) has been used in medicine for centuries and it has been applied to at least seven different infectious skin conditions¹¹⁴. The first occurrence of the term "herpes zoster" was in the 1st century in the work of Celsus who described rounded skin lesions spreading in a serpentine fashion like a belt on the trunk¹¹⁴. The word "shingles" derives from the Latin word "cingulus", which means "belt or girdle", referring to the characteristic skin lesions often around the waist¹¹⁵. Because smallpox (variola) was also present in Roman society, the word varicella possibly originated as a diminutive form of variola, which means „speckled" in Latin¹¹⁶. The difference between smallpox and chickenpox was not recognized until Heberden described the clinical difference between the two contagious disorders in 1767¹¹⁷. In the 19th century many observations confirmed that smallpox and chickenpox w unrelated,

and chickenpox is not a milder form of the smallpox infection, rather it is a contagious agent, which is associated with damage of spinal ganglia^{118,119}. A Hungarian pediatrician János von Bókay recognized in 1892 and reported for first time in 1909 that herpes zoster (shingles) and varicella (chickenpox) are linked together¹²⁰. In the middle of the 20th century, modern techniques helped prove that both diseases are caused by the same virus^{121,122}.

Structure and biology of VZV

The VZV virion is approximately 80–120 nm in diameter and is composed of an icosahedral nucleocapsid surrounded by a tegument layer¹²³. The outer virion component is an envelope derived from the host cell membrane with incorporated viral glycoproteins¹²⁴. The entire VZV genome sequence was first published in 1986¹²⁵. It is composed of a ~125-kbp double-stranded DNA molecule with an average G+C content of 46%, which contains at least 70 annotated open reading frames (ORFs)¹²⁵. Many of these ORFs are arranged in 3'-coterminial families, resembling the gene organization of other Herpesviruses¹²⁶. The VZV genome contains two equimolar isomers, they consist of a unique long (UL) and a 5200 bp long unique short (US) region (*Figure 3*), in which the US region can have either of two orientations¹²⁷. Two 7300 bp long inverted repeats (IRs) surround the US region, while the terminal repeat (TRS) of the UL region is only 88 bp long. In comparison, the Herpes simplex virus has four isomers with longer internal and terminal repeat regions; hence both UL and US segments can undergo inversion^{126,128}. The origin of replication (ori) is located in the repeat region adjacent to the UL segment¹²⁹. Three genes (*orf62*, *orf63* and *orf64*) located at the IR regions are in duplicate and three ORFs (*orf66*, *orf67* and *orf68*) are situated in tandem orientation in the US region^{125,130}. The main transcriptional trans activator proteins of VZV are present in tegument and encoded by *orf4*, *orf10*, *orf62* and *orf63* genes^{131,132}.

Figure 3. Display of VZV genome architecture. TRS, terminal repeat short; TLR, terminal repeat long; UL, unique long; US, unique short; IRL, internal repeat long; IRS, internal repeat short; ori, origin of replication. Dark blue ORFs are transcribed from the sense strand; yellow ORFs are transcribed from the antisense strand In: D. Scott Schmid et al. 2010¹³³

Replication of Herpesviruses is principally controlled by the regulation of transcription, which is carried out through a sequential activation of viral genes during the viral life cycle. The immediate early genes (IE) are expressed first; their products control the expression of early (E) genes, which are followed by the activation of the late (L) kinetic class of genes¹³⁴. Early gene products are generally needed to control

viral replication, while gene products of the late kinetic class build up the virion. The IE protein ORF62 is the most important essential trans activator of VZV, which recruits the general transcriptional apparatus of the host cell and thereby controls the expression of other viral genes^{135,136}. Virions enter the cell through membrane fusion or receptor-mediated endocytosis. The conserved herpesvirus glycoprotein gB and the heterodimer gH-gL mediate virion envelope fusion with cell membranes during virus entry¹³⁷. Thereafter, when uncoated virions are delivered to the nuclear membrane by the cell's cytoskeleton, the genome will be injected into the nucleus¹²³. Tegument proteins, like ORF47, ORF66 and ORF62, form complexes with cellular regulatory proteins, such as *transcription factor specificity protein 1*, to trans activate viral promoters and recruit cellular RNA Polymerase II for viral transcription¹²³. Before viral replication, the linear genome of VZV circularizes and then bidirectional synthesis is initiated via the origin of DNA replication binding protein, which is encoded by the *orf51* gene. VZV DNA starts to replicate with a theta-like structure whereafter the process switches to rolling circle mode¹³⁸.

Finally, viral replication is followed by the production of structural proteins and newly assembled nucleocapsids are released from the nucleus into the periplasmic space by budding – the capsids are simply too large to exit through the nuclear pore complex - and acquiring the tegument (containing e.g. ORF9-12 proteins), and they enter the Golgi network to undergo maturation^{123,139}. The glycoproteins of VZV are targeted to the Golgi to complete coating of intracellular virions¹⁴⁰. Enveloped virions reach the cell membrane by vesicular transport trafficking¹⁴¹. VZV differs from other herpesviruses in that assembled virions typically remain highly cell-associated, because of the presence of viral glycoproteins (e.g. gE, gB and gM) on the cell membrane^{134,142}. They induce fusion of neighboring cells, producing syncytia, which enhance virus spreading, thus it is difficult to analyze the viral life cycle *in vitro*, due to the lack of synchronous infection¹⁴².

Pathogenesis of VZV

VZV, as well VACV, have no known animal reservoir. There are no animal cell cultures or models for virus propagation, because VZV is highly human cell associated. The only animal model for investigating the pathogenesis of VZV infection is a human skin and thymus-liver or dorsal root ganglia xenograft implanted in mice with severe combined immunodeficiency (SCID)^{143–145}. VZV virions are spread from infected patients to susceptible persons by aerosols or direct contact with the vesicular fluid^{146,147}. Virions infect epithelial cells of the respiratory mucosa by inhalation of viral particles or the virus may reach tracheobronchial lymphoid tissues¹⁴⁸. The virions reach tonsils and other local lymph nodes, from where VZV can spread via the bloodstream within infected T cells. Subsequently, VZV may infect dermal perivascular cells, then epidermal and hair follicle keratinocytes via cell-to-cell spread and cause vesicular skin lesions throughout the body¹⁴⁹. Promoting cell fusion enables VZV to overcome innate immune boundaries and accelerates viral spread. The viral glycoproteins gB, gH and gL play roles in the cell-cell fusion¹²³. Viral replication in skin is probably linked with keratinocyte

differentiation causing viremia, as RNAseq analysis of host and VZV gene expression has shown¹⁵⁰. The next step of VZV pathogenesis is that virions find free nerve endings of sensory neurons in skin, located close to the cutaneous capillaries and hair follicles, and VZV travels via afferent axons to the neurons of sensory ganglia (dorsal root ganglia, trigeminal ganglia, vestibular ganglia) and autonomic ganglia, where lifelong latency is established^{149,151}.

Latency

VZV most likely has a long coevolutionary history with the human species¹⁵², so after the primary childhood infection a healthy individual rarely experiences its presence. Latency is a long-term, inactive, non-replicative, dormant period in a viral life cycle. It should not be confused with clinical latency, which is the incubation period in which the virus is active, but symptoms have not yet appeared¹⁵³. There is a lack of good *in vivo* models for the analysis of VZV latency, such experiments are restricted to the investigation of the non-human, primate simian varicella virus (SVV).^{154,155} VZV can only be obtained post mortem from human trigeminal ganglia¹⁵⁶. New *in vitro* models of Herpesvirus latency are based on the analysis of neurons obtained from induced pluripotent stem cells¹⁵⁷. Markus et al. recently reported, that latency can be established in induced neurons using microfluidic chamber devices, where soma and neuronal axons are artificially separated, imitating their natural situation^{158,159}. These experiments have demonstrated that the VZV genome is circularized during latency and forms an episome in the nucleus¹⁵¹. The copy number of VZV genomic DNA is at least ten times lower than that of Herpes simplex, and according to many analyses, the genome copy number averages six per neuron^{151,160}. Preserved gene activity during latency is unclear. Early studies reported that a limited number of protein coding genes (namely *orf* 4, 21, 29, 62, 63 and 66) express proteins in latency, but others have reported that only the immediate early ORF63 gene product is present in post mortem autopsy generated trigeminal neurons^{112,161}. Other studies have shown that transactivator protein coding gene promoters may be associated with histone proteins, thus epigenetic factors maintain the VZV genome in a heterochromatic state¹⁶². For many years, the clue to VZV latency was that, unlike other alpha herpesviruses, VZV does not express a latency associated transcript (LAT)¹⁶². Nevertheless, the discovery of the multisplliced latency-associated VZV transcript (VLT), which is expressed antisense to ORF61, fundamentally formed a new picture about the latent model of VZV inasmuch as it resembles the general alpha herpesvirus scenario, namely HSV and PRV express a similar transcript, which maintains latency in neurons^{163,164}.

Reactivation from latency

The exact molecular basis of VZV reactivation is not (yet) well known. VZV reactivation occurs 15-25 times more frequently in immunocompromised people and in the elderly, than in uninfected populations¹⁶⁵. It is known that innate immune response and antibody-driven humoral immunity, supplemented with CD4 and CD8 T-cell response, control VZV infected neurons in ganglia. However,

VZV has developed several mechanisms to inhibit the activation of both inherited and adaptive immune responses¹⁶⁶. Viral ORF47, ORF61, ORF62 block the NF- κ B pathway, while Jak/STAT2 signaling is arrested by ORF63, and ORF4 can modulate natural killer cells' cytotoxicity¹⁶⁷. Age-related decline in the frequency of VZV-specific T-cells is thought to allow the virus to reactivate and in monkeys reactivation can be experimentally induced by immunosuppression and stress^{151,168}. Although the exact molecular mechanism of reactivation remains uncertain, in a new model, proposed by Laemmle and colleagues, only a subpopulation of nerve cells allow the awakening of the silenced genome, an effect they termed *reactivable latency*¹⁵⁸. As a result of reactivation, viral particles are trafficked by retrograde axonal transport from ganglia to dermatomal epithelial cells, where VZV causes secondary infection in the form of a painful rash, called shingles.

Clinical complications, prevention and therapy

Varicella is a common benign infection in childhood. Complications are rare, however, they can be severe (cerebellitis, encephalitis, meningitis, necrotising fasciitis, pneumonia)^{169,170}. Long term complications can occur after the reactivation of VZV in immunosuppressed conditions. Until zoster blisters scab over, open sores are contagious; therefore, immunosuppressed people, pregnant women and newborns should avoid contact with people who have shingles. Zoster can lead to *post-herpetic neuralgia*, which is a rare, chronic, painful disorder¹⁷¹. Zoster in adulthood can cause an increased risk of encephalitis, possible vision loss or even death^{172,173}. Active immunization with attenuated VZV vaccine has become a routine childhood immunization in several countries to prevent chickenpox. Commonly used varicella vaccines contain a live-attenuated virus of the Oka strain¹⁷⁴. Varicella vaccination became available in the United States 25 years ago¹⁷⁵, but use varies widely between European countries¹⁷⁶. Varicella is associated with substantial clinical burden¹⁷⁷; on average 40 000 children under the age of 12 are infected each year and 90% of the population under the age of 15 are estimated to have been infected in Hungary¹⁷⁸. Therefore, Hungary announced the introduction of a two-dose funded, universal varicella vaccination in 2019¹⁷⁹.

Transcriptome research of VZV

Northern-blot hybridization technique using a library of recombinant fragments was used in the first effort to map particular parts of the VZV transcriptome¹⁸⁰. The same laboratory determined the direction of VZV mRNA and further expanded the number of known VZV transcripts to 77 using Northern-blotting¹⁸¹. Although a couple of very large viral RNAs (e.g. longer than 6.5 kb) have been detected in this study, their position was not elucidated in the publication. In the '80s the genetic map and sequence of VZV was determined, nevertheless technical limitations hampered the precise mapping of viral transcripts^{125,182}. Subsequently, Northern-blot analysis combined with more accurate techniques, like S1 nuclease digestion, primer extension and later Rapid Amplification of cDNA Ends (RACE) studies were

able to identify the exact 5' and 3' ends of some viral mRNA, like ORF1, ORF2, ORF4, ORF10, ORF14, ORF21, ORF61, ORF62, ORF63, ORF67, ORF68^{183–190}.

Hybridization-based microarray and second-generation short-read sequencing (SRS) techniques have revolutionized transcriptome research of VZV¹⁹¹. The first whole transcriptome PCR-based transcriptional array determined the general relative expressivity of VZV genes during lytic infection in BSC-1 cell line, in which the most abundant VZV transcripts mapped to ORFs 9, 64, 33, and 49 genes¹⁹². This study also showed that short intergenic regions can be transcriptionally active. Transcriptome profile of pOka and vOka strains were compared by macroarray analysis¹⁹³. Transcriptional response of fibroblasts and T cells to VZV infection was analyzed by microarray *in vitro* and in human skin xenografts in SCID-hu mice *in vivo*¹⁹⁴. Quantitative differences in viral gene expression were detected by microarray between two different cell types (melanoma and human astrocyte cell line) during lytic infection. SRS, including Pyrosequencing and Illumina sequencing, was used to compare VZV transcriptomes, as well as host cell responses, in neurons and fibroblasts and identified differentially expressed gene patterns between different cell types^{191,195}. Differentially expressed microRNAs were discovered in primary human fibroblast cells after wildtype and vaccine VZV infection. It was later demonstrated that the virus, itself, produces microRNAs^{196–198}. Up until now, 24 small non-coding RNAs (sncRNA), including miRNA, have been discovered in the VZV transcriptome. Some of them are mapped to 5' UTRs of regulatory proteins and seem to affect viral replication¹⁹⁹.

Other Illumina RNA-seq analysis demonstrated changes in the host transcriptome of VZV-infected aging cells and found a potential new VZV sensing receptor²⁰⁰. It has long been thought that VZV lacks a latency-associated transcript, which is well known to be expressed in other Alpha-Herpesviruses during latency. In 2018, Depledge and colleagues discovered a peculiar, multi-exonic transcript, later called VZV latency-associated transcript (VLT), using Illumina SRS technique in human trigeminal ganglia isolated shortly post mortem¹⁶³. This study clearly showed that it is worth investing in transcriptome research of VZV, because the application of other NGS techniques could provide a new level in the detection of transcriptional events, since microarrays and SRS techniques perform poorly in the detection of multiple introns of overlapping transcripts, transcript UTR variants and multicistronic RNA molecules²⁰¹.

Characteristics of VACV

The Poxviridae virus family contains a broad spectrum of poxviruses. They mainly infect vertebrates (Chordopoxvirinae subfamily), including birds (Avipoxviruses) and reptiles (Crocodylidpoxviruses). They also infect mammals including domestic animals, like sheep (Capripoxviruses) and swine (Suipoxviruses). Humans can be infected with members of Orthopoxviruses and Molluscipoxviruses²⁰². The poxviruses are among the largest of all animal viruses, not surprisingly they were the first virus visualized under light microscope^{203–205}. Common features of Poxviruses are their large, bricked-shaped

and membrane-coated virions of about 200-300 nm diameter, and large, linear, dsDNA genome, which encodes viral DNA- and viral RNA polymerases, allowing viral replication in the host cell's cytoplasm^{206,207}. Poxviruses share a group of ancient, conservative genes of Nucleocytoplasmic large DNA viruses (like Mimiviruses), which were probably acquired by lateral gene transfer from the eukaryotic host cell²⁰⁸. These features place Poxviruses phylogenetically close to Adenoviruses and to African swine fever virus (Asfarviridae)²⁰⁹. Interestingly, they also share a common hypothetical ancestor with bacteriophages, therefore it seems, that Poxviruses are probably one of the members of ancient viruses of eukaryotes²⁰⁸.

History and origin of Vaccinia virus

Smallpox (*Variola vera*) has been noted throughout history as a deadly dangerous infection. It has killed more people throughout history than any other infectious disease²¹⁰. The first known evidence of smallpox was found on the mummified body of Pharaoh Ramesses V. of Egypt, who died in 1157 B.C.²¹¹. As trading increased, smallpox spread from Egypt through Asia, and reached Europe 1000 years ago²⁰². After the discovery of the New World, European empires conquered Native Americans not only with weapons, but diseases like measles and smallpox. Aztec and Inka villages were devastated and later the North American population suffered similar devastation, because they had not developed immunity against smallpox²¹². According to contemporary descriptions 40-80% mortality rates among Native American populations were not rare²¹³. Variolation was the only more or less effective immunological prevention against smallpox. This process was an introduction of powdered smallpox scab into healthy individuals, however severe side effects occurred in some cases²¹⁴. Variolation had been practiced for centuries in Asia and it was introduced in Europe in the early 1700's by Lady Mary W. Montagu²¹⁵. This method was rather unreliable and dangerous, because variolated people could infect others, and symptoms were often fatal. However, variolation indisputably decreased the high mortality rate of smallpox endemics by an order of magnitude²¹⁶.

The smallpox vaccine has probably been responsible for more lives than any other vaccine, nevertheless it was the first step toward the development of modern vaccines²¹⁷. In 1790 Edward Jenner observed that milkmaids who had contracted a mild, pox-forming disease from infected cows were later immune to smallpox. Jenner applied his observation in practice with the intradermal introduction of cowpox virus ("*variolae vaccinae*") into a boy, afterwards infecting him with wildtype smallpox and, in agreement with his theory, the disease did not develop²¹⁸. With similar human trials vaccination proved to be safer and more efficient than variolation had ever been²⁰². Despite the contemporary "anti-vaccination" propaganda, Jenner's method spread quickly to many countries; Jenner used the vaccines on his own children and he sent samples all over the world. His booklet was translated into several languages, "vaccination institutes" were established and vaccine samples were transported from Europe to the world²¹⁹. At that time there was no globally organized vaccination program, therefore many physicians used strains from unknown sources with an unknown passage history for vaccination. The origin and

natural host(s) of vaccinia virus are unknown. Probably the virus originated in the XIX. century, either from vaccinated people by “arm-to-arm” vaccination, and/or from cows or horses (England, America), or from buffaloes in India, or even camels in Egypt^{202,218}. The oldest known vaccine sample from 1902 was recently sequenced and its phylogenetic analysis revealed that the core genome of the virus used as a vaccine has the highest degree of similarity (99.7%) to the horsepox virus²²⁰. This finding provides evidence of the suspected role of horsepox in the origin of the smallpox vaccine and probably “equination” was frequently used as vaccination²²¹

Etymology of smallpox and vaccinia

The word *variola* was commonly used for smallpox and was first documented by Bishop Marius of Avenches (Switzerland) in AD 570. It is probably derived from the Latin word *varius*, which means “stained,” or from *varus*, meaning “mark on the skin”²²². The virus left remarkable scars on faces of people who had smallpox, but survived. The term *small pockes* (*pocke* meaning sac) was first used in England at the end of the 15th century to distinguish the disease from syphilis, which was then known as the great pockes²²³. After Edward Jenner privately published a small booklet entitled “*An Inquiry into the Causes and Effects of the Variolae Vaccinae, a disease discovered in some of the western counties of England, particularly Gloucestershire and Known by the Name of Cow Pox*”²²⁴. The Latin word for cow is *vacca*, thus cowpox is *variola vaccinia*²²². Later every immunization procedure conducted by inoculation became universally known as *vaccination* and every immunizing agent termed as *vaccine* in honor of Jenner.

Vaccines used for the eradication of smallpox

In 1967, the WHO began a global program to eradicate smallpox. That year, more than 10 million cases of the disease and 2 million deaths had occurred in 42 countries, but in the following year more than 100 million people were vaccinated under the program²²⁵. Thanks to the heroic vaccination work, the WHO’s smallpox eradication committee declared in 1980 that smallpox, the deadliest virus had been eradicated²²⁶. Nearly 200 years after Jenner’s introduction of vaccination, his treatment saved more lives than anyone else in history. Many vaccinia strains from different sources were used as vaccines in the global smallpox eradication effort. In the United Kingdom, the Lister strain was used for vaccination beginning in 1892 and this strain was used throughout Europe, the strain NYCBH, originally derived from England (1856), was used in the USA and the strain Temple of Heaven (Tian Tan) in China^{227,228}. In laboratories the most widely used vaccinia strain is the Western Reserve (WR) strain. It has a long history as it was first cultured in rabbits, followed by mice and recently in laboratory cell cultures²²⁸. It was rarely subjected to clonal purification, therefore it originally contained a mixture of viruses⁶⁹. The VACV strain WR is adapted to neuronal cells, making it unsuitable for use as a vaccine^{229,230}, nevertheless, it has long been used as a model virus in the research of poxviruses²⁰².

Genome and gene organization of VACVs

VACV strains have linear, large (almost 200 kbp long), double-stranded DNA (dsDNA) genomes. The two strands of the DNA molecule are covalently cross-linked at both termini²⁰⁶. In addition, the VACV genomes contain very long inverted terminal repetitions of approximately 10 kbp pairs. These ends are evolutionally hypervariable, containing many deletions, truncated ORFs and complex, symmetrical sequence rearrangements²⁰⁶. The genomes of VACV strains differ from the genome of variola virus mainly in the inverted terminal repeat (ITR) regions (

Figure 4), where non-essential protein coding genes are located, determining the host range and immunogenicity of Orthopoxviruses^{231,232}. In contrast, the central region of the genome is highly conserved among Orthopoxviruses^{206,233}. This conserved core region comprises approximately 75% of the complete sequence of cowpox virus, and cowpox virus strains contain every gene present in all other Orthopoxvirus strains. Viral genes encoded in the central portion of the poxvirus genome generally code for proteins involved in viral replication, transcription, assembly and release of virions, supporting a reductive evolution of VACV²³⁴. Interestingly the GC-content of Orthopoxviruses is generally low, including VACV, which is 33.4% in total²³⁵. Genes are arranged in a tight, tandem orientation within the genome and separated by short intergenic sequences²³⁶.

Figure 4 Structure and organization of the core genome region of vaccinia virus. ITR, inverted terminal repetitions. From Christopher et al. 2016²³⁷

Protein coding and hypothetical protein coding genes are annotated in the reference genome, but current GenBank annotation does not contain cis-regulatory elements and untranslated regions (UTRs) (GenBank accession no. [NC_006998](#)). The nomenclature of genes follows the VACV Copenhagen strain's nomenclature, namely letters and numbers symbolize the HindIII digested fragments of the genome followed by L or R symbolizing the direction of a given ORF (e.g. D7R or D8L, Figure 4).

VACV genomes contain at least 200 ORFs, the exact number of annotated ORFs varies between 201-263 depending on the strain²³⁸.

Life cycle and replication of VACV

VACV produces two infectious forms during the assembly of virions: the intracellular mature virus (IMV) and the extracellular enveloped virus (EEV). Both particles enclose an identical DNA-containing core, which contain one or two viral membranes, respectively²⁰⁷. IMV is surrounded by a membrane from the trans-membrane Golgi network, while EEV's outer membrane is derived from the plasma membrane of host cell and contains viral structural proteins. The enveloped IMV enables efficient virus dissemination to the cell surface on microtubules, while EEV is responsible for cell-to-cell spread of the virus²³⁹. VACV particles are internalized into host cells by virus-induced micropinocytosis and the final step of entry is the release of the viral core into the host cell cytosol by a membrane fusion event²⁴⁰. After gaining entry into the cytoplasm, vaccinia virus begins transcribing early mRNAs²⁴¹, meanwhile the protein content of lateral bodies are proteolytically dissolved and the core undergoes an expansion to prepare for genome activation and an immediate immunomodulation of host cell antiviral response²⁴⁰.

While other mammalian DNA viruses, like Herpesviruses, must reach the nucleus to replicate, poxviruses, together with African swine fever virus (Asfarviridae), replicate in the cytoplasm within compartments called viral factories (VFs)²⁴². VFs are initially compact structures within the cytoplasm, which are surrounded by endoplasmic reticulum membranes to provide appropriate place for transcription and translation of viral mRNAs in addition to virion assembly²⁴³. Each of these factories are developed from a single virus particle after entry, therefore recombination between two viral genomes is impeded²⁴⁴. VACV genome encodes a viral DNA Polymerase. DNA starts to replicate about 3h p.i., which is a prerequisite of late gene expression²⁴³. The ITR sequence is only 0,7 kbp long in variola virus, which likely represents the minimal genetic element necessary for proper DNA replication²⁴⁵. By comparison ITR of VACV is 10 kbp long, it contains tandem repeat sequences at the very end of genome, which are only about 0.5 kbp long and help to build a hairpin-like structure. These terminal repeat sequences are thought to be the origin of replication of VACV²⁴⁶.

Transcriptomics of VACV

Biology of VACV is well documented in the literature. Smallpox, and later VACV, was used as a recombinant vector in oncolytic cancer therapy, because it replicates in the cytoplasm and its genome permits cloning large foreign sequences^{247–250}. Early studies on VACV transcriptome were performed with Northern-blot^{251,252} combined with *in vitro* expression studies^{253–257} and resolved expression profiles of single genes²⁵⁸, while microarray data were used to analyze the time course changes of genes expression of all VACV ORFs and host genes in parallel²⁵⁹. Other microarray data showed a new class of immediate early (IE) expressed genes⁴¹. Illumina sequencing showed that VACV incorporates viral

transcripts into virions during maturation –like Herpesviruses - and these RNAs may play important roles in immune evasion upon arrival into host cells²⁶⁰. SOLiD and Illumina deep sequencing platforms were also used to determine the precise 5' and 3' ends of early VACV RNAs^{261,262}. Late RNAs showed pervasive transcription initiation and termination²⁶³ and these results, obtained with SRS techniques, suggest extensive read-through, particularly in late genes²⁶². Several articles include SRS technique-detected distinctive cis regulatory elements of early, intermediate and late promoters. Early transcripts have a TTTTNT consensus box followed by a pyrimidine rich motif for transcription termination at their 3' end, which is not recognized late in infection²⁶⁴. A characteristic feature of early transcripts is their exact 3' ends in contrast to late transcripts, which delineate extensive polymorphism, therefore transcriptional read-through is frequently seen between the adjacent, tandemly oriented genes^{261,262}. Two unusual features place poxviruses in a distinct group among DNA viruses. In contrast to their host cell, splicing is totally lacking in VACV transcriptome and the presence of poly(A) leader sequences at the 5'-end of transcripts promote translation of viral²⁶⁵ transcripts, and regulate host mRNA decay²⁶⁵. A/T composition situated upstream to post-replicative promoters determines which intermediate genes show mixed expression activity, they can enhance transcription at a late time point of infection suggesting multistaging in promoter usage²⁶⁶.

Aims of the study

Genome wide SRS studies on VACV transcriptome have determined hundreds of TSS, TES and unusual TSSs originated within ORFs, however these TSS and TES positions could not be linked together only by SRS technique, because of the transcriptional complexity of VACV. Our major goals were the following:

- genome reconstruction based on full-genome-covering, overlapping cDNAs
- 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 VACV RNAs by linking TSS and TES positions together;
- annotation of transcript variants, detection of transcript isoforms.

VZV is a well-known human pathogen Herpesvirus, its transcriptome had been analyzed by SRS technique, however mRNAs and non-coding RNAs of VZV have not been studied by long read sequencing techniques. Exact positions of transcripts were known in some cases, however the majority of mRNAs of VZV remained undefined. Our major aim was to characterize and reannotate the transcripts of VZV produced in lytic infection and provide a new, comprehensive transcriptomic landscape of the virus. Our aims were therefore 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 VACV RNAs by linking TSS and TES positions together;
- annotation of transcript variants, detection of isoforms.

Nanopore sequencing enables the detection of complex network of intro-exon junctions, therefore new, alternative splicing detection in VZV transcriptome was another goal of the study. Two transcriptomes of DNA viruses (VZV & VACV) were analyzed in detail by long read sequencing.

Methods

Propagation of VZV and purification of RNA

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), the cells were harvested, when displaying 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%. Total RNA was isolated using the Nucleospin RNA Kit (Macherey-Nagel) according to the manufacturer's protocol. gDNA was digested with the RNase-free rDNase solution. To eliminate residual DNA contamination a subsequent TURBO DNase treatment (Thermo Fisher Scientific) was conducted. From 35 µg of total RNA the poly(A)⁺ RNA fraction was isolated using the Oligotex mRNA Mini Kit (Qiagen). RNA samples were stored at -80°C until use.

Oxford Nanopore MinION oligo(d)T sequencing and barcoding of VZV's cDNA

The cDNA library was prepared using the Ligation Sequencing kit (ONT SQK-LSK108) according to the modified 1D strand switching cDNA by ligation protocol. In short, the polyA(+) RNA fraction was reverse transcribed (RT) using an oligo(d)T-containing primer [(VN)T20 (Bio Basic)]. The RT reaction was carried out using SuperScript IV enzyme (Life Technologies); a strand-switching oligo [containing three O-methyl-guanine RNA bases (PCR_Sw_mod_3G; Bio Basic)] was added to the sample. The cDNA was amplified by using KAPA HiFi DNA Polymerase (Kapa Biosystems) and Ligation Sequencing Kit Primer Mix. End repair was carried out on the samples using NEBNext End repair/dA-tailing Module (New England Biolabs), which was followed by “barcoding” (ONT PCR Barcoding Kit 96; EXP-PBC096) according to the 1D PCR barcoding (96) gDNA protocol. The barcoded sample was amplified by PCR using KAPA HiFi DNA Polymerase. The PCR product was 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.

Cap selection, cDNA synthesis and sequencing

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 (5' - 3':CTCAGGCGTTTTTTTTTTTTTTTTTTT). The RT product was purified by using spin column-based method (Lexogen kit). 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 by the TeloPrime Kit's manual. The generation of the sequencing-ready library from this sample was based on the 1D strand switching cDNA by ligation protocol from ONT (Ligation Sequencing kit (SQK-LSK108) as described above. We used barcoding for the better *in silico* identification of the transcripts' ends. The library was sequenced on the abovementioned flow cell, basecalling was carried out using Albacore v2.1.10.

Targeted sequencing of NTO transcripts

Validation of the TSSs of three 'near the transcription origin' (NTO) transcripts, and the spliced isoforms of the region was carried out using targeted long-read sequencing. The cDNA library was prepared using the ONT 1D sequencing and barcoding protocol with the following modifications: we used ribodepletion instead of poly(A) selection, and target-specific primers (ordered from IDT) were applied instead of oligo(dT)-priming for the cDNA preparation. The target-specific primers used in this study are listed in *Supplementary Supplementary Table*.

Data analysis and alignment

Reads resulting from ONT sequencing were aligned to the reference genome of VZV (GenBank accession: [NC_001348.1](#)) and the host cell genome (Homo sapiens - GRCh37, BioProject number: [PRJNA31257](#)) using GMAP v2017-04-24²⁶⁷. 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)²⁶⁸. The 5'-ends of the rare long reads were inspected individually using the Integrative Genome Viewer (IGV)²⁶⁹. Poly(A) tails were defined using the same algorithm and parameters, by mapping 15nt 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 positions 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. The computational pipeline is outlined on the *Supplementary Figure 1*.

Annotation of transcripts and splice junctions

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. 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 ± 10 nt-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 ²⁷⁰, with an expected threshold of 10.

Analysis of hyperediting

To evaluate the effect of hyper-editing on the secondary structure of RNAs, we used the 'RNAstructure Software' suite²⁷¹. To simulate the presence of inosine, we changed the edited adenines to guanine. Reads were visualized using the Geneious ²⁷² software suite and IGV.

Cell culture and propagation of VACV; RNA purification of VACV

African green monkey kidney fibroblast (CV-1) cells were infected with 3 ml of VACV at 10 MOI/cell and were incubated at 37°C for 1,2,3,4,6,8,12 and 16 hours in a humidified 5% CO₂ atmosphere. The experiment was performed in duplicate. Total RNA was extracted from the VACV infected CV-1 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). The whole experimental design is described in the *Supplementary Table 4, Supplementary Figure 7*.

PacBio RSII & Sequel sequencing

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 pi) 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 from 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. SMRTbell library preparation and sequencing was briefly the following: primer annealing and polymerase binding reactions were carried out using the DNA

Sequencing Reagent Kit 4.0 v2 with DNA Polymerase P6 for RSII platform, whereas the Sequel Sequencing Kit 2.1 and Sequel DNA Polymerase 2.0 were applied for sequencing on the Sequel. The PacBio's MagBead Kit (v2) was used for the reaction. 240 or 600 minute movies were captured using the RSII and Sequel machines, respectively. One movie was recorded for each SMRTcell. The PacBio library preparation is depicted on the *Supplementary Figure 8*.

ONT MinION platform – cDNA sequencing of polyA(+) RNAs

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, and the NEBNext End repair/dA-tailing Module NEB Blunt/TA Ligase Master Mix (New England Biolabs), according to the kits' manuals, as it is described in the VZV Methods section. PCR products derived from the mixed RNA sample were size-selected manually and then run on Ultrapure Agarose gel (Thermo Fischer Scientific), followed by the isolation of 500 bp+ fragments with the Zymoclean Large Fragment DNA Recovery Kit (Zymo Research). The individually sequenced cDNAs were barcoded with the combination of two ONT protocols: the 1D protocol was used until the first end-preparation step, which was then followed by the barcode ligation step using the ONT PCR Barcoding Kit 96 (EXP-PBC096): the barcode adapters were ligated to the end-prepped samples by using the Blunt/TA Ligase Master Mix (New England Biolabs).

ONT MinION platform – Library preparation from Cap-selected samples

For more accurate definition of the TSSs of the full-length RNA molecules, a 5'-Cap-specific cDNA generation protocol was followed and combined with the ONT 1D cDNA library preparation method. The cDNAs were produced from a mixture of total RNA samples (1, 2, 3, 4, 6, 8, 12, and 16 hours' pi) by using the TeloPrime Full-Length cDNA Amplification Kit (Lexogen). The amplified PolyA- and Cap-selected samples were used for the library preparation following the ONT's 1D strand-switching cDNA by a ligation method (ONT Ligation Sequencing 1D kit) as it is described in the VZV Methods section. The ONT cDNA libraries and the Cap-selected samples were loaded onto 3 and 2 ONT R9.4 SpotON Flow Cells for sequencing, respectively. The runs were carried out using MinKNOW application (ONT), *.fast5* files were basecalled both Albacore v2.1.10 and checked by Guppy2.3 software also.

Analysis of reads, mapping and annotation of TSS and TES positions and transcripts

The PacBio consensus reads of insert (ROI) reads were created from the RSII raw data using the RS_ReadsOfInsert protocol (SMRT Analysis v2.3.0), The ROIs from the Sequel dataset were generated using SMRT Link5.0.1.9585. Bioinformatical steps were performed in Ubuntu/Linux operational

system in python3. The ONT's Albacore software v.2.0.1 (Albacore, RRID:SCR_015897) was used for the MinION base calling. The newest release of the ONT's Guppy (Guppy v.3.6.0) was also used for basecalling with the aim to validate the TSS and TES positions. The PacBio ROIs and the ONT raw reads were aligned to the reference genome of the VACV ([LT966077.1](#)) and to the host cell (*Chlorocebus sabaeus*) (GenBank RefSeq assembly accession: [GCA_000409795.2](#) [latest] using minimap2 aligner (version 2.13) with options -ax splice -Y -C5 -cs. Mapping was performed with minimap2 tool²⁷³ with the following command: `minimap2 -ax splice -Y --cs /mnt/reference_genom.fasta /mnt/input/file.fastq > /mnt/output/file.sam`. Mapping was followed by the application of the LoRTIA toolkit (<https://github.com/zsolt-balazs/LoRTIA>) for the determination the 5' and 3' ends of transcripts, as well as for detecting the full-length reads with the following commands: `#LoRTIA -5 TGCCATTAGGCCGGG --five_score 14 --check_in_soft 15 -3 AAAAAAAAAAAAAAAAAA --three_score 14 -s poisson -f True /mnt/d/Data/MinION/Mappings/input_pA.sam /mnt/d/Data/MinION/Mappings/reference.fa` and in case of cap-selected sample: `#LoRTIA -s poisson -f True /mnt/d/Data/MinION/Mappings/input_cap.sam /mnt/RefGenomes/reference.fa`. The output bed files from cDNA sequencing were visualized by Circos plot²⁷⁴. The computational pipeline is drafted on the *Supplementary Figure 9*.

Nomenclature of transcripts

The novel ORFs and transcripts were named according to the common (Copenhagen) HindIII fragment letter/number-based nomenclature (e.g. A1R)^{275,276}. Previously described ORFs were not renamed and our scheme allowed for future additions of novel transcripts. Multicistronic transcripts were named for all contributed ORFs, the most upstream ORF was listed first. The most abundant transcript isoform specified by a given gene named as the canonical (main) transcript. The 5' length isoforms were represented with an "s-", or "l-" prefix (short and long respectively). Variations in 3' ends were represented with "-AT" meaning alternative termination. Internal, or embedded, possible protein coding genes were named with a ".X" affix (X=1-9), where ".5" variant represented about the half of the main transcript. Isoform categories are presented on the *Supplementary Figure 17*.

Prediction of cis-regulatory sequences

Sequences at 100 nt upstream of a given TSS were extracted and an in house script based on the GPMIner promoter prediction tool^{277,278} was used for analysis of upstream cis-regulatory sequences of novel transcript. We used the eukaryotic promoter database of GPMIner and the Vaccinia early promoter motif^{261,262} and late promoter motif²⁶³ was picked up and visualized by the GeneiousR10 Motif Finder tool²⁷⁹ allowing one substitution in the string. Promoter statistics was calculated by Mann-Whitney U test with two tailed p-value.

Results of VZV analysis

Structure of VZV transcriptome; number of sequenced reads in VZV transcriptome

In order to get a comprehensive picture about the lytic transcriptome of VZV we isolated oligod(T) selected RNA from human fibroblast cells infected with vOKA strain of VZV. Then we prepared and three different sequencing libraries, a cap-selected, a non-cap selected and a target sequencing library respectively on MinION LRS platform. *Table 1.* summarizes the number of reads obtained from the sequenced samples. For mapping of viral reads, the VZV reference genome (GenBank Accession: NC_001348.1) was used.

Sequencing library	Cap-selected (bp)	Non-cap selected (bp)	Targeted (bp)
all reads	6918054	619687	1792059
mapped viral reads	614192	66455	328118
average read length	294	1349	651
average coverage	1200	625	1710

Table 1. Sequenced read number in three different sequencing library of VZV transcriptome

Targeted sequencing was used only for confirming some special transcripts at the Ori region, therefore it was not included in the transcript annotation process. This technique yielded 1,792,059 reads of which 18% mapped to the VZV genome. The 2.41% of the targeted reads mapping to the VZV genome included the target-specific primers and mapped to the correct position, whereas, 24.112% of the reads included the primer but mapped off-target, while on the rest of reads the target-specific primer was missing. These latter two categories could be a result of nonspecific priming or template switching.

Sequencing of the cap-selected samples yielded 6,918,054 reads, of which 614,192 mapped to the viral genome, with an average coverage of 1,200, and 6,303,862 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 5.a*) Intriguingly, we obtained the same poor result with other herpesviruses, including PRV and HSV-1^{280,281}, however, the cap-selection technique performed very well in VACV²⁸². This phenomenon could be elucidated by the higher GC-content of Herpesviruses, than Orthopoxviruses (PRV: 73%, HSV-1: 68%, VACV: 33%), however VZV has only 46% GC-content. These results together with an analysis of the nucleotide distribution at the vicinity of the 5'-ends in the cap-selected datasets (*Figure 5.b*) demonstrate that the read length of cap-selected RNA-s is not determined by the GC-content, but by a yet unknown factor.

Figure 5. The read length distribution of cap-selected and non-cap-selected RNAs. *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 fraction 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

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. Data are listed online in *Przasák et. al. (2018 Supplementary tables 1-4)*²⁸³. 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 (*Table 2*). 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.

Filtering method	Non-cap-selected		Cap-selected	
	5' ends	3' ends	5' ends	3' ends
None	1124	255	1428	279
Peak analysis and Poisson probability	151	92	202	116
Template switching/false priming	102	76	181	100

Table 2. The number of read ends following each step of filtering.

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.

Transcription machinery needs cis-acting sequences for effective transcription. Therefore the sequences upstream of the TSSs and TESs detected by LoRTIA tool were analyzed *in silico* to extract genomic regions to look for putative GC-, CCAAT- and TATA-box motifs, and Poly(A) signals (PASs), by motif searching (*Table 3*).

	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

Table 3. Putative promoter sequences and Poly(A) signals (PASs) of the VZV transcripts. 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.

An enrichment of As and Us in an interval of up to -10 nt upstream of the putative TES, and an abundant GT-rich region downstream in the +10 nt interval was also detected. After transcript annotation, we could annotate altogether 240 novel transcripts of which 143 were confirmed by at least three reads. Additionally, we mapped the TSS and TES of 37 previously known transcripts - described by others - lacking former mapping of RNA end positions (these are listed in the original article's [online supplement](#)²⁸⁴).

New putative mRNAs of VZV transcriptome

Coterminal transcriptional units representing embedded genes with common polyA signals are characteristic features of alpha herpesviruses. Transcripts embedded into 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 of 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^{125,130} one of which (the orf33.5) is a tORF. Our analysis reported the identification of 25 embedded viral transcripts containing tORFs. We could identify promoters for 10 of these transcripts (*Table 4*). Three of the 25 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. Eighteen of these *possibly protein-coding* VZV transcripts contain a canonical PAS. We also determined the TSSs and TESs of 25 mRNAs whose ORFs had earlier been described (*Supplementary*

Figure 2). The TSS and TES of these monocistronic mRNAs were not characterized before by any other means. Common promoter sequence motif can be found in cases of 12 of these transcripts and 19 have a canonical PAS upstream their TESs (Table 4.).

Non-coding transcripts of VZV transcriptome

The ncRNAs are transcripts without the ability to encode proteins. They lack ORF and cis-acting elements for translation. ncRNAs can overlap coding sequences of their host genes in the same orientation or in the opposite orientation [antisense (as)RNAs], or they can be located between genes, in intergenic region. The 5'-truncated transcripts have their own promoters but share the PASs and TESs with the host transcript, while the 3'-truncated RNAs are controlled by the same promoters as the canonical transcript in which they are embedded but have no in-frame ORFs. 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, all of them being controlled by their own putative promoters. In total, 17 of the novel ncRNAs contain a canonical promoter sequence, while 21 have a canonical PAS. The non-coding transcripts ORF42-43-AS and ORF35-AS overlap multiple canonical mRNAs. The transcript 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 (Supplementary Figure 2). The LAT RNA has been described in every member of the alpha herpesvirus subfamily^{285,286}, including VZV¹⁶³. 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 (Supplementary Figure 3).

5'- and 3'-UTR isoforms of VZV transcriptome

Twenty-two and five transcript isoforms were identified altogether with the non-cap-selected, and cap-selected dataset, respectively with a relative abundance higher than 1%. The 5'-UTR isoforms (TSS variants) start upstream or downstream of the TSS of the earlier annotated main transcripts, and their expression is regulated by their own promoters. 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 18 novel 5'-UTR length variants, 7 being shorter and 11 being longer than the earlier annotated transcript isoforms. We found canonical promoter sequences in 10 of the 5'-UTR isoforms. Additionally, we detected eight 3'-UTR length isoform, all with a canonical PAS. An intriguing finding is the putative TSS at the genome position +4 belonging to the rare transcripts 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 (TRs) of the genome. This putative promoter is supposed to be active in the circular form of genome during replication.

	Promoter mean distance upstream from TSS (nt)	± SD	PAS mean distance upstream from TES (nt)	± SD
putative mRNAs	93.85	32.18.	19.44	8.91
monogenic mRNAs	51.57	33.31	14.96	11.59
Non-coding RNAs	60.4	35.19	16.35	8.87
5' UTR isoforms	94.93	38.58	-	-
3' UTR isoforms	-	-	22.75	7.65

Table 4. Summary of cis regulatory elements of the novel RNAs of VZV

Splice sites and splice isoforms in VZV transcriptome

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 low abundancy splice sites and reads potentially generated by PCR byproducts. From the initial set of more than 1000 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. Splice events are thought to be rare in VZV transcriptome compared to other Herpesviruses. 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. 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^{287,288} (*Figure 6*). Furthermore, we found ten (!) novel splice junctions of ORF63 splice variants coinciding the recently described VLT_{ly} splice variants¹⁶³, thus they were labeled as VLT-ORF63-C (*Figure 6*). These splice junctions were confirmed using targeted sequencing as well. It is noted that one of the splice donor sites, present in both NTO1v1 and NTO1v2 is GC, which differs from the canonical splice donor sequence (GT). We detected all of the three ORF50 splice isoforms (ORF50A, B, and C) previously described¹⁴², but ORF50C occurs in relatively low abundance in our dataset, below our acceptance threshold. However, we detected another new splice variant in low abundance in the ORF50 cluster and named it ORF50D. All of the newly detected splice variants are listed in the end of this work in the *Supplementary Table 3*. In four detected novel isoforms, splicing affects the sizes of the ORFs, all producing hypothetical proteins truncated near the N-terminal. Roviš and colleagues²⁸⁹ have identified the majority of VZV proteins using monoclonal antibodies. In some

cases, lower-size bands appeared in their Western blots, which they explained to be the result of putative post-translational modification or proteolytic cleavage. In view of our long read transcriptomics results, we can raise the question of whether splicing is a possible option for these nonspecific protein isoforms, listed in *Table 5*.

Transcript	Host protein length (aa)	Spliced protein length (aa)	Intron position	Splice donor position	Splice acceptor position
ORF5-4-SP	ORF4: 452	452	5' UTR	4141	2783
ORF12-13-SP	ORF12: 661	233	in frame, 5'truncated	16214	17088
ORF24-SP	ORF24: 269	137	in frame	44021	43286
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

Table 5. Proteins and hypothetical proteins produced by the novel spliced transcripts.

Figure 6. a. Transcripts neighboring OriS. The ORF63 of VZV, similarly to the HSV-1 US1 has a two-exon-containing splice isoform. The target-specific primers used to confirm the TSSs of VZV's NTO2,

NTO3, NTO4 are shown as green arrows b. VZV expresses non-overlapping non-coding RNAs in the proximity of OriS (blue arrows).

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

Our research group reported that PRV expresses transcripts located near the replication origins⁸⁶. These transcripts have been named CTO family –which means “*close to Ori*”- (including CTO-S, CTO-S-AT, CTO-M, and CTO-L) at the OriL and the PTOs (PTO and PTO-US1)⁸⁵ at the OriS. Expression of nroRNAs has also been described in HSV-1 (OriS-RNA:²⁹⁰) and in HCMV (OriLyt:²⁹¹; RNA4.9:²⁹²). VZV genome contains OriS, and probably lacks the replication origin at the UL segment – although Khalil et al. showed that OriS mutations do not completely impede VZV replication in cell culture, suggesting an alternative, hidden OriL²⁹³. 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, while the NTO1v2, NTO1v4, and NTO1v6 are the spliced long TSS variant of ORF64 (*Figure 6a*). NTO2 starts at the same TSS as NTO1v1 but terminates 30 bp downstream of its splice donor site. The 5' end of the reads of NTO3 and NTO4 are positioned downstream of NTO2. These transcripts were present in our cap-selected but not in the non-cap-selected sample presumably because of their low abundance. In order to augment the detection of these transcripts, we used target-specific primers (*Supplementary Table 1.*) for sequencing library preparation, followed by MinION cDNA sequencing. The starting positions detected in the cap-selected sample were not confirmed by the targeted sequencing of these transcripts, thus we hypothesize that NTO3 and NTO4 starts at the closest upstream confirmed TSS, that of NTO1v1. The arrangement of the NTO transcripts is similar to that observed in PRV, where the TSS of the PTO-US1 (positionally similar to the NTO1v1) is located at the same genetic locus as PTO (locationally similar to the NTO2, NTO3 and NTO4, but not homologous) (*Figure 6b*). *In silico* analysis revealed that the poly(A) cleavage sites of NTO2, NTO3, NTO4 are located within a canonical CA element, downstream from the TES and their polyA signals are either canonical or a CTAAAA, or an ATATAA respectively^{294, 295}. No enrichment of adenines was observed in the genomic region of the TES-s in any of the three NTO transcripts, which would be the source of strand switching or false priming.

Polycistronic and complex transcripts

A major issue of SRS, as well as microarray and quantitative PCR approaches is that they have severe limitations in distinguishing between monocistronic 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. We identified 33 hitherto unknown, long polycistronic RNAs in VZV transcriptome, including four complex transcripts. Complex transcripts are polycistronic RNAs that contain one or more genes in opposite orientations, therefore antisense sequences on the complex transcripts are unable to encode proteins. 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 VLT_{ly} isoforms (*Figure 7*).

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 7*). The N-terminal overhangs of the putative proteins show no homology with any other known proteins in online databases.

Figure 7. Splice isoforms of ORF63 and ORF64 with a similar splicing pattern as the VLT_{ly}

Transcriptional overlaps

Transcripts can overlap each other in parallel (tandem or head-to-tail), convergent (tail-to-tail) or divergent (head-to-head) manner (*Supplementary Figure 4*). RNAs identified and annotated with a certain TSS and TES in this work form a total of 481 overlaps, of which 15 are head-to-head, 450 are head-to-tail and 16 are tail-to-tail. The transcriptional overlaps can be the result of alternative promoter usage (TSS isoforms) or transcriptional read-through (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 (*Supplementary Figure 4*).

Upstream ORFs

Using *in silico* methods, we have detected 44 potential (u)ORFs on the 5'-UTRs of 81 VZV transcripts (**Supplementary Table 9**). Five of the longer TSS variants contain more than one uORFs, while their shorter isoform does not. We had previously described this phenomenon in HCMV⁸⁸ and PRV²⁸⁰. The average size of an uORF was 54.9 ± 44.64 nt (mean $\pm$ SD, median=35nt), while the average distance of the uORF stop codon from the protein coding ORF's AUG was 174.84 ± 143.58 nt (mean $\pm$ SD,

median=110nt). This space between the two uORF and the canonical ORF is enough for a potential reinitiation event. We identified Kozak consensus sequences in five uORFs (*Supplementary Table 2*).

RNA editing

The sequencing reads of one of the replication origo associated NTO3 transcript show a very high frequency of A to G substitution, which is interestingly not present in the overlapping reads of other transcripts. We found that 58% of all substitutions are A->G (*Supplementary Figure 5a and 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) (*Supplementary Figure 5*) making 22.07% of all As of the transcript edited. The targeted sequencing dataset revealed a similar A->G substitution pattern in the NTO3 overlapping reads as the cap-selected data. This suggests a hyper-editing event in NTO3. Using the MEME software suite²⁹⁶ we could detect a slight enrichment of the GU dinucleotide preceding the editing site resulting in a GUAG base editing motif (the editing site is underlined) (*Supplementary Figure 5d*).

Using the 'RNAstructure' software suite²⁷¹, we predicted the secondary structure with the lowest free energy of the NTO3 ssRNA (*Supplementary Figure 6*) and of the NTO3-ORF62-5' dsRNA hybrid, using both the unedited and edited forms of the ssRNA (*Supplementary Figure 6*). 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.

Results of VACV analysis

Genome reconstruction of VACV

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. Generally, the PacBio sequencing reads are more accurate, than ONT, which is prone to generate mismatches in individual reads, although the consensus sequences can be

fairly accurate if coverage is high⁸³. 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%, which is in agreement with other Chordopoxviruses with low GC-content²³⁵. We annotated 218 protein coding genes. Our data showed that the sequenced WR strain differs in 163 point mutations from the reference genome, mainly in the F region. The most mutated protein coding sequence was the C9L. Not surprisingly the mutations mainly occurred in genes playing role in host defense modulation and transcription regulation. The genome sequence of this WR strain of VACV is available from the European Nucleotide Archive under the accession number [LT966077](#).

Long-read transcriptome sequencing of VACV

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. For nanopore sequencing, we used the library preparation approach provided by the ONT company and the Teloprime Cap-selection protocol (Cap-Seq) by Lexogen. In general, the used sequencing techniques achieved high coverage to determine both TSSs and TESs of viral transcripts with base-pair precision. However, identification of transcript ends (especially, TSSs) was extremely challenging compared to Herpesviruses due to the tremendous complexity of the VACV transcriptome. We used a transcriptome annotation tool suited for long reads, called LoRTIA²⁹⁷ developed by our group, to obtain viral TSS and TES coordinates. In total, about 1,115,000 reads of inserts (R

Figure 8. Box plot presentation of the average of aligned read lengths obtained from the applied sequencing methods. The reads were mapped to the VACV genome, and the average lengths were plotted with the standard deviation values.

OIs) were generated using the PacBio platform (Supplementary Table 1). 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 PacBioSMRT and 619 ± 148.2 SD in the ONT Minion samples), however both of the ONT MinION 1D cDNA and the 8h Sequel libraries could produce the longest aligned reads (>6000 bp) (*Supplementary Table 6*). Average read lengths in each libraries are depicted on the *Figure 8*. For MinION sequencing the cap-seq approach was used to identify 5' ends of mRNAs. For 3' end detection the oligo(dT)-based sequencing protocol was used. We also prepared random primed non-polyA(+) RNA fractions to detect and validate TSSs. Our combined sequencing approach detected transcriptional activity at every nucleotide of the VACV genome in late time points of infections (*Figure 10*).

Transcription start and end sites (TSSs & 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^{298–301}. 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 (*Supplementary Figure 10*). Altogether 898 TSSs were detected in the 'regular' regions of VACV transcriptome (*Figure 9, Figure 10*).

Figure 9. Jitter plot visualization of transcriptional start site distribution across the VACV genome; distribution of TSSs of our annotated transcripts are shown at the top (red dots); the TSSs detected using LoRTIA are depicted in the middle (orange dots). TSS positions described by others^{262,263,300,301} are located at the bottom of the figure (blue dots).

The TESs of early transcripts were increasingly heterogenic at late time points of infection, potentially because transcription termination signal recognition depends on different factors in early- and PR phases^{298–301}. Due to the extremely high TES diversity, we analyzed their positions within a ± 30 -nt interval. We found that 24% of our annotated TESs matched the positions of the polyadenylation sites described by Yang et al.³⁰⁰ (*Supplementary Figure 10*). Our analysis revealed that VACV genes express transcripts with multiple shared TSS and TES positions (*Figure 10B*). Distributions of TSSs and TESs along the VACV genome are presented in, where chaotic genomic segments show an accumulation of TESs and TSSs (*Figure 9*). Moreover, most genes express multiple transcript isoforms (median = 4, mean = 6.95) and genes with only one transcript isoform are rare. In the extreme cases of N1L, O2L,

I1L, and A12L regions, more than 30 transcript isoforms of a single ORF were expressed (Suppl. Fig 16.). Moreover, no TSS positions belonging to 30 annotated ORFs, particularly in the A8R-A17L region, have been annotated in our and others' studies (*Suppl. Note 2.*).

The novel ORFs, transcripts, and their isoforms were named according to the common (Copenhagen) HindIII fragment letter/number-based nomenclature (e.g. A1R). The detailed naming scheme is written in the Methods section. We found 8191 unique putative viral transcripts using the in house developed LoRTIA pipeline. From the 218 annotated ORFs 184 were expressed alone on separate RNA molecules. To exclude failed annotation these transcripts were further processed under strict criteria: only the transcripts occurring by two different techniques and/or in three independent experiments were considered for detailed analysis, thus the number of VACV's transcripts decreased to 1480. Because of the large genome of VACV, it is not possible to depict the whole transcriptome, but all VACV transcripts (identified by the LoRTIA pipeline) are presented in a *Geneious* file available for download at *figshare server*, see *Suppl. Note 1*.

Figure 10. A. Depths of VACV read coverage across the viral life cycle; transcripts were obtained at six time points and were sequenced using PacBio Sequel (inner radius) and Oxford nanopore technology (ONT) MinION (outer radius) platforms and were visualized in a Circos plot as concentric circles; each circle represents an individual time point with early to later time points represented as inner to outer circles, respectively. B. Complexity of VACV ‘chaotic regions’; above: high expression of antisense transcripts within the A18R locus; below: extremely high expression levels of antisense RNAs and other non-coding transcripts at the A10L-A12L region; d. enormous numbers of non-coding transcripts and transcript isoforms within the O1L– I3L genomic segment. All transcripts were identified using the LoRTIA toolkit.

TSS and/or TES position, therefore transcript annotation was almost insurmountable. In two genomic regions many starts or ends of RNAs were occurred within annotated ORFs or on complementary DNA strand (e.g. O1L and A11R ORF regions are filled up with antisense reads) contributing in an extreme high level of read-through transcription. Therefore, we named these regions ‘chaotic’ and for the others we used the term ‘regular’ transcriptional region of the viral genome (*Suppl. Note 2*). At the ‘regular’ genomic segments viral genes produce transcript with more or less precise TSSs and TESs –like in Herpesviruses-, whereas at the ‘chaotic’ regions (such as the genomic loci of O1L, O2L, I1L, I2L, I3L, A10L, A11R, A12L, A18R genes) a large amount of TSSs and TESs diversity are produced (*Figure 10.B*) Some VACV gene clusters exhibited a genomic organization that resembled that of herpesviruses, with tandem genes producing overlapping 3’co-terminal RNA molecules, e.g. transcripts of the following genes A51R-A52R, B12R-B13R, B13R-B15R, F16L-F15L-F14L, F15L-F14L or F5L-F4L-F3L, F4L-F3L.

Novel mono-, bi- and polycistronic RNAs

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. Similar to other viruses and prokaryotes, but unlike the eukaryotic organisms, VACV reportedly express bi- and polycistronic transcripts^{49,260,302,303}. Yet although all genes are translated from prokaryotic polycistronic mRNAs, experimental evidence is lacking for translation of downstream genes from polycistronic mRNAs of VACV. 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 (see online supplementary files S3 in [Tombácz and Prazsák et al. 2020](#)). Yet not all of these VACV polycistronic transcripts are ‘regular’ because many are initiated or terminated within the ORFs.

Novel putative protein-coding genes

We identified novel genes in new intergenic positions, embedded genes (5’-truncated in-frame ORFs within larger canonical ORFs), and short upstream ORFs (uORFs, preceding the main ORFs).

(a) 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. Both of them are situated between two divergently oriented genes. They exhibit short ORFs with 51bps. Their TSS positions have been already reported in previous publications^{299,301} and our analysis showed that these ORFs are expressed alone in separate transcripts and/or as upstream ORFs (uORFs) located 5’ of canonical ORFs in longer transcripts (*Figure 11*).

(b) Embedded genes A characteristic feature of the VACV genome is that it contains several shorter in-frame ORFs. These have been proved by ribosome profiling experiments that they are able to be translated^{299,301}, because several internal translational initiation sites (TISs) occurred within the VACV

ORFs. Our full-length sequencing identified 49 novel putative embedded genes with 5'-truncated in-frame ORFs which were shorter than the *in silico* annotated canonical ORFs. This study demonstrated that these putative genes specify more than 300 transcript isoforms. We initially assumed that the TISs of these internal ORFs were the closest in-frame ATGs to canonical host ORFs. However, this is not necessarily the case, a large number of TISs were detected by others using ribosome profiling²⁹⁸. We identified 22 unique promoters associated with these TISs using our in-house scripts but we detected only six transcripts that contain in-frame ORFs.

Figure 11. Schematic representation of two examples of the identified putative new 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.

(c) **Upstream ORFs (uORFs)** are considered common regulatory RNA elements in mammalian transcripts³⁰⁴. 5'-UTRs of uORF-containing transcripts are reportedly ≥ 75 bp in length³⁰⁴. Several *in silico* and experimental studies show that 40–50% of human and mouse mRNAs contain at least one uORF^{305,306}. In the dataset obtained and filtered by the LoRTIA toolkit we revealed that 25 previously annotated VACV genes express longer TSS variants than the canonical transcripts. On average, 5'-UTR of the main transcript isoforms were 101 bp (*Supplementary Figure 13.*) whereas the longer variants were 258 bp (*Supplementary Table 9*). Most of these transcript isoforms (64%) contain at least one ORF preceding the main ORF with an average 5'-UTR of 331 nts. We set the minimal size limit of these uORFs to 9 nts comprising at least one triplet between the start and stop codons, as described by Scholz and colleagues³⁰⁷. We also set a maximal ORF length of 90 nts and considered transcripts with longer upstream ORFs as bicistronic. Eleven transcripts with 5'-UTR variants were found to contain multiple uORFs. G5.5R was the only gene in which the longer transcript isoform contained a single uORF. The VACV transcriptome is distinguished by high number of TSS and TES positions and combinations per ORF. The same TSS can be used by the longer variants of downstream genes and by a 5'-truncated, potentially protein coding transcript of the upstream gene. In many such cases, the overlapping part of the two transcripts (the long 5'-UTR of the downstream gene and the 5'-truncated version of the upstream gene) contains one or more ORFs that can be considered uORF (within the L variant) and/or

a novel potential protein-coding genes, as for A1L, A34R, and H3L regions, and C5L. The present novel transcripts of possible uORFs are listed in *Supplementary Table 9*.

Non-coding transcripts

In the LoRTIA analyses presented here, we identified 356 novel putative long non-coding RNAs (lncRNAs) of longer than 200 bps. We also identified 13 potential short ncRNAs (sncRNAs) with lengths varying between 121 and 193 bps. These ncRNAs are classified as antisense RNAs, 3'-truncated transcripts, replication-associated transcripts and complex transcripts.

Antisense RNAs

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, that is why the LoRTIA tool was unable to identify these molecules probably. After applying strict annotation criteria, exactly 100 antisense transcripts were located altogether at the A5R ORF a single antisense transcript, at the A18R ORF ten, and at the A11R ORF 89 antisense transcripts (*Figure 10B*). The latter transcripts are various combinations of six TSS and 40 TES positions, whereas antisense A18R transcripts have the same TSSs but vary in their termination sites. We confirmed asRNAs of A18R with qPCR as well (*Supplementary Figure 11*). Some of these transcripts contain small ORFs, which might indicate coding potential (*Supplementary Note 2*).

Intragenic non-coding transcripts

After filtering transcripts with strict criteria, 260 3'-truncated transcripts were determined from the entire dataset. They were considered non-coding because they lacked an in-frame stop codon. These transcripts were found within 41 VACV genes and 36.15% of them were located within the O1L region. The A37R (8.85%) and A10L (5.38%) regions and the F12L-F13L gene cluster (11.15%) also contained substantial numbers of 3'-truncated RNAs, and cumulatively, 38,47% of the 3'-truncated ncRNAs were mapped within five ORFs (A11R, A12L, A14L, A16L, A18R,).

Replication-associated transcripts

The VACV genome telomeric region contains multiple replication origins (Oris) located near the genomic termini in the terminal repeat region, called concatamer junction, as it has been demonstrated by SRS technique³⁰⁸. Earlier studies proved that VACV telomeres contain promoters for late RNA production, which may involve in concatemer resolution or in replication. Some of these late telomeric transcripts bridge the entire Ori region^{255,309}. Our raw data showed that Oris are overlapped by many non-coding transcripts that are expressed in very low abundance. The TSSs and TESs of these ncRNAs were not identified by LoRTIA during the transcript annotation because the transcripts vary widely in

size. Nonetheless, we demonstrate that these RNAs are expressed from 4h and accumulated up to 8h post infection and they may play identical role as replication-associated RNAs in Herpesviruses³¹⁰.

Miscellaneous, complex transcripts

Complex RNAs (cxRNAs) are defined as a transcripts containing at least one oppositely oriented ORF on a multigenic transcripts. These transcripts has been also found in Herpesviruses³¹¹ presumably other members of various virus families are producing this new type of RNA as well. Because 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 Herpesviruses³¹². In VACV, the LoRTIA annotated 30 cxRNA molecules (*Supplementary Table 10*) localized within nine genomic regions. We considered six of these cxRNAs as 3'UTR isoforms of the A12L–A11R and four of these as alternative TES variants of VACVWR_161-A38L RNA. Five as TSS and TES combinations within the A14L-A13L-A12L-A11R region. Two cxRNAs (c-A21L-A20R, c-G3L-G2R) were considered non-coding because their upstream genes were antisense on the transcript. Completely overlapping convergent cxRNAs were also detected at E10R-E11L and A8R-A9L genomic regions (*Supplementary Figure 14*). The remaining complex transcripts were categorized as mRNAs, because their first ORFs were in the sense direction (*Supplementary Table 10*)

Cis-regulatory sequence prediction of novel transcripts

Yang and colleagues described a 15-nucleotide consensus promoter core motif (AAAANTGAAAANNNA) upstream of E genes. We used *in silico* promoter predictions to detect potential TATA- and CAAT-box consensus sequences of the present TSSs. We identified TSSs for 90% of the promoters described by Yang et al., (2010), with the exceptions of G2R, J6R, A18R, A20R, and five hypothetical genes located in the repeat region of the VACV genome. The used promoter prediction tool identified the VACV early promoter motif at nine novel positions upstream (C9.5L, F4L, F7L, J3R, A15R, VACWR_161, and A51R) and inside (A36R and VACWR_169) of ORFs. Promoter elements have not previously been annotated within these regions. We analyzed 146 TSSs without previously annotated promoters and found TATA-boxes in 88, and CAAT-boxes in 5 of these (*Figure 12*). The median distance between TSS and putative TATA-box was 54 nts $\pm$ 24 SD. Yet, no significant differences in promoter–TSS distances were identified between 5'-truncated and canonical transcript isoforms, suggesting that these cis-regulatory structures were similar in these two types of VACV genes (*Figure 12*).

Figure 12. *In silico* analysis of promoter elements; a. the bar chart shows average distances between the TATA boxes and TSSs. b. The WebLogo shows the consensus sequence of the identified TATA boxes.

Additionally, most TESs of novel transcripts are located about 50 nucleotides downstream of UUUUUNU termination signals, as described by Yang et al.³⁰¹ Those TESs which were not preceded by this motif might use an alternative mechanism for polyadenylation-site selection. Matching TSSs with TESs can only be carried out using LRS techniques, which are therefore crucial for the transcriptome analysis, especially in VACV which exhibits a very complex gene expression profile. We found that in certain DNA segments of 'chaotic' regions almost all base positions can function as TESs.

Transcriptional overlaps

Transcript isoforms include 5'- and 3'-UTR length variants of RNA molecules. Using RNA-Seq tag-based methods with SOLiD and Illumina sequencing platforms, Yang and co-workers determined 5'- and 3'-ends of VACV early mRNAs with base pair precision^{299,301} and mapped the 3' termini of postreplicative (PR) transcripts²⁶³. These 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 ORFs 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. We also detected 106 parallel transcriptional overlaps between tandem genes. Seventeen of the 32 convergent gene pairs formed UTR-UTR overlaps, whereas only 15 gene pairs formed UTR-ORF overlaps. This huge number of overlaps hamper to categorize VACV transcripts in a usually way, because it is up against the usual transcriptome structure of alpha Herpesviruses^{280,313} and only partly resembles it³¹⁴.

Discussion

Genome reconstruction

We used the PacBio cDNA data to reconstruct the VACV genome sequence. The length of the reconstructed VACV genome sequence of the WR strain was composed of 194,888 base pairs. The average GC content was 33.3%, which corresponds to the other WR strains. We annotated 218 protein-coding genes, including 3 novel ORFs (B25.5R/C19.5L, C9.5L) in the intergenic region. Our data showed that the sequenced WR strain differs from the reference genome in 163 point mutations, mainly in the F region. The F region contains core genes with diverse functions from anti-apoptotic modulators to structural proteins of the mature virion^{315,316}. Most of the amino acid substitutions occurred in the C9L gene. The putative product of this gene is required to resist the human interferon-induced state and counteracts interferon-stimulated genes³¹⁷. The annotated new genes (B25.5R/C19.5L, C9.5L) lay in the highly variable genome region²³³ and they probably code for ankyrin-like proteins, which mediate virulence through the regulation of the NFkB pathway^{245,318}. The complete and annotated genome sequence of this VACV WR strain has been deposited in the European Nucleotide Archive under the accession number: [LT966077](#).

Transcriptome annotation

It is a major advantage that LRS methods can perform end-to-end sequencing of entire RNA or cDNA molecules, allowing identification of full-length transcripts. We used cDNA libraries prepared from VZV and VACV infected cells to sequence and depict the transcriptomic landscape in lytic life cycle of these viruses. Twenty-five different preparation methods were used to obtain VACV RNAs and three different methods for VZV transcriptome sequencing. The PacBio SMRT and ONT MinION long-read sequencing platforms were used for VACV and the ONT MinION platform for VZV transcriptome analysis. Both of the chosen techniques are capable of sequencing full-length RNAs in the form of cDNA. The nanopore sequencer can analyze native RNA as well. Since SRS fails to detect full-length transcripts, theoretically every RNA molecule identified by LRS techniques can be considered to be novel in VACV. Our transcriptome analysis was based mainly on reads derived from the poly(A)-selected and Cap-selected libraries. Despite the relatively high inaccuracy of sequenced reads - compared to the Illumina platform – we could detect start and end points of transcripts (TSS and TES respectively), and even splice junction positions with base-pair precision.

We applied a novel statistical algorithm for VZV analysis to detect significant read accumulations alongside the mapped read population. This new method was later implemented by Zsolt Balázs as a python3 program and has become a complete bioinformatical pipeline for general transcriptome analysis. We used his toolkit for the annotation of VACV transcriptome. There are more than 350 tools and scripts available mainly for genomic analysis of long-read data, but only 36 for transcriptome analysis, therefore our dataset and transcriptome annotation pipeline are pioneering in the rapidly

developing area of transcriptomics⁸⁰. The dataset we have generated is one of the few openly accessible long-read RNA materials. All of the raw sequencing data presented here were deposited in the European Nucleotide Archive (ENA) under the project accession numbers [PRJEB25401](#) (Long-read Sequencing Dataset of Varicella-zoster virus), [PRJEB26434](#) (Characterization of the Vaccinia virus transcriptome), and [PRJEB26430](#) (Dynamic characterization of the Vaccinia virus transcriptome) and can be used for further transcriptomics studies.

We obtained shorter average read lengths in our VACV ONT dataset compared to PacBio SMRT sequencing. This can be explained by the differences in preparation procedures. The PacBio technique's library preparation protocol prefers RNAs longer than 500bp, while the nanopore protocol does not differentiate in the size of RNAs. Similar results were observed in other viral studies^{281,319}.

Although these results are mainly descriptive, functional transcriptomic questions cannot be answered without structural transcriptomic knowledge. Detailed genome annotations are available for model organisms and humans, but rarely for viruses. As a result, alternative RNA isoforms or overlapping transcription structures are underrepresented, therefore the complexity of transcriptomes remains undetected²⁰¹. Most often only ORFs are annotated and not entire genes, although genes can be inferred by the determination of transcript boundaries. LRS techniques are able to determine alternative TSSs, poly(A) sites, and alternative exon usage. By contrast, general short-read RNA-seq protocols rarely detect TSSs and poly(A) sites and usually only recover splice donor and acceptor sites.

Earlier results of various experiments (primer extension, S1 nuclease assay, and Illumina sequencing) determined the TSS or TES of 37 VZV transcripts^{187,196,320–323}. These transcripts mainly encode regulatory proteins, playing roles in viral DNA replication (e.g. ORF0, ORF19) or in latency and reactivation (e.g. ORF21, ORF62, ORF63). We confirmed these ends and also determined the unknown 5' or 3' end sites of these transcripts. Using the ONT sequencing platform, we confirmed 18 previously known TSSs and nine TESs. Additionally, we annotated 124 new putative TSSs and 71 TESs. Altogether we identified 114 novel transcripts in VZV, including mRNAs, non-coding RNAs, polycistronic RNAs, and complex transcripts.

Previous studies have detected large numbers of TSSs, TESs, and TISs of VAVC transcripts using SRS^{261,298,300,324}. However, these analyses have not identified full-length transcripts because SRS techniques are unable to match TSS and TES positions.

A distinctive advantage of long-read sequencing is that it can detect overlapping transcriptional units as well, especially short, 5' truncated transcripts, whose common poly(A) signal is shared with the host gene. Both Cap- and poly(A)-cDNA sequencing detected in VZV and VACV transcriptomes numerous embedded, 5' truncated transcripts which is in concordance with others' findings among herpesviruses^{109,325,326} and in other viruses, including baculoviruses and asfarviruses^{327,328}. Our results suggest that these embedded transcripts may contribute to a new group of possible protein-coding

transcripts in DNA viruses. Although we can only speculate whether the short ORFs encoded by these embedded transcripts are translated or not, ribosome profiling analysis of the VACV transcriptome found several ‘anomalous’ ribosome binding sites³²⁴ which could correspond to our results extending the protein-coding capacity of DNA viruses. Furthermore, we found essential transcription regulator motifs located upstream to the embedded TSSs. The average promoter distance to TSSs of embedded, 5'-truncated VZV transcripts (*type .5* transcripts) was longer than in the case of non-embedded, 5'-truncated VZV transcripts. However, this was not true for VACV embedded, 5'-truncated transcripts, where the promoter found upstream of their TSSs was not situated significantly farther than their isoform's TSS. It is worth noting that we could not find canonical promoters for all *.5 type* VZV or VACV transcripts, therefore our findings may not represent the entire population of this transcript family. Our transcript annotation pipeline identified 195 TSSs and 704 TESs altogether, which contributed to 1480 novel transcripts in VACV transcriptome.

Although long-read sequencing is superior for overlapping transcripts, the technique itself suffers limitations. The library preparation protocol of the PacBio SMRT technique selectively prefers longer reads: reads shorter than 600bp are practically missing from the libraries, while the nanopore sequencer has no limitation on size of the sequenced molecules. Another problem during library preparation of LRS protocols is that PCR may introduce amplification bias, and false priming of oligo(dT) primers and adapter sequences can also occur, resulting in non-valid transcript end positions³²⁹. We have considered these errors in our transcript annotation pipeline. Moreover, recent nanopore sequencers can identify native RNA molecules without using the reverse transcription step, eliminating artifacts of library preparation³³⁰.

Multigenic transcripts

A distinct feature of the VACV transcriptome is that the transcriptional apparatus selectively recognizes the UUUUUUNU termination signal, therefore early genes are prevented to produce longer transcripts, while late genes are not terminated after the PAS in late time points of infection, resulting in extensive transcriptional read-through into the neighboring downstream gene(s). This phenomenon was already known in the VACV transcriptome^{26,299,300}. We confirmed not only the read-through of late genes, but also identified 286 polycistronic transcripts and transcript isoforms. While eukaryotic mRNAs are still not accepted widely as transcripts with multiple functional ORFs, growing evidence reveals that eukaryotes and their viruses can use two protein-coding genes on an mRNA for effective translation³³¹.

The non-canonical mechanism of translation initiation has been studied extensively in viruses and mammalian cells, as well. Initiation of translation at a downstream ORF is usually achieved in RNA viruses by conserved structural sequence motifs called internal ribosome entry sites (IRES), however, other viruses use non-cap dependent translation initiation. These processes still need further elucidation.

The longest, annotated transcript of VACV was the F5L-F4L-F3L-F2L-F1L with 4616 bp, and the longest, non-annotated, complex transcript, overlapping seven CDSs (by a single read) was 5192 bp long in the E2L-F15L region (data not shown). The longest transcript in VZV was the ORF28-27-26-25-24-C at 8340 bp. Altogether, thirty-six polycistronic transcripts were found. Although these transcripts are rare, long multicistronic transcripts may only be produced in cell culture propagated viruses as transcriptional noise, thus further studies need to prove their functional role.

Overlapping coterminal transcription units with common PAS are characteristic of herpesviruses. In the VACV transcriptome the large number of 3' UTR isoforms suggests a different transcriptional structure. We detected long 5'-UTR isoforms containing uORFs in most VACV genes, moreover, some transcripts contain multiple uORFs, and most isoforms contain unique combinations of ORFs and uORFs.

We can ask what is the function of length in these transcript isoforms. Despite the general supposition that only the first ORF is translated on long eukaryotic mRNA, emerging evidence suggests, that ribosomes can use upstream and downstream ORFs as templates on multicistronic mRNA. This scenario indicates a transcriptional-translational regulatory mechanism, namely the 5' UTR isoform of a given transcript, comprising uORF, may regulate the translation of a downstream oriented ORF, as is described in Kaposi's sarcoma-associated herpesvirus (KSHV)³³². Another study suggests that leaky scanning allows primary ORF expression, while a reinitiation mechanism after the translation of the uORFs enables the expression of bicistronic downstream ORFs in KSHV³³³. While this is a rare example of uORF playing a role in enabling viral polycistronic translation, it is possible that a similar mechanism may regulate a subset of the herein described mRNA isoforms containing uORFs.

Our results suggest that at least some of the uORFs could play a role in the production of N-terminally truncated proteins, while most of them have sufficiently large space between their stop codons and the protein-coding ORF's AUG for initiation events, thus resulting in unaltered protein translation³³⁴. Further studies of ribosome profiling and mRNA and protein expression studies are needed to determine the precise function of these uORFs. Nevertheless, the use of alternate promoters for producing TSS variants with or without uORFs seems to play a crucial role in human cytomegalovirus (HCMV), providing differential control of translation at distinct stages of the viral lifecycle¹⁰⁹.

The widespread occurrence of adjacent genes in chimeric transcripts suggests that transcriptional read-through followed by cis-splicing may occur. Thus, the existence of transcriptional read-throughs is not just restricted to the herpesviruses, but they may represent a general phenomenon in VACV. Furthermore, the antisense RNAs produced from their own promoters or by transcriptional read-throughs may hybridize with their sense counterparts thereby initiating RNA interference³³⁵⁻³³⁷. Very long complex transcripts form a distinct category among multigenic transcripts because of their oppositely oriented ORFs and increased size. Similarly to other herpesviruses^{109,311} they are present in

very low abundance in VZV, however, their existence is ambiguous, as they can be formed during reverse transcription by template switching³³⁸.

The enrichment of *As* and *Us* upstream of PASs in mammalian systems is well-established^{339,340}. These homopolymer *A* stretches may also cause false priming and template switching events during reverse transcription, which results in false TESs³⁴¹. In our work, we excluded these artifacts by using our analysis pipeline and demonstrating the verity of the annotated TESs by the presence of the canonical GU-rich region in the + 10 interval downstream of TESs³⁴². The majority of the length isoforms were 3' UTR variants of transcripts, especially in VACV transcriptome. Such a large number of transcriptional read-throughs have been detected by others in late time points of VACV infection^{262,300,324}.

The function of the detected novel transcript isoforms is still unknown. The majority of the transcripts contain protein-coding sequences, thus they probably code for proteins. Translatome and proteome analyses could answer the question of which transcript isoforms contribute to viral proteins. VACV has been analyzed by ribosome profiling and it was found that 5' truncated transcripts bind ribosomes at their downstream ATG and VACV probably produces short, anomalous proteins²⁹⁸. Whole proteome analysis of VACV identified 136 proteins during early and late of infection and 162 proteins in the mature virion^{343,344}. These findings suggest that a remarkable number of transcripts are not expressed as proteins.

The high number of transcript isoforms may also play an important role in viral transcription regulation. One explanation could be that different isoforms are expressed by different kinetics, based on their alternative promoter usage. In VACV early and late genes have distinct promoter motifs, however such a conservative regulatory is yet not known in VZV. Another explanation could be that neighboring genes constitute transcriptional units, and length isoforms -especially 3' end isoforms – affect downstream gene expression. There is a hypothesis that viruses use transcriptional interference for fine-tuning their temporarily expressed genes, and the whole viral transcriptome builds a transcriptional interference network (TIN)³³⁷.

Splicing

Splicing and alternative splicing are major mechanisms increasing the complexity of gene expression in viruses. Interestingly VACV mRNAs are not synthesized by splicing mechanisms and introns are completely missing in poxviruses²⁰⁶. Splicing is rare in the VZV transcriptome. In earlier studies, only five genes of 70 seemed to be spliced¹²⁵, however splicing is much more significant in other alphaherpesviruses^{314,345,346}. Although splice events are thought to be rare in VZV, in light of LRS techniques they appear to be underestimated¹⁶⁴. In this work, we enriched the list of spliced transcripts in the lytic phase of viral infection, and we identified novel combinations of splice sites in ORF42/45. We also detected many more introns with LRS. It is known that the library preparation procedure can introduce false introns and trans-splicing, therefore we applied strict criteria and implemented a

template-switching filter into the annotation pipeline^{347,348}. The scarcity of splice isoforms using techniques capable of detecting rare transcripts provides further validation. A very new study published during the preparation of the thesis confirmed our findings by direct RNA sequencing and increased the number of splice sites by another one in VZV³⁴⁹.

RNA editing

More than one hundred RNA base modifications have been documented to date. These modifications have functional roles in mRNA stability or translational efficiency⁸⁰. Hyper-editing of given positions of mRNAs was previously observed in members of herpesviruses^{350–352}. It probably plays a crucial role in the cell's innate immunity³⁵³, while it can be hijacked by some viruses to evade inactivation³⁵⁴. Additionally, in *hepatitis delta virus*, hyper-editing is indispensable for replication³⁵⁵. It is also known that 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 homologues³⁵⁶. We did not detect hyper-editing in any of the VACV transcripts, but it was observed in the novel NTO3 VZV transcript, which is expressed in the vicinity of replication origo. Our *in-silico* analysis showed 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. Rather it affects the formation of I·U wobble pairs, which are significantly more resistant to possible Dicer cleavage inhibiting RNA interference³⁵⁶. 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)^{357,358}. Our conclusion is that this type of RNA editing is special in VZV and further investigation is needed to elucidate the significance of hyper-editing in transcriptomics studies.

nroRNAs

Replication of VZV starts with the theta-type process, followed by the rolling circle mechanism, similar to the replication of bacteriophages, however, some evidence suggests that the replication of herpesviruses generally may be more complex and concatemers may be formed by recombination³⁵⁹. Nevertheless, replication-associated transcripts have also been identified in viruses, bacteria, and eukaryotes. One of the main outcomes of our genome-wide transcriptome analysis is the identification of replication origo (Ori) associated transcripts in herpesviruses. Several years ago our research group published the existence of CTO and PTO non-coding transcripts in PRV³⁶⁰. During the analysis of HSV-1, and VZV transcriptome become clear that RNAs are expressed proximal to the Ori of these viruses as well, or they completely overlap it³¹⁰. The same pattern is observable in other herpesviruses, e.g. HCMV³⁶¹, KSHV³⁶², and EBV³⁶³.

It was previously thought that the OriS (Ori of the US genomic region) of VZV lacks any overlapping transcripts¹²⁵, although Davison and colleagues (1986) hypothesized the existence of non-coding RNA in the intergenic region between *orf62* and *orf63*¹²⁵. 32 years later, our results confirmed Davison's

hypothesis. We detected several transcripts overlapping OriS, and small non-coding RNAs (the NTOs) between the aforementioned genes. Moreover, this region of the viral transcriptome is structurally and positionally similar to the PRV³⁶⁴, HSV³²⁵, and EHV transcriptome (unpublished data), however, the sequences of these transcripts are non-homologous.

The gene *orf63* in VZV corresponds to the gene *us1* in PRV and its gene product is a homolog to the IE ICP22 protein of HSV-1. Our earlier research proved that *us1*-deleted PRV showed delayed replication in lytic infection³⁶⁵. This protein is one of the main transcriptional regulators in both HSV-1 and VZV. Gene expression is highly restricted in the latent phase of VZV¹⁶¹, only ORF63 was found to be co-expressed with the recently discovered multi-exonic VZV latency transcript (VLT)^{163,164}. We confirmed that isoforms of VLT are also expressed in the lytic phase¹⁶³, furthermore, we found long (up to 6kbp), multi-spliced polycistronic complex transcripts and isoforms, which we named VLT_{ly}-ORF63-C(1-6). A recently published study independently confirmed the findings that we proposed the first time (Prazsák et al. 2018, Fig 7.): this long, multi-spliced complex transcript forms a translatable fusion mRNA, which probably acts as a transcriptional switch in reactivation from latency^{164,349}. *In vitro* expression assays have shown that our *in silico* prediction, that it codes at least two protein isoforms of VLT-ORF63 fusion transcript, is possible³⁶⁶. An *in vivo* immunoblotting experiment illustrated that N-terminal length variation in ORF63 proteins encoded by different VLT-ORF63 transcript isoforms allocated differentially in the infected cell's nucleus and cytoplasm³⁶⁶, suggesting their distinct functions.

VACV DNA synthesis is initiated near the genome terminus, where a repeat region forms a hairpin end that acts as a conserved replication initiation site. SRS analysis has revealed multiple potential replication start sites. It has been demonstrated that VACV telomeres contain promoters for late RNA production, which may be involved in concatemer resolution or in replication^{309,367}. Some of these late telomeric transcripts (lateRNAs) bridge the entire Ori region. The 3' ends of the lateRNAs vary in length, and some of them encompass the entire telomeric hairpin, which contains the Ori. It has been proposed that ncRNAs expressed within the concatemer junction might play a role in opening up the DNA duplex during the initiation of the replication³⁰⁸ and nroRNAs are also present in orthopoxviruses³¹⁰.

Conclusion

The genetic information of viruses and their host cells is encoded in their genome and in their transcriptome. By decoding this information, we see new faces of a particular pathogen: how they work, what genes are expressed when. With accurate transcriptome annotation, RNA-seq experiments can be performed to investigate the differential expression of viral genes or predict which proteins are present in a host cell during the viral life cycle. The discovery of new protein-coding, non-coding genes and transcript isoforms helps us in the fight against pathogens. Long-read sequencing technology overcomes the limitations of short-read sequencing technology, therefore it is an ideal tool to annotate viral transcriptomes.

The human genome and genomes of model organisms are well annotated, but more often than not, precise annotation is missing or oversimplified for viruses. Although NGS techniques revolutionized genomics and transcriptomics, technical limitations have hampered the inference of viral genomes,. SRS produces high coverage, so highly accurate variant detection is possible and many bioinformatics tools and pipelines support SRS⁷⁴. Natural RNAs can be eight orders of magnitude longer than amplified fragments during sequencing, which complicates the task of reconstructing their original features. As a result, the presence of complex, overlapping transcripts remain undetected, and impair subsequent interpretations of viral biology^{201,368}. VZV and VACV genomes contain duplicated regions with repeats which can result in short or long-read sequences being automatically discarded. Long-read sequencing can thus improve *de novo* assembly, mapping certainty, transcript isoform detection, and identification of structural variants, and base modifications without amplification bias^{80,201}. Applying LRS techniques for the first time in the analysis of VZV and VACV, we have broadened the number of possible protein-coding transcripts and non-coding transcripts of VZV and detected more than one thousand transcripts of VACV. New and alternative introns were also found in VZV, increasing the coding capacity of this human pathogen. A new family of RNAs, the replication origo associated transcripts were also described in VZV for the first time.

With the development of LRS techniques, a new door has been opened in the field of transcriptomics, and we have laid new slabs on the path with reannotation of known and annotation of never seen transcriptomes.

In my dissertation, I undertook to perform the first comprehensive transcriptome analysis of the human pathogenic Varicella-zoster and the Vaccinia virus with long-read sequencing, in order to better understand and represent the fundamentals of viral gene expression.

Acknowledgements

Foremost, I thank my advisors, Prof. Zsolt Boldogkői and Dr. Dóra Tombácz, for their encouragement and help with my research; they have not only supported my work, but formed my attitude towards science. I am incredibly grateful to my colleagues at the Department of Medical Biology and to the entire staff of the 3G research group for their support of my work, which would not have gone nearly as far without them. I kindly thank Dr. Norbert Moldován, for his invaluable advice, Dr. Zsolt Csabai, Dr. Zsolt Balázs, Dr. Attila Szűcs and Dr. Péter Oláh who are coauthors in the publications and made it possible to evaluate the experiments; the huge amount of data could not have been prepared without their hard work. I thank Dr. Klára Megyeri (Department of Medical Microbiology, University of Szeged) and Dr. Béla Dénes (National Food Chain Safety Office, Budapest) for their assistance and excellent virological work, which enabled the RNA extractions. I thank Prof. Michael Snyder for the PacBio experiments at the University of Stanford.

I thank Jenifer Tusz for her critical reading of the thesis. I thank to Romuáld Máthé OSB for igniting my interest in the world of cell biology as a student. I thank my family and friends (foremost to KLIK and ASK) their support at weekends. Lastly, I wish to thank my wife and my children their support and infinite patience during the years of my research work. I wish I could share this achievement with my father and mother.

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.

Supplementary documents of the publication “Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus”

All figures and additional tables, containing data which are too large to print out are available at: <https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-5267-8>. Here the printable figures and tables are represented.

Primer name	Sequence
pNTO2	ACTTGCCTGTCGCTCTATCTTCCGGTCAGGATCT
pNTO3	ACTTGCCTGTCGCTCTATCTTCGTA CTGTTGGA ACT
pNTO4	ACTTGCCTGTCGCTCTATCTTCATGTGAGACAACC
pNTO1	ACTTGCCTGTCGCTCTATCTTCTCGTGTGACG

Supplementary Table 1

Target-specific primers used for the detection of NTO transcripts. The primers are composed of a sequence complementary to the MinION cDNA adapter and a target-specific sequence (in italics).

Supplementary Figure 1.

The procedure of finding novel transcriptional isoforms. The flowchart describes the pipeline used to detect novel TSSs and TES-s.

Supplementary Figure 2. The VZV transcriptome.

The transcriptome of the varicella zoster virus was analyzed using long-read cDNA sequencing. The reads were annotated using our custom script (<https://github.com/zsolt-balazs/LoRTIA>). Transcriptional start sites and transcriptional end sites of the annotated transcripts are the 5' and 3'-ends with significantly higher abundance than any other read end in their ± 50 vicinity, according to Poisson-probability, and those that are not a result of strand switching or false priming

55

103935	103970	35	137	-	in 5'UTR	
104176	104256	80	378	-	in 5'UTR	
104176	104232	56	378	-	in 5'UTR	
109234	109263	29	101	-	in 5'UTR	*
109234	109356	122	101	-	in 5'UTR	
109284	109349	65	151	-	in 5'UTR	
111461	111496	35	69	+	in 5'UTR	
112676	112714	38	36	-	in 5'UTR	
112676	112735	59	36	-	in 5'UTR	
112731	112763	32	91	-	in 5'UTR	
115743	115775	32	33	+	in 5'UTR	
118401	118436	35	69	-	in 5'UTR	
23788	23835	47	-6	-	in ORF	
59650	59856	206	-90	+	in ORF	*
59668	59856	188	-90	+	in ORF	
62131	62229	98	-7	-	in ORF	
82562	82600	38	-31	-	in ORF	
101168	101260	92	-51	-	in ORF	

*Supplementary Table 2. Upstream open reading frames (uORFs) of the VZV transcripts. We detected a total of 44 uORFs in silico, 38 of which are starting and terminating in the transcript's 5' untranslated region (5'UTR), while six are terminating in the ORF of the mRNA, the later resembling upstream ATGs. The uORFs with Strong Kozak consensus sequences are marked by *.*

Supplementary Figure 4. Types of overlaps between VZV transcripts. Reads were visualized using IGV. Blue arrow-rectangles represent transcript annotations. The overlaps are framed in orange. a. Parallel overlaps of the ORF5-4-3 cluster. b. Divergent overlaps of the ORF8 and ORF9 and ORF9A-9 transcripts. c. Convergent overlaps of the ORF17 and ORF18-AT transcripts

Supplementary Figure 5

*A to I hyper-editing of NTO3. a. Reads of NTO3 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 NTO3 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 NTO3 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 suite with an E-value of 0.58 and a log likelihood ratio of 46. The edited adenosine is marked with a **

Supplementary Figure 6

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 NTO3 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

Intron Start	Intron End	Strand	Intron Length	Donor Site	Acceptor Site	Accepted as novel	Intron depth	Transcript name	References
585	714	+	129	GT	AG		45,62%	ORF0-1-C-SP	369
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	125
81537	79053	-	2484	GT	AG	✓	12,72%	ORF42/ORF45-SP-2	
87759	86782	-	977	GT	AG		4,15%	ORF50B	142
87759	86982	-	777	GT	AG		2,33%	ORF50A	142
87436	86782	-	654	GT	AG		0,62%	ORF50C	142
86982	87436	-	454	GT	AG		0,33%	ORF50D	
101728	102420	+	692	GT	AG		87,50%	VLT _{ly}	163
102484	102853	+	369	GT	AG		100,00%	VLT _{ly}	163
102983	103827	+	844	GT	AG		2,42%	VLT _{ly}	163
103924	104293	+	369	GT	AG		3,99%	VLT _{ly}	163
104433	104768	+	335	GT	AG		5,22%	VLT _{ly}	163
104824	110509	+	5685	GT	AG	✓	2,18%	ORF63-SP2-C	
108830	110509	+	1679	GC	AG	✓	11,12%	NT01v1 (ORF63-SP 1)	
114920	115050	+	130	GT	AG	✓	1,04%	ORF67-SP	
121067	119388	-	1679	GC	AG	✓	10,65%	NT01v1 (ORF63-SP 1)	

Supplementary Table 3. Splice junction sites of the VZV transcriptome. 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 6. The TSS and TES belonging to the transcript with splice sites in row 7 could not be determined with certainty.

Supplementary documents of the publication “Long-read Assays Shed New Light on the Transcriptome Complexity of a Viral Pathogen”

All figures and additional tables, containing data which are too large to print out are available at: <https://www.nature.com/articles/s41598-020-70794-5#Abs1> Here the printable figures and tables are represented.

Supplementary Note 1. Terminology of the listed transcripts.

X2R= a given CDS

X2R – the most abundant transcript, probably the main transcript of a given CDS

X2.5R – genes in novel genomic position

X2R.5 - new embedded, potential protein coding transcript

X2R-AT – transcript isoform with alternative 3’ end, resulting a longer 3’ UTR (may contains dORF)

X2R-l – long transcript isoform with alternative TSS, resulting a longer 5’ UTR than the main transcript (these may include uORF)

X2R-s – short transcript isoform with alternative TSS, resulting a shorter 5’ UTR than the main transcript

nc-X2R – non-coding transcript isoform terminating before the stop codon of a known ORF or transcript without ORF

tr-X2R – truncated transcript isoform within a CDS containing no start codon

c-X2R-X1L - complex transcript isoform, spanning oppositely oriented ORFs

as-X2R - antisense transcript isoform, oriented oppositely to the given ORF

1. Supplementary Note 2. Information for downloading and using the Geneious file.

The whole transcriptome profile of the VACV available at **FigShare**: 10.6084/m9.figshare.10191230

The dataset can be view by using Geneious software which is available at <https://www.geneious.com/free-trial/>

Features found in literature were transferred from the VACV reference genome (NC_006998) to the LT966077 with the Geneious 10.2.6 Annotation Transfer tool.

The following abbreviations and terms are used in the file:

Anomalous TSS: RNA 5' ends were mapped within ORFs and antisense to ORFs [15,19]

Cat box: CCAAT-box, upstream regulatory sequence of TSS [42]

CDS: coding sequence;

CDS-corr: correction of annotated ORF; based on the ribosome profiling data [12]

CDS-d: downstream frameshifting ORF; based on the ribosome profiling data [12]

CDS-n: previous unannotated non-coding region ORF; based on the ribosome profiling data [12]

CDS-u: upstream ORF; based on the ribosome profiling data [12]

CDS-tr: downstream truncated ORF or upstream non-frame shifting ORF; based on the ribosome profiling data [12]

Early Promoter motif: sequence motif (AAAANTGAAAANNA) of early gene's promoters [14]

Early PAS motif: sequence motif (TTTTTNT) for polyadenylation site of early genes [15]

mRNA: LoRTIA annotated transcript

Late Promoter motif: sequence motif (TNNNNNNNNNTAAATG) of late gene's promoters [19]

LoRTIA: Long-read RNA-Seq Transcript Isoform Annotator toolkit [36]

PAS: polyadenylation site

TIS: translational initiation site

TSS: transcriptional start site

VACV: Vaccinia Virus

Total RNA isolation	PolyA selection	Ribodepletion	Reverse transcription & dsDNA production	cdNA synthesis by PCR	Size selection	Library preparation kit	Barcoding kit	Sequencing chemistry	Instrument
Macherey-Nagel RNA	Qiagen Oligotex mRNA mini Kit	Epicentre Ribo-Zero Magnetic Kit HMR	Clontech SMARTer PCR cDNA Synthesis Kit	KAPA HiFi PCR Kit	- SageScience BluePippin 0.75% Dye-Free Agarose Gel Cassettes	PacBio Template Preparation Kit	-	P6-C4 (DNA Sequencing Reagent Kit 4.0 v2 and DNA Polymerase P6)	RSII
			SuperScript IV	KAPA HiFi PCR Kit	- Thermo Fischer Ultrapore Agarose & ZymoClean Large Fragment DNA Recovery Kit	Ligation Sequencing Kit 1D	ONT PCR Barcoding Kit 96	NA	MinION
			SuperScript IV						
			SuperScript III						
			L exogen Teloprime Kit enzymes & reagents	L exogen Teloprime PCR mix	-	Direct RNA Sequencing Kit	-	NA	MinION
			-	-					

Supplementary Table 4. Summary table of the reagents and chemistries used for the sequencing

Supplementary Figure 7. Draft workflow of the study. Infection, sequencing, and data integration strategy for transcript annotation.

Supplementary Figure 8. Comprehensive experimental workflow of the PacBio and MinION sequencing.

Sequencing method	Library prep step	Name, availability	Catalog No.	Sequence (5' → 3')
PacBio amplified PolyA	RT	3' SMART CDS primer II A -SMARTer PCR cDNA Synthesis kit (Clontech)	634 925 & 634 926	AAGCAGTGGTATCAA CGCAGAGTAC(T)30V N
PacBio amplified Random	RT	Custom-made (IDT DNA)	–	AAGCAGTGGTATCAA CGCAGAGTACNNNNN N (G: 37%; C: 37%; A: 13%; T: 13%)
MinION cDNA	RT	Poly(T)-containing anchored primer [(VN)T20 - ONT recommended, custom-made (Bio Basic)	–	5phos/ACTTGCCTGTC GCTCTATCTTC(T)20V N
MinION CAP-Seq	RT	TeloPrime Full-Length cDNA Amplification Kit (Lexogen)	013.08 & 013.24	TCTCAGGCGTTTTTTT TTTTTTTTTTT
MinION dRNA	RT	RT adapter—Direct RNA Sequencing Kit (Oxford Nanopore Technologies)	SQK-RNA001	GAGGCGAGCGGTCAA TTTTCTAAGAGCAA GAAGAAGCCTTTTTT TTTT
MinION CAP-Seq	test qPCR	D1R fw—custom-made (IDT DNA)	–	CGAACTAGAGGACCG TTGGG
MinION CAP-Seq	test qPCR	D1R rev—custom-made (IDT DNA)	–	TTTCCAGGTCAGCAC CGTTT
MinION cDNA barcoded	barcoding	A1/>BC01/(ONT PCR Barcoding Kit 96)	EXP-PBC096	AAGAAAGTTGTCGGT GTCTTTGTG
MinION cDNA barcoded	barcoding	A2/>BC02/(ONT PCR Barcoding Kit 96)	EXP-PBC096	TCGATTCCGTTTGTAG TCGTCTGT
MinION cDNA barcoded	barcoding	A3/>BC03/(ONT PCR Barcoding Kit 96)	EXP-PBC096	GAGTCTTGTGTCCCA GTTACCAGG
MinION cDNA barcoded	barcoding	A4/>BC04/(ONT PCR Barcoding Kit 96)	EXP-PBC096	TTCGGATTCTATCGTG TTTCCCTA
MinION cDNA barcoded	barcoding	A5/>BC05/(ONT PCR Barcoding Kit 96)	EXP-PBC096	CTTGTCCAGGGTTTGT GTAACCTT
MinION cDNA barcoded	barcoding	A6/>BC06/(ONT PCR Barcoding Kit 96)	EXP-PBC096	TTCTCGCAAAGGCAG AAAGTAGTC
MinION cDNA barcoded	barcoding	A7/>BC07/(ONT PCR Barcoding Kit 96)	EXP-PBC096	GTGTTACCGTGGGAA TGAATCCTT
PacBio	adapter ligation	PacBio blunt adapter (PacBio Template Prep Kit 1.0)	PN 100–222-300	ATCTCTCTCTTTTCCT CCTCCT- CCGTTGTTGTTGTTGA GAGAGAT
MinION	adapter ligation	5' adapter (ONT Ligation Sequencing 1D kit)	SQK-LSK108	GGTGCTG
MinION	adapter ligation	3' adapter (ONT Ligation Sequencing 1D kit)	SQK-LSK108	TTAACCT

Supplementary Table 5. Primers sequences used in this study for the reverse transcription reactions. The table also contains the sequence of the gene-specific primer pair used for the amplification of D1R gene of VACV, as well as the sequencing adapters and barcodes.

Supplementary Figure 9. Bioinformatic pipeline for VACV study

A	Run #	Major specificities of libraries	Median Read-length	Median Aligned Read-length	Average Read-length $\pm$ SE	Average Aligned Read-length $\pm$ SE	Average insertion frequency $\pm$ SE	Average deletion frequency $\pm$ SE	Average mismatch frequency $\pm$ SE
	1	RSII mix no size selection	1331.5	1310	1385.66 $\pm$ 46.40	1361.5 $\pm$ 46.58	0.04 $\pm$ 0.0096	0.21 $\pm$ 0.0603	0.02 $\pm$ 0.0056
	2	RSII mix random primed	660	572	799.74 $\pm$ 88.58	714.87 $\pm$ 89.19	0.09 $\pm$ 0.0447	0.13 $\pm$ 0.0634	0.02 $\pm$ 0.0117
	3	RSII mix BluePippin size selection: 0.8-5kb+	1008	885	1152.67 $\pm$ 9.28	937.94 $\pm$ 3.12	0.25 $\pm$ 0.0084	0.30 $\pm$ 0.0056	0.11 $\pm$ 0.0047
	4	RSII BluePippin size selection: 0.8-2kb	1055	938	1103.85 $\pm$ 10.87	978.35 $\pm$ 9.58	0.07 $\pm$ 0.0070	0.15 $\pm$ 0.0136	0.07 $\pm$ 0.0291
	5	RSII BluePippin size selection: 2-3kb	998	897	1062.55 $\pm$ 5.78	947.92 $\pm$ 5.23	0.12 $\pm$ 0.0071	0.11 $\pm$ 0.0059	0.03 $\pm$ 0.0053
	6	RSII BluePippin size selection: 3-5kb	1053	942	1063.17 $\pm$ 2.89	946.56 $\pm$ 2.57	0.14 $\pm$ 0.0038	0.16 $\pm$ 0.0042	0.04 $\pm$ 0.0032
	7	RSII BluePippin size selection: 5kb+	1082	990	1118.95 $\pm$ 4.70	1022.00 $\pm$ 4.45	0.11 $\pm$ 0.0050	0.19 $\pm$ 0.0062	0.04 $\pm$ 0.0039
	8	Sequel 1h	902	701	1149.16 $\pm$ 40.59	774.48 $\pm$ 22.65	0.44 $\pm$ 0.0909	0.19 $\pm$ 0.0544	0.55 $\pm$ 0.1298
	9	Sequel 2h	882	725	1054.95 $\pm$ 29.44	821.98 $\pm$ 23.19	0.19 $\pm$ 0.0453	0.09 $\pm$ 0.0120	0.28 $\pm$ 0.0925
	10	Sequel 3h	988.5	828.5	1125.82 $\pm$ 19.84	909.89 $\pm$ 15.13	0.20 $\pm$ 0.0459	0.10 $\pm$ 0.0090	0.25 $\pm$ 0.0597
	11	Sequel 4h	840	676	1128.46 $\pm$ 36.37	727.83 $\pm$ 19.74	0.34 $\pm$ 0.0672	0.13 $\pm$ 0.0141	0.45 $\pm$ 0.0942
	12	Sequel 4h 2nd	1006	885	1227.22 $\pm$ 10.32	969.31 $\pm$ 6.94	0.50 $\pm$ 0.0213	0.17 $\pm$ 0.0057	0.49 $\pm$ 0.0390
	13	Sequel 6h	953	753	1116.97 $\pm$ 13.58	861.71 $\pm$ 10.00	0.23 $\pm$ 0.0184	0.09 $\pm$ 0.0049	0.24 $\pm$ 0.0337
	14	Sequel 8h	996	878	1177.50 $\pm$ 9.61	981.88 $\pm$ 7.95	0.21 $\pm$ 0.0133	0.10 $\pm$ 0.0041	0.12 $\pm$ 0.0141
	15	Sequel 8h 2nd	1089	907.5	1277.91 $\pm$ 6.85	1057.00 $\pm$ 5.46	0.26 $\pm$ 0.0101	0.13 $\pm$ 0.0031	0.14 $\pm$ 0.0089
	16	MinION 1D cDNA Manual size selection: 500bp+	606	479	792.48 $\pm$ 1.89	655.79 $\pm$ 1.63	3.26 $\pm$ 0.0063	6.04 $\pm$ 0.0068	6.63 $\pm$ 0.0091
	17	MinION dRNA	465	408	537.17 $\pm$ 9.76	524.67 $\pm$ 10.47	2.01 $\pm$ 0.0537	8.84 $\pm$ 0.0832	5.57 $\pm$ 0.0620
	18	MinION Cap-selection	783	637	965.34 $\pm$ 1.36	688.73 $\pm$ 0.74	3.29 $\pm$ 0.0047	5.28 $\pm$ 0.0042	6.74 $\pm$ 0.0066
	19	MinION 1D cDNA barcoded 1h	485	286	589.68 $\pm$ 2.19	392.62 $\pm$ 2.00	2.96 $\pm$ 0.0156	4.54 $\pm$ 0.0169	4.61 $\pm$ 0.0186
	20	MinION 1D cDNA barcoded 2h	462	281	518.55 $\pm$ 0.69	334.78 $\pm$ 0.59	3.04 $\pm$ 0.0071	4.57 $\pm$ 0.0073	4.66 $\pm$ 0.0081
	21	MinION 1D cDNA barcoded 3h	466	283	521.58 $\pm$ 1.32	337.82 $\pm$ 1.18	2.80 $\pm$ 0.0138	4.39 $\pm$ 0.0142	4.51 $\pm$ 0.0163
	22	MinION 1D cDNA barcoded 4h	469	284	526.01 $\pm$ 1.01	342.85 $\pm$ 0.91	3.01 $\pm$ 0.0123	4.46 $\pm$ 0.0105	4.64 $\pm$ 0.0120
	23	MinION 1D cDNA barcoded 6h	487	291	570.35 $\pm$ 1.29	383.47 $\pm$ 1.19	2.84 $\pm$ 0.0109	4.47 $\pm$ 0.0107	4.88 $\pm$ 0.0118
	24	MinION 1D cDNA barcoded 8h	513	303	648.63 $\pm$ 1.70	447.74 $\pm$ 1.55	2.92 $\pm$ 0.0109	4.37 $\pm$ 0.0098	4.70 $\pm$ 0.0109
	25	MinION 1D cDNA barcoded 12h	469	284	526.76 $\pm$ 0.85	346.08 $\pm$ 0.79	2.85 $\pm$ 0.0092	4.47 $\pm$ 0.008	4.55 $\pm$ 0.0101
B	Run #	Major specificities of libraries	Median Read-length	Median Aligned Read-length	Average Read-length $\pm$ SE	Average Aligned-read Length $\pm$ SE	Average insertion frequency $\pm$ SE	Average deletion frequency $\pm$ SE	Average mismatch frequency $\pm$ SE
	1	RSII mix no size selection	1393.5	1354	1446.21 $\pm$ 23.62	1374.1 $\pm$ 24.84	0.29 $\pm$ 0.0968	0.66 $\pm$ 0.063	1.24 $\pm$ 0.0957
	2	RSII mix random primed	847	744	945.83 $\pm$ 9.04	840.11 $\pm$ 9.32	0.22 $\pm$ 0.0399	0.46 $\pm$ 0.0200	1.03 $\pm$ 0.0524
	3	RSII mix BluePippin size selection: 0.8-5kb+	1081	933	1132.69 $\pm$ 2.05	933.03 $\pm$ 2.40	0.60 $\pm$ 0.0171	0.57 $\pm$ 0.0895	1.27 $\pm$ 0.0099
	4	RSII BluePippin size selection: 0.8-2kb	1090	956	1199.54 $\pm$ 4.16	1016.52 $\pm$ 4.77	0.57 $\pm$ 0.0299	0.49 $\pm$ 0.0159	1.22 $\pm$ 0.0173
	5	RSII BluePippin size selection: 2-3kb	1238	1141	1257.92 $\pm$ 4.21	1129.90 $\pm$ 4.83	0.32 $\pm$ 0.0255	0.52 $\pm$ 0.0143	1.45 $\pm$ 0.0241
	6	RSII BluePippin size selection: 3-5kb	1015	789	1151.05 $\pm$ 2.51	881.73 $\pm$ 2.33	0.68 $\pm$ 0.0151	0.73 $\pm$ 0.0085	1.82 $\pm$ 0.0102
	7	RSII BluePippin size selection: 5kb+	1033	830	1092.51 $\pm$ 3.52	852.45 $\pm$ 4.33	0.68 $\pm$ 0.0348	0.65 $\pm$ 0.0180	2.30 $\pm$ 0.0221
	8	Sequel 1h	1341	922	1488.88 $\pm$ 4.38	1063.58 $\pm$ 4.33	0.53 $\pm$ 0.0131	0.29 $\pm$ 0.0063	1.49 $\pm$ 0.0138
	9	Sequel 2h	1112	997	1318.41 $\pm$ 4.40	1191.07 $\pm$ 4.09	0.38 $\pm$ 0.0081	0.18 $\pm$ 0.0026	1.41 $\pm$ 0.0115
	10	Sequel 3h	1317	1196	1453.95 $\pm$ 3.34	1304.42 $\pm$ 2.95	0.51 $\pm$ 0.0070	0.22 $\pm$ 0.0024	1.15 $\pm$ 0.0075
	11	Sequel 4h	1240	1088	1448.78 $\pm$ 5.32	1185.50 $\pm$ 4.01	0.68 $\pm$ 0.0114	0.28 $\pm$ 0.0041	1.13 $\pm$ 0.0096
	12	Sequel 4h 2nd	1796	1340	1865.73 $\pm$ 1.69	1302.18 $\pm$ 1.96	0.84 $\pm$ 0.0071	0.26 $\pm$ 0.0022	0.71 $\pm$ 0.0031
	13	Sequel 6h	1403	1268	1530.68 $\pm$ 2.86	1368.60 $\pm$ 2.57	0.53 $\pm$ 0.0065	0.22 $\pm$ 0.0020	1.25 $\pm$ 0.0064
	14	Sequel 8h	1531	1394	1642.16 $\pm$ 3.03	1505.74 $\pm$ 3.00	0.50 $\pm$ 0.0067	0.21 $\pm$ 0.0020	1.85 $\pm$ 0.0078
	15	Sequel 8h 2nd	1656	1058	1720.46 $\pm$ 3.48	1152.01 $\pm$ 3.82	1.04 $\pm$ 0.0208	0.33 $\pm$ 0.0056	1.16 $\pm$ 0.0079
	16	MinION 1D cDNA Manual size selection: 500bp+	775	481	1127.15 $\pm$ 2.36	763.42 $\pm$ 1.65	4.25 $\pm$ 0.0120	5.63 $\pm$ 0.0072	6.62 $\pm$ 0.0073
	17	MinION dRNA	510	376	606.26 $\pm$ 4.10	495.05 $\pm$ 3.93	3.06 $\pm$ 0.0453	7.86 $\pm$ 0.0402	5.92 $\pm$ 0.0305
	18	MinION Cap-selection	664	476	815.83 $\pm$ 0.87	502.38 $\pm$ 0.48	3.81 $\pm$ 0.0091	5.60 $\pm$ 0.0053	7.02 $\pm$ 0.0061
	19	MinION 1D cDNA barcoded 1h	712	478	914.17 $\pm$ 2.18	674.09 $\pm$ 2.18	3.35 $\pm$ 0.0125	5.11 $\pm$ 0.0117	5.72 $\pm$ 0.0105
	20	MinION 1D cDNA barcoded 2h	593	389	775.40 $\pm$ 0.77	558.52 $\pm$ 0.76	3.30 $\pm$ 0.0047	5.15 $\pm$ 0.0045	5.64 $\pm$ 0.0040
	21	MinION 1D cDNA barcoded 3h	596	395	755.74 $\pm$ 1.62	539.09 $\pm$ 1.62	3.00 $\pm$ 0.0113	5.03 $\pm$ 0.0103	5.37 $\pm$ 0.0094
	22	MinION 1D cDNA barcoded 4h	595	390	743.67 $\pm$ 1.25	520.05 $\pm$ 1.26	3.03 $\pm$ 0.0096	5.06 $\pm$ 0.0086	5.39 $\pm$ 0.0078
	23	MinION 1D cDNA barcoded 6h	610	396	769.19 $\pm$ 1.49	535.50 $\pm$ 1.49	3.04 $\pm$ 0.0109	5.06 $\pm$ 0.0099	5.42 $\pm$ 0.0088
	24	MinION 1D cDNA barcoded 8h	645	390	852.11 $\pm$ 2.46	543.46 $\pm$ 2.54	3.02 $\pm$ 0.0214	4.49 $\pm$ 0.165	4.92 $\pm$ 0.0142
	25	MinION 1D cDNA barcoded 12h	497	45	611.37 $\pm$ 1.81	252.51 $\pm$ 2.07	2.93 $\pm$ 0.0352	3.26 $\pm$ 0.0242	3.76 $\pm$ 0.0198

Supplementary Table 6. Summary statistics of the sequencing reads which mapped to the viral genome (A) and to the host reference genome (B) from each run. SE: standard error. * The difference between the yield of the size-selected and non-size-selected samples might be caused by the

underloading of the SMRT Cell and it is independent from the size-selection step. In some cases, PacBio run results in low output, for which the possible reason is the underloading of the Cells

Run #	Samples	<i>Chlorocebus sabaeus</i>	VACV
Run 1	RSII mix no size selection	82.31	17.69
Run 2	RSII mix random primed	98.94	1.06
Run 3	RSII mix Bluepippin: 0.8-5kb+	74.28	25.72
Run 4	RSII Bluepippin: 0.8-2kb	87.21	12.79
Run 5	RSII Bluepippin: 2-3kb	75.14	24.86
Run 6	RSII Bluepippin: 3-5kb	71.40	28.60
Run 7	RSII Bluepippin: 5kb+	64.42	35.58
Run 8	Sequel 1h	98.82	1.18
Run 9	Sequel 2h	98.64	1.36
Run 10	Sequel 3h	98.46	1.54
Run 11	Sequel 4h	98.12	1.88
Run 12	Sequel 4h 2nd	98.10	1.90
Run 13	Sequel 6h	97.10	2.90
Run 14	Sequel 8h	89.73	10.27
Run 15	Sequel 8h 2nd	92.10	7.90
Run 16	MinION 1D cDNA Manual size selection: 500bp+	76.54	23.46
Run 17	MinION dRNA	92.13	7.87
Run 18	MinION Cap-selection	67.78	32.22
Run 19	MinION 1D cDNA barcoded 1h	80.20	19.80
Run 20	MinION 1D cDNA barcoded 2h	83.43	16.57
Run 21	MinION 1D cDNA barcoded 3h	79.99	20.01
Run 22	MinION 1D cDNA barcoded 4h	76.27	23.73
Run 23	MinION 1D cDNA barcoded 6h	71.77	28.23
Run 24	MinION 1D cDNA barcoded 8h	51.32	48.68
Run 25	MinION 1D cDNA barcoded 12h	35.88	64.12

Supplementary Table 7. Summary statistics of the viral and host reads from each run.

Supplementary Figure 10. TSS and TES positions detected by LoRTIA share the TSS and TES positions obtained from other studies (Supplementary Table S9) with different frequency. Bar charts represent the fall of TSS and TES positions within a ± 10 – 20 – 30 sliding window.

2. *Supplementary Note List of transcripts with unknown or non-validated TSSs.*

ORFs without annotated TSS: VACVWR_00030, VACVWR_00040, VACVWR_00050, VACVWR_00130, F10L, I5L, I8R, G1L, G6R, G7L, L3L, H2R, H4L, D2L, D3R, D6R, D10R, A7L, A11L, VACVWR_01470, A30L, A32L, A38L, VACVWR_01640, VACVWR_01870, B4R, VACVWR_01940, VACVWR_02050, VACVWR_02140, VACVWR_02150, VACVWR_02160

Only Moss-lab: VACVWR_00340, F3L, J5L, A24R

Only our data: VACVWR_00120, C3L, C8L, VACVWR_00430, F9L, F13L, F17R, E7R, E8R, E10R (?c) E11L, O2L, I1L, I2L, I6L, I7L, G2R (c), G8R, G9R, L1R, L4R, L5R, J1R, H1L, H3L, H6R, H7R, D11L, D13L, A1L, A2L, A2.5L, A3L, A6L, A9L, A10L, A12L, A13L, A14L, A14.5L, A15L, A16L (??A17L.5), A18R, A19L, A20R (?c), A22R, VACVWR_01440, VACVWR_01450, VACVWR_01460, A26L, A27L, A28L, A31R, A34R, A39R, A43R, VACVWR_01840, B7R, VACVWR_01990, B18R, VACVWR_02040, VACVWR_0207, C13L, B25R, B28R

3. *Supplementary Note Quantitative reverse transcription PCR*

Quantitative reverse transcription PCR [RT-qPCR; Rotor-Gene Q (Qiagen)] was carried out for checking the specificity of the cDNA products derived from the Cap-Seq library preparation, as well as for the validation of the antisense expression within the A11R and A18R genomic loci. 20 ng total RNA, SuperScript IV reverse transcriptase and oligo(dT) primer was used for the RT, while ABsolute qPCR SYBR Green Mix (Thermo Fisher Scientific) and gene specific primers (**Supplementary Table 5**) were applied for the amplification of the RT products.

Supplementary Figure II. Detection of antisense RNA expression from the complementary DNA strands of two VACV genes using qRT-PCR. (a) Polymerase chain reaction (PCR) products of the A11R transcript and its antisense partner. Abbreviations: M: molecular weight marker; A: A11R antisense RNA; B: A11R mRNA; C: A11R antisense RNA—NO-RT control; D: A11R mRNA—NO-RT control; E: A18R antisense RNA F: A18R mRNA; G: A18R antisense RNA—NO-RT control; H: A18R mRNA—NO-RT control. (b) amplification curves of the A18R gene and its antisense transcripts. (c)

Amplification curves of the A18R gene and its antisense transcripts. (d–g) Melting curves are shown to demonstrate the specificity of the amplification. The curves of transcripts cross the threshold line within 19.5–23.4 cycles, whereas the curves of NO-RT controls remain flat. (d) A11R mRNA and antisense RNA. (e) A11R DNA control. (f) A18R mRNA and antisense RNA (g) A18R DNA control.

Transcript name	Distance from the previously published TSS	TSS ref	Potential protein (TIS)	Predicted promoter distance	Early promoter motif distance
C9.5L	2nt	Yang et al (2010)	nORF - Yang et al (2015)	-	88
C19.5L	3nt	Yang et al (2011a)	nORF - Yang et al (2015)	45	-
B25.5R	3nt	Yang et al (2011a)	nORF - Yang et al (2015)	45	-

Supplementary Tble 8. List of the novel intergenic genes. from the predicted promoters are also summarized. Previously published corresponding TSS positions, ORF information, and the distance of the TSSs of novel genes

Gene	Length of ORFs in different frames	5' UTR length	Number of ORFs
C19L/B25R	f1: C19.5L (17AS, nORF), 4AS, k2: 5AS, 20AS	131	4
C10L	f1: C9.5 (17AS, nORF), 7AS (nORF), 20/21/22AS, k2: 3AS, 12AS, 18AS	399	6
VACWR_15	-	151	-
N1L	-	66	-
K1L	f1: K1.5L (22AS, dORF), K1.2L (5AS, nORF), 4AS, 18AS, f2: 8AS, 15AS, 5AS	285	7
F5L	-	68	-
A1L	f1: 32AS (A2L.2-A1L bic) , f2: ouf 8AS (dORF), 4AS, 9AS, f3: 11AS,	593	5
A8R	f1: 3AS (nORF), f2: A8AS, f3: 3AS	168	3
A12L	-	84	-
VACVWR_161	-	28	-
VACWR_183	-	560	-
B13R	f1: 19AS f2: 4AS (nonMet, uORF)	233	3
F17R	f1: 9AS, 8AS, 9AS, 6AS, 3AS, f2: 14AS	270	6
G5.5R	f1: 3AS	187	1
H6R	f1: 89AS (H5R.5 bic)	280	1
A34R	f1: 416AS (A33R.7?? bic)	541	1
A43R	f1: 28AS (A42R.5?)	181	1
B7R	f1: 106AS (B6R.5 bic?)	359	1
C13L	-	106	-
M1L	f1: 96AS (M2L.5 bic?) f2: 39AS, 6AS/3AS, f3: 5AS, 3AS, 3AS	273	6
F14.5L	f1: 3AS, f2: 16AS	95	2
H3L	f1: 80AS (H4L.5 bic), f2: 6AS, f3: 10AS	364	3
A2L	-	50	-
A14.5L	-	55	-
C5L	f1:198AS (C4L.5 bic) , f2: 12AS, f3: 6AS	931	3
		mean	median
	all	258	187
	mean of potential uORF containing 1 variants	331	277
	mean of variants without extra ORF	130	68

Supplementary Table 9. List of the VACV genes with longer TSS isoforms than the canonical transcripts. Length of ORFs within the longer UTR variants are presented with differences between main and longer isoforms and numbers of ORFs within longer variants.

5' UTR length distribution of I variants of transcripts

Supplementary Figure 13 S2. 5'-untranslated region (UTR) length distributions of transcript variants. The column diagram and the violin plot show 5'-UTR lengths of VACV transcripts ranging from 2 to 400 nts.

Supplementary Table. List of the complex RNAs.	
ORF	Name
A12L	c-A12L-A11R-AT5
A12L	c-A12L-A11R-AT4
A12L	c-A12L-A11R-AT3
A12L	c-A12L-A11R-AT2
A12L	c-A12L-A11R-AT1
A12L	c-A12L-A11R
A14L	c-A14L-A13L-A12L-A11R-AT2
A14L	c-A14L-A13L-A12L-A11R-AT1
A14L	c-A14L-A13L-A12L-A11R
A14L.1	c-A14L.1-A13L-A12L-A11R-AT
A14L.1	c-A14L.1-A13L-A12L-A11R
A19L	c-A19L-A18R-AT
A19L	c-A19L-A18R
A8R	c-A8R-A9L-l
A8R	c-A8R-A9L
A9L	c-A9L-A8R
nc	c-A21L-A20R
E10R	c-E10R-E11L
E11L	c-E11L-E10R
F17R	c F17R-E1L
F17R	c F17R-E1L-AT
nc	c-G3L-G2R
G4L	c-G4L-G2R
G4L	c-G4L-G2R-AT
H3L.5	c-H3L.5-H2R
H3L.5	c-H3L.5-H2R-AT
VACVWR_161	c-VACVWR_12-A38L
VACVWR_161	c-VACVWR_161-l-A38L
VACVWR_161	c-VACVWR_161-A38L
A37R	c-A37R-VACVWR_161-A38L

Supplementary Table 10. List of the complex RNAs

Supplementary Figure 15 S4. Shared transcription start and end site positions across the VACV genome. (a) TSS sites shared by ≥ 1 transcript isoforms; the X axis represents genome positions and the Y axis shows the number of transcript isoforms. (b) TES positions shared by ≥ 1 RNA isoforms; the X axis represents the genome positions and the Y axis shows the number of transcript isoforms.

Supplementary Figure17. S7. The transcript isoform categories used in this study and their abbreviations are shown.

References

1. Mummert, A., Esche, E., Robinson, J. & Armelagos, G. J. Stature and robusticity during the agricultural transition: Evidence from the bioarchaeological record. *Econ. Hum. Biol.* **9**, 284–301 (2011).
2. Taylor, M. W. *Viruses and man: A history of interactions. Viruses and Man: A History of Interactions* (2015). doi:10.1007/978-3-319-07758-1
3. Enard, D., Cai, L., Gwennap, C. & Petrov, D. A. Viruses are a dominant driver of protein adaptation in mammals. *Elife* **5**, (2016).
4. Virgin, H. W. The virome in mammalian physiology and disease. *Cell* **157**, 142–150 (2014).
5. Haig, D. Retroviruses and the placenta. *Current Biology* (2012). doi:10.1016/j.cub.2012.06.002
6. Chuong, E. B. Retroviruses facilitate the rapid evolution of the mammalian placenta. *BioEssays* (2013). doi:10.1002/bies.201300059
7. Bachiller, S., del-Pozo-Martín, Y. & Carrión, Á. M. L1 retrotransposition alters the hippocampal genomic landscape enabling memory formation. *Brain. Behav. Immun.* (2017). doi:10.1016/j.bbi.2016.12.018
8. Bodea, G. O., McKelvey, E. G. Z. & Faulkner, G. J. Retrotransposon-induced mosaicism in the neural genome. *Royal Society Open Science* **8**, (2018).
9. Mattick, J. S. & Makunin, I. V. Non-coding RNA. *Human molecular genetics* (2006). doi:10.1093/hmg/ddl046
10. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
11. Boldogkői, Z., Moldován, N., Balázs, Z., Snyder, M. & Tombácz, D. Long-Read Sequencing – A Powerful Tool in Viral Transcriptome Research. *Trends in Microbiology* (2019). doi:10.1016/j.tim.2019.01.010
12. Baltimore, D. Expression of animal virus genomes. *Bacteriol. Rev.* (1971). doi:10.1128/mmbr.35.3.235-241.1971
13. Holley, R. W. *et al.* Structure of a ribonucleic acid. *Science* (80-.). (1965). doi:10.1126/science.147.3664.1462
14. Wu, R. & Kaiser, A. D. Mapping the 5'-terminal nucleotides of the DNA of bacteriophage lambda and related phages. *Proc. Natl. Acad. Sci. U. S. A.* (1967). doi:10.1073/pnas.57.1.170
15. Fiers, W. *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene. *Nature* (1976). doi:10.1038/260500a0
16. Jou, W. M., Haegeman, G., Ysebaert, M. & Fiers, W. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* (1972). doi:10.1038/237082a0
17. Sanger, F. *et al.* Nucliotide sequence of bacteriophage phi X174 DNA. *Nature* (1977).
18. Baltimore, D. Viral RNA-dependent DNA polymerase: RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* (1970). doi:10.1038/2261209a0
19. Haddad, F. & Baldwin, K. M. Reverse transcription of the ribonucleic acid: the first step in RT-PCR assay. *Methods Mol. Biol.* (2010). doi:10.1007/978-1-60761-629-0_17
20. Mullis, K. *et al.* Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. *Cold Spring Harb. Symp. Quant. Biol.* (1986). doi:10.1101/sqb.1986.051.01.032

21. Mullis, K. B. Process for amplifying nucleic acids. *US Pat. App. 12* (1985).
22. Sheridan, G. E. C., Masters, C. I., Shallcross, J. A. & Mackey, B. M. Detection of mRNA by reverse transcription-PCR as indicator of viability in *Escherichia coli* cells. *Appl. Environ. Microbiol.* (1998). doi:10.1128/aem.64.4.1313-1318.1998
23. Wang, A. M., Doyle, M. V. & Mark, D. F. Quantitation of mRNA by the polymerase chain reaction. *Proc. Natl. Acad. Sci. U. S. A.* (1989). doi:10.1073/pnas.86.24.9717
24. Boone, R. F. & Moss, B. Sequence complexity and relative abundance of vaccinia virus mRNA's synthesized in vivo and in vitro. *J. Virol.* (1978).
25. Wittek, R., Cooper, J. A., Barbosa, E. & Moss, B. Expression of the vaccinia virus genome: Analysis and mapping of mRNAs encoded within the inverted terminal repetition. *Cell* (1980). doi:10.1016/0092-8674(80)90485-7
26. Yuen, L. & Moss, B. Multiple 3' ends of mRNA encoding vaccinia virus growth factor occur within a series of repeated sequences downstream of T clusters. *J. Virol.* (1986). doi:10.1128/jvi.60.1.320-323.1986
27. Cohrs, R. J. *et al.* Varicella-zoster virus (VZV) transcription during latency in human ganglia: construction of a cDNA library from latently infected human trigeminal ganglia and detection of a VZV transcript. *J. Virol.* (1994). doi:10.1128/jvi.68.12.7900-7908.1994
28. Rich, A. & Davies, D. R. A New Two Stranded Helical Structure: Polyadenylic Acid and Polyuridylic Acid. *Journal of the American Chemical Society* (1956). doi:10.1021/ja01595a086
29. Rich, A. A HYBRID HELIX CONTAINING BOTH DEOXYRIBOSE AND RIBOSE POLYNUCLEOTIDES AND ITS RELATION TO THE TRANSFER OF INFORMATION BETWEEN THE NUCLEIC ACIDS. *Proc. Natl. Acad. Sci.* (1960). doi:10.1073/pnas.46.8.1044
30. Varshavsky, A. Discovering the RNA Double Helix and Hybridization. *Cell* (2006). doi:10.1016/j.cell.2006.12.008
31. Crampton, J., Humphries, S., Woods, D. & Williamson, R. The isolation of cloned cDNA sequen which are differentially expressed in human lymphocytes and fibroblasts. *Nucleic Acids Res.* (1980). doi:10.1093/nar/8.24.6007
32. Gress, T. M., Hoheisel, J. D., Lennon, G. G., Zehetner, G. & Lehrach, H. Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. *Mamm. Genome* (1992). doi:10.1007/BF00352477
33. Landry, M. L. & Fong, C. K. Y. Nucleic acid hybridization in the diagnosis of viral infections. *Clinics in Laboratory Medicine* (1985). doi:10.1016/s0272-2712(18)30857-6
34. Lenoir, T. & Giannella, E. The emergence and diffusion of DNA microarray technology. *J. Biomed. Discov. Collab.* (2006). doi:10.1186/1747-5333-1-11
35. Fodor, S. P. A. *et al.* Light-directed, spatially addressable parallel chemical synthesis. *Science* (80-.). (1991). doi:10.1126/science.1990438
36. Lipshutz, R. J. *et al.* Using oligonucleotide probe arrays to access genetic diversity. *BioTechniques* (1995). doi:10.1007/978-3-0348-8817-2_9
37. Kellam, P. Post-genomic virology: The impact of bioinformatics, microarrays and proteomics on investigating host and pathogen interactions. *Reviews in Medical Virology* (2001). doi:10.1002/rmv.328
38. Guerra, S. *et al.* Microarray Analysis Reveals Characteristic Changes of Host Cell Gene Expression in Response to Attenuated Modified Vaccinia Virus Ankara Infection of Human HeLa Cells. *J. Virol.* (2004). doi:10.1128/jvi.78.11.5820-5834.2004

39. Balogh, A. *et al.* Gene expression profiling in PC12 cells infected with an oncolytic Newcastle disease virus strain. *Virus Res.* (2014). doi:10.1016/j.virusres.2014.03.003
40. Kennedy, P. G. E. *et al.* Transcriptomal analysis of varicella-zoster virus infection using long oligonucleotide-based microarrays. *J. Gen. Virol.* (2005). doi:10.1099/vir.0.80946-0
41. Assarsson, E. *et al.* Kinetic analysis of a complete poxvirus transcriptome reveals an immediate-early class of genes. *Proc. Natl. Acad. Sci. U. S. A.* (2008). doi:10.1073/pnas.0711573105
42. Harrison, A. *et al.* Physico-chemical foundations underpinning microarray and next-generation sequencing experiments. *Nucleic Acids Research* (2013). doi:10.1093/nar/gks1358
43. Git, A. *et al.* Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *RNA* (2010). doi:10.1261/rna.1947110
44. Hurd, P. J. & Nelson, C. J. Advantages of next-generation sequencing versus the microarray in epigenetic research. *Briefings Funct. Genomics Proteomics* (2009). doi:10.1093/bfpg/elp013
45. Adams, M. D. *et al.* Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* (80-.). (1991). doi:10.1126/science.2047873
46. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* (80-.). (1995). doi:10.1126/science.270.5235.484
47. Saha, S. *et al.* Using the transcriptome to annotate the genome. *Nat. Biotechnol.* (2002). doi:10.1038/nbt0502-508
48. Kenzelmann, M. & Muöhlemann, K. Transcriptome analysis of fibroblast cells immediate-early after human cytomegalovirus infection. *Journal of Molecular Biology* (2000). doi:10.1006/jmbi.2000.4271
49. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* (2008). doi:10.1038/nmeth.1226
50. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* (2008). doi:10.1038/ng.259
51. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* (2009). doi:10.1038/nmeth.1315
52. Sanger, F. Sequences, Sequences, and Sequences. *Annu. Rev. Biochem.* (1988). doi:10.1146/annurev.bi.57.070188.000245
53. RajBhandary, U. L. & Köhrer, C. Early days of tRNA research: Discovery, function, purification and sequence analysis. in *Journal of Biosciences* (2006). doi:10.1007/BF02705183
54. Barciszewska, M. Z., Perrigue, P. M. & Barciszewski, J. tRNA-the golden standard in molecular biology. *Molecular BioSystems* (2016). doi:10.1039/c5mb00557d
55. Pierrel, J. An RNA Phage Lab: MS2 in Walter Fiers' Laboratory of Molecular Biology in Ghent, from Genetic Code to Gene and Genome, 1963-1976. *J. Hist. Biol.* (2012). doi:10.1007/s10739-010-9267-z
56. Fields, S. Would Fred Sanger get funded today? *Genetics* (2014). doi:10.1534/genetics.114.165134
57. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* (1977). doi:10.1073/pnas.74.12.5463

58. The ultimate goal: Sequencing DNA.
59. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* (2016). doi:10.1016/j.ygeno.2015.11.003
60. Naidoo, N., Pawitan, Y., Soong, R., Cooper, D. N. & Ku, C. S. Human genetics and genomics a decade after the release of the draft sequence of the human genome. *Human Genomics* (2011). doi:10.1186/1479-7364-5-6-577
61. Hood, L. & Rowen, L. The human genome project: Big science transforms biology and medicine. *Genome Med.* (2013). doi:10.1186/gm483
62. Keith Bradnam (The Institute of Cancer Research, L. Next-generation sequencing must die! Available at: <http://www.acgt.me/blog/2014/3/7/next-generation-sequencing-must-die>.
63. Nyrén, P., Pettersson, B. & Uhlén, M. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal. Biochem.* (1993). doi:10.1006/abio.1993.1024
64. Noonan, J. P. *et al.* Sequencing and analysis of Neanderthal genomic DNA. *Science* (80-.). (2006). doi:10.1126/science.1131412
65. Fusco, G., Amoroso, M. G., Gesualdi Montesano, N. & Viscardi, M. Development of a pyrosequencing assay for the typing of alphaherpesviruses. *MethodsX* (2015). doi:10.1016/j.mex.2015.01.001
66. Spatz, S. J. & Rue, C. A. Sequence determination of a mildly virulent strain (CU-2) of Gallid herpesvirus type 2 using 454 pyrosequencing. *Virus Genes* (2008). doi:10.1007/s11262-008-0213-5
67. Palmeira, L., MacHiels, B., Lété, C., Vanderplasschen, A. & Gillet, L. Sequencing of bovine herpesvirus 4 v.test strain reveals important genome features. *Virol. J.* (2011). doi:10.1186/1743-422X-8-406
68. Kim, D. N., Song, Y. J. & Lee, S. K. The role of promoter methylation in epstein-barr virus (EBV) micro RNA expression in EBV-infected B cell lines. *Exp. Mol. Med.* (2011). doi:10.3858/emm.2011.43.7.044
69. Qin, L., Upton, C., Hazes, B. & Evans, D. H. Genomic Analysis of the Vaccinia Virus Strain Variants Found in Dryvax Vaccine. *J. Virol.* (2011). doi:10.1128/jvi.05779-11
70. *Solexa: second-gen genetic sequencing.*
71. Barba, M., Czosnek, H. & Hadidi, A. Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses* (2013). doi:10.3390/v6010106
72. Check Hayden, E. Technology: The \$1,000 genome. *Nature* (2014). doi:10.1038/507294a
73. National Center for Biotechnology Information, (NCBI). We want to hear from you about changes to NIH's Sequence Read Archive data format and storage. Available at: <https://ncbiinsights.ncbi.nlm.nih.gov/tag/sra/>. (Accessed: 8th January 2021)
74. Ambardar, S., Gupta, R., Trakroo, D., Lal, R. & Vakhlu, J. High Throughput Sequencing: An Overview of Sequencing Chemistry. *Indian Journal of Microbiology* (2016). doi:10.1007/s12088-016-0606-4
75. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* (2008). doi:10.1038/nature07517
76. Buermans, H. P. J. & den Dunnen, J. T. Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta - Molecular Basis of Disease* (2014). doi:10.1016/j.bbadis.2014.06.015

77. Cornelis, S. Forensic Lab-on-a-Chip DNA analysis. (Ghent University, 2019).
78. Lahens, N. F. *et al.* A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression. *BMC Genomics* (2017). doi:10.1186/s12864-017-4011-0
79. Garrido-Cardenas, J. A., Garcia-Maroto, F., Alvarez-Bermejo, J. A. & Manzano-Agugliaro, F. DNA sequencing sensors: An overview. *Sensors (Switzerland)* (2017). doi:10.3390/s17030588
80. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* (2020). doi:10.1186/s13059-020-1935-5
81. McCarthy, A. Third Generation DNA Sequencing: Pacific Biosciences' Single Molecule Real Time Technology. *Chem. Biol.* **17**, 675–676 (2010).
82. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics. Proteomics Bioinformatics* **13**, 278–89 (2015).
83. Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* **6**, 100 (2017).
84. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* (2019). doi:10.1038/s41587-019-0217-9
85. Tombácz, D. *et al.* Full-Length Isoform Sequencing Reveals Novel Transcripts and Substantial Transcriptional Overlaps in a Herpesvirus. *PLoS One* **11**, e0162868 (2016).
86. Tombácz, D. *et al.* Characterization of novel transcripts in pseudorabies virus. *Viruses* **7**, 2727–44 (2015).
87. Moldován, N. *et al.* Multi-platform sequencing approach reveals a novel transcriptome profile in pseudorabies virus. *Front. Microbiol.* (2018). doi:10.3389/fmicb.2017.02708
88. 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* **4**, 170194 (2017).
89. Tombácz, D. *et al.* Dynamic transcriptome profiling dataset of vaccinia virus obtained from long-read sequencing techniques. *Gigascience* (2018). doi:10.1093/gigascience/giy139
90. Bayley, H. Nanopore sequencing: From imagination to reality. *Clin. Chem.* **61**, 25–31 (2015).
91. The long view on sequencing. *Nature Biotechnology* (2018). doi:10.1038/nbt.4125
92. de Lannoy, C., de Ridder, D. & Risse, J. A sequencer coming of age: De novo genome assembly using MinION reads. *F1000Research* (2017). doi:10.12688/f1000research.12012.1
93. Haque, F., Li, J., Wu, H. C., Liang, X. J. & Guo, P. Solid-state and biological nanopore for real-time sensing of single chemical and sequencing of DNA. *Nano Today* (2013). doi:10.1016/j.nantod.2012.12.008
94. Pennisi, E. Search for pore-fection. *Science* (2012). doi:10.1126/science.336.6081.534
95. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nature Biotechnology* (2016). doi:10.1038/nbt.3423
96. Wang, Y., Yang, Q. & Wang, Z. The evolution of nanopore sequencing. *Front. Genet.* (2014). doi:10.3389/fgene.2014.00449
97. Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. in *Proceedings of the National Academy of Sciences of the United States of America* (1996). doi:10.1073/pnas.93.24.13770

98. Ashkenasy, N., Sánchez-Quesada, J., Bayley, H. & Ghadiri, M. R. Recognizing a single base in an individual DNA strand: A step toward DNA sequencing in nanopores. *Angew. Chemie - Int. Ed.* (2005). doi:10.1002/anie.200462114
99. Houghtaling, J., List, J. & Mayer, M. Nanopore-Based, Rapid Characterization of Individual Amyloid Particles in Solution: Concepts, Challenges, and Prospects. *Small* (2018). doi:10.1002/smll.201802412
100. Barkova, D. V., Andrianova, M. S., Komarova, N. V. & Kuznetsov, A. E. Channel and Motor Proteins for Translocation of Nucleic Acids in Nanopore Sequencing. *Moscow Univ. Chem. Bull.* (2020). doi:10.3103/S0027131420030025
101. Laszlo, A. H., Derrington, I. M. & Gundlach, J. H. MspA nanopore as a single-molecule tool: From sequencing to SPRNT. *Methods* (2016). doi:10.1016/j.ymeth.2016.03.026
102. de Lannoy, C., de Ridder, D. & Risse, J. The long reads ahead: De novo genome assembly using the MinION. *F1000Research* (2017). doi:10.12688/f1000research.12012.2
103. Bowden, R. *et al.* Sequencing of human genomes with nanopore technology. *Nat. Commun.* (2019). doi:10.1038/s41467-019-09637-5
104. Laver, T. *et al.* Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* **3**, 1–8 (2015).
105. Oxford Nanopore Technologies. NANOPORE SEQUENCING ACCURACY. Available at: <https://nanoporetech.com/accuracy>. (Accessed: 27th October 2020)
106. Mikheyev, A. S. & Tin, M. M. Y. A first look at the Oxford Nanopore MinION sequencer. *Mol. Ecol. Resour.* (2014). doi:10.1111/1755-0998.12324
107. Moldován, N. *et al.* Multi-platform analysis reveals a complex transcriptome architecture of a circovirus. *Virus Res.* **237**, 37–46 (2017).
108. Tombácz, D. *et al.* Long-read isoform sequencing reveals a hidden complexity of the transcriptional landscape of herpes simplex virus type 1. *Front. Microbiol.* (2017). doi:10.3389/fmicb.2017.01079
109. Balázs, Z. *et al.* Long-Read Sequencing of Human Cytomegalovirus Transcriptome Reveals RNA Isoforms Carrying Distinct Coding Potentials. *Sci. Rep.* (2017). doi:10.1038/s41598-017-16262-z
110. Tombácz, D., Balázs, Z., Csabai, Z., Snyder, M. & Boldogkoi, Z. Long-read sequencing revealed an extensive transcript complexity in herpesviruses. *Frontiers in Genetics* (2018). doi:10.3389/fgene.2018.00259
111. SONIAK, M. What Do Chickens Have to Do With Chicken Pox? (2012). Available at: <https://www.mentalfloss.com/article/30635/what-do-chickens-have-do-chicken-pox>. (Accessed: 18th February 2020)
112. Ann M. Arvin, A. A. G. *Varicella-Zoster Virus: Virology and Clinical Management*. (Cambridge University Press, 2000). doi:https://doi.org/10.1017/CBO9780511601194
113. Aronson, J. When I use a word . . . : Chickenpox. *BMJ* (2000). doi:10.1136/bmj.321.7262.682
114. Beswick, T. S. L. The origin and the use of the word herpes. *Med. Hist.* (1962). doi:10.1017/S002572730002737X
115. Galetta, K. M. & Gilden, D. Zeroing in on zoster: A tale of many disorders produced by one virus. *Journal of the Neurological Sciences* (2015). doi:10.1016/j.jns.2015.10.004
116. C., T. E., J. The first description of chickenpox as a disease, by William Heberden the Elder (1710-1801) in 1767. *Pediatrics* (1970).

117. Lian, T. Y. & Lim, K. K. T. The legacy of William Heberden the Elder (1710-1801). *Rheumatology* (2004). doi:10.1093/rheumatology/keg007
118. Rosenow, E. C. & Oftedal, S. The etiology and experimental production of herpes zoster with plates 8 to 19. *J. Infect. Dis.* (1916). doi:10.1093/infdis/18.5.477
119. Head, H. & Campbell, A. W. The pathology of herpes zoster and its bearing on sensory localisation. *Brain* (1900). doi:10.1093/brain/23.3.353
120. Bókay, J. von. Über den aetiologischen Zusammenhang der Varicellen mit gewissen Fällen von Herpes Zoster. *Wien. Klin. Wochenschr.* **22**, 1323–1326 (1909).
121. Taylor-Robinson, D., and Downie, A. W. Chicken-pox and Herpes Zoster I. Complement Fixation Studies. *Br. J. Exp. Pathol.* **40**, 398–409 (1959).
122. WELLER, T. H. & STODDARD, M. B. Intranuclear inclusion bodies in cultures of human tissue inoculated with varicella vesicle fluid. *J. Immunol.* (1952).
123. Zerboni, L., Sen, N., Oliver, S. L. & Arvin, A. M. Molecular mechanisms of varicella zoster virus pathogenesis. *Nat Rev Micro* **12**, 197–210 (2014).
124. Maresova, L., Pasieka, T. J., Homan, E., Gerday, E. & Grose, C. Incorporation of Three Endocytosed Varicella-Zoster Virus Glycoproteins, gE, gH, and gB, into the Virion Envelope. *J. Virol.* **79**, 997–1007 (2005).
125. Davison, A. J. & Scott, J. E. The complete DNA sequence of varicella-zoster virus. *J. Gen. Virol.* **67**, 1759–1816 (1986).
126. Davison, A. J. *Comparative analysis of the genomes In Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis.* (Cambridge University Press, Cambridge, 2007). doi:10.1017/CBO9780511545313.003
127. Cohen, J. I. Genomic structure and organization of varicella-zoster virus. *Contributions to microbiology* (1999). doi:10.1159/000060313
128. Davison, A. J. Structure of the genome termini of varicella-zoster virus. *J. Gen. Virol.* (1984). doi:10.1099/0022-1317-65-11-1969
129. JI, C. VZV: molecular basis of persistence (latency and reactivation). in *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis* (ed. Arvin A, Campadelli-Fiume G, Mocarski E, et al.) (Cambridge University Press, 2007).
130. Tillieux, S. L. *et al.* Complete DNA sequences of two oka strain varicella-zoster virus genomes. *J. Virol.* **82**, 11023–44 (2008).
131. Perera, L. P., Mosca, J. D., Sadeghi-Zadeh, M., Ruyechan, W. T. & Hay, J. The Varicella-Zoster virus immediate early protein, IE62, can positively regulate its cognate promoter. *Virology* **191**, 346–354 (1992).
132. Kinchington, P. R., Bookey, D. & Turse, S. E. 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.* **69**, 4274–82 (1995).
133. Schmid, D. S. & Jumaan, A. O. Impact of varicella vaccine on varicella-zoster virus dynamics. *Clinical Microbiology Reviews* (2010). doi:10.1128/CMR.00031-09
134. Reichelt, M., Brady, J. & Arvin, A. M. 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.* **83**, 3904–3918 (2009).
135. Khalil, M. I. *et al.* Mutational analysis of varicella-zoster virus (VZV) immediate early protein (IE62) subdomains and their importance in viral replication. *Virology* **492**, 82–91 (2016).

136. Cohrs, R. J., Hurley, M. P. & Gilden, D. H. Array analysis of viral gene transcription during lytic infection of cells in tissue culture with varicella-zoster virus. *J. Virol.* **77**, 11718–11732 (2003).
137. Xing, Y. *et al.* A site of varicella-zoster virus vulnerability identified by structural studies of neutralizing antibodies bound to the glycoprotein complex gHgL. *Proc. Natl. Acad. Sci. U. S. A.* (2015). doi:10.1073/pnas.1501176112
138. Chen, D. & Olivo, P. D. Expression of the varicella-zoster virus origin-binding protein and analysis of its site-specific DNA-binding properties. *J. Virol.* (1994). doi:10.1128/jvi.68.6.3841-3849.1994
139. Wang, Z.-H. *et al.* Essential Role Played by the C-Terminal Domain of Glycoprotein I in Envelopment of Varicella-Zoster Virus in the trans-Golgi Network: Interactions of Glycoproteins with Tegument. *J. Virol.* (2001). doi:10.1128/jvi.75.1.323-340.2001
140. Wild, P., Kaech, A., Schraner, E. M., Walser, L. & Ackermann, M. Endoplasmic reticulum-to-Golgi transitions upon herpes virus infection. *F1000Research* (2018). doi:10.12688/f1000research.12252.2
141. Lyman, M. G. & Enquist, L. W. Herpesvirus Interactions with the Host Cytoskeleton. *J. Virol.* (2009). doi:10.1128/jvi.01718-08
142. Sadaoka, T., Yanagi, T., Yamanishi, K. & Mori, Y. Characterization of the varicella-zoster virus ORF50 gene, which encodes glycoprotein M. *J. Virol.* **84**, 3488–3502 (2010).
143. Moffat, J. F. *et al.* Attenuation of the Vaccine Oka Strain of Varicella-Zoster Virus and Role of Glycoprotein C in Alphaherpesvirus Virulence Demonstrated in the SCID-hu Mouse. *J. Virol.* (1998). doi:10.1128/jvi.72.2.965-974.1998
144. Zerboni, L., Ku, C. C., Jones, C. D., Zehnder, J. L. & Arvin, A. M. Varicella-zoster virus infection of human dorsal root ganglia in vivo. *Proc. Natl. Acad. Sci. U. S. A.* (2005). doi:10.1073/pnas.0501045102
145. Moffat, J. *et al.* Functions of the C-Terminal Domain of Varicella-Zoster Virus Glycoprotein E in Viral Replication In Vitro and Skin and T-Cell Tropism In Vivo. *J. Virol.* (2004). doi:10.1128/jvi.78.22.12406-12415.2004
146. Leclair, J. M., Zaia, J. A., Levin, M. J., Congdon, R. G. & Goldmann, D. A. Airborne Transmission of Chickenpox in a Hospital. *N. Engl. J. Med.* (1980). doi:10.1056/nejm198002213020807
147. Grose, C. Variation on a theme by Fenner: the pathogenesis of chickenpox. *Pediatrics* (1981).
148. Arvin, A. M. & Gilden, D. Varicella-zoster virus. in *Fields Virology: Sixth Edition* (2013).
149. Thornton, E. E. *et al.* Spatiotemporally separated antigen uptake by alveolar dendritic cells and airway presentation to T cells in the lung. *J. Exp. Med.* (2012). doi:10.1084/jem.20112667
150. Jones, M. *et al.* RNA-seq Analysis of Host and Viral Gene Expression Highlights Interaction between Varicella Zoster Virus and Keratinocyte Differentiation. *PLoS Pathog.* (2014). doi:10.1371/journal.ppat.1003896
151. Ouwendijk, W. J. D. & Verjans, G. M. G. M. Pathogenesis of varicelloviruses in primates. *J. Pathol.* (2015). doi:10.1002/path.4451
152. Grose, C. Pangaea and the Out-of-Africa Model of Varicella-Zoster Virus Evolution and Phylogeography. *J. Virol.* (2012). doi:10.1128/jvi.00357-12
153. Grinde, B. Herpesviruses: latency and reactivation – viral strategies and host response. *J. Oral Microbiol.* (2013). doi:10.3402/jom.v5i0.22766

154. Messaoudi, I. *et al.* Simian varicella virus infection of rhesus macaques recapitulates essential features of varicella zoster virus infection in humans. *PLoS Pathog.* (2009). doi:10.1371/journal.ppat.1000657
155. Gilden., R. M. and D. H. Simian varicella virus. in *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis* (eds. Arvin, A. et al.) (Cambridge University Press, 2007). doi:10.1017/CBO9780511545313
156. Kennedy, P. G. E., Grinfeld, E. & Gow, J. W. Latent varicella-zoster virus is located predominantly in neurons in human trigeminal ganglia. *Proc. Natl. Acad. Sci. U. S. A.* (1998). doi:10.1073/pnas.95.8.4658
157. Lee, K. S. *et al.* Human Sensory Neurons Derived from Induced Pluripotent Stem Cells Support Varicella-Zoster Virus Infection. *PLoS One* (2012). doi:10.1371/journal.pone.0053010
158. Laemmle, L., Goldstein, R. S. & Kinchington, P. R. Modeling Varicella Zoster Virus Persistence and Reactivation – Closer to Resolving a Perplexing Persistent State. *Front. Microbiol.* (2019). doi:10.3389/fmicb.2019.01634
159. Mahalingam, R. *et al.* Current in vivo models of varicella-zoster virus neurotropism. *Viruses* (2019). doi:10.3390/v11060502
160. Kinchington, P. R., Leger, A. J. S., Guedon, J.-M. G. & Hendricks, R. L. Herpes simplex virus and varicella zoster virus, the house guests who never leave. *Herpesviridae* **3**, 5 (2012).
161. Ouwendijk, W. J. D. *et al.* Restricted Varicella-Zoster Virus Transcription in Human Trigeminal Ganglia Obtained Soon after Death. *J. Virol.* **86**, 10203–10206 (2012).
162. Jeffrey I. Cohen. VZV: molecular basis of persistence (latency and reactivation). in *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis* (eds. Rvin, A. et al.) (Cambridge University Press, 2007). doi:10.1017/CBO9780511545313
163. Depledge, D. P. *et al.* A spliced latency-associated VZV transcript maps antisense to the viral transactivator gene 61. *Nat. Commun.* **9**, 1167 (2018).
164. Depledge, D. P., Sadaoka, T. & Ouwendijk, W. J. D. Molecular aspects of varicella-zoster virus latency. *Viruses* **10**, (2018).
165. Berencsi Görgy. *Orvosi Molekuláris Virologia*. (Convention Budapest Kft., 2004).
166. Laing, K. J., Ouwendijk, W. J. D., Koelle, D. M. & Verjans, G. M. G. M. Immunobiology of Varicella-Zoster Virus Infection. *The Journal of infectious diseases* (2018). doi:10.1093/infdis/jiy403
167. De Pelsmaeker, S., Romero, N., Vitale, M. & Favoreel, H. W. Herpesvirus Evasion of Natural Killer Cells. *J. Virol.* (2018). doi:10.1128/jvi.02105-17
168. Mahalingam, R. *et al.* Simian varicella virus reactivation in cynomolgus monkeys. *Virology* (2007). doi:10.1016/j.virol.2007.06.025
169. Cameron, J. C. *et al.* Severe complications of chickenpox in hospitalised children in the UK and Ireland. *Arch. Dis. Child.* (2007). doi:10.1136/adc.2007.123232
170. Ziebold, C., von Kries, R., Lang, R., Weigl, J. & Schmitt, H. J. Severe complications of varicella in previously healthy children in Germany: a 1-year survey. *Pediatrics* (2001).
171. Sampathkumar, P., Drage, L. A. & Martin, D. P. Herpes zoster (Shingles) and postherpetic neuralgia. *Mayo Clinic Proceedings* (2009). doi:10.4065/84.3.274
172. Ahmed, A. M., Brantley, J. S., Madkan, V., Mendoza, N. & Tyring, S. K. Managing herpes zoster in immunocompromised patients. *Herpes* (2007).

173. Liese, J. G. *et al.* The burden of varicella complications before the introduction of routine varicella vaccination in Germany. *Pediatr. Infect. Dis. J.* (2008). doi:10.1097/INF.0b013e3181586665
174. Varicella and herpes zoster vaccines: WHO position paper, June 2014 - Recommendations. *Vaccine* (2016). doi:10.1016/j.vaccine.2014.07.068
175. National Center for Immunization and Respiratory Diseases. Chickenpox Vaccination: What Everyone Should Know. (2019). Available at: <https://www.cdc.gov/vaccines/vpd/varicella/public/index.html>.
176. Bonanni, P. *et al.* Varicella vaccination in Europe - Taking the practical approach. *BMC Medicine* (2009). doi:10.1186/1741-7015-7-26
177. Meszner, Z. *et al.* Economic burden of varicella in children 1-12 Years of age in Hungary, 2011-2015. *BMC Infect. Dis.* (2017). doi:10.1186/s12879-017-2575-6
178. *Varicella vaccination in the European Union.* doi:10.2900/633956
179. Mészner, Z. *et al.* Burden of varicella in Central and Eastern Europe: findings from a systematic literature review. *Expert Review of Vaccines* (2019). doi:10.1080/14760584.2019.1573145
180. Ostrove, J. M. *et al.* Transcription mapping of the varicella-zoster virus genome. *J. Virol.* **56**, 600–6 (1985).
181. Ostrove, J. M. *et al.* Transcription mapping of the varicella-zoster virus genome. *J. Virol.* (1985). doi:10.1128/jvi.56.2.600-606.1985
182. Straus, S. E. *et al.* Molecular cloning and physical mapping of varicella zoster virus DNA. *Proc. Natl. Acad. Sci. U. S. A.* (1982). doi:10.1073/pnas.79.4.993
183. Cohen, J. I. & Seidel, K. E. Varicella-zoster virus open reading frame 1 encodes a membrane protein that is dispensable for growth of VZV in vitro. *Virology* (1995). doi:10.1006/viro.1995.1006
184. Sato, H., Pesnicak, L. & Cohen, J. I. Varicella-Zoster Virus Open Reading Frame 2 Encodes a Membrane Phosphoprotein That Is Dispensable for Viral Replication and for Establishment of Latency. *J. Virol.* (2002). doi:10.1128/jvi.76.7.3575-3578.2002
185. Cohrs, R. J., Barbour, M. & Gilden, D. H. Varicella-Zoster Virus Gene 21: Transcriptional Start Site and Promoter Region. *J. Virol.* (1998). doi:10.1128/jvi.72.1.42-47.1998
186. Moriuchi, H., Moriuchi, M., Smith, H. A., Straus, S. E. & Cohen, J. I. Varicella-zoster virus open reading frame 61 protein is functionally homologous to herpes simplex virus type 1 ICP0. *J. Virol.* (1992). doi:10.1128/jvi.66.12.7303-7308.1992
187. Kost, R. G., Kupinsky, H. & Straus, S. E. Varicella-Zoster Virus Gene 63: Transcript Mapping and Regulatory Activity. *Virology* **209**, 218–224 (1995).
188. Moriuchi, H., Moriuchi, M., Straus, S. E. & Cohen, J. I. Varicella-zoster virus open reading frame 10 protein, the herpes simplex virus VP16 homolog, transactivates herpesvirus immediate-early gene promoters. *J. Virol.* (1993). doi:10.1128/jvi.67.5.2739-2746.1993
189. Kinchington, P. R., Vergnes, J. P., Defechereux, P., Piette, J. & Turse, S. E. Transcriptional mapping of the varicella-zoster virus regulatory genes encoding open reading frames 4 and 63. *J. Virol.* (1994). doi:10.1128/jvi.68.6.3570-3581.1994
190. McKee, T. A., Disney, G. H., Everett, R. D. & Preston, C. M. Control of expression of the varicella-zoster virus major immediate early gene. *J. Gen. Virol.* (1990). doi:10.1099/0022-1317-71-4-897

191. Baird, N. L., Bowlin, J. L., Cohrs, R. J., Gilden, D. & Jones, K. L. Comparison of Varicella-Zoster Virus RNA Sequences in Human Neurons and Fibroblasts. *J. Virol.* **88**, 5877–5880 (2014).
192. Cohrs, R. J., Hurley, M. P. & Gilden, D. H. Array Analysis of Viral Gene Transcription during Lytic Infection of Cells in Tissue Culture with Varicella-Zoster Virus. *J. Virol.* (2003). doi:10.1128/jvi.77.21.11718-11732.2003
193. Cohrs, R. J., Gilden, D. H., Gomi, Y., Yamanishi, K. & Cohen, J. I. Comparison of Virus Transcription during Lytic Infection of the Oka Parental and Vaccine Strains of Varicella-Zoster Virus. *J. Virol.* (2006). doi:10.1128/jvi.80.5.2076-2082.2006
194. Jones, J. O. & Arvin, A. M. Microarray Analysis of Host Cell Gene Transcription in Response to Varicella-Zoster Virus Infection of Human T Cells and Fibroblasts In Vitro and SCIDhu Skin Xenografts In Vivo. *J. Virol.* (2003). doi:10.1128/jvi.77.2.1268-1280.2003
195. Markus, A., Waldman Ben-Asher, H., Kinchington, P. R. & Goldstein, R. S. Cellular Transcriptome Analysis Reveals Differential Expression of Pro- and Antiapoptosis Genes by Varicella-Zoster Virus-Infected Neurons and Fibroblasts. *J. Virol.* (2014). doi:10.1128/jvi.00500-14
196. Peters, G. A. *et al.* The Attenuated Genotype of Varicella-Zoster Virus Includes an ORF0 Transitional Stop Codon Mutation. *J. Virol.* (2012). doi:10.1128/jvi.01067-12
197. Markus, A. *et al.* Varicella-Zoster Virus Expresses Multiple Small Noncoding RNAs. *J. Virol.* **91**, e01710-17 (2017).
198. Oh, S. J. *et al.* Whole transcriptome analyses reveal differential mRNA and microRNA expression profiles in primary human dermal fibroblasts infected with clinical or vaccine strains of varicella zoster virus. *Pathogens* (2019). doi:10.3390/pathogens8040183
199. Bisht, P., Das, B., Kinchington, P. R. & Goldstein, R. S. Varicella-Zoster Virus (VZV) Small Noncoding RNAs Antisense to the VZV Latency-Encoded Transcript VLT Enhance Viral Replication. *J. Virol.* (2020). doi:10.1128/jvi.00123-20
200. Kim, J. A., Park, S. K., Kumar, M., Lee, C. H. & Shin, O. S. Insights into the role of immunosenescence during varicella zoster virus infection (shingles) in the aging cell model. *Oncotarget* (2015). doi:10.18632/oncotarget.6117
201. Depledge, D. P., Mohr, I. & Wilson, A. C. Going the Distance: Optimizing RNA-Seq Strategies for Transcriptomic Analysis of Complex Viral Genomes. *J. Virol.* (2018). doi:10.1128/jvi.01342-18
202. Frank Fenner, Riccardo Wittek, K. R. D. *The Orthopoxviruses*. (Academic Press Inc, 1989).
203. Diven, D. G. An overview of poxviruses. *Journal of the American Academy of Dermatology* (2001). doi:10.1067/mjd.2001.109302
204. Fenner, F. Adventures with poxviruses of vertebrates. *FEMS Microbiology Reviews* (2000). doi:10.1016/S0168-6445(00)00027-9
205. Gammon, D. B. *et al.* Vaccinia virus-encoded ribonucleotide reductase subunits are differentially required for replication and pathogenesis. *PLoS Pathog.* (2010). doi:10.1371/journal.ppat.1000984
206. Wittek, R. Organization and expression of the poxvirus genome. *Experientia* (1982). doi:10.1007/BF01949349
207. Moss, B. Poxviridae. in *Fields Virology: Sixth Edition* (2013). doi:10.1016/b978-012511340-3/50016-7
208. Filée, J., Pouget, N. & Chandler, M. Phylogenetic evidence for extensive lateral acquisition of

- cellular genes by Nucleocytoplasmic large DNA viruses. *BMC Evol. Biol.* (2008). doi:10.1186/1471-2148-8-320
209. Adrian J. Gibbs, Charles H. Calisher, F. G.-A. Evolution of Poxviruses and african swine viruses. in *Molecular Basis of Virus Evolution* (Cambridge University Press, 1997).
 210. Reza Rahnemon, M., Javad Sobhaniefar, M. & Reza Haghshenas, M. Smallpox and bioterrorism. *Journal of Mazandaran University of Medical Sciences* (2013). doi:10.3949/ccjm.69.5.360
 211. STROUHAL, E. Traces of a smallpox epidemic in the family of Ramesses V of the Egyptian 20th dynasty. *Anthropol.* (1996).
 212. Patterson, K. B. & Runge, T. Smallpox and the Native American. *American Journal of the Medical Sciences* (2002). doi:10.1097/00000441-200204000-00009
 213. Carlos, A. M. & Lewis, F. D. Smallpox and Native American mortality: The 1780s epidemic in the Hudson Bay region. *Explor. Econ. Hist.* (2012). doi:10.1016/j.eeh.2012.04.003
 214. Stewart, A. J. & Devlin, P. M. The history of the smallpox vaccine. *J. Infect.* (2006). doi:10.1016/j.jinf.2005.07.021
 215. Grundy, I. Montagu's variolation. *Endeavour* (2000). doi:10.1016/S0160-9327(99)01244-2
 216. Wilkins, G. V. W. and. *Microbiology and Infectious Diseases.* (1997).
 217. WILSON, G. S. *The hazards of immunization.* (Based on UNIVERSITY OF LONDON HEATH CLARK LECTURES 1966 delivered at The London School of Hygiene and Tropical Medicine The Hazards of Immunization By SIR GRAHAM S. WILSON M.D., LL.D., F.R.G.P., D.P.H. Honorary Lecturer in the Department of Bacteriology at th, 1967). doi:https://doi.org/10.4269/ajtmh.1969.18.336
 218. Jenner, E. An inquiry into the causes and effects of the variolae vaccinae, known by the name of the cow-pox. in *The Conquest of Disease* (1798).
 219. Huygelen, C. Jenner's cowpox vaccine in light of current vaccinology. *Verh. K. Acad. Geneesk. Belg.* (1996).
 220. Schrick, L. *et al.* An Early American Smallpox Vaccine Based on Horsepox. *N. Engl. J. Med.* **377**, 1491–1492 (2017).
 221. Esparza, J., Lederman, S., Nitsche, A. & Damaso, C. R. Early smallpox vaccine manufacturing in the United States: Introduction of the “animal vaccine” in 1870, establishment of “vaccine farms”, and the beginnings of the vaccine industry. *Vaccine* (2020). doi:10.1016/j.vaccine.2020.05.037
 222. Riedel, S. Edward Jenner and the History of Smallpox and Vaccination. *Baylor Univ. Med. Cent. Proc.* (2005). doi:10.1080/08998280.2005.11928028
 223. Moore, J. C. *The History of the Small Pox.* (Longman, Hurst, Rees, Orme, and Brown, 1815).
 224. Baxby, D. Edward Jenner's Inquiry after 200 years. *BMJ* (1999). doi:10.1136/bmj.318.7180.390
 225. Henderson, D. A. Frank Fenner (1914-2010). *Nature* (2011). doi:10.1038/469035a
 226. Fenner, Frank, Henderson, Donald A, Arita, Isao, Jezek, Zdenek, Ladnyi, I. D. et al. *Smallpox and its eradication.* (World Health Organization, 1988).
 227. Parrino, J. & Graham, B. S. Smallpox vaccines: Past, present, and future. *J. Allergy Clin. Immunol.* (2006). doi:10.1016/j.jaci.2006.09.037
 228. Jacobs, B. L. *et al.* Vaccinia virus vaccines: Past, present and future. *Antiviral Research* (2009).

229. Bengali, Z., Townsley, A. C. & Moss, B. Vaccinia virus strain differences in cell attachment and entry. *Virology* (2009). doi:10.1016/j.virol.2009.04.012
230. de Freitas, L. F. D. *et al.* The Virulence of Different Vaccinia Virus Strains Is Directly Proportional to Their Ability To Downmodulate Specific Cell-Mediated Immune Compartments In Vivo . *J. Virol.* (2018). doi:10.1128/jvi.02191-18
231. Paez, E. & Esteban, M. Stability of vaccinia virus DNA during persistent infections: Accumulation of left-end deletions and of tandem repeats at both ends of the viral genome and prevention by interferon. *Virology* (1988). doi:10.1016/0042-6822(88)90241-3
232. DeLange, A. M. & McFadden, G. Efficient resolution of replicated poxvirus telomeres to native hairpin structures requires two inverted symmetrical copies of a core target DNA sequence. *J. Virol.* (1987). doi:10.1128/jvi.61.6.1957-1963.1987
233. Qin, L., Favis, N., Famulski, J. & Evans, D. H. Evolution of and Evolutionary Relationships between Extant Vaccinia Virus Strains. *J. Virol.* **89**, 1809–1824 (2015).
234. Hendrickson, R. C., Wang, C., Hatcher, E. L. & Lefkowitz, E. J. Orthopoxvirus genome evolution: The role of gene loss. *Viruses* (2010). doi:10.3390/v2091933
235. Li, Y., Meyer, H., Zhao, H. & Damon, I. K. GC content-based pan-pox universal PCR assays for poxvirus detection. *J. Clin. Microbiol.* (2010). doi:10.1128/JCM.01697-09
236. Goebel, S. J. *et al.* The complete DNA sequence of vaccinia virus. *Virology* (1990). doi:10.1016/0042-6822(90)90294-2
237. Christopher J. Burrell, C. R. H. and F. A. M. Fenner and White's Medical Virology. in *Principles of Virology* 604 (Academic Press Inc, 2016).
238. Garcel, A., Crance, J. M., Drillien, R., Garin, D. & Favier, A. L. Genomic sequence of a clonal isolate of the vaccinia virus Lister strain employed for smallpox vaccination in France and its comparison to other orthopoxviruses. *J. Gen. Virol.* (2007). doi:10.1099/vir.0.82708-0
239. Smith, G. L., Vanderplasschen, A. & Law, M. The formation and function of extracellular enveloped vaccinia virus. *Journal of General Virology* (2002). doi:10.1099/0022-1317-83-12-2915
240. Schmidt, F. I. *et al.* Vaccinia virus entry is followed by core activation and proteasome-mediated release of the immunomodulatory effector VH1 from lateral bodies. *Cell Rep.* (2013). doi:10.1016/j.celrep.2013.06.028
241. Vermeer, P. D. *et al.* Vaccinia Virus Entry, Exit, and Interaction with Differentiated Human Airway Epithelia. *J. Virol.* (2007). doi:10.1128/jvi.00601-07
242. Walsh, D. Poxviruses: Slipping and sliding through transcription and translation. *PLoS Pathogens* (2017). doi:10.1371/journal.ppat.1006634
243. Moss, B. Poxvirus DNA replication. *Cold Spring Harb. Perspect. Biol.* **5**, (2013).
244. Paszkowski, P., Noyce, R. S. & Evans, D. H. Live-Cell Imaging of Vaccinia Virus Recombination. *PLoS Pathog.* (2016). doi:10.1371/journal.ppat.1005824
245. Shchelkunov, S. N. *et al.* Comparison of the genetic maps of variola and vaccinia viruses. *FEBS Lett.* (1993). doi:10.1016/0014-5793(93)81013-P
246. Senkevich, T. G. *et al.* Mapping vaccinia virus DNA replication origins at nucleotide level by deep sequencing. *Proc. Natl. Acad. Sci. U. S. A.* (2015). doi:10.1073/pnas.1514809112
247. Dénes, B., Fodor, N., Obenaus, A. & Fodor, I. Engineering Oncolytic Vaccinia Viruses for

- Non-Invasive Optical Imaging of Tumors. *Open Biotechnol. J.* (2014). doi:10.2174/1874070700802010252
248. Hung, C. F. *et al.* Vaccinia virus preferentially infects and controls human and murine ovarian tumors in mice. *Gene Ther.* (2007). doi:10.1038/sj.gt.3302840
 249. Chard, L. S. *et al.* A vaccinia virus armed with interleukin-10 Is a promising therapeutic agent for treatment of murine pancreatic cancer. *Clin. Cancer Res.* (2015). doi:10.1158/1078-0432.CCR-14-0464
 250. Stamatos, N. M., Chakrabarti, S., Moss, B. & Hare, J. D. Expression of polyomavirus virion proteins by a vaccinia virus vector: association of VP1 and VP2 with the nuclear framework. *J. Virol.* (1987). doi:10.1128/jvi.61.2.516-525.1987
 251. Baldick, C. J. & Moss, B. Characterization and temporal regulation of mRNAs encoded by vaccinia virus intermediate-stage genes. *J. Virol.* (1993). doi:10.1128/jvi.67.6.3515-3527.1993
 252. Wittek, R. & Moss, B. 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. *J. Virol.* (1982). doi:10.1128/jvi.42.2.447-455.1982
 253. Venkatesan, S., Gershowitz, A. & Moss, B. Complete Nucleotide Sequences of Two Adjacent Early Vaccinia Virus Genes Located Within the Inverted Terminal Repetition. *J. Virol.* (1982). doi:10.1128/jvi.44.2.637-646.1982
 254. Bajszár, G., Wittek, R., Weir, J. P. & Moss, B. Vaccinia virus thymidine kinase and neighboring genes: mRNAs and polypeptides of wild-type virus and putative nonsense mutants. *J. Virol.* (1983). doi:10.1128/jvi.45.1.62-72.1983
 255. Ahn, B. Y., Jones, E. V. & Moss, B. 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. *J. Virol.* (1990).
 256. Ahn, B. Y., Rosel, J., Cole, N. B. & Moss, B. Identification and expression of rpo19, a vaccinia virus gene encoding a 19-kilodalton DNA-dependent RNA polymerase subunit. *J. Virol.* (1992). doi:10.1128/jvi.66.2.971-982.1992
 257. Amegadzie, B. Y., Ahn, B. Y. & Moss, B. Identification, sequence, and expression of the gene encoding a M(r) 35,000 subunit of the vaccinia virus DNA-dependent RNA polymerase. *J. Biol. Chem.* (1991). doi:10.1016/s0021-9258(18)92757-2
 258. Cooper, J. A. & Moss, B. In Vitro translation of immediate early, early, and late classes of RNA from vaccinia virus-infected cells. *Virology* (1979). doi:10.1016/0042-6822(79)90095-3
 259. Rubins, K. H. *et al.* Comparative analysis of viral gene expression programs during poxvirus infection: A transcriptional map of the vaccinia and monkeypox genomes. *PLoS One* (2008). doi:10.1371/journal.pone.0002628
 260. Grossegessle, M., Doellinger, J., Haldemann, B., Schaade, L. & Nitsche, A. A next-generation sequencing approach uncovers viral transcripts incorporated in poxvirus virions. *Viruses* (2017). doi:10.3390/v9100296
 261. Yang, Z., Bruno, D. P., Martens, C. A., Porcella, S. F. & Moss, B. Genome-Wide Analysis of the 5' and 3' Ends of Vaccinia Virus Early mRNAs Delineates Regulatory Sequences of Annotated and Anomalous Transcripts. *J. Virol.* (2011). doi:10.1128/jvi.00428-11
 262. Yang, Z., Bruno, D. P., Martens, C. A., Porcella, S. F. & Moss, B. Simultaneous high-resolution analysis of vaccinia virus and host cell transcriptomes by deep RNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* (2010). doi:10.1073/pnas.1006594107
 263. Yang, Z., Martens, C. A., Bruno, D. P., Porcella, S. F. & Moss, B. Pervasive initiation and 3'-

- end formation of poxvirus postreplicative RNAs. *J. Biol. Chem.* (2012). doi:10.1074/jbc.M112.390054
264. Yuen, L. & Moss, B. Oligonucleotide sequence signaling transcriptional termination of vaccinia virus early genes. *Proc. Natl. Acad. Sci. U. S. A.* (1987). doi:10.1073/pnas.84.18.6417
 265. Ink, B. S. & Pickup, D. J. Vaccinia virus directs the synthesis of early mRNAs containing 5' poly(A) sequences. *Proc. Natl. Acad. Sci. U. S. A.* (1990). doi:10.1073/pnas.87.4.1536
 266. Yang, Z., Maruri-Avidal, L., Sisler, J., Stuart, C. A. & Moss, B. Cascade regulation of vaccinia virus gene expression is modulated by multistage promoters. *Virology* (2013). doi:10.1016/j.virol.2013.09.007
 267. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–75 (2005).
 268. Prazsák, I. *et al.* Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus. *BMC Genomics* **19**, 873 (2018).
 269. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–92 (2013).
 270. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–402 (1997).
 271. Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).
 272. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–9 (2012).
 273. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty191
 274. Krzywinski, M. *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res.* (2009). doi:10.1101/gr.092759.109
 275. Cabrera, C. V., Esteban, M., McCarron, R., McAllister, W. T. & Holowczak, J. A. Vaccinia virus transcription: Hybridization of mRNA to restriction fragments of vaccinia DNA. *Virology* (1978). doi:10.1016/0042-6822(78)90011-9
 276. Rosel, J. L., Earl, P. L., Weir, J. P. & Moss, B. Conserved TAAATG sequence at the transcriptional and translational initiation sites of vaccinia virus late genes deduced by structural and functional analysis of the HindIII H genome fragment. *J. Virol.* (1986).
 277. Lee, T. Y., Chang, W. C., Hsu, J. B. K., Chang, T. H. & Shien, D. M. GPMiner: An integrated system for mining combinatorial cis-regulatory elements in mammalian gene group. in *Series on Advances in Bioinformatics and Computational Biology* (2012). doi:10.1186/1471-2164-13-S1-S3
 278. Narang, V., Sung, W.-K. & Mittal, A. Computational modeling of oligonucleotide positional densities for human promoter prediction. *Artif. Intell. Med.* **35**, 107–119 (2005).
 279. Biomatters development team. Geneious. (2005).
 280. Moldován, N. *et al.* Multi-Platform Sequencing Approach Reveals a Novel Transcriptome Profile in Pseudorabies Virus. *Front. Microbiol.* **8**, 2708 (2018).
 281. Boldogkői, Z. *et al.* Transcriptomic study of herpes simplex virus type-1 using full-length sequencing techniques. *Sci. Data* (2018). doi:10.1038/sdata.2018.266

282. Tombácz, D. *et al.* Long-read assays shed new light on the transcriptome complexity of a viral pathogen. *Sci. Rep.* (2020). doi:10.1038/s41598-020-70794-5
283. Prazsák, I. *et al.* Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus. *BMC Genomics* (2018). doi:10.1186/s12864-018-5267-8
284. Prazsak, I. *et al.* Long-read Sequencing Uncovers a Complex Transcriptome Topology in Varicella Zoster Virus. *bioRxiv* 399048 (2018). doi:10.1101/399048
285. Rock, D. L. *et al.* Detection of latency-related viral RNAs in trigeminal ganglia of rabbits latently infected with herpes simplex virus type 1. *J. Virol.* **61**, 3820–6 (1987).
286. Jin, L. & Scherba, G. Expression of the pseudorabies virus latency-associated transcript gene during productive infection of cultured cells. *J. Virol.* **73**, 9781–8 (1999).
287. Tombácz, D. *et al.* Long-Read Isoform Sequencing Reveals a Hidden Complexity of the Transcriptional Landscape of Herpes Simplex Virus Type 1. *Front. Microbiol.* **8**, 1079 (2017).
288. Tombacz, D. *et al.* Strain Kaplan of Pseudorabies Virus Genome Sequenced by PacBio Single-Molecule Real-Time Sequencing Technology. *Genome Announc.* **2**, e00628-14-e00628-14 (2014).
289. Lenac Roviš, T. *et al.* Comprehensive analysis of varicella-zoster virus proteins using a new monoclonal antibody collection. *J. Virol.* **87**, 6943–54 (2013).
290. Voss, J. H. & Roizman, B. Properties of two 5'-coterminial RNAs transcribed part way and across the S component origin of DNA synthesis of the herpes simplex virus 1 genome. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 8454–8 (1988).
291. Huang, L., Zhu, Y. & Anders, D. G. The variable 3' ends of a human cytomegalovirus oriLyt transcript (SRT) overlap an essential, conserved replicator element. *J. Virol.* **70**, 5272–81 (1996).
292. Kotenko, S. V. *et al.* Human cytomegalovirus harbors its own unique IL-10 homolog (cmvIL-10). *Proc. Natl. Acad. Sci.* **97**, 1695–1700 (2000).
293. Khalil, M. I., Sommer, M. H., Hay, J., Ruyechan, W. T. & Arvin, A. M. Varicella-zoster virus (VZV) origin of DNA replication oriS influences origin-dependent DNA replication and flanking gene transcription. *Virology* (2015). doi:10.1016/j.virol.2015.02.049
294. Kainov, Y. A., Aushev, V. N., Naumenko, S. A., Tchevkina, E. M. & Bazykin, G. A. Complex Selection on Human Polyadenylation Signals Revealed by Polymorphism and Divergence Data. *Genome Biol. Evol.* **8**, 1971–9 (2016).
295. Pinto, P. A. B. *et al.* RNA polymerase II kinetics in polo polyadenylation signal selection. *EMBO J.* **30**, 2431–44 (2011).
296. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
297. Balázs, Z. LoRTIA - Long-read RNA-Seq Transcript Isoform Annotator toolkit. (2019).
298. Yang, Z. *et al.* Deciphering Poxvirus Gene Expression by RNA Sequencing and Ribosome Profiling. *J. Virol.* (2015). doi:10.1128/jvi.00528-15
299. Yang, Z., Bruno, D. P., Martens, C. A., Porcella, S. F. & Moss, B. Simultaneous high-resolution analysis of vaccinia virus and host cell transcriptomes by deep RNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* (2010). doi:10.1073/pnas.1006594107
300. Yang, Z. *et al.* Expression Profiling of the Intermediate and Late Stages of Poxvirus Replication. *J. Virol.* (2011). doi:10.1128/jvi.05446-11

301. Yang, Z., Bruno, D. P., Martens, C. A., Porcella, S. F. & Moss, B. Genome-Wide Analysis of the 5' and 3' Ends of Vaccinia Virus Early mRNAs Delineates Regulatory Sequences of Annotated and Anomalous Transcripts. *J. Virol.* (2011). doi:10.1128/jvi.00428-11
302. Smith, G. L., Chan, Y. S. & Kerr, S. M. Transcriptional mapping and nucleotide sequence of a vaccinia virus gene encoding a polypeptide with extensive homology to DNA ligases. *Nucleic Acids Res.* (1989). doi:10.1093/nar/17.22.9051
303. Brown, C. K., Turner, P. C. & Moyer, R. W. Molecular characterization of the vaccinia virus hemagglutinin gene. *J. Virol.* (1991).
304. Somers, J., Pöyry, T. & Willis, A. E. A perspective on mammalian upstream open reading frame function. *International Journal of Biochemistry and Cell Biology* (2013). doi:10.1016/j.biocel.2013.04.020
305. Calvo, S. E., Pagliarini, D. J. & Mootha, V. K. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl. Acad. Sci. U. S. A.* (2009). doi:10.1073/pnas.0810916106
306. Matsui, M., Yachie, N., Okada, Y., Saito, R. & Tomita, M. Bioinformatic analysis of post-transcriptional regulation by uORF in human and mouse. *FEBS Lett.* (2007). doi:10.1016/j.febslet.2007.07.057
307. Scholz, A. *et al.* UORF-Tools-Workflow for the determination of translation-regulatory upstream open reading frames. *PLoS One* (2019). doi:10.1371/journal.pone.0222459
308. Senkevich, T. G. *et al.* Mapping vaccinia virus DNA replication origins at nucleotide level by deep sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 10908–13 (2015).
309. Parsons, B. L. & Pickup, D. J. Transcription of orthopoxvirus telomeres at late times during infection. *Virology* **175**, 69–80 (1990).
310. Boldogkői, Z., Balázs, Z., Moldován, N., Prazsák, I. & Tombácz, D. Novel classes of replication-associated transcripts discovered in viruses. *RNA Biology* (2019). doi:10.1080/15476286.2018.1564468
311. Tombácz, D. *et al.* Long-read isoform sequencing reveals a hidden complexity of the transcriptional landscape of herpes simplex virus type 1. *Front. Microbiol.* **8**, (2017).
312. Tombácz, D. *et al.* Full-length isoform sequencing reveals novel transcripts and substantial transcriptional overlaps in a herpesvirus. *PLoS One* **11**, (2016).
313. Tombácz, D. *et al.* Meta-analytic approach for transcriptome profiling of herpes simplex virus type 1. *Sci. Data* (2020). doi:10.1038/s41597-020-0558-8
314. Gábor Torma , Dóra Tombácz , Zsolt Csabai , Dániel Göbhardt , Zoltán Deim , Michael Snyder, Z. B. An Integrated Sequencing Approach for Updating of Pseudorabies Virus Transcriptome. *Preprints* (2020). doi:10.20944/preprints202012.0616.v1
315. Bahar, M. W. *et al.* How vaccinia virus has evolved to subvert the host immune response. *Journal of Structural Biology* (2011). doi:10.1016/j.jsb.2011.03.010
316. Morales, I. *et al.* The vaccinia virus F11L gene product facilitates cell detachment and promotes migration. *Traffic* (2008). doi:10.1111/j.1600-0854.2008.00762.x
317. Liu, R. & Moss, B. Vaccinia Virus C9 Ankyrin Repeat/F-Box Protein Is a Newly Identified Antagonist of the Type I Interferon-Induced Antiviral State. *J. Virol.* (2018). doi:10.1128/jvi.00053-18
318. Shchelkunov, S. N. Orthopoxvirus genes that mediate disease virulence and host tropism. *Advances in Virology* (2012). doi:10.1155/2012/524743

319. Tombácz, D. *et al.* Characterization of the Dynamic Transcriptome of a Herpesvirus with Long-read Single Molecule Real-Time Sequencing. *Sci. Rep.* **7**, 43751 (2017).
320. Kinchington, P. R., Vergnes, J. P., Defechereux, P., Piette, J. & Turse, S. E. Transcriptional mapping of the varicella-zoster virus regulatory genes encoding open reading frames 4 and 63. *J. Virol.* **68**, 3570–81 (1994).
321. Kemble, G. W. *et al.* Open Reading Frame S/L of Varicella-Zoster Virus Encodes a Cytoplasmic Protein Expressed in Infected Cells. *J. Virol.* (2000). doi:10.1128/jvi.74.23.11311-11321.2000
322. Meier, J. L. & Straus, S. E. Varicella-zoster virus DNA polymerase and major DNA-binding protein genes have overlapping divergent promoters. *J. Virol.* **67**, 7573–81 (1993).
323. Stevenson, D., Colman, K. L. & Davison, A. J. Characterization of the varicella-zoster virus gene 61 protein. *J. Gen. Virol.* (1992). doi:10.1099/0022-1317-73-3-521
324. Yang, Z. *et al.* Deciphering Poxvirus Gene Expression by RNA Sequencing and Ribosome Profiling. *J. Virol.* **89**, 6874–6886 (2015).
325. Tombácz, D. *et al.* Multiple Long-Read Sequencing Survey of Herpes Simplex Virus Dynamic Transcriptome. *Front. Genet.* (2019). doi:10.3389/fgene.2019.00834
326. Moldován, N. *et al.* Time-course profiling of bovine alphaherpesvirus 1.1 transcriptome using multiplatform sequencing. *Sci. Rep.* (2020). doi:10.1038/s41598-020-77520-1
327. Torma, G. *et al.* Combined short and long-read sequencing reveals a complex transcriptomic architecture of African swine fever virus. *bioRxiv* 2020.07.18.202820 (2020). doi:10.1101/2020.07.18.202820
328. Moldován, N. *et al.* Third-generation Sequencing Reveals Extensive Polycistronism and Transcriptional Overlapping in a Baculovirus. *Sci. Rep.* **8**, 8604 (2018).
329. Oikonomopoulos, S. *et al.* Methodologies for Transcript Profiling Using Long-Read Technologies. *Frontiers in Genetics* (2020). doi:10.3389/fgene.2020.00606
330. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).
331. Karginov, T. A., Pastor, D. P. H., Semler, B. L. & Gomez, C. M. Mammalian Polycistronic mRNAs and Disease. *Trends in Genetics* (2017). doi:10.1016/j.tig.2016.11.007
332. Kronstad, L. M., Brulois, K. F., Jung, J. U. & Glaunsinger, B. A. Dual Short Upstream Open Reading Frames Control Translation of a Herpesviral Polycistronic mRNA. *PLoS Pathog.* (2013). doi:10.1371/journal.ppat.1003156
333. Kronstad, L. M., Brulois, K. F., Jung, J. U. & Glaunsinger, B. A. Reinitiation after Translation of Two Upstream Open Reading Frames (ORF) Governs Expression of the ORF35-37 Kaposi's Sarcoma-Associated Herpesvirus Polycistronic mRNA. *J. Virol.* (2014). doi:10.1128/jvi.00202-14
334. Poyry, T. A. A., Kaminski, A. & Jackson, R. J. What determines whether mammalian ribosomes resume scanning after translation of a short upstream open reading frame? *Genes Dev.* **18**, 62–75 (2004).
335. Katayama, S. *et al.* Antisense Transcription in the Mammalian Transcriptome. *Science* (80-.). **309**, 1564–1566 (2005).
336. Boldogkői, Z., Tombácz, D. & Balázs, Z. Interactions between the transcription and replication machineries regulate the RNA and DNA synthesis in the herpesviruses. *Virus Genes* (2019). doi:10.1007/s11262-019-01643-5

337. Boldogkői, Z. Transcriptional interference networks coordinate the expression of functionally related genes clustered in the same genomic loci. *Front. Genet.* (2012). doi:10.3389/fgene.2012.00122
338. Yuan, C., Liu, Y., Yang, M. & Liao, D. J. 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.* **10**, 957–967 (2013).
339. Tian, B., Hu, J., Zhang, H. & Lutz, C. S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* **33**, 201–12 (2005).
340. Tian, B. & Graber, J. H. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip. Rev. RNA* **3**, 385–96 (2012).
341. Balázs, Z. *et al.* Template-switching artifacts resemble alternative polyadenylation. *BMC Genomics* (2019). doi:10.1186/s12864-019-6199-7
342. Hu, J., Lutz, C. S., Wilusz, J. & Tian, B. Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *RNA* **11**, 1485–93 (2005).
343. Chou, W., Ngo, T. & Gershon, P. D. An Overview of the Vaccinia Virus Infectome: a Survey of the Proteins of the Poxvirus-Infected Cell. *J. Virol.* (2012). doi:10.1128/jvi.06084-11
344. Doellinger, J., Schaade, L. & Nitsche, A. Comparison of the cowpox virus and vaccinia virus mature virion proteome: Analysis of the species-and strain-specific proteome. *PLoS One* (2015). doi:10.1371/journal.pone.0141527
345. Tang, S., Patel, A. & Krause, P. R. Hidden regulation of herpes simplex virus 1 pre-mRNA splicing and polyadenylation by virally encoded immediate early gene ICP27. *PLoS Pathog.* (2019). doi:10.1371/journal.ppat.1007884
346. Moldován, N. *et al.* Time-course profiling of bovine herpesvirus type 1 and host cell transcriptomes using multiplatform sequencing. *bioRxiv* (2020). doi:10.1101/2020.05.25.114843
347. Cocquet, J., Chong, A., Zhang, G. & Veitia, R. A. Reverse transcriptase template switching and false alternative transcripts. *Genomics* **88**, 127–131 (2006).
348. Houseley, J. & Tollervey, D. Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. *PLoS One* (2010). doi:10.1371/journal.pone.0012271
349. Braspenning, S. E. *et al.* Decoding the architecture of the varicella-zoster virus transcriptome. *MBio* (2020). doi:10.1128/mBio.01568-20
350. 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.* **97**, 2973–2988 (2016).
351. Gandy, S. Z. *et al.* RNA editing of the human herpesvirus 8 kaposin transcript eliminates its transforming activity and is induced during lytic replication. *J. Virol.* **81**, 13544–51 (2007).
352. Iizasa, H. *et al.* Editing of Epstein-Barr Virus-encoded BART6 MicroRNAs Controls Their Dicer Targeting and Consequently Affects Viral Latency. *J. Biol. Chem.* **285**, 33358–33370 (2010).
353. Mannion, N. M. *et al.* The RNA-Editing Enzyme ADAR1 Controls Innate Immune Responses to RNA. *Cell Rep.* **9**, 1482–1494 (2014).
354. Zahn, R. C., Schelp, I., Utermöhlen, O. & von Laer, D. A-to-G hypermutation in the genome of lymphocytic choriomeningitis virus. *J. Virol.* **81**, 457–64 (2007).
355. Wong, S. K. & Lazinski, D. W. Replicating hepatitis delta virus RNA is edited in the nucleus

- by the small form of ADAR1. *Proc. Natl. Acad. Sci.* **99**, 15118–15123 (2002).
356. Scadden, A. D. & Smith, C. W. RNAi is antagonized by A-to-I hyper-editing. *EMBO Rep.* **2**, 1107–11 (2001).
 357. Scadden, A. D. J. & Smith, C. W. Specific cleavage of hyper-edited dsRNAs. *EMBO J.* **20**, 4243–4252 (2001).
 358. Scadden, A. D. J. The RISC subunit Tudor-SN binds to hyper-edited double-stranded RNA and promotes its cleavage. *Nat. Struct. Mol. Biol.* **12**, 489–496 (2005).
 359. Sandri-Goldin, R. M. Replication of the herpes simplex virus genome: Does it really go around in circles? *Proceedings of the National Academy of Sciences of the United States of America* (2003). doi:10.1073/pnas.1432875100
 360. Oláh, P. *et al.* Characterization of pseudorabies virus transcriptome by Illumina sequencing. *BMC Microbiol.* **15**, 130 (2015).
 361. Tai-Schmiedel, J. *et al.* Human cytomegalovirus long noncoding RNA4.9 regulates viral DNA replication. *PLoS Pathog.* (2020). doi:10.1371/journal.ppat.1008390
 362. Wang, Y. *et al.* Kaposi's sarcoma-associated herpesvirus ori-Lyt-dependent DNA replication: cis-acting requirements for replication and ori-Lyt-associated RNA transcription. *J. Virol.* **78**, 8615–29 (2004).
 363. Fixman, E. D., Hayward, G. S. & Hayward, S. D. Replication of Epstein-Barr virus oriLyt: lack of a dedicated virally encoded origin-binding protein and dependence on Zta in cotransfection assays. *J. Virol.* **69**, 2998–3006 (1995).
 364. Tombácz, D. *et al.* Full-length isoform sequencing reveals novel transcripts and substantial transcriptional overlaps in a herpesvirus. *PLoS One* (2016). doi:10.1371/journal.pone.0162868
 365. Takács, I. F. *et al.* The ICP22 protein selectively modifies the transcription of different kinetic classes of pseudorabies virus genes. *BMC Mol. Biol.* **14**, 2 (2013).
 366. Ouwendijk, W. J. D. *et al.* Varicella-zoster virus VLT-ORF63 fusion transcript induces broad viral gene expression during reactivation from neuronal latency. *Nat. Commun.* (2020). doi:10.1038/s41467-020-20031-4
 367. Hu, F. Q. & Pickup, D. J. Transcription of the terminal loop region of vaccinia virus DNA is initiated from the telomere sequences directing DNA resolution. *Virology* **181**, 716–720 (1991).
 368. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biology* (2016). doi:10.1186/s13059-016-0881-8
 369. Kemble, G. W. *et al.* Open reading frame S/L of varicella-zoster virus encodes a cytoplasmic protein expressed in infected cells. *J. Virol.* **74**, 11311–21 (2000).

OPEN

Long-read assays shed new light on the transcriptome complexity of a viral pathogen

Dóra Tombácz^{1,4}, István Prazsák^{1,4}, Zsolt Csabai¹, Norbert Moldován¹, Béla Dénes², Michael Snyder³ & Zsolt Boldogkői^{1✉}

Characterization of global transcriptomes using conventional short-read sequencing is challenging due to the insensitivity of these platforms to transcripts isoforms, multigenic RNA molecules, and transcriptional overlaps. Long-read sequencing (LRS) can overcome these limitations by reading full-length transcripts. Employment of these technologies has led to the redefinition of transcriptional complexities in reported organisms. In this study, we applied LRS platforms from Pacific Biosciences and Oxford Nanopore Technologies to profile the vaccinia virus (VACV) transcriptome. We performed cDNA and direct RNA sequencing analyses and revealed an extremely complex transcriptional landscape of this virus. In particular, VACV genes produce large numbers of transcript isoforms that vary in their start and termination sites. A significant fraction of VACV transcripts start or end within coding regions of neighbouring genes. This study provides new insights into the transcriptomic profile of this viral pathogen.

Members of the *Poxviridae* family infect various vertebrate and invertebrate host species¹. Most notably, the variola virus is the causative agent of smallpox². VACV has roughly 90% sequence homology with variola. The VACV virion contains 195 kilobase pairs (kbp) of double-stranded DNA comprising at least 200 open reading frames (ORFs), twelve of which are present in terminal repeats^{3–5}. This viral genome encodes enzymes for DNA and RNA synthesis, transcription factors, and for enzymes that cap⁶ and polyadenylate⁷ RNA molecules. All of these proteins allow VACV replication in the cytoplasm of host cell.

Viral gene expression is regulated by stage-specific transcription factors that specifically bind to promoters of early (E), intermediate (I), and late (L) genes^{8–11}. The complete transcription machinery is already packaged in the VACV virion allowing E genes to be expressed immediately after entering the cell, when the viral genome is still encapsidated⁴. Subsequently, DNA replication occurs which is followed by the expression of I and then L genes classes. Synthesis of I mRNAs is dependent on de novo expression of E viral proteins, whereas synthesis of mRNAs from L genes requires the expression of certain E and I genes. During the last stage of infection, newly assembled virus particles egress from host cells. E genes encode proteins that synthesize DNA and RNA molecules, and others that play roles in virus-host interactions; whereas post-replicative (PR) genes (I and L genes) mainly specify structural elements of the virus¹². VACV genes belonging in the same kinetic class have been shown to exhibit a strong preference for co-localization¹³. Namely, E genes are reportedly clustered near genomic termini and are transcribed in the same direction¹⁴, whereas I and L genes are clustered at the central part of the viral DNA¹⁵. Using genome tiling arrays, Assarsson and colleagues demonstrated that 35 viral genes are expressed in immediate-early (IE) kinetics¹³. However, this terminology has not become widely accepted. In other DNA viruses, such as herpesviruses¹⁶ and baculoviruses¹⁷, IE RNAs are expressed in the absence of de novo protein synthesis, whereas E and L RNAs are dependent on the synthesis of IE proteins. According to this classification, all VACV E transcripts should belong to the IE kinetic class¹⁸. An alternative classification by Yang et al. categorized early VACV gene transcripts into two classes, E1.1 and E1.2¹⁹, and defined the ensuing expression kinetics according to higher affinity of transcription factors for E1.1 than for E1.2 promoters. The three classes of genes have distinctive promoters¹⁰. The consensus transcription termination sequence of E transcripts (UUUUUNU) is non-functional in PR transcripts, which mostly have polymorphic 3' ends^{20,21}.

¹Department of Medical Biology, Faculty of Medicine, University of Szeged, Szeged 6720, Hungary. ²Veterinary Diagnostic Directorate of the National Food Chain Safety Office, Budapest 1143, Hungary. ³Department of Genetics, School of Medicine, Stanford University, Stanford, CA 94305, USA. ⁴These authors contributed equally: Dóra Tombácz and István Prazsák. ✉email: boldogkoi.zsolt@med.u-szeged.hu

Figure 1. Draft workflow of the study. Infection, sequencing, and data integration strategy for transcript annotation.

Many PR and a few E transcripts have 5' poly(A) tails (PAT). PATs are supposedly generated by RNA polymerase slippage onto adjacent thymines on the DNA strand²². In addition to genome tiling analyses²³, RNA-Seq¹⁹, ribosome profiling^{12,24}, and microarrays^{25,26} have also been used for VACV transcriptome profiling. Yang and colleagues mapped 118 E genes and 93 PR genes using a short-read sequencing (SRS) technique¹⁶, and in a later study, they distinguished 53 I and 38 L genes among the PR genes¹⁷. Ribosome profiling analysis also confirmed canonical translational initiation sites (TISs) and demonstrated that additional TISs occur mostly within the ORFs. These TISs were also detected in 5'-untranslated (UTRs) and intergenic regions, as well as on complementary DNA strands¹². However, because most of these ORFs were short, the authors doubted their biological relevance. In another study by Yang and colleagues, transcript abundance was highly correlated with ribosome-protected reads, suggesting that translation is largely regulated by mRNA abundance¹². The techniques used in the above mentioned studies cannot detect full-length RNA molecules, and hence fail to provide a comprehensive assessment of the viral transcriptome. Although deep SRS provides satisfactory sequencing depth and coverage for global transcriptome profiling, the resulting RNA assemblies are incomplete, leading to inadequate annotations.

The major limitations in VACV transcriptome profiling relate to multiple transcriptional read-throughs that generate a complex meshwork of overlapping RNAs and to the large variation in transcriptional end sites (TESs). Additionally, the extensive use of alternative promoters leads to a large diversity of transcription start sites (TSSs). Hence, conventional techniques that fail to read entire RNA molecules have considerably underestimated the complexity of the poxvirus transcriptome. Earlier, we used two long-read sequencing (LRS) platform Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT)] to profile the transcriptomes of herpesviruses, including pseudorabies virus^{27,28}, herpes simplex virus type 1^{29,30}, varicella-zoster virus³¹, and human cytomegalovirus^{32,33}. These analyses identified polycistronic RNAs, transcript isoforms, and transcriptional overlaps, and facilitated the kinetic characterization of viral transcripts^{34,35}.

Herein, we report the transcriptomic analyses of the Western Reserve (WR) strain of VACV using PacBio RSII and Sequel platforms, and ONT MinION platform for cDNA, direct RNA (dRNA), and Cap-sequencing (Cap-Seq).

Results

Long-read transcriptome sequencing of vaccinia virus. In this study, we performed polyadenylation sequencing techniques for the investigation of the VACV transcriptome. We used PacBio isoform sequencing (Iso-Seq) template preparation protocol for the RSII and Sequel platforms, and the ONT MinION device to sequence cDNAs and native RNAs. For ONT sequencing, we used the company's own library preparation approach (1D-Seq), or the Teloprime Cap-selection protocol from Lexogen, which was adapted for the MinION sequencer. The obtained datasets were then used to define and validate viral TSS and TES coordinates using the LoRTIA pipeline (<https://github.com/zsolt-balazs/LoRTIA>) that was developed by our research group³⁶. The workflow for library preparation and analysis is shown in Fig. 1.

PCR amplification was used in all our sequencing techniques with the exception of dRNA sequencing where no amplification occurred. Direct RNA sequencing offers an alternative to cDNA sequencing because it lacks the recoding and amplification biases inherent in reverse transcription (RT) and PCR-based methods. We also used random primer-based RT for some ONT sequencing analyses. All methods generated high-coverage LRS data and determined both TSSs and TESs of viral transcripts with base-pair precision (see below). LRS can be used to detect individual RNA molecules and identify alternatively transcribed and processed RNAs, polycistronic transcripts, and transcriptional overlaps. Despite this straightforward utility, identification of transcript ends (especially of TSSs) was extremely challenging due to the tremendous complexity of the VACV transcriptome. The amplified SMRT Iso-Seq technique utilizes a switching mechanism at the 5'-end of the RNA template, thereby generating full-length cDNAs³⁷.

An advantageous feature of the PacBio technique is that any error that arises can be easily corrected due to its high consensus accuracy³⁸. In this study, we used a workflow and pipeline for transcriptome profiling of long-read RNA sequencing data that was developed in our laboratory^{30,32,39}. In total, about 1,115,000 reads of inserts (ROIs) were generated using the PacBio platform (Supplementary Table S1). In addition to PATs, ROIs can potentially be produced non-specifically from A-rich regions of the RNA molecules³⁶. Thus, we excluded these products from further analysis using the LoRTIA toolkit. Non-specific binding of the adapter and PCR primer sequences to cDNAs can also lead to false positives and we excluded these artefacts using bioinformatic filtering. We also used the LoRTIA toolkit to confirm that the 5'- and 3'-termini of the sequencing reads represented real TSSs and TESs, respectively. When two or more sequencing reads with the same TSS contained different TESs, they were considered independent ROIs. Similarly, ROIs with different PAT lengths were regarded as independent. We also accepted those TSSs that have been described by others^{19,40–46}.

The advantages of the ONT MinION sequencing technique over the PacBio platform include cost-effectiveness, higher read output, and the ability to read sequences within the range of 200 to 800bps, in which PacBio and the SRS techniques are less effective⁴⁷. Although this method is hampered by high error rates, sequencing accuracy is not particularly detrimental to transcriptome analyses of well-annotated genomes, such as that of VACV⁴⁸. In addition to the oligo(dT)-based 1D protocol from ONT, we prepared a random hexanucleotide-primed RT to detect the non-polyA(+) RNA fractions and to validate TSSs. A Cap-Seq approach was used to identify TSS positions. Together, the nanopore sequencing methods yielded approximately 535,000 viral sequencing reads.

VACV ORFs are relatively short (Supplementary Table S2) and many of them overlap with each other. Because the PacBio MagBead loading protocol selectively removes DNA fragments shorter than 1kb⁴⁹, potential monocistronic transcripts with a short ORFs are underrepresented or missing in these datasets. Nonetheless, in analyses of very long polycistronic and complex transcripts, PacBio has better sequencing precision than ONT. Thus, the combined use of these LRS methods eliminates the shortcomings of both. Native RNA sequencing can circumvent spurious transcription reads that are generated in PCR and RT reactions by false priming, template switching, or second-strand synthesis. The major disadvantage of dRNA sequencing is that a few bases are always missing from the 5'-ends of transcripts, and in many cases also from the 3'-ends⁵⁰. Under these sequencing conditions, it is assumed that the lacking 5'-end nucleotides are the consequence of perturbed base calling due to premature release of the mRNA from the motor protein, which hasten the progress of RNA molecules across pores. The frequent absence of PATs in sequencing reads is explained by the miscalling of adapter nucleotides that are ligated downstream of PATs on the RNA molecule, thus muddling of the raw signal of the downstream 'A' homopolymer. The present dRNA-Seq analysis returned a total of approximately 195,000 reads.

This study demonstrates an extremely complex transcription pattern along the entire viral genome. Our sequencing approaches detected transcriptional activity at every nucleotide of the VACV genome. The number of annotated ORFs vary between 201 and 263^{3,51} depending on the strain and study. We identified 218 ORFs in the WR strain (GenBank accession number is LT966077.1) using *in silico methods*. 184 out of 218 ORFs were expressed either as exclusively monocistronic or as both mono- and polycistronic RNA molecules. We detected 8,191 unique putative transcripts using the LoRTIA pipeline (Supplementary Table S3a). When we applied a criterion of accepting only those transcripts which were identified by two different techniques and/or in three independent experiments, this number decreased to 1480 transcripts (Supplementary Table S3b), including potential mRNAs, non-coding transcripts, or RNA isoforms. Transcripts annotated by LoRTIA represented 175 ORFs (Supplementary Table S3b, c). Altogether 99.32% of TSSs, and 98.62% of TES positions of the annotated 1,480 transcripts were reobtained by Guppy basecalling (Supplementary Table S3b). Our dataset contains full-length transcripts for about 20 additional ORFs (Supplementary Table S3d), but no full-length reads were detected for about 25 ORFs. These regions contain several transcriptional reads without exact TSS and/or TES positions.

We introduce the concepts of 'regular' and 'chaotic' genomic regions. At the 'regular' genomic segments viral genes produce transcript with relatively exact TSSs and TESs (Supplementary Fig. S1). However, at the 'chaotic' genomic loci the transcripts exhibit high heterogeneity with respect to the exact positions of TSSs and TESs (Fig. 2). Transcripts with highly diverse transcription initiation and termination sites are encoded by the I and L genes (such as O1L, O2L, I1L, I2L, I3L, A10L, A11R, A12L, A18R genes) at particularly the late time points of the infection. This phenomenon creates considerable difficulty in identifying individual transcripts.

Putative protein-coding genes. Herein, we identified (a) genes in new genomic positions, (b) embedded genes (5'-truncated in-frame ORFs within larger canonical ORFs), and (c) short upstream ORFs (uORFs, preceding the main ORFs). We considered the most abundant transcript isoform specified by a given gene as the canonical (main) transcript. (Note S1). The novel ORFs and transcripts were named according to the common (Copenhagen) HindIII fragment letter/number-based nomenclature⁴⁵. All VACV transcripts (identified by the LoRTIA pipeline) are presented in a Geneious file available at FigShare (Supplementary Note S2).

Genes at novel genomic positions. We detected two novel putative mRNAs with short ORFs (both 51bps) that are located in the intergenic genomic regions (B25.5R/C19.5L, C9.5L) of two divergently oriented genes. The TSS positions of these transcripts have previously been reported^{14,19} (Supplementary Table S4). Our analysis showed that these ORFs are expressed both alone in separate transcripts and as upstream ORFs (uORFs) located 5' of canonical ORFs in longer transcripts.

Embedded genes. A characteristic feature of the VACV genome is that it contains several short in-frame ORFs, called embedded genes, many of which are translated^{12,40}. Using full-length RNA sequencing we identified 49 novel putative embedded genes with 5'-truncated in-frame ORFs, which were shorter than the *in silico* annotated canonical ORFs (Fig. 3). This study demonstrated that these putative genes specify more than 300 tran-

Figure 2. Complexity of VACV ‘chaotic regions’. **(a)** high expression of antisense transcripts within the A18R locus; **(b)** transcription start and end sites within the open reading frames (ORFs) of L genes; **(c)** extremely high expression levels of antisense RNAs and other non-coding transcripts at the A10L–A12L region; **(d)** enormous numbers of non-coding transcripts and transcript isoforms within the O1L–I3L genomic segment. All transcripts were identified using the LoRTIA toolkit.

Figure 3. Schematic 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.

script isoforms. We initially assumed that the TISs of these internal ORFs were the closest in-frame ATGs to canonical host ORFs. However, this is not necessarily the case, a large number of TISs were detected by others using ribosome profiling¹². We identified 22 unique promoters associated with these TISs using our in-house scripts but we detected only six transcripts that contain in-frame ORFs (Supplementary Table S5).

Upstream ORFs. Upstream ORF-containing transcripts are considered to play a regulatory role in eukaryotic gene expression⁵². Translation of uORFs typically inhibit downstream expression of canonical ORFs. The 5'-UTR of uORF-containing transcripts are reportedly ≥ 75 bp in length⁵². Several in silico and experimental studies have shown that 40–50% of human and mouse mRNAs contain at least one uORF^{53,54}. We applied strict criteria for filtering data obtained by the LoRTIA toolkit and thereby revealed that 25 previously annotated VACV genes express longer TSS variants than the canonical transcripts. On average, the 5'-UTRs of the main transcript isoforms were 101 bp (Supplementary Fig. S2), whereas the longer variants were 258 bp long (Supplementary Table S6). Most of these transcript isoforms (64%) contain at least one ORF preceding the main ORF with an average 5'-UTR length of 331 nts. We set the minimal size limit of these uORFs to 9 nts comprising at least one triplet between the putative start and stop codons, as described by Scholz and colleagues⁵⁵. We also set a maximal ORF length of 90 nts and considered transcripts with longer upstream ORFs as bicistronic. Eleven VACV transcripts with 5'-UTR variants were found to contain multiple uORFs. G5.5R was the only gene in which the longer transcript isoform contained a single uORF. The VACV transcriptome is distinguished by a high number of TSS and TES positions and.

Non-coding transcripts. In this work, we identified 356 novel putative long non-coding RNAs (lncRNAs; ≥ 200 bps) using the LoRTIA toolkit (Supplementary Table S3b). We also identified 13 potential short ncRNAs (sncRNAs; ≤ 200 bps) with lengths varying between 121 and 193 bps.

Antisense RNAs. With few exceptions, we detected transcriptional activity from both DNA strands along the entire viral genome. The ratio between the mRNAs and antisense RNAs (asRNAs) varied between genomic locations. After applying the strict criteria to the LoRTIA tool, exactly 100 antisense transcripts were located at the genomic loci A5R (a single transcript), A11R (89 transcripts), and A18R (10 transcripts) (Fig. 4). The A11R antisense transcripts are various combinations of six TSS and 40 TES positions, whereas antisense A18R transcripts have the same TSSs but vary in their termination sites. Some of these transcripts contain small ORFs, which might indicate coding potentials (Note S2). In most cases, the level of asRNA expression was relative low, that is why the LoRTIA tool was unable to identify these molecules.

Intragenic non-coding transcripts. After applying our strict filtering criteria, we identified 260 3'-truncated transcripts from the LoRTIA dataset. We considered these as non-coding transcripts because they lacked in-frame stop codons. However, we cannot exclude the possibility that they utilize an out-of-frame ORF for protein coding. These transcripts were detected to be expressed from 41 viral genes (36.15% of them are located within the O1L region). The A37R (8.85%) and A10L (5.38%) regions and the F12L-F13L gene cluster (11.15%) also expressed a substantial number of 3'-truncated RNAs. Altogether 62% of the 3'-truncated ncRNAs are located within five VACV ORFs (Supplementary Table S3b).

Figure 4. Detection of antisense RNA expression from the complementary DNA strands of two VACV genes using qRT-PCR. **(a)** Polymerase chain reaction (PCR) products of the A11R transcript and its antisense partner. Abbreviations: M: molecular weight marker; A: A11R antisense RNA; B: A11R mRNA; C: A11R antisense RNA—NO-RT control; D: A11R mRNA—NO-RT control; E: A18R antisense RNA; F: A18R mRNA; G: A18R antisense RNA—NO-RT control; H: A18R mRNA—NO-RT control. **(b)** amplification curves of the A18R gene and its antisense transcripts. **(c)** Amplification curves of the A18R gene and its antisense transcripts. **(d–g)** Melting curves are shown to demonstrate the specificity of the amplification. The curves of transcripts cross the threshold line within 19.5–23.4 cycles, whereas the curves of NO-RT controls remain flat. **(d)** A11R mRNA and antisense RNA. **(e)** A11R DNA control. **(f)** A18R mRNA and antisense RNA **(g)** A18R DNA control.

Replication-associated transcripts. The VACV genome contains multiple replication origins (Oris)⁵⁶ located near the genomic termini. Oris are overlapped by many low-abundance ncRNAs; however, due to their large variation in length, their TSSs and TESs were not identified by either LoRTIA or other methods. We demonstrate that these RNAs are expressed from 4 h post infection (p.i.), and they may play similar roles as the replication-associated RNAs of other herpesviruses⁵⁷.

Novel mono-, bi- and polycistronic RNAs. In ‘regular’ genomic regions we detected 135 monocistronic coding transcripts of previously annotated ORFs using the LoRTIA software suit, and 12 additional mRNAs that did not meet the LoRTIA criteria due to their low abundance (Supplementary Table S7). Similar to other viruses and prokaryotes, but unlike the eukaryotic organisms, VACV reportedly express bi- and polycistronic transcripts^{58–60}. Our study detected altogether 43 bicistronic, 137 tricistronic, 92 tetracistronic, 15 pentacistronic, and 5 hexacistronic RNA molecules (Supplementary Table S3b, c).

Complex transcripts. We term complex RNAs (cxRNAs) those multigenic transcripts which contain at least two ORFs in opposite orientations. This study detected 30 cxRNA molecules (Supplementary Table S3b, c, d, S8, Note S2) encoded by nine genomic segments. Six of these are TES isoforms of the A12L–A11R complex transcript, five cxRNAs are the combinations of TSSs and TESs within the A14L–A13L–A12L–A11R region, and four transcripts are alternative TES variants of VACVWR_161–A38L RNA. Three additional complex transcripts were excluded by LoRTIA analyses due to the low coverage at this genomic region. Since most VACV genes stand in tandem orientation with each other, and only a few convergently and divergently positioned gene pairs are present, the number cxRNAs is low compared to herpesviruses. Two cxRNAs (c-A21L–A20R, c-G3L–G2R) were considered non-coding because their upstream genes stand in an antisense orientation on the transcript. The remaining complex transcripts were categorized as mRNAs, because their first ORFs stand in the sense direction. We detected cxRNAs with convergently-oriented genes at E10R–E11L and A8R–A9L genomic regions. Fully overlapping convergent cxRNAs were also detected in the E10R–E11L region (Supplementary Fig. S3).

Figure 5. Jitter plot visualization of transcriptional start site distribution across the VACV genome. (a) Distribution of TSSs of our annotated transcripts are shown at the top (red dots); the TSSs detected using LoRTIA are depicted in the middle (orange). TSS positions described by others^{14,15,19,40} are located at the bottom of the figure (blue). (b) TSS and TES positions detected by LoRTIA share the TSS and TES positions obtained from other studies (Supplementary Table S9) with different frequency. Bar charts represent the fall of TSS and TES positions within a ± 10 – 20 – 30 sliding window.

LRS reveals highly complex isoform heterogeneity. Transcript isoforms include 5'- and 3'-UTR length variants of RNA molecules. Using tag-based RNA-Seq methods with SOLiD and Illumina sequencing platforms, Yang and co-workers determined 5'- and 3'-ends of VACV E mRNAs with high precision^{14,19} and also mapped terminal sequences of I and L transcripts⁴⁰. These authors also described the expression profiles of I and L RNA-products¹⁶. These studies identified several hundreds of TSSs and PASs; but they were unable to ascertain the precise 3'-ends of RNAs due to the high transcriptional complexity¹⁴.

Transcription start sites. We annotated 1073 TSSs of VACV transcripts, 987 of which have not previously been detected (Supplementary Table S9). Eighteen percent of TSSs matched with base-pair precision to those described by Yang and colleagues^{14,15,19,40}. 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 (Fig. 5a). Altogether 898 TSSs were detected in 'regular' regions. The previously published TSS positions are depicted in FigShare (Supplementary Note S2).

Transcription end sites. The TESs of early transcripts were increasingly heterogenic at late time points of infection, potentially because transcription termination signal recognition depends on different factors in early- and PR phases^{12,14,15,19}. Due to the extremely high TES diversity (Supplementary Fig. S4), we analysed their positions within a ± 30 -nt interval. We found that 24% of our annotated TESs matched the positions of the polyadenylation sites described by Yang et al.¹⁵ (Fig. 5b). Moreover, most genes express multiple transcript isoforms (median = 4, mean = 6.95), up to 30 in extreme cases, such as in regions.N1L, O2L, I1L, and A12L (Supplementary Table S3b; Supplementary Fig. S5). Moreover, no TSS positions belonging to 30 annotated ORFs, particularly in the A8R-A17L region, have been annotated in our and others' studies (Note S3).

Cis-regulatory elements. Yang and colleagues described a 15-nucleotide consensus promoter core motif upstream of E genes, but these sequences were also present at other genomic locations. We identified this motif at nine novel positions upstream of (C9.5L, F4L, F7L, J3R, A15R, VACWR_161, and A51R) and inside (A36R and VACWR_169) ORFs. Cis-regulatory elements have not previously been annotated within these regions (Note S2). However, these core motifs are not present in the chaotic genomic regions. Additionally, most TESs of novel transcripts are located about 50 nucleotides downstream of UUUUUNU termination signals, as described by Yang et al.¹⁹. Those TESs which were not preceded by this motif might use an alternative mechanism for polyadenylation-site selection. Matching TSSs with TESs can only be carried out using LRS techniques, which are therefore crucial for the transcriptome analysis, especially in VACV which exhibits a very complex gene expression profile. We found that in certain DNA segments of 'chaotic' regions almost all base positions can function as TESs. Transcription initiation was also found to be stochastic at these regions. Accordingly, our analysis revealed that VACV genes express transcripts with multiple shared TSSs (Fig. 2a–c, Supplementary Fig. S4a) and TES positions (Fig. 2c, d, Supplementary Fig. S3b). Distributions of TSSs and TESs along the VACV genome are presented in Supplementary Fig. S6. We identified TSSs for 90% of the promoters described by Yang et al., (2010), with the exception of G2R, J6R, A18R, A20R, and five hypothetical genes located in the repeat region of the VACV genome. We used in silico promoter predictions to detect potential TATA- and CAAT-box consensus sequences of the present TSSs. We analysed 146 TSSs without previously annotated promoters and found TATA-boxes in 88 (Fig. 6, Supplementary Table S10), and CAAT-boxes in 5 of them (Supplementary Table S10). The median distance between a TSS and a putative TATA-box was 54 nts. Yet, no significant differences in promoter–TSS distances were identified between 5'-truncated and canonical transcript isoforms, suggesting that these cis-regulatory structures were similar in these two types of VACV genes (Fig. 6).

Transcriptional overlaps. Using multiplatform analyses, we revealed an extremely complex meshwork of transcriptional overlaps, which were expressed by all viral genes. Cumulatively, 154 tandem, 32 divergent, and 32 convergent gene pairs were found in the VACV genome, with 21 overlapping ORFs 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. We also detected 106 parallel transcriptional overlaps between tandem genes (Supplementary Table S11). Seventeen of the 32 convergent gene pairs formed UTR-UTR overlaps (Sup-

Figure 6. In silico analysis of promoter elements. **(a)** The bar chart shows average distances between the TATA boxes and TSSs. **(b)** The WebLogo shows the consensus sequence of the identified TATA boxes.

plementary Table S11), whereas only 15 gene pairs formed UTR-ORF overlaps. Twelve of the 16 divergently-oriented gene pairs were also found to form transcriptional overlaps with each other (Supplementary Table S11). Polycistronic and complex transcripts are also the results of transcriptional overlaps (Supplementary Table S12).

Discussion

Short-read sequencing has become a standard technique for the characterization of transcriptomes^{61,62}. Yet comprehensive annotation of transcripts from SRS data remains challenging⁶³. Auxiliary methods, such as Northern-blotting and rapid amplification of cDNA ends analyses are too laborious for investigations of large transcriptomes. Long-read sequencing allows the determination of full-length transcripts in single reads without computational inference, and thus distinguishes between transcript isoforms, polycistronic RNAs, and transcript overlaps with ease⁶⁴. Previous studies using Illumina sequencing have detected large numbers of TSSs, TESs, and TISs of VAVC transcripts^{12,15,19}. However, these analyses have not identified transcripts because SRS techniques are unable to match TSSs and TESs.

In this study, we employed two LRS platforms (PacBio and ONT) and sequenced cDNAs and native RNAs using various library preparation methods for surveying the temporal transcriptional activities of VAVC genes. We developed pipelines for profiling long-read RNA sequencing data. In actuality, RNA molecules expressed by the virus may be much more numerous than we annotated. This may be due to a high heterogeneity of transcript ends, especially in 'chaotic' regions in the viral genome, which are difficult to analyse by bioinformatic techniques. The intermediate and late transcripts are particularly polymorphic in length, but the variation in TSS and TES positions of early transcripts is also increased at late stages of infection. Generally, canonical transcripts are significantly overrepresented among isoforms encoded in 'regular' genomic regions. Apart from canonical transcripts, VAVC genes typically encompass shorter genes with N-terminal truncated ORFs, which may encode shorter polypeptides, possibly with altered effector functions. Moreover, longer transcript variants often incorporate uORFs, which may regulate the expression of downstream genes³².

The genomic organization of poxviruses differs from that of herpesviruses in many ways. Poxvirus E genes are gathered at the genome termini, whereas I and L genes are located in central parts of the DNA molecule. In addition, unlike herpesviruses, VAVC DNA encodes numerous TSS and TES isoforms. While prototype herpesvirus transcripts are organized into tandem gene clusters generating overlapping transcripts with co-terminal RNA molecules, adjacent pox genes produce large numbers of alternative TESs. In VAVC, the heterogeneity of TESs exceeds that of TSSs, whereas the opposite is the case in herpesviruses. The general presence of within-ORF TSSs is also unique to VAVC. Yang and colleagues, using ribosome profiling assays, have also identified several downstream methionine residues and mapped these within main ORFs¹².

We detected numerous TSSs upstream of TISs, thereby providing transcriptional evidence for the existence of potential truncated proteins. We also detected long 5'-UTR isoforms containing uORFs in most VAVC genes, providing a potential source of micro peptides with regulatory functions, as demonstrated previously in Kaposi's sarcoma virus⁶⁵. The observed transcript diversity of VAVC may, however, represent transcriptional noise from cryptic promoters and/or error-prone transcriptional machinery. Moreover, the majority of transcripts contain multiple translationally active ORFs, and most isoforms contain unique combinations of ORFs and uORFs, suggesting that this transcriptional diversity is functional. Hence, further studies are needed to demonstrate the biological significance of these RNA molecules.

We provide evidence that poly(A) signals are frequently transmitted by RNA polymerase without downstream cleavage of transcripts. This process results in the production of polycistronic RNAs from tandem genes and asRNAs from convergent gene pairs. The ensuing stochastic transcription initiation and termination in poxviruses represents a unique gene expression system, which has not yet been described in other well-characterized organisms. No previous studies show translation of polycistronic RNA molecules from downstream genes in

poxviruses. To explain this phenomenon and read-through transcription in herpesviruses^{66,67}, we hypothesize that transcriptional overlaps are sites of transcriptional interference. These interactions have been suggested to be a part of a genome-wide regulatory system, coined transcriptional interference networks (TINs)⁶⁸. We report sources of potential error in RT, PCR, and bioinformatic techniques, and propose strategies for addressing these. In particular, we applied very strict criteria for accepting sequencing reads as transcripts. However, a large fraction of excluded reads could represent existing transcripts.

Methods

Cells and viruses. CV-1 African green monkey kidney fibroblast cells were obtained from the American Type Culture Collection. Cells were plated at a density of 2×10^6 cells per 150-cm² tissue culture flask and were cultured in RPMI 1640 medium (Sigma-Aldrich) supplemented with 10% foetal bovine serum and antibiotic–antimycotic solution (Sigma-Aldrich). CV-1 cells were incubated at 37 °C in a humidified atmosphere containing 5% CO₂ until confluent. Subsequently, about 2.6×10^7 cells were rinsed with serum free RPMI medium immediately prior to infection with the vaccinia virus WR strain diluted in serum free RPMI medium.

Infection. Cells were infected with 3-ml aliquots of VACV at a multiplicity of infection of 10/cell, and were incubated at 37 °C in an atmosphere containing 5% CO₂ for 1 h with brief agitation at 10-min intervals to redistribute virions. Three-mL aliquots of complete growth medium (RPMI + 10% FBS) were then added to tissue culture flasks and the cells were incubated at 37 °C for 1, 2, 3, 4, 6, 8, 12, and 16 h in a humidified atmosphere containing 5% CO₂. Following incubation, media were removed and cells were rinsed with serum free RPMI 1,640 medium and were subjected to three cycles of freeze–thawing. Cells were finally scraped into 2-ml aliquots of PBS and were stored at –80 °C until use.

RNA purification. Total RNA was extracted from infected cells at various stages of viral infection using Macherey–Nagel RNA kit according to the manufacturer's instructions. Polyadenylated RNA fractions were isolated from total RNA samples using Oligotex mRNA Mini Kit (Qiagen) following the Spin-Column Protocol. We used Ribo-Zero Magnetic Kit H/M/R (Illumina) to remove rRNAs from total RNA samples for analyses of non-polyadenylated RNAs.

PacBio RSII and sequel sequencing. *cDNA synthesis.* Copy DNAs were generated from polyA(+) RNA fractions using SMARTer PCR cDNA Synthesis Kits (Clontech) according to 'PacBio Isoform Sequencing (Iso-Seq) using Clontech SMARTer PCR cDNA Synthesis Kit and No Size Selection' protocols. Samples from different infection time points (1-, 4-, 8-, and 12-h p.i.) were mixed for RSII library preparation, whereas 1-, 2-, 3-, 4-, 6-, and 8-h p.i. samples were mixed for Sequel sequencing. Samples of cDNA were prepared from rRNA-depleted RNA mixtures from 1-, 4-, 8-, and 12-h time points using modified random hexamer primers (Supplementary Table S13) instead of using the oligo(d)T-containing primer provided in the SMARTer Kit. Samples were used for SMRTbell template preparation using the PacBio DNA Template Prep Kit 1.0.

SMRTbell library preparation and sequencing. The detailed version of the template preparation protocol is described in our earlier publication⁶⁹. Briefly, primer annealing and polymerase binding reactions were performed using the DNA Sequencing Reagent Kit 4.0 v2 with DNA Polymerase P6 for the RSII platform, whereas the Sequel Sequencing Kit 2.1 and Sequel DNA Polymerase 2.0 were applied for sequencing on the Sequel. Polymerase–template complexes were bound to magbeads prior to loading into the PacBio instrument. Reactions were then performed using The PacBio's MagBead Kit (v2). Finally, 240- or 600-min movies were captured using the RSII and Sequel machines, respectively. One movie was recorded for each SMRTcell.

ONT MinION platform–cDNA sequencing. *1D cDNA library preparation.* PolyA(+) RNA fractions were used for cDNA sequencing on the MinION device. RNAs from different infection time points were converted to cDNAs according to the ONT 1D strand-switching cDNA ligation protocol (Version: SSE_9011_v108_revS_18Oct2016). An RNA mixture containing equal amounts of RNA from 1-, 2-, 3-, 4-, 6-, 8-, 12-, and 16-h p.i. were prepared for sequencing. Libraries were generated using the above mentioned 1D ligation kit and protocol, the Ligation Sequencing 1D kit (SQK-LSK108, ONT), and the NEBNext End repair/dA-tailing Module, NEB Blunt/TA Ligase Master Mix (New England Biolabs) according to the manufacturer's instructions. Briefly, polyA(+)-selected RNAs were converted to cDNAs using Poly(T)-containing anchored primers [(VN)T20; Bio Basic, Canada], dNTPs (10 mM, Thermo Scientific), SuperScript IV Reverse Transcriptase Kit (Life Technologies), RNaseOUT (Life Technologies), and strand-switching oligonucleotides with three O-methyl-guanine RNA bases (PCR_Sw_mod_3G; Bio Basic, Canada). The resulting double-stranded cDNAs were amplified by PCR using KAPA HiFi DNA Polymerase (Kapa Biosystems), Ligation Sequencing Kit Primer Mix (from the ONT 1D Kit), and a Veriti Thermal Cycler (Applied Biosystems). The NEBNext End repair/dA-tailing Module was used to blunt and phosphorylate cDNA ends, and the NEB Blunt/TA Ligase Master Mix was used for adapter (supplied in the 1D kit) ligation.

Size selection: PCR products from mixed RNA samples were size-selected manually and were then run on Ultrapure Agarose gels (Thermo Fischer Scientific). Subsequently, fragments of greater than 500-bp were isolated using Zymoclean Large Fragment DNA Recovery Kits (Zymo Research).

Barcoding: Individually sequenced cDNAs were barcoded using a combination of the following ONT protocols: the 1D protocol was used until the first end-preparation step, then we switched to the 1D PCR barcoding (96) genomic DNA (SQK-LSK108) protocol (version: PBGE96_9015_v108_revS_18Oct2016, updated

25/10/2017), followed by the barcode ligation step using the ONT PCR Barcoding Kit 96 (EXP-PBC096). Barcode adapters were ligated to end-prepped samples using Blunt/TA Ligase Master Mix.

Library preparation from cap-selected samples. For more accurate definitions of TSSs of full-length RNA molecules, we followed the 5'-Cap-specific cDNA generation protocol combined with the ONT 1D-seq library preparation method. We produced cDNAs from a mixture of total RNA samples (1-, 2-, 3-, 4-, 6-, 8-, 12-, 16-h p.i.) using TeloPrime Full-Length cDNA Amplification Kits (Lexogen). Amplified PolyA- and Cap-selected samples were then used for library preparation following the ONT 1D strand-switching cDNA ligation method (ONT Ligation Sequencing 1D kit). Samples were then end-repaired (NEBNext End repair/dA-tailing Module) and ligated to ONT 1D adapters (NEB Blunt/TA Ligase Master Mix).

Sequencing on the MinION device. ONT 1D-cDNA and Cap-selected libraries were loaded onto three and two ONT R9.4 SpotON Flow Cells for sequencing, respectively. Sequencing runs were performed using MinKNOW.

ONT MinION platform-dRNA sequencing. To avoid probable amplification-based biases, we used the ONT PCR-free direct RNA (dRNA) sequencing protocol (Version: DRS_9026_v1_revM_15Dec2016). A mixture containing PolyA(+) RNAs from eight time points (1-, 2-, 3-, 4-, 6-, 8-, 12-, 16-h p.i.) was used to generate the library. RNA samples, the oligo(dT)-containing adapter (ONT Direct RNA Sequencing Kit; SQK-RNA001), and T4 DNA ligase (2 M U/mL; New England BioLabs) were mixed and incubated for 10 min, and first-strand cDNAs were then prepared using SuperScript III Reverse Transcriptase enzyme (Life Technologies) according to the dRNA protocol. The RMX sequencing adapter was ligated to the samples with NEBNext Quick Ligation Reaction Buffer and T4 DNA ligase. The dRNA library was run on a R9.4 SpotON Flow Cell. Runs were performed using MinKNOW.

Purification of libraries. Samples were purified using Agencourt AMPure XP magnetic beads (Beckman Coulter) after each enzymatic steps during PacBio and ONT library preparation. Beads were handled before use with RNaseOUT (40 U/ μ L; 2U enzyme/1 μ L bead) for dRNA libraries.

Data analysis and visualization. *Generation of consensus sequences from the PacBio dataset.* ROI reads were created from the RSII raw data using the RS_ReadsOfInsert protocol (SMRT Analysis v2.3.0) with the following settings: Minimum Full Passes = 1, Minimum Predicted Accuracy = 90, Minimum Length of Reads of Insert = 1, Maximum Length of Reads of Insert = No Limit. ROIs from the Sequel dataset were generated using SMRT Link5.0.1.9585.

ONT dataset-basecalling. MinION base calling was performed using the ONT Albacore software v2.0.1. The newest release of the ONT's Guppy (Guppy v3.6.0) was also used for basecalling with the aim to validate the TSS and TES positions of the annotated transcripts using a ± 10 bp window.

Mapping and annotation of TSS and TES positions and transcripts. PacBio ROIs and ONT raw reads were aligned to the reference genome of the virus (LT966077.1); RefSeq assembly accession: GCF_000409795.2 [latest]) using minimap2 aligner (version 2.13) with the options -ax splice -Y -C5-cs. After applying the LoRTIA toolkit (<https://github.com/zsolt-balazs/LoRTIA>), the determined 5'- and 3'-ends of transcripts and detected full-length reads were mapped.

Prediction of cis-regulatory sequences. Sequences at 100-nt upstream of a given TSS were extracted and an in-house script based on the GPMIner promoter prediction tool^{70,71} was used for analyses of upstream cis-regulatory sequences of novel transcripts. The general settings of the algorithm were as follows: the eukaryotic promoter database of GPMIner was used to search exact matches without gaps or substitutions. The Vaccinia early promoter motif^{4,15,19} and late promoter motif⁴¹ were picked and visualized using the Geneious R10 Motif Finder tool, allowing one substitution in the string. Promoter statistics were calculated using Mann-Whitney U tests with two-tailed p-values.

Transcript naming scheme. Previously described (Copenhagen nomenclature⁴⁵) ORFs were not renamed. Our scheme allowed for future additions of novel transcripts. Multicistronic transcripts were named for all contributed ORFs, the most upstream ORF was listed first. Non-coding transcripts are identified with the prefix "nc," and complex transcripts carry the prefix "c." When a non-coding transcript is antisense to the ORF for which it was named, an "as" prefix was added. Names of length isoforms, e.g., TSS and TES isoforms, end with "l" for long, "s" for short, or "AT" for alternative termination (Note S1). Isoform categories are presented in Supplementary Fig. S7.

The details for validation of specificity, quality and quantity of libraries are found in Supplementary Methods. The detailed protocols and the raw dataset were published⁷².

Ethics declaration. All experiments using VACV were conducted under biosafety level 2. Neither human nor animal experiments were applied in this study.

Code availability

Relevant data are within the manuscript and its supporting information files. In addition, long-read sequencing datasets are available in European Nucleotide Archive (ENA) under the accession number PRJEB26434 and PRJEB26430. The complete map of the VACV transcriptome is available at FigShare: https://figshare.com/articles/Vaccinia_virus_transcriptome/12196275?fbclid=IwAR0WXkaqAjX4JpIrFJobkBM515H8y40ek6_fxz1X019Lz aigRdUMgSbUaY as a Geneious file.

Received: 8 February 2020; Accepted: 3 August 2020

Published online: 14 August 2020

References

1. Moss, B. Poxviridae in *Fields Virology* (eds. Knipe, D. M., et al.) 2129–2159 (Lippincott Williams & Wilkins, 2013).
2. Esposito, J. J. et al. Genome sequence diversity and clues to the evolution of variola (smallpox) virus. *Science* **313**, 807–812. <https://doi.org/10.1126/science.1125134> (2006).
3. Garcel, A., Crance, J., Drillien, R., Garin, D. & Favier, A. Genomic sequence of a clonal isolate of the vaccinia virus Lister strain employed for smallpox vaccination in France and its comparison to other orthopoxviruses. *J. Gen. Virol.* **88**, 1906–1916. <https://doi.org/10.1099/vir.0.82708-0> (2007).
4. Moss, B. Poxviridae: The viruses and their replication. In *Fields Virology* (eds Knipe, D. M. & Howley, P. M.) 2905–2946 (Lippincott Williams & Wilkins, Philadelphia, 2007).
5. Yang, Z. & Moss, B. Decoding poxvirus genome. *Oncotarget* **30**, 28513–28514. <https://doi.org/10.18632/oncotarget.5892> (2015).
6. Wei, C. M. & Moss, B. Methylated nucleotides block 5'-terminus of vaccinia virus mRNA. *Proc. Natl. Acad. Sci. U.S.A.* **72**, 318–322. <https://doi.org/10.1073/pnas.72.1.318> (1975).
7. Kates, J. & Beeson, J. Ribonucleic acid synthesis in vaccinia virus. I. The mechanism of synthesis and release of RNA in vaccinia cores. *J. Mol. Biol.* **50**, 1–18. [https://doi.org/10.1016/0022-2836\(70\)90100-2](https://doi.org/10.1016/0022-2836(70)90100-2) (1970).
8. Davison, A. J. & Moss, B. Structure of vaccinia virus early promoters. *J. Mol. Biol.* **210**, 749–769. [https://doi.org/10.1016/0022-2836\(89\)90107-1](https://doi.org/10.1016/0022-2836(89)90107-1) (1989).
9. Davison, A. J. & Moss, B. Structure of vaccinia virus late promoters. *J. Mol. Biol.* **210**, 771–784. [https://doi.org/10.1016/0022-2836\(89\)90108-3](https://doi.org/10.1016/0022-2836(89)90108-3) (1989).
10. Baldick, C. J. Jr., Keck, J. G. & Moss, B. Mutational analysis of the core, spacer, and initiator regions of vaccinia virus intermediate-class promoters. *J. Virol.* **66**, 4710–4719 (1992).
11. Broyles, S. Vaccinia virus transcription. *J. Gen. Virol.* **84**, 2293–2303. <https://doi.org/10.1099/vir.0.18942-0> (2003).
12. Yang, Z. et al. Deciphering poxvirus gene expression by RNA sequencing and ribosome profiling. *J. Virol.* **89**, 6874–6886. <https://doi.org/10.1128/JVI.00528-15> (2015).
13. Assarsson, E. et al. Kinetic analysis of a complete poxvirus transcriptome reveals an immediate-early class of genes. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 2140–2145. <https://doi.org/10.1073/pnas.0711573105> (2008).
14. Yang, Z., Bruno, D. P., Martens, C. A., Porcella, S. F. & Moss, B. Simultaneous high-resolution analysis of vaccinia virus and host cell transcriptomes by deep RNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 11513–11518. <https://doi.org/10.1073/pnas.1006594107> (2010).
15. Yang, Z. et al. Expression profiling of the intermediate and late stages of poxvirus replication. *J. Virol.* **85**, 9899–9908. <https://doi.org/10.1128/JVI.05446-11> (2011).
16. Honess, R. W. & Roizman, B. Regulation of herpesvirus macromolecular synthesis: Sequential transition of polypeptide synthesis requires functional viral polypeptides. *Proc. Natl. Acad. Sci. U.S.A.* **72**, 1276–1280. <https://doi.org/10.1073/pnas.72.4.1276> (1975).
17. Ross, L. & Guarino, L. A. Cycloheximide inhibition of delayed early gene expression in baculovirus-infected cells. *Virology* **232**, 105–113. <https://doi.org/10.1006/viro.1997.8557> (1997).
18. Cooper, J. A. & Moss, B. In vitro translation of immediate early, early, and late classes of RNA from vaccinia virus-infected cells. *Virology* **96**, 368–380. [https://doi.org/10.1016/0042-6822\(79\)90095-3](https://doi.org/10.1016/0042-6822(79)90095-3) (1979).
19. Yang, Z., Bruno, D. P., Martens, C. A., Porcella, S. F. & Moss, B. Genome-wide analysis of the 5' and 3' ends of vaccinia virus early mRNAs delineates regulatory sequences of annotated and anomalous transcripts. *J. Virol.* **85**, 5897–5909. <https://doi.org/10.1128/JVI.00428-11> (2011).
20. Yuen, L. & Moss, B. Oligonucleotide sequence signaling transcriptional termination of vaccinia virus early genes. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 6417–6421. <https://doi.org/10.1073/pnas.84.18.6417> (1987).
21. Shuman, S. & Moss, B. Factor-dependent transcription termination by vaccinia virus RNA polymerase. Evidence that the cis-acting termination signal is in nascent RNA. *J. Biol. Chem.* **263**, 6220–6225 (1988).
22. Ahn, B. Y., Jones, E. V. & Moss, B. 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. *J. Virol.* **64**, 3019–3024 (1990).
23. Rubins, K. H. et al. Comparative analysis of viral gene expression programs during poxvirus infection: a transcriptional map of the vaccinia and monkey pox genomes. *PLoS ONE* **3**, e2628. <https://doi.org/10.1371/journal.pone.0002628> (2008).
24. Stern-Ginossar, N. et al. Decoding human cytomegalovirus. *Science* **338**, 1088–1093. <https://doi.org/10.1126/science.1227919> (2012).
25. Guerra, S. et al. Cellular gene expression survey of vaccinia virus infection of human HeLa cells. *J. Virol.* **77**, 6493–6506. <https://doi.org/10.1128/jvi.77.11.6493-6506.2003> (2003).
26. Brum, L. M., Lopez, M. C., Varela, J. C., Baker, H. V. & Moyer, R. W. Microarray analysis of A549 cells infected with rabbitpox virus (RPV): a comparison of wild-type RPV and RPV deleted for the host range gene, SPI-1. *Virology* **315**, 322–334. [https://doi.org/10.1016/S0042-6822\(03\)00532-4](https://doi.org/10.1016/S0042-6822(03)00532-4) (2003).
27. Tombácz, D. et al. Full-length isoform sequencing reveals novel transcripts and substantial transcriptional overlaps in a herpesvirus. *PLoS ONE* **11**, e0162868. <https://doi.org/10.1371/journal.pone.0162868> (2016).
28. Moldován, N. et al. Multi-platform sequencing approach reveals a novel transcriptome profile in pseudorabies virus. *Front. Microbiol.* **8**, 2708. <https://doi.org/10.3389/fmicb.2017.02708> (2018).
29. Tombácz, D. et al. Long-read isoform sequencing reveals a hidden complexity of the transcriptional landscape of herpes simplex virus type 1. *Front. Microbiol.* **8**, 1079. <https://doi.org/10.3389/fmicb.2017.01079> (2017).
30. Tombácz, D. et al. Multiple long-read sequencing survey of herpes simplex virus dynamic transcriptome. *Front. Genet.* **10**, 834. <https://doi.org/10.3389/fgene.2019.00834> (2019).
31. Prazsák, I. et al. Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus. *BMC Genom.* **19**, 873. <https://doi.org/10.1186/s12864-018-5267-8> (2018).
32. Balázs, Z. et al. Long-read sequencing of human cytomegalovirus transcriptome reveals RNA isoforms carrying distinct coding potentials. *Sci. Rep.* **7**, 15989. <https://doi.org/10.1038/s41598-017-16262-z> (2017).
33. 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* **4**, 170194. <https://doi.org/10.1038/sdata.2017.194> (2017).

34. Tombácz, D. *et al.* Characterization of the dynamic transcriptome of a herpesvirus with long-read single molecule real-time sequencing. *Sci. Rep.* **7**, 43751. <https://doi.org/10.1038/srep43751> (2017).
35. Boldogkői, Z., Moldován, N., Balázs, Z., Snyder, M. & Tombácz, D. Long-read sequencing—a powerful tool in viral transcriptome research. *Trends Microbiol.* **27**, 578–592. <https://doi.org/10.1016/j.tim.2019.01.010> (2019).
36. Balázs, Z. *et al.* Template-switching artefacts resemble alternative polyadenylation. *BMC Genom.* **20**, 824. <https://doi.org/10.1186/s12864-019-6199-7> (2019).
37. Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R. & Siebert, P. D. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* **30**, 892–897. <https://doi.org/10.2144/01304pf02> (2001).
38. Miyamoto, M. *et al.* Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. *BMC Genom.* **15**, 699. <https://doi.org/10.1186/1471-2164-15-699> (2014).
39. Tombácz, D. *et al.* Characterization of novel transcripts in pseudorabies virus. *Viruses* **7**, 2727–2744. <https://doi.org/10.3390/v7052727> (2015).
40. Yang, Z., Martens, C. A., Bruno, D. P., Porcella, S. F. & Moss, B. Pervasive initiation and 3'-end formation of poxvirus postreplicative RNAs. *J. Biol. Chem.* **287**, 31050–31060. <https://doi.org/10.1074/jbc.M112.390054> (2012).
41. Engelstad, M., Howard, S. T. & Smith, G. L. A constitutively expressed vaccinia gene encodes a 42-kDa glycoprotein related to complement control factors that forms part of the extracellular virus envelope. *Virology* **188**, 801–810. [https://doi.org/10.1016/0042-6822\(92\)90535-w](https://doi.org/10.1016/0042-6822(92)90535-w) (1992).
42. Lee-Chen, G. J., Bourgeois, N., Davidson, K., Condit, R. C. & Niles, E. G. Structure of the transcription initiation and termination sequences of seven early genes in the vaccinia virus HindIII D fragment. *Virology* **163**, 64–79. [https://doi.org/10.1016/0042-6822\(88\)90234-6](https://doi.org/10.1016/0042-6822(88)90234-6) (1988).
43. Schmitt, J. F. & Stunnenberg, H. G. Sequence and transcriptional analysis of the vaccinia virus HindIII I fragment. *J. Virol.* **62**, 1889–1897 (1988).
44. Tengelsen, L. A., Slabaugh, M. B., Bibler, J. K. & Hruby, D. E. Nucleotide sequence and molecular genetic analysis of the large subunit of ribonucleotide reductase encoded by vaccinia virus. *Virology* **164**, 121–131. [https://doi.org/10.1016/0042-6822\(88\)90627-7](https://doi.org/10.1016/0042-6822(88)90627-7) (1988).
45. Rosel, J. L., Earl, P. L., Weir, J. P. & Moss, B. Conserved TAAATG sequence at the transcriptional and translational initiation sites of vaccinia virus late genes deduced by structural and functional analysis of the HindIII H genome fragment. *J. Virol.* **60**, 436–449 (1986).
46. Broyles, S. S. & Moss, B. Homology between RNA polymerases of poxviruses, prokaryotes, and eukaryotes: nucleotide sequence and transcriptional analysis of vaccinia virus genes encoding 147-kDa and 22-kDa subunits. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 3141–3145. <https://doi.org/10.1073/pnas.83.10.3141> (1986).
47. Larkin, J., Henley, R. Y., Jadhav, V., Korch, J. & Wanunu, M. Length-independent DNA packing into nanopore zero-mode waveguides for low-input DNA sequencing. *Nat. Nanotechnol.* **12**, 1169–1175. <https://doi.org/10.1038/nnano.2017.176> (2017).
48. Prazsák, I. *et al.* Full genome sequence of the western reserve strain of vaccinia virus determined by third-generation sequencing. *Genome Announc.* **6**, e01570-e1617. <https://doi.org/10.1128/genomeA.01570-17> (2018).
49. Ardui, S., Ameur, A., Vermeesch, J. R. & Hestand, M. S. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res.* **46**, 2159–2168. <https://doi.org/10.1093/nar/gky066> (2018).
50. Moldován, N. *et al.* Third-generation sequencing reveals extensive polycistronism and transcriptional overlapping in a baculovirus. *Sci. Rep.* **8**, 8604. <https://doi.org/10.1038/s41598-018-26955-8> (2018).
51. Goebel, S. J. *et al.* The complete DNA sequence of vaccinia virus. *Virology* **179**, 247–266. [https://doi.org/10.1016/0042-6822\(90\)90294-2](https://doi.org/10.1016/0042-6822(90)90294-2) (1990).
52. Somers, J., Pöyry, T. & Willis, A. E. A perspective on mammalian upstream open reading frame function. *Int. J. Biochem. Cell Biol.* **45**, 1690–1700. <https://doi.org/10.1016/j.biocel.2013.04.020> (2013).
53. Calvo, S. E., Pagliarini, D. J. & Mootha, V. K. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 7507–7512. <https://doi.org/10.1073/pnas.0810916106> (2009).
54. Matsui, M., Yachie, N., Okada, Y., Saito, R. & Tomita, M. Bioinformatic analysis of post-transcriptional regulation by uORF in human and mouse. *FEBS Lett.* **581**, 4184–4188. <https://doi.org/10.1016/j.febslet.2007.07.057> (2007).
55. Scholz, A. *et al.* uORF-tools-workflow for the determination of translation-regulatory upstream open reading frames. *PLoS ONE* **14**, e0222459. <https://doi.org/10.1371/journal.pone.0222459> (2019).
56. Senkevich, T. G. *et al.* Mapping vaccinia virus DNA replication origins at nucleotide level by deep sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 10908–10913. <https://doi.org/10.1073/pnas.1514809112> (2015).
57. Boldogkői, Z., Balázs, Z., Moldován, N., Prazsák, I. & Tombácz, D. Novel classes of replication-associated transcripts discovered in viruses. *RNA Biol.* **16**, 166–175. <https://doi.org/10.1080/15476286.2018.1564468> (2019).
58. Smith, G. L., Chan, Y. S. & Kerr, S. M. Transcriptional mapping and nucleotide sequence of a vaccinia virus gene encoding a polypeptide with extensive homology to DNA ligases. *Nucleic Acids Res.* **17**, 9051–9062. <https://doi.org/10.1093/nar/17.22.9051> (1989).
59. Brown, C. K., Turner, P. C. & Moyer, R. W. Molecular characterization of the vaccinia virus hemagglutinin gene. *J. Virol.* **65**, 3598–3606 (1991).
60. Grossegeisse, M., Doellinger, J., Haldemann, B., Schaade, L. & Nitsche, A. A next-generation sequencing approach uncovers viral transcripts incorporated in poxvirus virions. *Viruses* **9**, 296. <https://doi.org/10.3390/v9100296> (2017).
61. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628. <https://doi.org/10.1038/nmeth.1226> (2008).
62. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108. <https://doi.org/10.1038/nature11233> (2012).
63. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10**, 1177–1184. <https://doi.org/10.1038/nmeth.2714> (2013).
64. Rhoads, A. & Au, K. F. PacBio sequencing and its applications. *Genom. Proteom. Bioinform.* **13**, 278–289. <https://doi.org/10.1016/j.gpb.2015.08.002> (2015).
65. Kronstad, L. M., Brulois, K. F., Jung, J. U. & Glaunsinger, B. A. Dual short upstream open reading frames control translation of a herpesviral polycistronic mRNA. *PLoS Pathog.* **9**, e1003156. <https://doi.org/10.1371/journal.ppat.1003156> (2013).
66. Tombácz, D., Balázs, Z., Csabai, Z., Snyder, M. & Boldogkői, Z. Long-read sequencing revealed an extensive transcript complexity in herpesviruses. *Front. Genet.* **9**, 259. <https://doi.org/10.3389/fgene.2018.00259> (2018).
67. Boldogkői, Z., Tombácz, D. & Balázs, Z. Interactions between the transcription and replication machineries regulate the RNA and DNA synthesis in the herpesviruses. *Virus Genes* **55**, 274–279. <https://doi.org/10.1007/s11262-019-01643-5> (2019).
68. Boldogkői, Z. Transcriptional interference networks coordinate the expression of functionally related genes clustered in the same genomic loci. *Front. Genet.* **3**, 122. <https://doi.org/10.3389/fgene.2012.00122> (2012).
69. Tombácz, D. *et al.* Transcriptome-wide survey of pseudorabies virus using next- and third-generation sequencing platforms. *Sci. Data* **5**, 180119. <https://doi.org/10.1038/sdata.2018.119> (2018).
70. Lee, T. Y., Chang, W. C., Hsu, J. B., Chang, T. H. & Shien, D. M. GPMIner: an integrated system for mining combinatorial cis-regulatory elements in mammalian gene group. *BMC Genom.* **13**, S3. <https://doi.org/10.1186/1471-2164-13-S1-S3> (2012).
71. Narang, V., Sung, W. K. & Mittal, A. Computational modeling of oligonucleotide positional densities for human promoter prediction. *Artif. Intell. Med.* **35**, 107–119. <https://doi.org/10.1016/j.artmed.2005.02.005> (2005).

72. Tombácz, D. *et al.* Dynamic transcriptome profiling dataset of vaccinia virus obtained from long-read sequencing techniques. *Gigascience* 7, 139. <https://doi.org/10.1093/gigascience/giy139> (2018).

Acknowledgements

This study was supported by National Research, Development and Innovation Office (Hungary) Grants K 128247 to ZBo and FK 128252 to DT. DT was also supported by the Eötvös Scholarship of the Hungarian State. IP was supported by the UNKP-19-3-SZTE-243 New National Excellence Program of the Ministry for Innovation and Technology (Hungary). The project was also supported by the National Institutes of Health (USA) Centers of Excellence in Genomic Science Center for Personal Dynamic Regulomes (Grant No: 5P50HG00773502) to MS. The publications cost was covered by the University of Szeged Open Access Fund (Grant No: 4607). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

D.T., M.S. and Z.B. designed the study. N.M. provided additional input into study design. D.T. carried out the PacBio sequencing. D.T. and Z.C. performed the MinION sequencing. D.T., I.P., B.D., and Z.C. performed all other experiments. D.T., I.P., N.M., and Z. B. analysed the data. D.T., I.P., N.M. and Z.B. drafted the manuscript. D.T., I.P. and N.M. made the figures. D.T. and Z.B. wrote the final version of the manuscript. All authors read and approved the final paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-70794-5>.

Correspondence and requests for materials should be addressed to Z.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

ISSN: 1547-6286 (Print) 1555-8584 (Online) Journal homepage: <https://www.tandfonline.com/loi/krnb20>

Novel classes of replication-associated transcripts discovered in viruses

Zsolt Boldogkői, Zsolt Balázs, Norbert Moldován, István Prazsák & Dóra Tombácz

To cite this article: Zsolt Boldogkői, Zsolt Balázs, Norbert Moldován, István Prazsák & Dóra Tombácz (2019) Novel classes of replication-associated transcripts discovered in viruses, RNA Biology, 16:2, 166-175, DOI: [10.1080/15476286.2018.1564468](https://doi.org/10.1080/15476286.2018.1564468)

To link to this article: <https://doi.org/10.1080/15476286.2018.1564468>

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 15 Jan 2019.

Submit your article to this journal [↗](#)

Article views: 934

View related articles [↗](#)

View Crossmark data [↗](#)

Citing articles: 3 View citing articles [↗](#)

REVIEW

Novel classes of replication-associated transcripts discovered in viruses

Zsolt Boldogkői , Zsolt Balázs, Norbert Moldován , István Prazsák , and Dóra Tombác

Department of Medical Biology, Faculty of Medicine, University of Szeged, Szeged, Hungary

ABSTRACT

The role of RNA molecules in the priming of DNA replication and in providing a template for telomerase extension has been known for decades. Since then, several transcripts have been discovered, which play diverse roles in governing replication, including regulation of RNA primer formation, the recruitment of replication origin (Ori) recognition complex, and the assembly of replication fork. Recent studies on viral transcriptomes have revealed novel classes of replication-associated (ra)RNAs, which are expressed from the genomic locations in close vicinity to the Ori. Many of them overlap the Ori, whereas others are terminated close to the replication origin. These novel transcripts can be both protein-coding and non-coding RNAs. The Ori-overlapping part of the mRNAs is generally either the 5'-untranslated regions (UTRs), or the 3'-UTRs of the longer isoforms. Several raRNAs have been identified in various viral families using primarily third-generation long-read sequencing. Hyper-editing of these transcripts has also been described.

ARTICLE HISTORY

Received 5 November 2018
Revised 3 December 2018
Accepted 18 December 2018

KEYWORDS

Viruses; replication origin; replication fork; DNA polymerase; replication-associated transcript; non-coding RNA; long-read sequencing

DNA replication

The first step of eukaryotic DNA replication is the identification of replication origin (Ori) through the origin recognition complex (ORC), which is a multi-subunit structure composed of Orc1-6p [1]. ORC serves as a landing platform for the assembly of the replication forks. The unwinding of the double stranded DNA molecule into two single strands is initiated at the Ori. This process is carried out by the DNA helicase, an enzyme disrupting the hydrogen bonds between the base pairs of the complementary DNA strands. While the bacterial genome contains a single Ori, the replication of eukaryotic DNA molecule is initiated at hundreds of thousands points. Depending on the species, viruses have a single or a few Oris. DNA viruses usually code for proteins that are responsible for viral DNA replication [2–4]. However, in some cases, such as the Epstein-Barr virus (EBV) during latency, viruses can use the host replication proteins for viral DNA synthesis [5]. The replication fork is the site for the assembly of the replisome, a complex structure composed of the DNA polymerase (DNP), DNA helicase, topoisomerase, primase, DNA gyrase, DNA ligase, and telomerase among others. First, a short (11 ± 1 bp long) RNA sequence is synthesized by the primase enzyme, then this transcript is removed by an endonuclease, which is followed by the elongation with the DNP away from the origin of replication. Due to the 5' to 3' directionality of the DNA synthesis, the leading strand is continuously extended, whereas the lagging strand is synthesized discontinuously from multiple RNA primers. The resulting Okazaki fragments are joined together by the DNA ligase forming a single unified strand. In *E. coli* the termination of DNA replication occurs at specific consensus sequences and results in the disassembly of the replisome [6], while termination in eukaryotes is mostly not sequence specific [7]. Since

eukaryotes have linear DNA molecules, the DNP is unable to synthesize the very ends of the chromosomes (telomeres), and it leads to the shortening of telomeres in each replication cycle. Telomeres act as protective caps to prevent chromosomal integrity. In germ-line cells and in stem cells the telomerase enzyme catalyzes the repair of the telomere sequences by carrying a short complementary RNA molecule used for the priming of this process.

Replication-associated transcripts in prokaryotes and eukaryotes

The genomes of eukaryotic organisms are mostly comprised of non-coding DNA. Transcriptomic studies have revealed that the major parts of these genomic regions are transcriptionally active, producing non-protein coding RNAs (ncRNAs) [8]. These transcripts possess a wide range of functionality at practically all levels of the genetic regulation, including epigenetics, transcription, and post-transcriptional processes [9]. Evidence is also emerging that certain ncRNAs participate in the regulation of DNA replication. Besides the RNA primers and the telomerase RNA component – which were discovered decades ago, many replication-associated (ra) RNAs have been identified in the past few years. A recent genome-wide analysis has demonstrated that 72% of mammalian ORC1s are associated with active promoters, 46.5% of which controls the expression of protein coding genes, whereas 53.5% controls ncRNA genes [10].

Regulation of RNA primer synthesis by raRNAs

The vast majority of bacterial plasmids encode the Rep protein, the function of which is to separate the two DNA strands at the Ori region [11]. An alternative mechanism based on the

regulation of replication by ncRNAs has been demonstrated in a few cases. It has been shown that the Ori region of ColE1 plasmids of *Enterobacteriaceae* encodes two ncRNAs: RNA I and RNA II [12,13]. RNA II acts as a pre-primer by hybridizing with the DNA followed by being processed with RNase H. The resulted RNA fragments tend to initiate the synthesis of the leader DNA strand. RNA I is complementary to a short region of RNA II, and functions as a modulator of RNA II by binding to it, and thereby preventing the formation of RNA II/DNA hybrid that is needed for the initiation of replication. Besides the ColE1 family, a very similar RNA-mediated replication initiation mechanism has also been described in the marine RNA-based (MRB) plasmid family of *Vibrionaceae* [14]. Intriguingly, the sequences of MRB RNA I and RNA II transcripts are evolutionarily unrelated to those of ColE1-regulatory transcripts.

Inhibition of Rep synthesis by raRNAs

The expression of Rep protein itself is regulated in order to control the copy number of the plasmids. The R1 plasmids use, among others, an antisense RNA, termed CopA, for this control. CopA acts at the post-transcriptional level through binding RepA mRNA. The resultant double-stranded RNA is digested by RNase III, thereby preventing RepA synthesis [15]. Antisense RNAs also participate in the control of replication in ColIb-p9 plasmids. For the efficient translation of Rep protein, a pseudoknot has to be formed on the Rep mRNA. The Inc RNA has a complementary structure to the Rep mRNA, and therefore, it can block the replication through hybridization to the Rep mRNA [16].

raRNAs mediate ORC recruitment to the Ori

The mechanism of ORC recruitment to the Ori in eukaryotes varies in the different species. The ORC lacks sequence-specific DNA binding motives (except in yeast [17]), and so far, it has been unclear what factors control ORC binding to the DNA. One of the candidates are the non-coding raRNAs as they can provide sequence-specificity for the origin formation. In the protozoa *Tetrahymena thermophile* ribosomal DNA (rDNA) is amplified ~9,000 times during development. The 26T RNA has been shown to mediate the recruitment of ORC to the Ori region during the rDNA amplification through base pairing with the rDNA Ori [18]. It has also been demonstrated in mammalian systems that G-rich RNA mediates ORC recruitment to telomeres and to AT-rich heterochromatin [19,20]. Vertebrate genomes express the evolutionarily conserved Y RNAs, which are non-coding stem-loop transcripts, and they play a role in the initiation of replication [21,22]. The precise mechanism through which Y RNAs mediate their effects is unknown at present; however, it has been shown that these transcripts are recruited to the chromatin by the ORC.

Replication control by miRNAs

In addition to the above mechanisms, microRNAs (miRNAs) have also been described to control DNA replication through

fine-tuning this process. The miRNAs normally target the complementary mRNAs for translational repression or degradation. For example, in human cells, miR-29a targets the mRNA of Cdc7/Dbf4 kinase, which plays an essential role in the initiation of DNA replication. DNA damage leads to the up-regulation of Cdc7/Dbf4, which is accompanied by the repression of miR-29a in order to maximize the efficiency of repair processes [23].

Replication-associated RNAs in viruses

Replication-associated transcripts have also been identified in viruses. For example, a small replication-regulating (sr)RNA has been recently described in human BK polyomavirus (BKV) isolated from murine mammary tumor cells [24]. This transcript binds simultaneously to both sense and antisense DNA strands within the Ori region of the virus. The srRNA dramatically inhibits the replication of the poliovirus through interfering with the RNA primer synthesis, which changes the structure of the initiation complex even when the regulatory RNA is ectopically expressed in human cells [24]. BKV has an additional RNA-based mechanism for the control of replication: a miRNA targets the mRNA of the large T antigen during the early stage of infection that helps to establish persistence in the host cells [25]. Another example for the viral raRNAs is a highly structured GC-rich transcript of EBV, the function of which is to help for the viral proteins EBNA1 and HMGA1a proteins in ORC recruitment, and the origin formation at various chromosomal locations [26]. The miR-BART2 is a miRNA encoded by the EBV genome, and it binds to the mRNA of BALF5, which is the catalytic subunit of the EBV DNP. Repression of DNP blocks lytic replication, and it this leads to the establishment of latency in the human [27].

nroRNAs – novel replication-associated transcripts in herpesviruses

Next-generation short-read sequencing (SRS), and recently, third-generation long-read sequencing (LRS) have identified several novel raRNA molecules that are expressed from the genomic regions mapped in close vicinity to the replication origins in various viruses. These transcripts designated as near-replication origin (nro)RNAs in herpesviruses and further raRNAs in other viruses are supposed to be produced by a mechanism that regulates the replication initiation and the orientation of replisome progression, which is based on the collision of the replication and transcription apparatuses [28]. The herpesvirus family is subdivided into three subfamilies: alphaherpesviruses, including the human pathogenic Herpes simplex virus type 1 (HSV-1) and the Varicella-zoster virus (VZV) as well as the veterinary pathogen Pseudorabies virus (PRV); betaherpesviruses, such as the Human cytomegalovirus (HCMV); and gammaherpesviruses, including the Epstein-Barr virus (EBV) and Kaposi's sarcoma-associated virus (KSHV). Alphaherpesviruses have two genera: Varicellovirus (e.g. VZV and PRV) and Simplexvirus (e.g. HSV-1). The genome of alpha- and betaherpesviruses are composed of a unique long (UL) and a unique short (US)

region, which are either both bracketed by inverted repeats (IRs) such as in HSV-1, VZV and HCMV, or only the US region is surrounded by IRs such as in PRV (Fig. 1). Alphaherpesviruses express a variety of *ncRNAs* from the DNA segment around their *Oris*. PRV contains three *Oris*, one in its UL region (*OriL*), and two on the two copies of the IRs (*OriS*). CTO-S (Fig. 2A), a 286bp *ncRNA*, is first expressed at the onset of DNA replication (at 4 h post-infection) [28,29]. Despite being the most abundant PRV transcript, CTO-S is practically not expressed at all during the early stage of viral infection [30], which is very rare even among late viral transcripts. CTO-S has a transcription end site (TES) isoform (CTO-AT) which overlaps the convergently-oriented longer UL22 TES variant. CTO-L is a very long TES isoform of UL21 transcript. CTO-M is transcribed using the poly(A) signal of UL21 as a promoter. Except the low-abundance CTO-AS, the rest of the CTO transcripts have a common transcript start site (TSS). The CTO-M and CTO-L transcripts overlap the *OriS*. The PTO and the PTO-US1 transcripts are encoded by the genomic segments located near the PRV *OriLs*. The PTO is an *ncRNA* and it does not overlap the *Ori*, whereas the PTO-US1 (which can be considered as a very long TSS isoform of US1, or alternatively, a TES variant of PTO) overlaps the *OriLs*. PTO-US1 contain an intact open reading frame (ORF) of *us1* gene, but it is unknown whether this coding potential is realized in translation or not.

VZV contains two *OriSs*, one in each IR region, but it lacks *OriL*. The VZV NTO transcripts are similar to those of the PRV PTOs regarding their locations; however, the sequences of these transcripts are non-homologous (Fig. 2B) [31]: for example, the transcripts NTO1-3 overlap the ORF62 mRNA in an antiparallel manner, which is not the case for PTO; it does not overlap the ORF62-homologue IE180. The NTO2-4 RNAs are all non-coding. The NTO2-3 transcripts do not overlap the *OriS*, whereas the spliced NTO1 (expressed in two TES isoforms) does overlap it. Furthermore, the longer TSS variants of ORF63 (ORF63-L) and ORF63-64 (ORF63-L-64) are initiated very closely to the *OriS* (ORF63 is homologous to the US1 gene of PRV and HSV-1). Intriguingly, two closely-related organisms have evolved in parallel a distinct set of transcripts around their *OriSs* with presumably identical functions. The sequence of the *ncRNAs* appears to be irrelevant, only the *Ori*-proximal location seems to be important. A peculiar feature of the NTO3 transcript is that it is A to I hyper-edited, which has not been observed in any other VZV RNAs [31]. Hyper-editing has been shown to play a role in inhibiting RNA interference through making the mRNAs resistant to Dicer cleavage [32]. The function of this nucleotide modification in VZV and the importance of its proximity to the viral replication origin remain to be explored.

Unlike in PRV, the HSV-1 *us1* gene is located in the US region but it also produces *OriS* overlapping transcripts: the

Figure 1. The genomic structure of various herpesviruses. The genomes of herpesviruses are composed of varying number of unique and repeat regions. The replication origins can be located either in the unique, or in the repeat regions, or in both.

Abbreviations: PRV: Pseudorabies virus; VZV: Varicella-zoster virus; HSV-1: Herpes simplex virus type 1; HCMV: Human cytomegalovirus; EBV: Epstein-Barr virus; KSHV: Kaposi's sarcoma herpesvirus; UL: unique long region; US: unique short region; TRL: terminal repeat of UL region; TRS: terminal repeat of US region; TR: terminal repeat; IR: internal repeat; LUR: long unique coding region; Ori: replication origin.

Coloring black rectangle: Ori; yellow arrow-rectangle: open reading frame (ORF); blue arrow-rectangle connected with a line: intron; blue arrow-rectangle: mRNAs; red arrow-rectangle: non-coding RNA; red dashed rectangle: uncertain TSS and red rectangle with three black dots: transcript end terminating out of view.

Abbreviations: PRV: Pseudorabies virus; VZV: Varicella-zoster virus; HSV-1: Herpes simplex virus type 1; HCMV: Human cytomegalovirus; EBV: Epstein-Barr virus; KSHV: Kaposi's sarcoma herpesvirus

5'-coterminal ncRNA pairs, the OriS-RNA₁ and the OriS-RNA₂ (Fig. 2C). The OriS-RNA₁ has been shown to be expressed with an early kinetics, whereas OriS-RNA₂ is a late transcript [33]. Moreover, the longer isoform of the US1 transcript partially overlaps the OriS with its 5'-UTR [33]. Similarly, the 5'-UTRs of the US12 transcript variants also overlap the OriS. The OriL is located at a different position than in the PRV, but this genomic location also expresses an nroRNA, the OriL-RNA. In HSV, some nroRNAs also overlap with each other in an antiparallel fashion.

The OriLyts of HCMV [34], EBV [35] and KSHV [36] contain binding sites for the transactivator proteins (IE2, Zta and Rta respectively). In HCMV, a bidirectional promoter in the OriLyt region has been shown to regulate replication (Fig. 2D). This promoter could be functionally substituted with an SV40 promoter. Further, IE2 activation has been found to be required both for promoter activity and for DNA replication [37]. The numerous 3'-isoforms of the small replicator transcript (SRT) overlap the essential pyrimidine-rich region of the OriLyt, and therefore, they have been implicated in the regulation of replication. SRT is expressed from 2 h post-infection, and its expression increases throughout the viral replication cycle. RNA4.9, one of the most abundant transcripts of HCMV, also originates in the OriLyt region [38]. It has recently been shown that RNA4.9 regulates lytic replication both in *cis* and in *trans* [39]. Two virion-associated RNA (vRNA) species, one 300bp and another 500bp long, have also been discovered in the OriLyt region forming RNA-DNA hybrids. Such molecules have been supposed to be essential for the initiation of viral replication [40,41].

Lytic gammaherpesvirus DNA replication is dependent on the transcription in the viral Ori. The binding of KSHV Rta to the Rta responsive element has been shown to be essential for viral replication [36]. The produced OriLyt (T1.5) transcript, an early polyadenylated transcript [42], is indispensable for DNA replication [43]. The replication of EBV has also been reported to be dependent on Zta-induced transcription in the OriLyt [35]. The promoters and the transcription start sites of both LF3 and BHLF1 are found in the right and left lytic origins, respectively; however, no study has examined whether the transcription of these specific transcripts are required for the DNA replication.

Replication-associated transcripts in other viruses

Baculoviruses

The baculovirus Oris reside in the AT-rich repeat sequences, which are homologous within the members of *Baculoviridae* family [44,45] (Fig. 3A). The number and position of these homologous regions (*hrs*) varies greatly between the taxa [46]. The most studied baculovirus, *Autographa californica multiple nucleopolyhedrovirus* (AcMNPV) carries nine *hrs* [44,45]. Although the *hrs* have been identified as the main initiators of viral replication, other studies suggest that *non-hr* sequences [47,48] and promoters [49] can also act as origins of the DNA synthesis. It has been previously shown that the whole

AcMNPV genome is transcriptionally active [50,51], resulting in considerable read-through activity across all the *hrs*. LRS studies have revealed that the main source of overlaps is the very long transcripts spanning multiple oppositely-oriented ORFs (complex transcripts): 15 out of the 29 overlaps are formed by complex transcripts, seven by the most abundant isoforms, four by polycistronic transcripts, two by longer 5'-UTR isoforms, while only a single overlap is formed by a longer 3'-UTR isoform [50,51] (Fig. 3). Despite being the most numerous among the rRNAs by sort, complex RNAs are represented in low abundance [51], which indicates a much lower frequency of read-through events than in core and polycistronic transcripts as well as in UTR isoforms. Fourteen of the transcripts overlap the *hrs* with their 3'-UTRs, eleven with their 5'-UTRs, whereas in five transcripts (ORF114, PIF3-ORF114, ORF57-C, FP5K-ORF60-59-58 and FP25-K), the first protein-coding ORF overlaps the *hr* [50,51].

Poxviruses

The DNA primase-helicase encoded by Vaccinia virus (VACV) has been suggested to use RNA primers for the initiation of lagging-strand synthesis [52,53], although former studies proposed a model of rolling hairpin mechanism for viral replication [54,55]. The VACV DNA synthesis is initiated near the genome termini, where a repeat region forms a hairpin end acts as a conserved replication initiation site [56,57]. SRS analysis have revealed multiple potential replication start sites. The start positions of the leading strand are located around the Apex of the junction of concatenated VACV telomeric region within a near 400bp region, and putative Okazaki fragments have been mapped throughout the entire genome [58]. It has been demonstrated that VACV telomeres contain promoters for late RNA production, which may involve in concatemer resolution or in replication [59,60]. Some of these late telomeric transcripts (lateRNAs) bridge the entire Ori region. The 3' ends of the lateRNAs vary in length, and some of them encompass the entire telomeric hairpin, which contains the Ori (Fig. 3B). It has been proposed that the late transcription within the concatemer junction might play a role in opening up the DNA duplex during the initiation of the replication [58].

Circoviruses

A stem-loop structure in the intergenic region serves as the origin of replication for the small circular genome of circoviruses [61]. In the porcine circovirus type 1 (PCV-1), the 5'-UTR region of the core CAP transcript and of its two length isoforms form a full overlap with the Ori. Similarly, the 5'-UTR region of the three REP length isoforms form a full overlap, whereas the eight REP isoforms and the core REP transcript overlap the Ori with 5bps. Additionally, the 3'-UTR of the Atr transcript, encoding the putative ORF3 protein, overlaps the Ori with 18bps, while the two very long non-coding CTR and CTR' transcripts fully overlap the Ori [62–65] (Fig. 3C).

Figure 3. Replication-associated transcripts in a baculovirus, a poxvirus and a circovirus. These transcripts, similarly to the *nroRNA*-s of herpesviruses, are located in the close proximity of the replication origin, or they can overlap the Ori. a. Both ncRNAs and mRNAs have been described to start or terminate relatively close to the Ori of AcMNPV and VACV, while many of these overlap the Ori. b. All transcript isoforms of the PCV-1, except the CTR, overlap the Ori with their 5'-UTRs. c. The concatemer junction region of the Vaccinia virus encloses multiple replication origins [58], which are overlapped by many non-coding transcripts expressed in low abundance. The TSS and TES of these ncRNAs are highly variable.

Coloring: black rectangle: Ori; yellow arrow-rectangle: open reading frame (ORF); blue arrow-rectangle connected with a line: intron; blue arrow-rectangle: mRNAs; red arrow-rectangle: non-coding RNA. The highly variable ncRNAs of VACV are shown by red arrow-rectangles.

Abbreviations: AcMNPV: Autographa californica Multiple Nucleopolyhedrosis Virus, VACV: Vaccinia virus, PCV-1: Porcine circovirus type 1; NR: nonrepeated sequences

Interactions between the transcription and replication machineries

It is debated whether eukaryotic cells efficiently separate transcription and replication [66–68]. Concurrent transcription and replication can lead to the confrontation of the replication fork and the RNA polymerase. Using an *in vitro* system, Liu et al. [69] have demonstrated that RNA polymerase can continue elongation meanwhile the replication fork progresses on the other strand, parallel to it. *In vivo* experiments on a replicating plasmid in *E. coli* have shown that head-to-tail collision does not impede the progression of the replication fork, whereas head on collision does [70]. It has been demonstrated that crashes of the replication fork with the transcription machinery cause genetic instability in yeast and bacteria [71,72]. It has been described that RNA:DNA hybrids formed at sites of transcription-replication clashes, and that RNase H1 acts to suppress the instability of DNA breakage hot spots known as common fragile sites (CFSs) [73]. The authors suggest that replication and transcription are spatially and temporally separated in eukaryotic cells in order to avoid the confrontation of the polymerase molecules. DNA instability has been shown to be dependent on the length of the genes at a given genomic location because transcription of the longer genes takes more time. Helmrich et al. [73] have demonstrated that interference between the replication fork and the transcription complex is inevitable in the longest genes because of the inability of the two processes to be separated at large transcription units during the replication. This results in the formation of CFS, which is prone to genomic instability. Altogether, the authors argue that the

ongoing replication negatively regulates the transcription initiation in the long genes. The co-directional arrangement of the replication and transcription in prokaryotic [74] and eukaryotic [75] genomes may also serve to preclude the collision of these machineries. An alternative explanation for the separation of the transcription and replication is that the interaction between these processes causes this phenomenon, that is, the two processes may be under a common control.

Transcription replication interference network (TRIN)

It has been observed that the overall transcription rate of the individual PRV genes declines following the onset of DNA replication [76]. Similar to other organisms, the expression kinetics of herpesvirus genes is mainly governed by transcription factors acting on the promoters of these genes. Nevertheless, it is also possible that the process of replication affects the gene expression, and *vice versa*, the transcription exerts an effect on the progress of the replication fork. The discoveries of rRNAs in herpesviruses, baculoviruses and circoviruses [28,29,33,50,51,62,77,78], along with the general repressive effect of DNA synthesis on the transcription suggest that the two processes interact with one another on both the replication origins and across the entire genome. We present the proposed mechanism of the effects of nroRNAs on the determination of the orientation of the replication fork progression as well as the facilitatory effect of these transcripts on the DNA synthesis (Fig. 4). It is assumed that the RNP transcribing the CTO-M and CTO-L collides with the replisome (at the OriL) progressing in two

Figure 4. Collision of the RNA polymerase transcribing the nroRNAs and the DNA polymerase during the replication of PRV DNA. A: schematic representation of the transcripts expressed in the region around the PRV OriL (marked by a black box). B: First the PRV DNA is replicated through a bidirectional (theta) replication. C: Transcription initiated at the promoter of the CTO-S transcript furthers replication by unwinding the DNA. D: The CTO-M and CTO-L transcripts exert a repressive effect on DNA replication in one direction and the transcription of CTO-S facilitates it in the other direction by unwinding the DNA strands. In this hypothetical model, the existence of theta-type replication is not necessary, since the proposed collision-based mechanism can prevent its initiation.

directions (θ -type manner), and as a result, it renders the replication to be unidirectional (σ -type). In the meantime, CTO-S transcription unwinds the DNA strands, thereby helping the unidirectionality of the DNA replication. The transcription of CTO-AT is supposed to reduce the expression of the convergent *ul22* gene, thereby eliminating a potential inhibitory effect of transcription on the replication and also advancing the DNA synthesis in a σ -type manner. All in all, the CTO transcripts are thought to have a role in the switch from the θ -type to the σ -type of replication. Alternatively, this mechanism may impede the initiation of bidirectional replication. In this scenario, no θ -type replication occurs in herpesviruses at all. We assume that the interactions between the transcriptional machinery and the replication fork form a transcription and replication interference network (TRIN), which regulates the gene expression globally and determines the rate of the DNA synthesis in a mutually interdependent manner. It is very likely that the *nroRNAs* and the related *raRNAs* are not mere byproducts of this interference mechanism, but they also have functions as transcripts. This hypothesis is supported by the polyadenylation of these RNA molecules, the general function of which is to protect the RNA integrity. The viral *raRNAs* may direct the replication by forming R-loops [79].

Conclusions

It has become evident by now that all of the examined viruses express transcripts or transcript isoforms, which overlap the replication origin, or alternatively they start or end in close vicinity of the Ori. Altogether, we can distinguish four types of *raRNAs* on the basis of their coding potency and position to the Ori: (1) non-coding transcripts that do not overlap the Ori (such as CTO-S, CTO-S-AT, and PTO of PRV; NTO2, NTO3 and NTO4 of VZV; as well as SRT and RNA4.9 of HCMV); (2) non-coding transcripts that do overlap the Ori (such as CTO-M of PRV); (3) mRNA isoforms with very long 3'-UTR (such as CTO-L of PRV); and (4) mRNA isoforms with very long 5'-UTR variant [such as PTO-US1 of PRV, OriS RNA1 of HSV-1, as well as NTO1v1 and NTO1v2 of VZV]. The baculovirus *raRNAs* are all coding, two of them exhibit ORFs-Ori overlapping. The VACV lateRNAs are all non-coding. The difference between the size, location, expression characteristic, and posttranscriptional modification of the various viral *raRNAs* do not necessarily means that they would influence viral replication through distinct mechanisms. On the contrary, the expression of non-homologous transcripts near the Ori suggests that the same function can be solved in different ways, which are the result of convergent evolution. The mechanistic details of how these replication-associated transcripts exert their effects on the DNA synthesis have not yet been ascertained. We hypothesize that these transcripts are produced as byproducts of a mechanism regulating the DNA synthesis through the interaction between the replication and transcription machineries, but these transcripts very likely also function as RNA molecules. The inhibition of replication initiation through intervening in *raRNA* synthesis can be a useful approach for developing effective antiviral therapies. Even though the proteins responsible for

replication differ in viruses and eukaryotes, the mechanistic interactions between the replisome and the transcriptosome may be universal.

Abbreviations

AcMNPV	autographa californica multiple nucleopolyhedrovirus
BKV	BK polyomavirus
CFS	common fragile site
DNP	DNA polymerase
EBV	Epstein-Barr virus
HCMV	Human cytomegalovirus
HSV-1	herpes simplex virus type 1
lateRNA	late telomeric RNA
LRS	long-read sequencing
miRNA	microRNA
ncRNA	non-coding RNA
nroRNA	near-replication-origin RNA
ORC	origin recognition complex
ORF	open reading frame
PRV	Pseudorabies virus
raRNA	replication-associated RNAs
RNP	RNA polymerase
srRNA	small replication-regulating (sr)RNA
SRS	short-read sequencing
TES	transcript end site
TSS	transcript start site
UTR	untranslated region
VACV	Vaccinia virus
VZV	Varicella-zoster virus

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This project was supported by NKFIH OTKA K 128247 grant to ZBo and by NKFIH OTKA FK 128252 to DT.

ORCID

Zsolt Boldogkői <http://orcid.org/0000-0003-1184-7293>
 Norbert Moldován <http://orcid.org/0000-0003-4138-586X>
 István Prazsák <http://orcid.org/0000-0003-3195-503X>
 Dóra Tombácz <http://orcid.org/0000-0001-5520-2978>

References

- [1] Santocanale C, Diffley JF. ORC- and Cdc6-dependent complexes at active and inactive chromosomal replication origins in *Saccharomyces cerevisiae*. *Embo J* [Internet]. 1996 Dec 2 [cited 2018 Nov 27];15(23):6671–6679. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/8978693>
- [2] Moss B. Poxvirus DNA Replication. *Cold Spring Harb Perspect Biol*. 2013 Sep;5(9):a010199–a010199.
- [3] Cheung AK. Porcine circovirus: transcription and DNA replication. *Virus Res*. 2012 Mar;164(1–2):46–53.
- [4] Weller SK, Coen DM. Herpes simplex viruses: mechanisms of DNA replication. *Cold Spring Harb Perspect Biol*. 2012 Sep;4(9):a013011.
- [5] Dhar SK, Yoshida K, Machida Y, et al. Replication from oriP of Epstein-Barr virus requires human ORC and is inhibited by geminin. *Cell*. 2001 Aug;106(3):287–296.
- [6] Neylon C, Kralicek AV, Hill TM, et al. Replication termination in *Escherichia coli*: structure and antihelicase activity of the Tus-Ter

- complex. *Microbiol Mol Biol Rev* [Internet]. 2005 Sep [cited 2018 Nov 27];69(3):501–526. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/16148308>
- [7] McGuffee SR, Smith DJ, Whitehouse I. Quantitative, genome-wide analysis of eukaryotic replication initiation and termination. *Mol Cell* [Internet]. 2013 Apr 11 [cited 2018 Nov 27];50(1):123–135. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23562327>
 - [8] Hangauer MJ, Vaughn IW, McManus MT. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet*. 2013 Jun;9(6):e1003569.
 - [9] Palazzo AF, Lee ES. Non-coding RNA: what is functional and what is junk? *Front Genet*. 2015 Jan;6:2.
 - [10] Dellino GI, Cittaro D, Piccioni R, et al. Genome-wide mapping of human DNA-replication origins: levels of transcription at ORC1 sites regulate origin selection and replication timing. *Genome Res*. 2013 Jan;23(1):1–11.
 - [11] Węgrzyn K, Fuentes-Perez ME, Bury K, et al. Sequence-specific interactions of Rep proteins with ssDNA in the AT-rich region of the plasmid replication origin. *Nucleic Acids Res*. 2014 Jul;42(12):7807–7818.
 - [12] Tomizawa J, Itoh T, Selzer G, et al. Inhibition of ColE1 RNA primer formation by a plasmid-specified small RNA. *Proc Natl Acad Sci U S A*. 1981 Mar;78(3):1421–1425.
 - [13] Masukata H, Tomizawa J. Control of primer formation for ColE1 plasmid replication: conformational change of the primer transcript. *Cell*. 1986 Jan;44(1):125–136.
 - [14] Le Roux F, Davis BM, Waldor MK. Conserved small RNAs govern replication and incompatibility of a diverse new plasmid family from marine bacteria. *Nucleic Acids Res*. 2011 Feb;39(3):1004–1013.
 - [15] Blomberg P, Nordström K, Wagner EG. Replication control of plasmid R1: repA synthesis is regulated by CopA RNA through inhibition of leader peptide translation. *Embo J*. 1992 Jul;11(7):2675–2683.
 - [16] Asano K, Mizobuchi K. Copy number control of IncIalpha plasmid ColIb-P9 by competition between pseudoknot formation and antisense RNA binding at a specific RNA site. *Embo J*. 1998 Sep;17(17):5201–5213.
 - [17] Bell SP, Stillman B. ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. *Nature* [Internet]. 1992 May 14 [cited 2018 Nov 29];357(6374):128–134. Available from: <http://www.nature.com/doi/10.1038/357128a0>
 - [18] Mohammad MM, Donti TR, Sebastian Yakisich J, et al. Tetrahymena ORC contains a ribosomal RNA fragment that participates in rDNA origin recognition. *Embo J*. 2007 Dec;26(24):5048–5060.
 - [19] Hoshina S, Yura K, Teranishi H, et al. Human origin recognition complex binds preferentially to G-quadruplex-preferable RNA and single-stranded DNA. *J Biol Chem*. 2013 Oct;288(42):30161–30171.
 - [20] Deng Z, Norseen J, Wiedmer A, et al. TERRA RNA Binding to TRF2 Facilitates Heterochromatin Formation and ORC Recruitment at Telomeres. *Mol Cell*. 2009 Aug;35(4):403–413.
 - [21] Christov CP, Gardiner TJ, Szuts D, et al. Functional Requirement of Noncoding Y RNAs for Human Chromosomal DNA Replication. *Mol Cell Biol*. 2006 Sep;26(18):6993–7004.
 - [22] Krude T, Christov CP, Hyrien O, et al. Y RNA functions at the initiation step of mammalian chromosomal DNA replication. *J Cell Sci*. 2009 Aug;122(Pt 16):2836–2845.
 - [23] Barkley LR, Santocanale C. MicroRNA-29a regulates the benzo[a]pyrene dihydrodiol epoxide-induced DNA damage response through Cdc7 kinase in lung cancer cells. *Oncogenesis*. 2013 Jul;2(7):e57–e57.
 - [24] Tikhanovich I, Liang B, Seoighe C, et al. Inhibition of human BK polyomavirus replication by small noncoding RNAs. *J Virol*. 2011 Jul;85(14):6930–6940.
 - [25] Broekema NM, Imperiale MJ. miRNA regulation of BK polyomavirus replication during early infection. *Proc Natl Acad Sci*. 2013 May;110(20):8200–8205.
 - [26] Norseen J, Thoma A, Sridharan V, et al. RNA-dependent recruitment of the origin recognition complex. *Embo J*. 2008 Nov;27(22):3024–3035.
 - [27] Barth S, Pfuhl T, Mamiani A, et al. Epstein-Barr virus-encoded microRNA miR-BART2 down-regulates the viral DNA polymerase BALF5. *Nucleic Acids Res* [Internet]. 2007 Nov 26 [cited 2018 Oct 16];36(2):666–675. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18073197>
 - [28] Tombácz D, Csabai Z, Oláh P, et al. Characterization of novel transcripts in pseudorabies virus. *Viruses* [Internet]. 2015 May 22 [cited 2017 Nov 14];7(5):2727–2744. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26008709>
 - [29] Tombácz D, Csabai Z, Oláh P, et al. Full-Length Isoform Sequencing Reveals Novel Transcripts and Substantial Transcriptional Overlaps in a Herpesvirus. *PLoS One* [Internet]. 2016 Sep 29 [cited 2016 Oct 6];11(9):e0162868. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27685795>
 - [30] Tombácz D, Balázs Z, Csabai Z, et al. Characterization of the Dynamic Transcriptome of a Herpesvirus with Long-read Single Molecule Real-Time Sequencing. *Sci Rep* [Internet]. 2017 Mar 3 [cited 2017 Mar 29];7:43751. Available from: <http://www.nature.com/articles/srep43751>
 - [31] Prazsák I, Moldován N, Balázs Z, Tombácz D, Megyeri K, Szűcs A, et al. Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus. *BMC Genomics*. 2018 Dec;19(1):873.
 - [32] Scadden AD, Smith CW. RNAi is antagonized by A-to-I hyper-editing. *EMBO Rep*. 2001 Dec;2(12):1107–1111.
 - [33] 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* [Internet]. 1988 Nov 1 [cited 2018 Aug 21];85(22):8454–8458. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/2847162>
 - [34] Zhu Y, Huang L, Anders DG. Human cytomegalovirus oriLyt sequence requirements. *J Virol*. 1998 Jun;72(6):4989–4996.
 - [35] Fixman ED, Hayward GS, Hayward SD. Replication of Epstein-Barr virus oriLyt: lack of a dedicated virally encoded origin-binding protein and dependence on Zta in cotransfection assays. *J Virol*. 1995 May;69(5):2998–3006.
 - [36] Wang Y, Li H, Chan MY, et al. Kaposi's sarcoma-associated herpesvirus ori-Lyt-dependent DNA replication: cis-acting requirements for replication and ori-Lyt-associated RNA transcription. *J Virol*. 2004 Aug;78(16):8615–8629.
 - [37] Xu Y, Cei SA, Rodriguez Huete A, et al. Human cytomegalovirus DNA replication requires transcriptional activation via an IE2- and UL84-responsive bidirectional promoter element within oriLyt. *J Virol*. 2004 Nov;78(21):11664–11677.
 - [38] Gatherer D, Seirafian S, Cunningham C, et al. High-resolution human cytomegalovirus transcriptome. *Proc Natl Acad Sci U S A*. 2011 Dec;108(49):19755–19760.
 - [39] Tai-Schmiedel J, Karniely S, Ezra A, et al. The virally encoded long non-coding RNA4.9 is controlling viral DNA replication. In: *International Herpesvirus Workshop 2018*. Vancouver: University of British Columbia; 2018. p. 2.32.
 - [40] Prichard MN, Jairath S, Penfold ME, et al. Identification of persistent RNA-DNA hybrid structures within the origin of replication of human cytomegalovirus. *J Virol*. 1998 Sep;72(9):6997–7004.
 - [41] Rennekamp AJ, Lieberman PM. Initiation of Epstein-Barr Virus Lytic Replication Requires Transcription and the Formation of a Stable RNA-DNA Hybrid Molecule at OriLyt. *J Virol*. 2011;85(6):2837–2850.
 - [42] Purushothaman P, Thakker S, Verma SC. Transcriptome analysis of Kaposi's sarcoma-associated herpesvirus during de novo primary infection of human B and endothelial cells. *J Virol*. 2015 Mar;89(6):3093–3111.
 - [43] Wang Y, Tang Q, Maul GG, et al. Kaposi's sarcoma-associated herpesvirus ori-Lyt-dependent DNA replication: dual role of replication and transcription activator. *J Virol*. 2006 Dec;80(24):12171–12186.

- [44] Kool M, van Den Berg PMMM, Tramper J, et al. Location of Two Putative Origins of DNA Replication of Autographa californica Nuclear Polyhedrosis Virus. *Virology* [Internet]. 1993 Jan [cited 2018 Sep 6];192(1):94–101. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/8517035>
- [45] Pearson M, Bjornson R, Pearson G, et al. The Autographa californica baculovirus genome: evidence for multiple replication origins. *Science* [Internet]. 1992 Sep 4 [cited 2018 Sep 6];257(5075):1382–1384. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/1529337>
- [46] van Oers MM, Vlak JM. Baculovirus genomics. *Curr Drug Targets* [Internet]. 2007 Oct [cited 2018 Sep 6];8(10):1051–1068. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17979665>
- [47] Lee HY, Krell PJ. Generation and analysis of defective genomes of Autographa californica nuclear polyhedrosis virus. *J Virol* [Internet]. 1992 Jul [cited 2018 Sep 6];66(7):4339–4347. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/1602548>
- [48] Pearson MN, Bjornson RM, Ahrens C, et al. Identification and Characterization of a Putative Origin of DNA Replication in the Genome of a Baculovirus Pathogenic for *Orgyia pseudotsugata*. *Virology* [Internet]. 1993 Dec [cited 2018 Sep 6];197(2):715–725. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/8249294>
- [49] Wu Y, Liu G, Carstens EB. Replication, integration, and packaging of plasmid DNA following cotransfection with baculovirus viral DNA. *J Virol* [Internet]. 1999 Jul [cited 2018 Sep 6];73(7):5473–5480. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/10364295>
- [50] Chen Y-R, Zhong S, Fei Z, et al. The Transcriptome of the Baculovirus Autographa californica Multiple Nucleopolyhedrovirus in *Trichoplusia ni* Cells. *J Virol* [Internet]. 2013 Jun 1 [cited 2017 Nov 13];87(11):6391–6405. Available from: <http://jvi.asm.org/cgi/doi/10.1128/JVI.00194-13>
- [51] Moldován N, Tombácz D, Szűcs A, et al. Third-generation Sequencing Reveals Extensive Polycistronism and Transcriptional Overlapping in a Baculovirus. *Sci Rep* [Internet]. 2018 Dec 5 [cited 2018 Jun 5];8(1):8604. Available from: <http://www.nature.com/articles/s41598-018-26955-8>
- [52] De Silva FS, Lewis W, Berglund P, et al. Poxvirus DNA primase. *Proc Natl Acad Sci*. 2007;104(47):18724–18729.
- [53] Sèle C, Gabel F, Gutsche I, et al. Low-Resolution Structure of Vaccinia Virus DNA Replication Machinery. *J Virol*. 2013;87(3):1679–1689.
- [54] Baroudy BM, Venkatesan S, Moss B. Structure and replication of vaccinia virus telomeres. *Cold Spring Harb Symp Quant Biol*. 1982;47(2):723–729.
- [55] Baroudy BM, Venkatesan S, Moss B. Incompletely base-paired flip-flop terminal loops link the two DNA strands of the vaccinia virus genome into one uninterrupted polynucleotide chain. *Cell*. 1982;28(2):315–324.
- [56] Traktman P. 27 Poxvirus DNA replication. In: *DNA replication in eukaryotic cells*. Cold Spring Harbor Monograph Archive. New York (NY): Woodbury; 1996. p. 775–798.
- [57] Moss B. Poxvirus DNA replication. *Cold Spring Harb Perspect Biol*. 2013;5(9).
- [58] Senkevich TG, Bruno D, Martens C, et al. Mapping vaccinia virus DNA replication origins at nucleotide level by deep sequencing. *Proc Natl Acad Sci U S A* [Internet]. 2015 Sep 1 [cited 2018 Oct 29];112(35):10908–10913. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26286988>
- [59] Parsons BL, Pickup DJ. Transcription of orthopoxvirus telomeres at late times during infection. *Virology*. 1990;175(1):69–80.
- [60] Hu FQ, Pickup DJ. Transcription of the terminal loop region of vaccinia virus DNA is initiated from the telomere sequences directing DNA resolution. *Virology*. 1991;181(2):716–720.
- [61] Mankertz A, Persson F, Mankertz J, et al. Mapping and characterization of the origin of DNA replication of porcine circovirus. *J Virol* [Internet]. 1997 Mar 15 [cited 2018 Sep 10];71(3):2562–2566. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/9032401>
- [62] Moldován N, Balázs Z, Tombácz D, et al. Multi-platform analysis reveals a complex transcriptome architecture of a circovirus. *Virus Res* [Internet]. 2017 Jun 2 [cited 2018 Jan 15];237:37–46. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/28549855>
- [63] Mankertz J, Buhk H-J, Blaess G, et al. Transcription Analysis of Porcine Circovirus (PCV). *Virus Genes*. 1998;16(3):267–276.
- [64] Mankertz A, Hillenbrand B. Analysis of transcription of Porcine circovirus type 1. *J Gen Virol* [Internet]. 2002 Nov [cited 2016 Oct 11];83(Pt 11):2743–2751. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12388810>
- [65] Cheung AK. Comparative analysis of the transcriptional patterns of pathogenic and nonpathogenic porcine circoviruses. *Virology* [Internet]. 2003 May 25 [cited 2016 Oct 11];310(1):41–49. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12788629>
- [66] Hassan AB, Errington RJ, White NS, et al. Replication and transcription sites are colocalized in human cells. *J Cell Sci*. 1994;107(2):425–434.
- [67] Wei X, Samarabandu J, Devdhar RS, et al. Segregation of transcription and replication sites into higher order domains. *Science* [Internet]. 1998 Sep 4 [cited 2018 Nov 29];281(5382):1502–1506. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/9727975>
- [68] MacAlpine DM. Coordination of replication and transcription along a Drosophila chromosome. *Genes Dev* [Internet]. 2004 Dec 15 [cited 2018 Nov 29];18(24):3094–3105. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/15601823>
- [69] Liu B, Wong ML, Alberts B. A transcribing RNA polymerase molecule survives DNA replication without aborting its growing RNA chain. *Proc Natl Acad Sci U S A*. 1994 Oct;91(22):10660–10664.
- [70] Mirkin EV, Mirkin SM. Mechanisms of transcription-replication collisions in bacteria. *Mol Cell Biol*. 2005 Feb;25(3):888–895.
- [71] Prado F, Aguilera A. Impairment of replication fork progression mediates RNA polII transcription-associated recombination. *Embo J*. 2005 Mar;24(6):1267–1276.
- [72] Torres JZ, Schnakenberg SL, Zakian VA. Saccharomyces cerevisiae Rrm3p DNA helicase promotes genome integrity by preventing replication fork stalling: viability of rrm3 cells requires the intra-S-phase checkpoint and fork restart activities. *Mol Cell Biol*. 2004 Apr;24(8):3198–3212.
- [73] Helmrich A, Ballarino M, Tora L. Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Mol Cell*. 2011 Dec;44(6):966–977.
- [74] Rocha EPC. The Organization of the Bacterial Genome. *Annu Rev Genet*. 2008 Dec;42(1):211–233.
- [75] Huvet M, Nicolay S, Touchon M, et al. Human gene organization driven by the coordination of replication and transcription. *Genome Res* [Internet]. 2007 Jul 25 [cited 2017 Nov 11];17(9):1278–1285. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17675363>
- [76] Takács IF, Tombácz D, Berta B, et al. The ICP22 protein selectively modifies the transcription of different kinetic classes of pseudorabies virus genes. *BMC Mol Biol*. 2013 Jan;14:2.
- [77] 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* [Internet]. 1996 Aug 1 [cited 2018 Aug 21];70(8):5272–5281. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/8764037>
- [78] Tombácz D, Csabai Z, Szűcs A, et al. Long-Read Isoform Sequencing Reveals a Hidden Complexity of the Transcriptional Landscape of Herpes Simplex Virus Type 1. *Front Microbiol* [Internet]. 2017 Jun 20 [cited 2017 Dec 18];8:1079. Available from: <http://journal.frontiersin.org/article/10.3389/fmicb.2017.01079/full>
- [79] Lombrana R, Almeida R, Álvarez A, et al. R-loops and initiation of DNA replication in human cells: a missing link? *Front Genet*. 2015;6:158.

DATA NOTE

Dynamic transcriptome profiling dataset of vaccinia virus obtained from long-read sequencing techniques

Dóra Tombácz ¹, István Prazsák ¹, Attila Szűcs ¹, Béla Dénes ², Michael Snyder ³ and Zsolt Boldogkői ^{1,*}

¹Department of Medical Biology, Faculty of Medicine, University of Szeged, Somogyi B. u. 4., 6720 Szeged, Hungary, ²Veterinary Diagnostic Directorate of the National Food Chain Safety Office, Tábornok u. 2., 1143 Budapest, Hungary and ³Department of Genetics, School of Medicine, Stanford University, 300 Pasteur Dr, Stanford, California, USA

*Correspondence address. Zsolt Boldogkői, Department of Medical Biology, Faculty of Medicine, University of Szeged, Somogyi B. u. 4., Szeged, 6720, Hungary. E-mail: boldogkoi.zsolt@med.u-szeged.hu <http://orcid.org/0000-0003-1184-7293>

Abstract

Background: Poxviruses are large DNA viruses that infect humans and animals. Vaccinia virus (VACV) has been applied as a live vaccine for immunization against smallpox, which was eradicated by 1980 as a result of worldwide vaccination. VACV is the prototype of poxviruses in the investigation of the molecular pathogenesis of the virus. Short-read sequencing methods have revolutionized transcriptomics; however, they are not efficient in distinguishing between the RNA isoforms and transcript overlaps. Long-read sequencing (LRS) is much better suited to solve these problems and also allow direct RNA sequencing. Despite the scientific relevance of VACV, no LRS data have been generated for the viral transcriptome to date. **Findings:** For the deep characterization of the VACV RNA profile, various LRS platforms and library preparation approaches were applied. The raw reads were mapped to the VACV reference genome and also to the host (*Chlorocebus sabaeus*) genome. In this study, we applied the Pacific Biosciences RSII and Sequel platforms, which altogether resulted in 937,531 mapped reads of inserts (1.42 Gb), while we obtained 2,160,348 aligned reads (1.75 Gb) from the different library preparation methods using the MinION device from Oxford Nanopore Technologies. **Conclusions:** By applying cutting-edge technologies, we were able to generate a large dataset that can serve as a valuable resource for the investigation of the dynamic VACV transcriptome, the virus-host interactions, and RNA base modifications. These data can provide useful information for novel gene annotations in the VACV genome. Our dataset can also be used to analyze the currently available LRS platforms, library preparation methods, and bioinformatics pipelines.

Keywords: poxvirus; vaccinia virus; long-read sequencing; full-length transcriptome; Pacific Biosciences; RS II system; Sequel system; Oxford Nanopore Technologies; MinION system; direct RNA sequencing

Data Description

Background

Poxviridae is a large virus family that infects vertebrates and invertebrates with highly pathogenic members, such as the Variola virus, which is the causative agent of smallpox [1]. Vaccinia virus (VACV) is the prototypic member of the Orthopoxvirus

genus. It is closely related to the Variola virus [2] that was eradicated as a result of a successful global vaccination program using live VACV.

It had generally been assumed that the virus in the smallpox vaccine, renamed vaccinia virus, is a cowpox virus. However, VACV differs from the cowpox virus and has no known natural hosts; its origin is still being investigated. It has been suggested

Received: 18 May 2018; Revised: 23 September 2018; Accepted: 12 November 2018

© The Author(s) 2018. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

that the smallpox vaccine was based on horsepox [3]. VACV has been extensively utilized as an expression and a gene delivery vector [4]. It also serves as a model system for the analysis of virus-host interactions, transcriptional regulation, and for other molecular biological studies [5].

Poxviruses are able to replicate in the cytoplasm of the host cell because they encode the proteins needed for DNA synthesis [6]. They have a relatively large (approximately 195 kbp) double-stranded DNA genome coding for about 220 proteins. The VACV genes are divided into three temporal classes: early (E), intermediate (I), and late (L) genes. A study characterized 35 VACV genes as immediate-early (IE) kinetics [7], but this categorization has not been widely accepted. The promoters of genes belonging to different kinetic classes are recognized by stage-specific transcription factors [8–11]. VACV genes belonging to the same kinetic group have been shown to be clustered in the genome [7]: E genes are located at the termini of the viral genome, while I and L genes are situated in the middle genomic region. Most of the adjacent VACV genes are oriented in the same direction, while convergent and divergent positioning is uncommon.

Although the extraordinary complexity of the VACV transcriptome has been thought to be well characterized [12–15], traditionally used techniques such as short-read sequencing (SRS), ribosome profiling, cap analysis of gene expression (CAGE), and genome tiling [16] are not able to span the entire transcript nor to distinguish between transcript isoforms, bi-, and polycistronic RNA variants, overlapping gene products, and embedded RNAs. Transcriptional overlaps generated by the read-through mechanism are very frequent in VACV and cause a major problem in the analysis of individual viral transcripts using traditional approaches. The transcription patterns of VACV genes exhibit an extreme stochasticity, which includes an enormous number of transcriptional start sites (TSSs) and transcription end sites (TESs) even within the open reading frames (ORFs). These features of transcription are uncommon even among large DNA viruses; it might represent a form of gene regulation that is unique to living organisms. Therefore, it is especially important to use full-length sequencing methods in order to match the transcript ends.

Previous studies have determined the precise TSSs and TESs of VACV transcripts [14, 17]. However, the methods that were applied were not suitable for detecting the entire transcripts at the single-molecule level, and therefore it was impossible to determine which TSSs are paired by certain TESs.

The Pacific Biosciences (PacBio) isoform sequencing (Iso-Seq) protocol (using oligo(dT) or random hexamer primers for the reverse transcription), the cDNA sequencing, and direct (d)RNA-sequencing (RNA-seq) methods from the Oxford Nanopore Technologies (ONT), as well as the Cap-selection (Cap-Seq) cDNA preparation method (Lexogen) are able to generate full-length transcripts, and thus they can circumvent the limitations of SRS techniques. By using these techniques for cDNA productions and library preparations with the PacBio Real-Time Sequencer (RS)II and Sequel, as well as the ONT MinION platforms, we were able to identify hundreds of novel RNA isoforms (e.g., TSS and TES variants, mono-, bi-, polycistronic transcripts), dozens of coding and non-coding RNAs, and numerous complex transcripts in various herpesviruses [18–23] and in a baculovirus [24] and we were also able to generate a comprehensive full-length transcript data catalog of VACV.

The PacBio Sequel and the ONT Cap-Seq methods yielded the highest number of full-length reads in our experiments (Fig. 1). The ratio between the complete and partial reads varies within the size-selected RSII samples. ONT 1D cDNA sequencing

yielded the lowest ratio of full-length transcripts but the highest number of read counts; therefore, full-length transcripts are also present in a large number in these samples. Even if a large proportion of the reads are incomplete, they can be utilized to, e.g., distinguish between the various transcript isoforms or to identify embedded transcripts, which is essential for the correct kinetic classification [13]. A large number of incomplete reads have been obtained from the dRNA-seq, which were consistent with our previous results [24]. The current method of this technique produces sequencing reads missing varying size short sequences from both ends. Random-primed reverse transcription (RT) - based sequencing rarely gains complete reads; the reason for which is that the primers seldom bind to exactly the 3'-ends of the transcripts. However, these samples provide further significant value to the dataset. For example, random-primed sequencing may result in novel, non-polyadenylated transcripts [25, 26], while direct RNA sequencing data may provide epitranscriptomic information by detecting base modifications (e.g., m7G). Furthermore, the dRNA-seq method is free of artifacts produced by RT and polymerase chain reaction (PCR) in cDNA sequencing.

The present report provides the first long-read, dynamic RNA profiling dataset from the family of Poxviruses and the host cell line (CV-1), which can redefine the VACV transcriptomic landscape. This study is a very large cohort of data from the currently available third-generation sequencing methods representing the forefront techniques for transcriptome research. As such, the data presented herein can prove to be useful not only at the molecular level and not just for virologists but also with respect to general genomics and bioinformatics.

Methods

A detailed workflow pertaining to the different library preparation strategies is presented in Fig. 2 and Table 1.

Cells and viruses

African green monkey (*Chlorocebus sabaeus*) kidney fibroblast cells (CV-1; American Type Culture Collection, [RRID:CVCL.022 9]) were cultured in RPMI 1640 medium (Sigma-Aldrich) supplemented with 10% fetal bovine serum (FBS) and antibiotic-antimycotic solution (Sigma-Aldrich) in a 150 cm² culture flask at 37°C in a humidified 5% CO₂ atmosphere until confluence was reached. The cells (~2.6 × 10⁷) were washed with serum-free medium before the infection. The highly virulent Western Reserve VACV strain was used in this study. The virus stock was diluted in serum-free RPMI 1640 medium and then used (3 mL virus solution; 10 MOI/cell) for the CV-1 infection. Samples were incubated at 37°C in 5% CO₂ atmosphere for 1 hour with brief agitation at 10-minute intervals to redistribute the virus. Three milliliters of complete growth medium (RPMI 1640 + 10% FBS) was added to the tissue culture flask, and the infected cells were further incubated for 1, 2, 4, and 8 hours for RSII sequencing; 1, 2, 3, 4, 6, and 8 hours for Sequel; or 1, 2, 3, 4, 6, 8, 12, and 16 hours for MinION sequencing (Table 1) at 37°C in a humidified 5% CO₂ atmosphere. After the incubation, the cells were rinsed with serum-free RPMI 1640 medium, which was followed by the application of three freeze-thaw cycles. Cells were scraped into 2 mL of phosphate-buffered saline and stored at –80°C until use.

Table 1: Summary table of the different wet lab approaches applied in this study.

Run No.	Platform	Sample collection strategy	Time points (h)	RNA sample	RT priming	Cap-selection	PCR	Size selection	Library prep	Barcodes	Base calling
1	RSII	Static	1, 2, 4, 8	PolyA(+)	Oligo(d)T	No	Yes	No	Iso-Seq	No	SMART Analysis v2.3.0
2	RSII	Static	1, 2, 4, 8	rRNA depletion	Random hexamer	No	Yes	No	Iso-Seq	No	SMART Analysis v2.3.0
3	RSII	Static	1, 2, 4, 8	PolyA(+)	Oligo(d)T	No	Yes	BluePippin 0.8kb-5kb+	Iso-Seq	No	SMART Analysis v2.3.0
4	RSII	Static	1, 2, 4, 8	PolyA(+)	Oligo(d)T	No	Yes	BluePippin 0.8-2kb	Iso-Seq	No	SMART Analysis v2.3.0
5	RSII	Static	1, 2, 4, 8	PolyA(+)	Oligo(d)T	No	Yes	BluePippin 2-3kb	Iso-Seq	No	SMART Analysis v2.3.0
6	RSII	Static	1, 2, 4, 8	PolyA(+)	Oligo(d)T	No	Yes	BluePippin 3-5kb	Iso-Seq	No	SMART Analysis v2.3.0
7	RSII	Static	1, 2, 4, 8	PolyA(+)	Oligo(d)T	No	Yes	BluePippin 5kb+	Iso-Seq	No	SMART Analysis v2.3.0
8	Sequel	Dynamic	1	PolyA(+)	Oligo(d)T	No	Yes	No	Iso-Seq	No	SMART Link v5.0.1.9585
9	Sequel	Dynamic	2	PolyA(+)	Oligo(d)T	No	Yes	No	Iso-Seq	No	SMART Link v5.0.1.9585
10	Sequel	Dynamic	3	PolyA(+)	Oligo(d)T	No	Yes	No	Iso-Seq	No	SMART Link v5.0.1.9585
11	Sequel	Dynamic	4	PolyA(+)	Oligo(d)T	No	Yes	No	Iso-Seq	No	SMART Link v5.0.1.9585
12	Sequel	Dynamic	4	PolyA(+)	Oligo(d)T	No	Yes	No	Iso-Seq	No	SMART Link v5.0.1.9585
13	Sequel	Dynamic	6	PolyA(+)	Oligo(d)T	No	Yes	No	Iso-Seq	No	SMART Link v5.0.1.9585
14	Sequel	Dynamic	8	PolyA(+)	Oligo(d)T	No	Yes	No	Iso-Seq	No	SMART Link v5.0.1.9585
15	Sequel	Dynamic	8	PolyA(+)	Oligo(d)T	No	Yes	No	Iso-Seq	No	SMART Link v5.0.1.9585
16	MinION	Static	1, 2, 3, 4, 6, 8, 12, 16	PolyA(+)	Oligo(d)T	No	Yes	Manual Gel 500bp+	1D cDNA	No	Albacore v2.0.1
17	MinION	Static	1, 2, 3, 4, 6, 8, 12, 16	Total RNA	Oligo(d)T	Yes	Yes	No	Teloprime&1D dRNA	No	Albacore v2.0.1
18	MinION	Static	1, 2, 3, 4, 6, 8, 12, 16	PolyA(+)	Oligo(d)T	No	No	No	1D cDNA	Yes	Albacore v2.0.1
19	MinION	Dynamic	1	PolyA(+)	Oligo(d)T	No	Yes	No	1D cDNA	Yes	Albacore v2.0.1
20	MinION	Dynamic	2	PolyA(+)	Oligo(d)T	No	Yes	No	1D cDNA	Yes	Albacore v2.0.1
21	MinION	Dynamic	3	PolyA(+)	Oligo(d)T	No	Yes	No	1D cDNA	Yes	Albacore v2.0.1
22	MinION	Dynamic	4	PolyA(+)	Oligo(d)T	No	Yes	No	1D cDNA	Yes	Albacore v2.0.1
23	MinION	Dynamic	6	PolyA(+)	Oligo(d)T	No	Yes	No	1D cDNA	Yes	Albacore v2.0.1
24	MinION	Dynamic	8	PolyA(+)	Oligo(d)T	No	Yes	No	1D cDNA	Yes	Albacore v2.0.1
25	MinION	Dynamic	12	PolyA(+)	Oligo(d)T	No	Yes	No	1D cDNA	Yes	Albacore v2.0.1

The dynamic transcriptome includes transcripts from various stages of viral infection (from 1 hour to 8 hours for Sequel and from 1 hour to 12 hours for MinION sequencing), while static transcriptome contain transcripts expressed at various time points of infection.

Figure 1: Distribution of sequencing reads. **(A)** The stacked bar chart of the proportion of full-length and partial reads from PolyA cDNA-sequencing shows large differences between the various library-preparation and sequencing methods. All of the PacBio methods and the Cap-selected ONT approach resulted in a higher percentage of full-length reads. The weakest ratios of complete/incomplete reads are from MinION 1D sequencing. The explanation for this result is the lack of size selection. In PacBio sequencing, even in the non-size-selected samples, the short RNA fragments were eliminated by the MagBead loading protocol. **(B)** The horizontal bar graph shows the proportion of full-length/partial reads derived from oligo(d)T-primed, non-size-selected cDNA sequencing, generated by the three different library preparation kits utilized in this study (the same cDNA kits were used for PacBio RSII and Sequel libraries). The sum of the read counts was taken from individual time points of Sequel and MinION 1D sequencing. In order to obtain a full set of transcripts, we mixed RNA samples obtained from various time points for the Cap-Seq analysis. No significant difference between the Sequel and the Cap-selected MinION libraries can be observed, while the MinION 1D-Seq produced much fewer complete sequencing reads. **(C)** This figure shows the methods that generated a very low amount of complete reads (<10%). The weak result of the non-size-selected RSII is not to be considered significant because of the very low yield of this run. However, due to technical reasons, this phenomenon is to be expected from the dRNA-seq and from the random primed sequencing.

Table 2: Summary table of the amount of RNA, cDNA, and library samples used for PacBio Sequel sequencing.

Run No.	Time points (h)	Amount of PolyA(+) RNA used for cDNA preparation (ng)	Concentration of PCR products (ng/ μ L)	Concentration of SMRTbell libraries (ng/ μ L)
1	mixed	16.2	86.1	5.8
2	mixed	10.0 ^a	109.1	5.2
3	mixed	14.6	75.5	6.9
4	mixed	14.6	98.2	7.2
5	mixed	14.6	84.1	7.1
6	mixed	14.6	89.4	9.1
7	mixed	14.6	77.9	7.9
8	1	27.3	410.0	8.1
9	2	10.0	112.0	9.0
10	3	18.9	87.8	11.1
11	4	51.8	460.0	12.1
12	4	19.9	98.3	6.8
13	6	20.3	95.0	7.8
14	8	39.2	460.0	12.0
15	8	19.6	120.0	6.1

^a Amount of rRNA-depleted RNA (for random-primed sequencing)

RNA

Total RNA was purified from the infected cells at various stages of viral infection from 1 to 16 hours post-infection (pi) using the NucleoSpin RNA Kit from Macherey-Nagel. Polyadenylated RNAs were purified from the total RNA using the Oligotex mRNA Mini Kit (Qiagen, Additional File 1). For the analysis of non-

polyadenylated RNAs, ribodepletion (Epicentre Ribo-Zero Magnetic Kit H/M/R) was carried out on the total RNA samples. RNAs were quantified (Table 2, row A) by Qubit 2.0 using the Qubit RNA BR Assay Kit for the total RNAs and the Qubit RNA HS Assay Kit for the polyA(+) RNAs (Life Technologies). The quality of

Figure 2: Comprehensive experimental workflow of the PacBio and MinION sequencing.

the samples was assessed with an Agilent 2100 Bioanalyzer. The samples used had RNA integrity numbers greater than 9.5.

Library preparation for PacBio RSII and Sequel sequencing

The cDNAs were generated from the polyA(+) RNA fractions in accordance with PacBio's recommendations for Iso-Seq method using the Clontech SMARTer PCR cDNA Synthesis Kit and 'No Size Selection' or the 'BluePippin size-selection' protocol (Fig. 2, Table 1). The samples collected at various time points (1, 4, 8, and 12 hours pi) were mixed together for the RSII sequencing; however, different time points (1, 2, 3, 4, 6, and 8 hours pi) were used individually for the production of cDNA libraries for the Sequel method. An rRNA-depleted sample mixture (1, 4, 8, and 12 hours) was converted to cDNA with modified random hexamer primers (Table 3) instead of the SMARTer Kit's oligo(d)T-containing oligo. The amounts of the PCR products were measured by Qubit (Table 2, row B). The detailed library preparation methods are described in our recent publication [23]. Briefly, SMRTbell Template Prep Kit 1.0 was used for SMRTbell library production (the libraries were quantified by Qubit [Table 2, row C]), followed by primer annealing using the DNA Sequencing Reagent Kit 4.0 v2 and polymerase (DNA Polymerase P6) binding for RSII sequencing, whereas the Sequel Sequencing Kit 2.1 and Sequel DNA Polymerase 2.0 were applied for the Sequel platform. Samples were bound to magbeads (MagBead Kit v2) for

loading onto the PacBio instruments. The RSII movie lengths were set for 240 minutes, while 600-minute movies were captured using the Sequel technique. A single movie was recorded for each SMRT Cell. Seventeen RSII SMRT Cells v3 and 8 Sequel SMRT Cells v2 (SMRT Cell 1M) were used for sequencing. The cDNA samples and the SMRTbell templates were quantified (Table 2) by Qubit using Qubit dsDNA HS (High Sensitivity) Assay Kit.

ONT MinION cDNA sequencing

The polyA(+) RNAs were used for cDNA sequencing on the MinION device. We prepared one library from the RNA mixture (RNA samples from the 1, 2, 3, 4, 6, 8, 12, and 16 hours pi); but the various time points were also sequenced individually (Table 4). For the library preparation, we used the ONT 1D strand-switching cDNA by ligation protocol (Version: SSE.9011.v108_revS.18Oct2016), the Ligation Sequencing 1D kit (SQK-LSK108, Oxford Nanopore Technologies), and the NEBNext End repair/da-tailing Module NEB Blunt/TA Ligase Master Mix (New England Biolabs), according to the manufacturers' recommendations. Briefly, 50 ng of the polyA(+)-selected RNA samples were subjected to RT using Poly(T)-containing anchored oligonucleotides [(VN)T20; ordered from Bio Basic, Canada], (Table 3), dNTPs (10 mM, Thermo Scientific), Superscript IV Reverse Transcriptase Kit (Life Technologies), RNase OUT (Life Technologies), and strand-switching oligonucleotides with three

Table 3: Primers sequences used in this study for the reverse transcription reactions.

Sequencing method	Library prep step	Name, availability	Catalog No.	Sequence (5' → 3')
PacBio amplified PolyA	RT	3' SMART CDS primer II A -SMARTer PCR cDNA Synthesis kit (Clontech)	634 925 & 634 926	AAGCAGTGGTATCAACGCAGAGTAC(T) ₃₀ VN
PacBio amplified Random	RT	Custom-made (IDT DNA)	–	AAGCAGTGGTATCAACGCAGAGTACNNNNNNN (G: 37%; C: 37%; A: 13%; T: 13%)
MinION cDNA	RT	Poly(T)-containing anchored primer [(VN)T20 - ONT recommended, custom-made (Bio Basic)	–	5phos/ACTTGCCTGTCTCTATCTTC(T) ₂₀ VN
MinION CAP-Seq	RT	TeloPrime Full-Length cDNA Amplification Kit (Lexogen)	013.08 & 013.24	TCTCAGCGGTTTTTTTTTTTTTTTTTT
MinION dRNA	RT	RT adapter—Direct RNA Sequencing Kit (Oxford Nanopore Technologies)	SQK-RNA001	GAGGCGAGCGGTCAATTTTCTTAAGAGCAAG AAGAAGCCTTTTTTTTTT CGAACTAGAGGACCGTTGGG TTTCCAGGTCAGCACCGTTT AAGAAAGTTGTCGGTGTCTTTGTG
MinION CAP-Seq	test qPCR	D1R fw—custom-made (IDT DNA)	–	CGAACTAGAGGACCGTTGGG
MinION CAP-Seq	test qPCR	D1R rev—custom-made (IDT DNA)	–	TTTCCAGGTCAGCACCGTTT
MinION cDNA barcoded	barcoding	A1/>BC01/(ONT PCR Barcoding Kit 96)	EXP-PBC096	AAGAAAGTTGTCGGTGTCTTTGTG
MinION cDNA barcoded	barcoding	A2/>BC02/(ONT PCR Barcoding Kit 96)	EXP-PBC096	TCGATTCCGTTTGTAGTCGTCTGT
MinION cDNA barcoded	barcoding	A3/>BC03/(ONT PCR Barcoding Kit 96)	EXP-PBC096	GAGTCTTGTGTCCCAGTTACCAGG
MinION cDNA barcoded	barcoding	A4/>BC04/(ONT PCR Barcoding Kit 96)	EXP-PBC096	TTCGATTCTATCGTGTTCCTTA
MinION cDNA barcoded	barcoding	A5/>BC05/(ONT PCR Barcoding Kit 96)	EXP-PBC096	CTTGTCCAGGGTTGTGTAACCTT
MinION cDNA barcoded	barcoding	A6/>BC06/(ONT PCR Barcoding Kit 96)	EXP-PBC096	TTCTCGCAAAGGCAGAAAGTAGTC
MinION cDNA barcoded	barcoding	A7/>BC07/(ONT PCR Barcoding Kit 96)	EXP-PBC096	GTGTTACCGTGGGAATGAATCCTT
PacBio	adapter ligation	PacBio blunt adapter (PacBio Template Prep Kit 1.0)	PN 100-222-300	ATCTCTCTCTTTTCTCTCTCT- CCGTTGTTGTTGTTGAGAGAGAT GGTGCTG
MinION	adapter ligation	5' adapter (ONT Ligation Sequencing 1D kit)	SQK-LSK108	TTAACCT
MinION	adapter ligation	3' adapter (ONT Ligation Sequencing 1D kit)	SQK-LSK108	

The table also contains the sequence of the gene-specific primer pair used for the amplification of D1R gene of VACV, as well as the sequencing adapters and barcodes.

Table 4: Summary of the amount of RNA, cDNA, and library samples used for ONT MinION sequencing.

Library	Starting RNA	Starting RNA amount (ng)	cDNA amount (PCR product, ng)	Library used for sequencing (ng)	Barcode no.	Number of flow cells
1D cDNA	polyA(+) mixed	29	253	65	–	1
1D cDNA	polyA(+) mixed	29	251	48	–	
1D cDNA	polyA(+) 1h	50	117	150	A1	1
1D cDNA	polyA(+) 2h	50	387		A2	
1D cDNA	polyA(+) 3h	50	360		A3	
1D cDNA	polyA(+) 4h	50	180		A4	
1D cDNA	polyA(+) 6h	50	207	300	A5	1
1D cDNA	polyA(+) 8h	50	103		A6	
1D cDNA	polyA(+) 12h	50	130		A7	
dRNA	polyA(+) mixed	60	no PCR	10.2	–	1
Cap-Seq	total RNA (1, 2, 3h)	2 µg	240	240	–	1
Cap-Seq	total RNA (4, 6, 8, 12, 16h)	2 µg	1125	320	–	1

O-methyl-guanine RNA bases (PCR.Sw.mod.3G; ordered from Bio Basic, Canada). First-strand cDNAs were generated at 50°C for 10 minutes of incubation, which was followed by the strand-switching step at 42°C for 10 minutes and a final inactivation step at 80°C for 10 minutes. Double-stranded cDNAs (5 µL from each) were amplified by using KAPA HiFi DNA Polymerase (Kapa Biosystems), Ligation Sequencing Kit Primer Mix (supplied by the 1D Kit), and a Veriti Thermal Cycler (Applied Biosystems).

The initial denaturation was conducted at 95°C for 30 seconds (1 cycle), the denaturation was carried out at 95°C for 15 seconds (15 cycles), the annealing step was set to 62°C for 15 seconds (15 cycles), while the elongation was set to 65°C for 4 minutes (15 cycles). The final extension step was carried out at 65°C for 1 minute. NEBNext End repair/dA-tailing Module (New England Biolabs) and the NEB Blunt/TA Ligase Master Mix (New England Biolabs) were used for end repair and for adapter lig-

Figure 3: Representation of the depth of viral read coverages generated from different long-read sequencing techniques. (A) Circos plot showing the genome-wide transcriptome profile of VACV. The colored boxes represent the genes that belong to different kinetic classes (red: early 1 [early]; green: early 2 [early-late]; yellow: postreplicative [late]; gray: unknown) [13]. Data derived from the five different library preparation and sequencing methods used in this study are shown on the histogram as follows: green: Sequel all data (data from different time points are mixed together); blue: RSII mixed sample; yellow: MinION 1D cDNA mixed sample; orange: MinION Cap-selected mixed sample; black: MinION 1D cDNA barcoded all data (data from different time points are mixed together). (B) Visualization of read coverage on VACV genome at individual time points. Six time points that were sequenced by PacBio Sequel (inner radius) and ONT MinION (outer radius) have been visualized in a segmented circos plot (every segment represents an individual time point). (C) Sashimi plot presentation of the dRNA-seq data across the VACV genome.

ations, respectively. The adapter sequences were provided by the 1D kit. Agencourt AMPure XP magnetic beads (Beckman Coulter) were used to purify the samples following each enzymatic step. The Qubit Fluorometer (Life Technologies Qubit 2.0) and the Qubit (ds)DNA HS Assay Kit were applied to measure the quantity of the libraries. Samples were loaded on R9.4 SpotON Flow Cells, and base calling was performed using Albacore v1.2.6. The PCR amplicons derived from the mixed RNA sample were size-selected manually and then run on Ultrapure Agarose gel (Thermo Fischer Scientific), followed by the isolation of 500 bp+ fragments using the Zymoclean Large Fragment DNA Recovery Kit (Zymo Research). The individually sequenced samples were labeled with barcodes applying a combination of two ONT protocols: first, the 1D protocol was used, but after the first end-prep step, we switched to the 1D PCR barcoding (96) genomic DNA (SQK-LSK108) protocol (version: PBGE96.9015.v108.revS.18Oct2016, updated 25/10/2017), which was then followed by the barcode ligation step using the ONT PCR Barcoding Kit 96 (EXP-PBC096): the barcode adapters (Table 3) were ligated to the end-prepped cDNA samples using the NEB Blunt/TA Ligase Master Mix (New England Biolabs), then they were amplified by PCR with Kapa HiFi DNA Polymerase. The quantities of the libraries were measured by Qubit 2.0 (Table 4).

ONT MinION cDNA-sequencing on cap-selected samples

For more precise identification of the 5'-ends of the full-length transcripts, a Cap-selection method was applied and combined with the ONT 1D cDNA library preparation protocol. The cDNAs were generated from a mixed total RNA sample (containing RNAs from 1, 2, 3, 4, 6, 8, 12, and 16 hours pi; Tables 1 and 4) by using the TeloPrime Full-Length cDNA Amplification Kit (Lexogen). The protocol contains a PCR amplification step. The specificity of the products was checked by qPCR (Rotor-Gene Q). A VACV gene-specific primer (D1R gene, Table 3) and Absolute qPCR SYBR Green Mix (Thermo Fisher Scientific) were used. The amplified PolyA(+) and Cap-selected samples were subjected to the ONT's 1D strand-switching cDNA by a ligation method (ONT Ligation Sequencing 1D kit); they were end-repaired, then ligated to the 1D adapters (NEBNext End repair/dA-tailing Module NEB Blunt/TA Ligase Master Mix).

ONT MinION—dRNA sequencing

In order to avoid the potential PCR biases, the amplification-free direct RNA sequencing (DRS) protocol (Version: DRS.9026.v1.revM.15Dec2016) from the ONTs was applied. The library was prepared from a PolyA(+) mixture of 8 time points (1, 2, 3, 4, 6, 8, 12, and 16 hours pi; Table 4). RNA was mixed with the RT (oligo(dT)-containing T10) adapter (provided by the ONT Direct RNA Sequencing Kit; SQK-RNA001) and

Table 5: Summary statistics of the sequencing reads that mapped to the viral and the host reference genomes from each run.

Run no.	Major specificities of libraries	VACV		<i>C. sabaeus</i>	
		Number of mapped reads ^a	Coverage	Number of mapped reads ^a	Coverage
1	RSII mix no size selection	110	0.7684	512	0.0002
2	RSII mix random primed	31	0.1137	2,905	0.0008
3	RSII mix BluePippin size selection: 0.8–5kb+	23,802	114.55	50,200	0.0167
4	RSII BluePippin size selection: 0.8–2kb	1,283	6.4407	15,206	0.0055
5	RSII BluePippin size selection: 2–3kb	5,029	24.46	8,752	0.0035
6	RSII BluePippin size selection: 3–5kb	20,103	97.639	68,766	0.0217
7	RSII BluePippin size selection: 5kb+	8,848	46.399	16,024	0.0048
8	Sequel 1h	455	1.8081	38,239	0.0145
9	Sequel 2h	527	2.2227	38,255	0.0163
10	Sequel 3h	1,068	4.9862	68,500	0.032
11	Sequel 4h	809	3.0213	42,379	0.018
12	Sequel 4h 2nd	4,522	22.49	233,709	0.109
13	Sequel 6h	3,031	13.401	101,745	0.0499
14	Sequel 8h	5,482	27.619	101,624	0.0548
15	Sequel 8h 2nd	11,628	63.066	63,987	0.0264
16	MinION 1D cDNA Manual size selection: 500bp+	89,778	302.1	293,048	0.0801
17	MinION dRNA	1,259	3.3894	14,757	0.0026
18	MinION Cap-selection	155,876	550.86	327,964	0.059
19	MinION 1D cDNA barcoded 1h	17,048	31.358	69,060	0.0166
20	MinION 1D cDNA barcoded 2h	94,125	147.62	474,008	0.0949
21	MinION 1D cDNA barcoded 3h	22,029	34.865	88,064	0.017
22	MinION 1D cDNA barcoded 4h	41,700	66.981	134,090	0.0249
23	MinION 1D cDNA barcoded 6h	42,082	75.602	106,989	0.0205
24	MinION 1D cDNA barcoded 8h	48,437	101.6	51,071	0.0099
25	MinION 1D cDNA barcoded 12h	57,039	92.483	31,924	0.0028

^aThe difference between the yield of the size-selected and nonsize-selected samples might be caused by the underloading of the SMRT Cell and it is independent from the size-selection step. In some cases, PacBio run results in low output, for which the possible reason is the underloading of the Cells.

T4 DNA ligase (2M U/mL; New England BioLabs). Following a 10-minute incubation, the first-strand cDNAs were generated with SuperScript III Reverse Transcriptase (Life Technologies), according to the DRS protocol, at 50°C for 50 minutes, then at 70°C for 10 minutes in a Veriti Thermal Cycler. Samples were purified by using Agencourt AMPure XP Beads (Beckman Coulter). XP Beads were handled before use with RNase OUT (40 U/μL; Life Technologies; 2 U enzyme/1 μL bead). Washed samples were eluted in Ambion Nuclease-Free Water (Thermo Fisher Scientific). An RMX sequencing adapter was ligated to the samples with NEBNext Quick Ligation Reaction Buffer (New England BioLabs) T4 DNA ligase. Samples were washed with RNase OUT-treated XP beads and Wash Buffer (part of the DRS Kit). Finally, the samples were eluted in 21 μL Elution Buffer (provided by the DRS Kit). The concentrations of the reverse-transcribed and adapter-ligated RNAs were measured using the Qubit 2.0 Fluorometer and Qubit dsDNA HS Assay Kit (Life Technologies). The ONT cDNA libraries, the Cap-selected samples, and the direct RNA libraries were loaded on 3, 2, and 1 ONT R9.4 SpotON Flow Cells for sequencing, respectively. The runs were carried out using MinKNOW. Voltage levels were set and reset in line with the suppliers' recommendations.

Data analysis and visualization

The PacBio RSII reads of insert (ROI) reads were generated using the RS.ReadsOfInsert protocol of the SMRT Analysis v2.3.0, with the following settings: Minimum Full Passes = 1, Minimum Predicted Accuracy = 90, Minimum Length of Reads of

Insert = 1, Maximum Length of Reads of Insert = No Limit. These consensus reads were mapped using GMAP (GMAP, [RRID:SCR.008992](#)) [27] (version 2017–09–30) with the default settings. GMAP was chosen in this work because we found it to be the best long-read aligner in our earlier publications [18–24]. Others have also found that GMAP produces the best alignment results [e.g., 28]. The ROIs from the Sequel data were created using SMRT Link5.0.1.9585. ONT's Albacore software v.2.0.1 (Albacore, [RRID:SCR.015897](#)) was then applied for the MinION base calling. This base caller is able to identify the nucleotide sequences directly from raw sequencing data. The reads were aligned with the GMAP program using the same setting as described above. The raw reads were aligned to the reference genome of the virus (LT966077.1) and the host cell (*Chlorocebus sabaeus*) (GenBank assembly accession: GCA.000409795.2 [latest]; RefSeq assembly accession: GCF.000409795.2 [latest]). In-house routines were used to acquire the quality information presented in this data note. The code has been archived on Github [29]. Bedtools genomecov software (BEDTools, [RRID:SCR.006646](#)) [30] was used to generate coverage files with the following parameters: -split—ibam. The output bed files from cDNA sequencing were visualized by Circos plot [31] (Fig. 3), while the low-coverage dRNA-seq data was shown using the Integrative Genomics Viewer (IGV, [RRID:SCR.011793](#)) [32].

Data Summary

The raw sequencing reads were mapped to both the VACV reference genome and to the host genome. In this study, we gen-

Figure 4: Polar plot representation of the percentages of virus-host read counts.

erated full-length transcripts of VACV and the CV-1 cells, yielding about 3.17 Gb of mapped sequencing data. Sequencing on the RSII and Sequel platforms yielded 86,728 and 850,803 ROIs aligned to the viral and the host genomes, respectively. The utilized nanopore-based cDNA sequencing approaches resulted in 413,497 VACV-specific reads (Table 5, Additional File 2), while we obtained 155,876 reads from the Cap-selected samples. The different MinION sequencing methods yielded 1,590,975 reads that mapped to the host genome. The ratio of viral transcripts is 21.9% on average in our samples. The exact ratio is dependent on the titer of the virus used for the infection, as well as on the stage of the viral life cycle at the examination period. The sequencing method affects the ratio of read counts between the virus and host cell; e.g., the MinION 1D-Seq method yields a higher amount of shorter reads compared to the PacBio Sequel technique. The VACV transcripts are relatively short compared to the host or to other large DNA viruses (such as herpesviruses and baculoviruses), which is assumed to result in the relatively high ratio of viral reads compared to the host reads in the MinION samples (Fig. 4, Additional File 3). The ratio of the viral reads in the RSII samples (with or without size selection) is lower than that of the MinION samples; however, this ratio is significantly higher than in the Sequel samples (without size selection). The

Sequel platform generated the same or longer read length than the RSII size-selected samples (Fig. 5, Additional File 2). In contrast to the Sequel and MinION samples where individual time points of viral infection were analyzed, we used mixed time-point samples for the RSII sequencing, therefore the comparison of the obtained results is not possible. The PacBio MagBead loading method and the overall yield of the given runs can also account for the generation of varying ratio of viral reads in the different samples.

The average lengths of ROIs aligning to the VACV genome were 1,098 bp for PacBio RSII and 1,157 bp for the Sequel. The MinION average read lengths were as follows: 557 bp for ONT barcoded cDNA sequencing, 792 bp for the cDNA-Seq, and 965 bp for the Cap-selected samples (Fig. 5, Additional File 2). The average read length produced by dRNA sequencing was 537 bp. It should be noted that the library preparation and size-selection methods resulted in different samples in terms of length; all library preparation methods resulted in longer average read length aligning to the host genome than to the viral genome (Fig. 5, Additional File 2). We have compared the average aligned read-length of cellular transcripts obtained in this and in other studies [18, 19, 21, 24, 33–35] in Fig. 6.

Figure 5: Box plot presentation of the average of aligned read lengths obtained from the applied sequencing methods. The reads were mapped to the VACV and to the host genome, and the average lengths were plotted with the standard deviation values.

The various sample preparation and sequencing techniques produced different read lengths, read numbers and precision, as well as different artifacts. There is a relatively large difference between the PacBio and ONT sequencing approaches concerning the quality of the sequencing reads; the PacBio technique produces fewer mismatches, insertions and deletions (INDELs) than nanopore sequencing. The various sequencing platforms recommend different cDNA production kits, which contain different enzymes and primers for both the RT and PCR. The various primers and library preparation conditions could produce artifacts; however, these can be easily filtered out if we compare the results of different methods. The PacBio MagBead loading selectively eliminates the short fragments (<1,000 bp). On the one hand, removal of incomplete cDNAs can be advantageous, at the same time, it is unfavorable, as we are unable to detect the shorter transcripts and RNA isoforms. Our data demonstrate that the ONT MinION sequencing resulted in higher error rates for both INDELs and mismatches in comparison to the PacBio systems (Additional File 2). The composition of the errors of the three platforms (RSII, Sequel, and MinION) and the various library preparation techniques (e.g., dRNA-seq, Cap-Seq, etc.) are different. Mismatches are the most common errors in ONT cDNA-Seq, which is consistent with others' data [36]. In agreement with the previously published data [36], our results also indicate that insertions are the least frequent errors in ONT MinION sequencing. In accordance with others' results [37], our dRNA reads have higher deletion error rates than either of the

cDNA datasets and lower than those of the ONT cDNA-Seq samples, which might be the result of the lower coverage of the dRNA-seq. In contrast to others' results [36], deletions are the major errors in our PacBio RSII dataset. The quality of the Sequel dataset shows "coverage-specificity"; mismatches are the major errors in the lower-coverage samples, which complies with others' data [36], while contrary to the same report in that the insertions are more frequent in the higher coverage samples in our dataset. The RSII and the Sequel platforms produce the same error rate. Conversely, our data show a somewhat higher error rate for the Sequel, which might be the result of the different library preparation approaches. In sum, the absolute error rate of both PacBio platforms is low, while the higher ONT error rate is "compensated" by the higher coverage. It is worth mentioning that read quality is not essential for transcriptome analysis if well-annotated genomes are available.

Our transcriptomic survey yielded extremely high coverage across the viral genome (Fig. 3): 290.1 fold for the RSII, 138.6 fold for the Sequel, 550.5 fold for the bar-coded MinION cDNA-Seq, 550.8 fold for the Cap-selected samples, and 302.1 fold for the cDNA sequencing (more detailed information, including quality information, is available in Table 5 and Additional Files 2 and 4). Our data show that the entire VACV genome is transcriptionally active, generating RNAs from both DNA strands. Our dataset also contains 1.56 Gb of raw data from Sequel sequencing, as well as from MinION dRNA-seq.

Figure 6: Comparison of the read lengths mapped to the host genome of this and other studies. It must be noted here that the analyzed cell lines are from different organisms and/or they were infected with different viruses using different incubation time points. (A–C) RSII and Sequel platforms provide relatively fixed read length (RSII: 800–1,400 bp; Sequel: 1,050–1,500 bp). The average read length of CV-1 samples is longer than those of the MRC-5 in the RSII PolyA-sequencing; however, the opposite result has been obtained with the random-primed RSII and the Sequel PolyA-Seq. (D–G) The MinION platform produces greater length variance (250–1,200 bp) except for the Cap-Seq approach, which shows a very small difference between the read lengths among the four different cell lines. Cell lines: African green monkey kidney fibroblast cells (CV-1) infected with VACV or herpes simplex virus type 1 (HSV-1); human lung fibroblast cells (MRC-5) infected with human cytomegalovirus (HCMV) or varicella-zoster virus (VZV); porcine kidney 15 (PK-15) cell line infected with pseudorabies virus (PRV); and Sf9 insect cell line infected with the baculovirus *Autographa californica* multiple nucleopolyhedrovirus (AcMNPV).

The read-length distributions for the dataset are shown in Fig. 7 (reads mapped to the VACV genome), as well as in Figs. 8 and 9 (data aligned to the VACV and to the host genome). Detailed information is available in Additional File 5.

The read counts aligned to the mRNAs have been calculated (Additional File 6). Most of the host-specific reads align to the coding region in this dataset (the values vary between 43% and 87% based on the read counts and between 35% and 85% if we compare the number of nucleotides).

We mapped the raw data to the VACV and to the host mRNAs. Ten viral and 10 host genes that are expressed at every examined time point were chosen for a heat map analysis (Fig. 10). Only the full-length transcripts were calculated for the analysis. A read was considered full length if it contained the 5' and 3' adapter sequences as well as the polyA-tail preceding the 3' adapter. Porechop software v.0.2.3 [38] was used to identify the 5' and 3' adapters. Reads lacking an adapter on either the 3'- or the 5'-end, or on both ends, and reads with 5' or 3' adapters on both ends were considered as non-full length reads. Reads that were categorized as full length were mapped to the reference sequences by GMAP (Additional File 7). A plus/minus 20 bp range was set to and from the previously annotated transcription start

and end sites, and the reads that belonged to this category were used for the analysis. The relative expression ratios of each examined transcript were calculated by dividing the obtained full-length read count of the transcript by the total read count in the given sample.

Conclusions and Reuse Potential

The present study generated data using state-of-art sequencing technologies (PacBio RSII and Sequel, as well as the ONT MinION platforms, applying a new protocol for barcoding the samples). These data allow a time-course look at the full-length transcriptome of VACV, as well as the CV-1 host cell line.

The dataset was primarily produced for the dynamic characterization of the VACV transcriptome. Another aim was to generate a deep coverage long-read dataset for the analysis of the different transcript isoforms, including length (5'-ends and 3'-ends) variants; mono-, bi-, and polycistronic transcripts; and also to define full-length transcripts produced by the various viral genes. This dataset is useful in understanding the complexity of the genetic regulation of VACV. The provided dataset can also

Figure 7: Bar chart representations of read-lengths distributions (depicted for 500 bp long bins, at log₁₀ scale). **(A) Sequel.** Most of the reads fall within the range of 501–1,000 bp at each time point. There are no substantial length differences between the samples within the first four intervals; however, the earlier time points disappear later; only the samples from 4, 6, 8, and 12 hours contain reads longer than 4,000 bp, while reads longer than 4,500 bp could be detected only within the 8-hour post-infection samples. **(B) MinION.** Most of the reads fall to the shortest range (1–500 bp), and very few reads are longer than 3,501 bp from the 4-, 6-, 8-, and 12-hour samples. **(C) RSII size-selected samples.** The shortest and longest reads are overrepresented in the 0.8–5 kb+ sample. The 0.8–2 kb sample represents the shortest read population; no reads are longer than 2 kb. There is no significant difference between the samples at the size ranges 2–3 kb, 3–5 kb, and 5 kb+; the largest amount of transcripts is within the 501–1,000 bp range. The reason for the relatively low read count within the higher size ranges may be that the length of the VACV transcripts are much shorter than, e.g., herpesviruses or baculoviruses. **(D) RSII no size selected PolyA vs. random-primed samples.** The shorter reads are overrepresented in the random primed sample (< 1,000 bp), while most of the reads from the PolyA-Seq sample fall within the 1,001–2,000 bp interval (this is the typical average read lengths of the PacBio RSII without size selection).

be used to investigate the effect of the viral infection on the gene expression of the host.

The provided binary alignment files contain reads already aligned to the VACV and to the host genome. These aligned reads can be further analyzed by comparing them to the results of various long-read aligners (e.g., BLASR [39], NGMLR [40], Min-

imap2 [41]) and bioinformatics tools (e.g., samtools [42] or bedtools [43]). Other long-read sequencing programs or pipelines (e.g., SQANTI [44]) can be tested using this dataset.

These data can be visualized by using different programs such as the Geneious [45], Artemis [46], or IGV [32]. Data can be useful for testing novel bioinformatics pipelines or to im-

Figure 8: Comparison of the read-length distributions between the VACV and the host (*Chlorocebus sabaeus*) transcripts within the utilized non-size-selected library preparation methods. Mapped read lengths are expressed in base pairs, and the distribution is shown for 100 bp long bins. The x-axis is only presented up to 4,000 base pairs, even though the longest read that was detected was up to as long as 9,000 bp; 99.86% of the alignments fall into this range. In most cases, the PacBio platforms generated longer reads than the ONT methods.

prove those already available. The files contain terminal polyA sequences as well as the 5' and 3' adapter sequences, which can be used to determine the orientations of the reads. The dataset contains the raw dataset from dRNA sequencing (fast5.tar.gz), which can be further analyzed by using the Tombo software package [47], which enables the detection and visualization of modified nucleotides, such as the 6-methyladenine, the most common internal mRNA modification described in eukaryotes [48–50] as well as in viruses [51–53], or the 5-methylcytosine,

which is another abundant modification recently confirmed in mRNA [54–57]. To the best of our knowledge, these modifications have not yet been shown in the *Poxviridae* family. The raw data provided from PacBio Sequel sequencing can be used to improve existing base caller algorithms or potentially to develop novel algorithms. Furthermore, the data contain the full set of quality values and kinetic measurements.

This dataset can be used to identify novel VACV and CV-1 transcripts and RNA isoforms including splice variants of the

Figure 9: Illustration of the read-length distributions of the VACV and the host (*Chlorocebus sabaeus*) transcripts within the utilized size-selected library preparation methods. Aligned read lengths are shown in base pairs per 100 bp intervals. The distribution patterns of the viral and host cell reads resemble one another in the size-selected RSII samples, especially in the 2–3 kb, 3–5 kb, and 5 kb+ samples. Samples reach their highest peaks around 1,000 bp; however, the peak shifts to the right according to the size selection. There is a significant peak in every sample within the shortest range (1–100 bp) in the host reads. The effect of size selection is the most dramatic in the MinION virus sample; the read counts drastically increase beyond 200 bp.

host transcripts, TSS and TES variants, as well as polycistronic transcripts of the virus and the host in order to examine the effect of VACV infection on the host gene expression at the different stages of viral life cycle, as well as for the comparison of the quality and length of the sequencing reads derived from different sequencing platforms. The various library preparation methods can also be compared with one another. The data provided could be used to better understanding the logic of gene expression control of Poxviruses and can also be used to design gene expression vectors.

Availability of source code and requirements

Project home page: <https://github.com/Szunyike/SAM-Statistic-2018>

Operating system: Windows

Programming language: VB.NET

Other requirements: NET framework

License: GPL v3

Availability of supporting data

All of the presented here data were deposited in the European Nucleotide Archive under the accession number PRJEB26434

(Characterization of the Vaccinia virus transcriptome) and PRJEB26430 (Dynamic characterization of the Vaccinia virus transcriptome). Alignments and other data are also available from the GigaScience GigaDB repository [58] (Additional file 8).

Additional files

Additional file 1: Summary table of the reagents and chemistries used for the sequencing

Additional file 2: Summary table of the virus-host ratios within the single runs

Additional file 3: Summary statistics of the sequencing reads which mapped to the viral genome (A) and to the host reference genome (B) from each run. SE: standard error. * The difference between the yield of the size-selected and non-size-selected samples might be caused by the underloading of the SMRT Cell and it is independent from the size-selection step. In some cases, PacBio run results in low output, for which the possible reason is the underloading of the Cells.

Additional file 4: Summary statistics of the viral and host reads from each run.

Additional file 5: Read-length distribution is depicted for 100bp long intervals.

Figure 10: Heat maps depict the relative expression ratios (the proportion of read counts to the total number of reads at a given time point; the values are expressed as percentages [%]). White indicates the lowest relative expression values, while the dark red (VACV) or dark blue (host) represents the highest values. The dynamic profiles of the examined viral genes differ in the two datasets derived from the different sequencing methods. This alteration can be explained by the different read-size preferences of the two methods (however, further data analysis is required for accurate kinetic findings). According to a previous study [13], the examined viral genes belong to the early kinetic class. This is evidenced by the fact that the relative expression values are higher at the early time points, especially in the MinION dataset. The majority of the examined cellular genes show constant expression level (mainly in MinION data), most of them belong to the housekeeping genes: https://hpcwebapps.cit.nih.gov/ESBL/Database/NephronRNAseq/Housekeeping_Genes.html. The expression patterns of the following genes were analyzed: VACV; 1. c11r: Epidermal growth factor-like protein (EGF-like protein); 2. c7l: Interferon antagonist C7 (host range protein 2); n2l: protein N2; m2l: protein M2; e3l: protein E3; h5r: Late transcription elongation factor H5; b8r: Soluble interferon gamma receptor B8; b19r: Ankyrin repeat protein B19; vacwr.4: Truncated CrmB protein. HOST; CST3: *C. sabaeus* cystatin C (XM.007961908.1); PSAP: *C. sabaeus* prosaposin, transcript variant X1 (XM.007963126.1); PKM: *C. sabaeus* pyruvate kinase PKM (LOC103217002) (XM.007964863.1); GAPDH: *C. sabaeus* glyceraldehyde-3-phosphate dehydrogenase (XM.007967342.1); ENO1: *C. sabaeus* enolase 1, (alpha), (XM.007980661.1); FTL: *C. sabaeus* ferritin, light polypeptide (XM.007997480.1); ATP5B: *C. sabaeus* ATP synthase, H+ transporting, mitochondrial F1 complex, beta polypeptide (XM.008003700.1); ACTG1: *C. sabaeus* actin, gamma 1 (XM.008013242.1); EEF1A1: *C. sabaeus* eukaryotic translation elongation factor 1 alpha 1 transcript variant X1 (XM.008013483.1); 60S: *C. sabaeus* 60S ribosomal protein L3-like (LOC103247496), mRNA (XM.008019639.1)].

Additional file 6: Statistics of the read counts mapped to the host genome versus the host mRNAs.

Additional file 7: GenBank accession numbers and URLs of the VACV and host genes selected for heatmap expression analysis.

Additional file 8: Correspondence between the file names of alignments deposited in ENA and the names that are used in this manuscript. The table also contains the ENA accession numbers of the study, of the experiments, samples, as well as the runs.

Abbreviations

m5C: 5-methyl cytosine; m6A: 6-methyl adenine; CV-1: African green monkey (*Chlorocebus sabaeus*) kidney fibroblast cells; ATCC: American Type Culture Collection; CAGE: cap analysis of gene expression; Cap-Seq: Cap-selection; dRNA: direct RNA; E: early; FBS: fetal bovine serum; I: intermediate; IE: immediate-early; IN-DEL: insertions and deletions; Iso-Seq: Isoform sequencing; L: late; LRS: long-read sequencing; ONT: Oxford Nanopore Technologies; ORF: open reading frame; PacBio: Pacific Biosciences; pi: post-infection; RNA-seq: RNA sequencing; ROI: RPMI 1640:

1640 Roswell Park Memorial Institute medium; RSII: Real-Time Sequencer II; SRS: short-read sequencing; TES: transcription end site; TSS: transcription start site; VACV: vaccinia virus;

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by the National Research, Development and Innovation Office (NKFIH OTKA K 128247) and by the Swiss-Hungarian Cooperation Programme (SH/7/2/8) to Zsolt Boldogkői; by the National Research, Development and Innovation Office (NKFIH OTKA FK 128252), by the Eötvös Scholarship of the Hungarian State and by a Bolyai János Scholarship of the Hungarian Academy of Sciences to Dóra Tombácz. The project was also supported by the National Institutes of Health, Centers of Excellence in Genomic Science Center for Personal Dynamic Regulomes (5P50HG00773502) to Michael Snyder

Author contributions

D.T., D.B., M.S., and Z.B. conceived and designed the experiments. D.B. propagated the cells and viruses. D.T. and I.P. prepared RNA samples and generated cDNAs. D.T. prepared the sequencing libraries and performed the PacBio and ONT sequencing. D.T., A.S., I.P., and Z.B. analyzed the data. D.T. and Z.B. wrote the manuscript. Z.B. supervised the project. All authors read and approved the final version of the manuscript.

References

- Yang Z, Cao S, Martens CA, et al. Deciphering poxvirus gene expression by RNA sequencing and ribosome profiling. *J Virol* 2015;**89**(13):6874–86.
- Pauli G, Blümel J, Burger R, et al. Orthopox viruses: infections in humans. *Transfus Med Hemother* 2010;**37**(6):351–64.
- Schrack L, Damaso CR, Esparza J, et al. An early American smallpox vaccine based on horsepox. *N Engl J Med* 2017;**377**:1491–2.
- Wyatt LS, Xiao W, Americo JL, et al. Novel nonreplicating vaccinia virus vector enhances expression of heterologous genes and suppresses synthesis of endogenous viral proteins. *mBio* 2017;**8**(3):e00790–17.
- Broyles SS. Vaccinia virus transcription. *J Gen Virol* 2003;**84**:2293–303.
- Schramm B, Locker JK. Cytoplasmic organization of POXvirus DNA replication. *Traffic* 2005;**6**:839–46.
- Assarsson E, Greenbaum JA, Sundström M, et al. Kinetic analysis of a complete poxvirus transcriptome reveals an immediate-early class of genes. *Proc Natl Acad Sci U S A* 2008;**105**(6):2140–5.
- Davison AJ, Moss B. Structure of vaccinia virus early promoters. *J Mol Biol* 1989;**210**(4):749–69.
- Davison AJ, Moss B. Structure of vaccinia virus late promoters. *J Mol Biol* 1989;**210**(4):771–84.
- Baldick CJ, Jr, Keck JG, Moss B. Mutational analysis of the core, spacer, and initiator regions of vaccinia virus intermediate-class promoters. *J Virol* 1992;**66**:4710–9.
- Broyles SS, Moss B. Homology between RNA polymerases of poxviruses, prokaryotes, and eukaryotes: nucleotide sequence and transcriptional analysis of vaccinia virus genes encoding 147-kDa and 22-kDa subunits. *Proc Natl Acad Sci U S A* 1986;**83**(10):3141–5.
- Witteck R, Cooper JA, Barbosa E, et al. Expression of the vaccinia virus genome: analysis and mapping of mRNAs encoded within the inverted terminal repetition. *Cell* 1980;**21**(2):487–93.
- Yang Z, Bruno DP, Martens CA, et al. Simultaneous high-resolution analysis of vaccinia virus and host cell transcriptomes by deep RNA sequencing. *PNAS* 2010;**107**(25):11513–8.
- Yang Z, Bruno DP, Martens CA, et al. Genome-wide analysis of the 5' and 3' ends of vaccinia virus early mRNAs delineates regulatory sequences of annotated and anomalous transcripts. *J Virol* 2011;**85**(12):5897–909.
- Yang Z, Maruri-Avidal L, Sisler J, et al. Cascade regulation of vaccinia virus gene expression is modulated by multistage promoters. *Virology* 2013;**447**(1–2):213–20.
- Rubins KH, Hensley LE, Bell GW, et al. Comparative analysis of viral gene expression programs during poxvirus infection: a transcriptional map of the vaccinia and monkey pox genomes. *PLoS One* 2008;**3**(7):e2628.
- Yang Z, Martens CA, Bruno DP, et al. Pervasive initiation and 3' end formation of poxvirus post-replicative RNAs. *J Biol Chem* 2012;**287**:31050–60.
- Tombácz D, Csabai Z, Oláh P, et al. Full-length isoform sequencing reveals novel transcripts and substantial transcriptional overlaps in a herpesvirus. *PLoS One* 2016;**11**(9):e0162868.
- Tombácz D, Csabai Z, Szűcs A, 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.
- Balázs Z, Tombácz D, Szűcs A, et al. Long-read sequencing of human cytomegalovirus transcriptome reveals rna isoforms carrying distinct coding potentials. *Sci Rep* 2017;**7**(1):15989.
- Balázs Z, Tombácz D, Szűcs A, et al. Long-read sequencing of the human cytomegalovirus transcriptome with the Pacific Biosciences RSII platform. *Sci Data* 2017;**4**:170194.
- Moldován N, Tombácz D, Szűcs A, et al. Multi-platform sequencing approach reveals a novel transcriptome profile in pseudorabies virus. *Front Microbiol* 2018;**8**:2708.
- Tombácz D, Sharon D, Szűcs A, et al. Transcriptome-wide survey of pseudorabies virus using next- and third-generation sequencing platforms. *Sci Data* 2018;**5**:180119.
- Moldován N, Tombácz D, Szűcs A, et al. Third-generation sequencing reveals extensive polycistronism and transcriptional overlapping in a baculovirus. *Sci Rep* 2018;**8**(1):8604.
- Yang L, Duff MO, Graveley BR, et al. Genome-wide characterization of non-polyadenylated RNAs. *Genome Biol* 2011;**12**(2):R16.
- Zhang Y, Yang L, Chen LL. Life without A tail: new formats of long noncoding RNAs. *Int J Biochem. Cell Biol* 2014;**54**:338–49.
- Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 2005;**21**(9):1859–75.
- Križanovic K, Echchiki A, Roux J, et al. Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics* 2018;**34**(5):748–54.
- Long-read sequencing data statistics. <https://github.com/Sunzyike/SAM-Statistic-2018>. Accessed: Sept. 2018.
- Quinlan AR. BEDTools: the Swiss-Army tool for genome feature analysis. *Curr Protoc Bioinformatics* 2014;**47**:11.12.1–34.
- Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. *Genome Res* 2009;**19**(9):1639–45.
- Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol* 2011;**29**:24–26.
- Prazsak I, Moldován N, Tombacz D, et al. Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus. *bioRxiv* 2018;399048; doi: <https://doi.org/10.1101/399048>.
- Balázs Z, Tombácz D, Szűcs A, et al. Dual platform long-read RNA-sequencing dataset of the human cytomegalovirus lytic transcriptome. *Front Genet* 2018;**9**:432.
- Tombácz D, Prazsák I, Moldován N, et al. Lytic transcriptome dataset of varicella zoster virus generated by long-read sequencing. *Front Genet* 2018;**9**:460.
- Weirather JL, de Cesare M, Wang Ye, et al. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. Version 2. *F1000Res* 2017;**6**:100.
- Garalde DR, Snell EA, Jachimowicz D, et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* 2018;**15**:201–6.
- <https://github.com/rrwick/Porechop>. GitHub. porechop v2.3 released on October 20, 2017.

39. Chaisson M, Tesler G. Mapping single molecule sequencing reads using Basic Local Alignment with Successive Refinement (BLASR): theory and application. *BMC Bioinformatics* 2012;13:238.
40. Sedlazeck FJ, Rescheneder P, Smolka M, et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* 2018;15:461–8.
41. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34(18):3094–100.
42. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25(16):2078–9.
43. Quinlan AR, Hall I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26(6):841–2.
44. Tardaguilla M, de la Fuente L, Marti C, et al. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res* 2018;28:396–411.
45. Kearse M, Moir R, Wilson A, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 2012;28(12):1647–9.
46. Rutherford K, Parkhill J, Crook J, et al. Artemis: sequence visualization and annotation. *Bioinformatics* 2010;16(10):944–5.
47. Stoiber MH, Quick J, Egan R, et al. De novo Identification of DNA modifications enabled by genome-guided nanopore signal processing. *bioRxiv* 2017;094672. doi: <https://doi.org/10.1101/094672>.
48. Desrosiers R, Friderici K, Rottman F. Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells. *Proc Natl Acad Sci U S A* 1974;71:3971–5.
49. Dominissini D, Moshitch-Moshkovitz S, Schwartz S, et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 2012;485(7397):201–6.
50. Liu J, Jia G. Methylation modifications in eukaryotic messenger RNA. *J Genet Genomics* 2014;41(1):21–33.
51. Fengchun Ye. RNA N6-adenosine methylation (m6A) steers epitranscriptomic control of herpesvirus replication. *Inflamm Cell Signal* 2017;4(3):e1604.
52. Kennedy EM, Courtney DG, Tsai K, et al. Viral epitranscriptomics. *J Virol* 2017;91(9):pii: e02263–16.
53. Tsai K, Courtney DG, Cullen BR. Addition of m6A to SV40 late mRNAs enhances viral structural gene expression and replication. *PLoS Pathog* 2018;14(2):e1006919.
54. Edelheit S, Schwartz S, Mumbach MR, et al. Transcriptome-wide mapping of 5-methylcytidine RNA modifications in bacteria, archaea, and yeast reveals m5C within archaeal mRNAs. *PLoS Genet* 2013;9(6):e1003602.
55. Khoddami V, Cairns BR. Identification of direct targets and modified bases of RNA cytosine methyltransferases. *Nat Biotechnol* 2013;31(5):458–64.
56. Amort T, Rieder D, Wille A, et al. Distinct 5-methylcytosine profiles in polyA RNA from mouse embryonic stem cells and brain. *Genome Biol* 2017;18(1):1.
57. Hussain S, Aleksic J, Blanco S, et al. Characterizing 5-methylcytosine in the mammalian epitranscriptome. *Genome Biol* 2013;14(11):215.
58. Tombácz D, Prazsák I, Szűcs A, et al. Supporting data for “Dynamic transcriptome profiling dataset of vaccinia virus obtained from long-read sequencing techniques.” *GigaScience Database* 2018 <http://dx.doi.org/10.5524/100525>.

Lytic Transcriptome Dataset of Varicella Zoster Virus Generated by Long-Read Sequencing

Dóra Tombácz, István Prazsák, Norbert Moldován, Attila Szűcs and Zsolt Boldogkői*

Department of Medical Biology, Faculty of Medicine, University of Szeged, Szeged, Hungary

Keywords: varicella zoster virus, long-read sequencing, transcriptome, full-length sequencing, nanopore-based sequencing, Oxford Nanopore Technologies

INTRODUCTION

Varicella zoster virus (VZV) belongs to the *Alphaherpesvirinae* subfamily of the *Herpesviridae* family. It is the etiological agent of chickenpox (varicella) caused by primary infection and shingles (zoster), which is due to reactivation of the virus from latency (Kennedy, 2002). Many countries have adopted recommendations for routine immunization of children and susceptible adults against VZV. The VZV virion is composed of an icosahedral nucleocapsid surrounded by a tegument layer, which is covered by an envelope derived from the host cell membrane with incorporated viral glycoproteins (Maresova et al., 2005). The genome of VZV consists of a linear double-stranded DNA molecule and is approximately 125 kbp in size, which contains more than 70 annotated open reading frames (ORFs) (Tyler et al., 2007). The transcription of the virus is strictly regulated by cascade-like processes. First, the immediate-early (IE) transcripts are expressed, which is then followed by the expression of the early (E), and then the late (L) kinetic classes of transcripts (Reichelt et al., 2009). The IE ORF62 gene of VZV encodes the major transactivator, which controls the expression of other viral genes. The viral E genes encode proteins that are used in DNA replication, while L genes code for the structural elements of the virus.

High-throughput short-read sequencing (SRS) techniques have revolutionized transcriptome research (Delseny et al., 2010). These techniques have also been utilized in the investigation of herpesvirus gene expression (e.g., Chambers et al., 1999; Ebrahimi et al., 2003; Baird et al., 2014; Oláh et al., 2015). However, the SRS approach has severe limitations in comparison to long-read sequencing (LRS), including Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) platforms. LRS techniques have been used before in transcriptome studies of the herpesviruses (O'Grady et al., 2016; Tombácz et al., 2016, 2017; Balázs et al., 2017a,b; Moldován et al., 2018). These studies uncovered a very complex transcriptome, which included the identification of a large number of novel RNA molecules and transcript isoforms (Tombácz et al., 2015, 2017; Balázs et al., 2017a). Moreover, an extended meshwork of overlaps between the transcripts was also detected by these studies (Tombácz et al., 2016; Moldován et al., 2018).

The presented data report is aimed toward providing a new, comprehensive transcript catalog of VZV using an LRS approach for the first time. In this study, we applied the ONT MinION device and various full-length cDNA sequencing protocols that capture the entire poly(A)-transcriptome of VZV.

OPEN ACCESS

Edited by:

Graziano Pesole,
Università degli Studi di Bari, Italy

Reviewed by:

Hyouta Himeno,
Hiroaki University, Japan
Mariana Boroni,
Instituto Nacional de Câncer (INCA),
Brazil
Matteo Chiara,
Università degli Studi di Milano, Italy

*Correspondence:

Zsolt Boldogkői
boldogkoi.zsolt@med.u-szeged.hu

Specialty section:

This article was submitted to
RNA,
a section of the journal
Frontiers in Genetics

Received: 30 April 2018

Accepted: 20 September 2018

Published: 16 October 2018

Citation:

Tombácz D, Prazsák I, Moldován N,
Szűcs A and Boldogkői Z (2018) Lytic
Transcriptome Dataset of Varicella
Zoster Virus Generated by Long-Read
Sequencing. *Front. Genet.* 9:460.
doi: 10.3389/fgene.2018.00460

VALUE OF THE DATA

1. Varicella zoster virus (VZV) is world-wide distributed human pathogenic alphaherpesvirus.
2. No long-read sequencing (LRS) transcriptome data from VZV has been published thus far. Here, we provide a dataset on the lytic polyadenylated transcriptome of the VZV generated by LRS from Oxford Nanopore Technologies.
3. LRS approaches have been reported to be superior to other methods, including short-read sequencing in the detection of embedded RNA molecules, polycistronic transcripts and transcriptional overlaps, as well as in distinguishing between RNA isoforms including splice and transcript end variants. These data will be useful for analyzing the complexity of VZV transcriptome, for identifying novel transcripts and for comparing the ONT platform with other cDNA sequencing approaches.

DATA

In this work, the ONT MinION sequencing technique was used for the analysis of the genome-wide expression of VZV genes using the 1D cDNA sequencing protocol. For the detection of full-length transcripts, a modified version of this approach was also carried out, starting with Cap-selection of the RNA samples, using a so called “all-in-one” Cap-selection protocol, but this technique produced very short average read lengths (Table 1). We obtained the same poor result with pseudorabies virus (PRV; Moldován et al., 2018; Tombácz et al., 2018) and Herpes simplex virus–1 (HSV; submitted). In contrast, Cap-selection performed very well in the analysis of the transcriptome of a baculovirus (unpublished) and vaccinia virus (VACV, unpublished). From these results, it has become apparent that not the AT-content (VACV: 66%, VZV: 54%, HCMV: 42%, HSV: 32%, PRV: 27%) accounts for the short read length in the Cap-selected samples, but rather an unknown factor that prevents the completion of reverse transcription in only the alphaherpesviruses.

ONT 1D cDNA sequencing yielded 57,888 VZV-specific sequencing reads with an average coverage of 649, while the sequencing on Cap-selected samples resulted in 509,531 reads (1,253-fold coverage) mapped to the VZV genome (NC_001348.1). The average read-lengths for the cDNA sequencing and for the Cap-sequencing were 1,470 and 427bp, respectively (Table 1, Figure 1A).

ONT sequencing is able to read full-length RNA molecules, but it falls short in accuracy. due to its high-throughput workflow. Additionally, we used barcoding for better identification of the transcripts' ends. ONT sequencing is afflicted by sample degradation, which in general can be eliminated by using Cap-selection. Reads were sorted by the Albacore v. 2.0.1 according to their q-score in two categories: pass and fail. Only the reads belong to the passed category were used in further analysis. Non-virus specific sequencing reads were filtered by mapping to the above mentioned viral genome using GMAP.

TABLE 1 | Summary of the obtained reads and read qualities mapped to the VZV genome (Unique: duplicates have been removed; Duplicates: duplicate reads are included in the statistics).

	1D cDNA	CAP seq
Reads ^A	57888	509531
Average ^B	1470,14	426,92
Median ^C	1033	258
Average ^D	1399,43	307,11
SD ^E	993,11	148,02
Coverage ^F	648,68	1253,03
Del ^G	3,74	4,16
SD ^H	1,74	2,18
Ins ^I	3,48	2,63
SD ^J	1,67	2,12
MM ^K	5,41	4,75
SD ^L	2,14	2,4

^Anumber of mapped reads; ^Baverage of read lengths; ^Cmedian of the aligned reads; ^Daverage of the aligned reads; ^ESD of average; ^Fread coverage across the VZV genome; ^Gpercentage of deletions; ^Hstandard deviations; ^Ipercentage of insertions; ^Jstandard deviations; ^Kpercentage of mismatches; ^Lstandard deviation values of mismatches.

MATERIALS AND METHODS

Cell Culture and Virus Growth

Human primary embryonic lung fibroblast cell line (MRC5) was obtained from the American Type Culture Collection and grown in Dulbecco's modified Eagle's minimal essential medium (Sigma-Aldrich, St. Louis, MO, USA) supplemented with 1% of an antibiotic/antimycotic (AB/AM) solution (Lonza, Walkersville, MD, USA) and 10% fetal bovine serum (GE Healthcare Life Sciences, Logan, UT, USA) at 37°C in a 5% CO₂ atmosphere. The live attenuated OKA/Merck strain of varicella zoster virus (VZV) (MSD Pharma Hungary Kft., Budapest, Hungary) was cultured at 37°C in MRC5 cell line, and the cells were harvested by trypsinization, when the monolayers had displayed near 80% specific cytopathic changes. For sub-passage of the virus or for experiments, the VZV-infected whole cells were used to inoculate MRC5 cultures previously grown to full confluence at a ratio of one infected cell to ten uninfected cells. The cultures were then incubated at 37°C for 5 days.

RNA Purification

Total RNA was isolated from the infected cell using the NucleoSpin[®] RNA kit (Macherey-Nagel) as was previously described (Tombácz et al., 2016, 2017; Balázs et al., 2017a). Total RNA samples were handled by Ambion[®] TURBO DNA-free[™] Kit (Thermo Fisher Scientific) to remove to potential gDNA contamination. For the 1D cDNA sequencing, the polyA(+) fraction was extracted from the total RNA samples by using the Qiagen Oligotex mRNA Mini Kit, following the “Spin Columns” protocol of the kit. Samples were quantified by Qubit 2.0 fluorimeter using the Qubit RNA BR and HS Assay Kits (Life Technologies) for the total RNA and PolyA(+)RNA measurement, respectively and then they were stored at –80°C until use.

FIGURE 1 | Read length and read length distributions. **(A)** The horizontal bar charts show the average read lengths of the two different libraries, as well as the median and the average read lengths mapped to the viral genome. **(B)** The average distribution of read lengths which align to the VZV genome; binned to 500bp intervals at log10 scale.

Generation of Sequencing Libraries

An ONT MinION device was used for the analysis of the full-length transcriptome profile of VZV. We applied two library preparation protocols. The 1D cDNA sequencing was carried out by using the 1D Strand switching cDNA by ligation protocol (Version: SSE_9011_v108_revS_18Oct2016). After the first End-prep step, a barcode (C11 barcode: ONT PCR Barcoding Kit 96; EXP-PBC096) was ligated to the cDNA samples following the relevant part of the 1D PCR barcoding (96) genomic DNA (SQK-LSK108) protocol, for better identification of the 5' end of the reads. We also applied a Cap-selection method combined with the 1D protocol for the detection of the 5'-ends of the transcripts. The TeloPrime Full-Length cDNA Amplification Kit (Lexogen) was used for the cDNA preparation. Total RNA was used for the reverse transcription (RT). The sample was mixed with RT buffer and a specific primer (both are part of the kit). The reaction started with incubation at 70°C for 30 s, which was then followed by a 1 min step at 37°C. The reverse transcriptase enzyme and the additional reagents (components of the kit) were mixed with the sample and then the reaction was contained at 37°C for 2 min. The next incubation step of the RT was carried out at 46°C for 50 min. The sample was purified by using the kit's Silica columns. The double-strand (ds) specific ligase enzyme (Lexogen

kit) was used to join the adapter to the cDNA. The ligation was done at 25°C overnight, then the sample was purified using the silica membranes of the kit. The ds cDNAs were generated by using the Enzyme Mix and the Second-Strand Mix (Lexogen kit). The cDNA generation was carried out in a Veriti thermal cycler, applying the following protocol: 98°C for 90 s, 62°C for 60 s, 72°C for 5 min (16 cycles), hold at 25°C. The library production from the dsDNA was based on the 1D protocol; the end-repair and the 1D adapter ligation (NEBNext End repair / dA-tailing Module NEB Blunt/TA Ligase Master Mix) steps were carried out by using the 1D protocol and kit. The ready libraries were run on an ONT R9.4 SpotON Flow Cells.

The concentration of both the ds cDNAs, as well as the libraries was detected by using Qubit 2.0 and Qubit dsDNA HS quantitation assay (Life Technologies).

Mapping, Data Processing, and Statistics

The ONT's Albacore software v2.0.1 was used for base calling. These sequencing reads were aligned using the GMAP (Wu and Watanabe, 2005) version 2017-09-30 with default setting to the genome NC_001348. Statistics about the read quality, such as insertions, deletions, and mismatches, as well as the coverages can be found in **Table 1**. In house scripts were used

to obtain the quality information presented in this data report (Github, doi:10.5281/zenodo.1034511). The basic statistic data about the FASTQ files are shown in **Supplementary Table 1**. The FASTQ analysis was carried out using the ea-utils package (Aronesty, 2011). This toolkit includes programs for calculating sequencing and alignment statistics, demultiplexing, and variant calling. The fastq-stats program from this software package was used to obtain data base quality scores, and additional basic information such as base composition, base count, read lengths. FastQC version 0.11.5 was used to generate quality reports (deposited in Figshare; https://figshare.com/articles/Varicella_Zoster_FastQC_data/7016372, **Supplementary Table 2**). The read length distribution of the samples is visualized in **Figure 1B**.

DATA AVAILABILITY

The sequencing data and the transcriptome assembly have been uploaded to the European Nucleotide Archive under the project accession number: PRJEB25401. The FASTQ and binary alignment (BAM) files have also been uploaded for each experiment to facilitate the usage of the dataset. The FASTQ files can be aligned to any reference genome, while the BAM files contain reads already mapped to the NC_001348.1. Sample SAMEA104667607, Run accession ERR2366789 is the polyA selected experiment, while sample SAMEA104667608, run accession ERR2366790 is the CAP selected experiment.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the European Nucleotide Archive PRJEB25401 [<https://www.ebi.ac.uk/ena/data/search?query=PRJEB25401>]; named as “Long-read Sequencing Dataset of Varicella Zoster Virus.”

REFERENCES

- Aronesty, E. (2011). *ea-utils: Command-Line Tools for Processing Biological Sequencing Data*. Available online at: <https://github.com/ExpressionAnalysis/ea-utils>
- Baird, N. L., Bowlin, J. L., Cohrs, R. J., Gilden, D., and Jones, K. L. (2014). Comparison of varicella-zoster virus RNA sequences in human neurons and fibroblasts. *J. Virol.* 88, 5877–5880. doi: 10.1128/JVI.00476-14
- Balázs, Z., Tombácz, D., Szucs, A., Csabai, Z., Megyeri, K., Petrov, A. N., et al. (2017a). Long-read sequencing of human cytomegalovirus transcriptome reveals RNA isoforms carrying distinct coding potentials. *Sci. Rep.* 7:15989. doi: 10.1038/s41598-017-16262-z
- Balázs, Z., Tombácz, D., Szucs, A., Snyder, M., and Boldogkoi, Z. (2017b). Long-read sequencing of the human cytomegalovirus transcriptome with the Pacific Biosciences RSII platform. *Sci. Data* 4:170194. doi: 10.1038/sdata.2017b.194
- Chambers, J., Angulo, A., Amaratunga, D., Guo, H., Jiang, Y., Wan, J. S., et al. (1999). DNA microarrays of the complex human cytomegalovirus genome: profiling kinetic class with drug sensitivity of viral gene expression. *J. Virol.* 73, 5757–5766.
- Delseny, M., Han, B., and Hsing, Y. I. (2010). High throughput DNA sequencing: the new sequencing revolution. *Plant Sci.* 179, 407–422. doi: 10.1016/j.plantsci.2010.07.019

AUTHOR CONTRIBUTIONS

DT carried out ONT sequencing, data analysis, and drafted the manuscript. IP take part in RNA purification, sequencing, and data analysis. NM participated in data analysis. AS carried out bioinformatics analysis. ZB conceived and designed the experiments, and wrote the manuscript. All authors read and approved the final paper.

FUNDING

This work was supported by the Swiss-Hungarian Cooperation Programme [SH/7/2/8] and by NKFIH OTKA K 128247 to ZB. The work was also supported by the Bolyai János Scholarship of the Hungarian Academy of Sciences to DT.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2018.00460/full#supplementary-material>

Supplementary Table 1 | Basic statistics of the reads.

^Anumber of reads in the given file; ^Bmean of the read lengths in given FASTQ; ^Cstandard deviation; ^DPhred scale used (Phred scale also known as the quality or Q scores of a base, which are often represented as ASCII characters. Tables converting between integer Q scores, ASCII characters and error probabilities are shown in the upper table in the following website: https://www.drive5.com/usearch/manual/quality_score.html. The ASCII_BASE 33 is now almost universally used for short – and long-read sequencing techniques); ^ENumber of reads used to generate duplicate read statistics (the total read count of a given sample was used for the duplicate analysis); ^Fbase Quality min; ^Gbase Quality max; ^Hbase Quality mean; ^Ibase Quality SD; ^JA content (%); ^KC content (%); ^LG content (%); ^MT content (%); ^Ntotal number of bases; (Unique: duplicates have been removed; Duplicates: duplicate reads are included in the statistics).

Supplementary Table 2 | This table summarizes the type of samples and the related FastQC file names, which have been deposited to the Figshare.

- Ebrahimi, B., Dutia, B. M., Roberts, K. L., Garcia-Ramirez, J. J., Dickinson, P., Stewart, J. P., et al. (2003). Transcriptome profile of murine gammaherpesvirus-68 lytic infection. *J. Gen. Virol.* 84, 99–109. doi: 10.1099/vir.0.18639-0
- Kennedy, P. G. (2002). Varicella-zoster virus latency in human ganglia. *Rev. Med. Virol.* 12, 327–334. doi: 10.1002/rmv.362
- Maresova, L., Pasiaka, T. J., Homan, E., Gerday, E., and Grose, C. (2005). Incorporation of three endocytosed varicella-zoster virus glycoproteins, gE, gH, and gB, into the virion envelope. *J. Virol.* 79, 997–1007. doi: 10.1128/JVI.79.2.997-1007.2005
- Moldován, N., Tombácz, D., Szucs, A., Csabai, Z., Snyder, M., and Boldogkoi, Z. (2018). Multi-platform sequencing approach reveals a novel transcriptome profile in pseudorabies virus. *Front. Microbiol.* 8:2708. doi: 10.3389/FMICB.2017.02708
- O'Grady, T., Wang, X., Höner zu Bentrup, K., Baddoo, M., Concha, M., and Flemington, E. K. (2016). Global transcript structure resolution of high gene density genomes through multi-platform data integration. *Nucleic Acids Res.* 44, e145–e145. doi:10.1093/nar/gkw629
- Oláh, P., Tombácz, D., Póka, N., Csabai, Z., Prazsák, I., and Boldogkoi, Z. (2015). Characterization of pseudorabies virus transcriptome by Illumina sequencing. *BMC Microbiol.* 15:130. doi: 10.1186/s12866-015-0470-0

- Reichelt, M., Brady, J., and Arvin, A. M. (2009). 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.* 83, 3904–3918. doi: 10.1128/JVI.02137-08
- Tombácz, D., Csabai, Z., Oláh, P., Balázs, Z., Likó, I., Zsigmond, L., et al. (2016). Full-length isoform sequencing reveals novel transcripts and substantial transcriptional overlaps in a herpesvirus. *PLoS ONE* 11:e0162868. doi: 10.1371/journal.pone.0162868
- Tombácz, D., Csabai, Z., Oláh, P., Havelda, Z., Sharon, D., Snyder, M., et al. (2015). Characterization of novel transcripts in pseudorabies virus. *Viruses* 7, 2727–2744. doi: 10.3390/v7052727
- Tombácz, D., Csabai, Z., Szucs, A., Balázs, Z., Moldován, N., Sharon, D., et al. (2017). Long-read isoform sequencing reveals a hidden complexity of the transcriptional landscape of herpes simplex virus type 1. *Front. Microbiol.* 8:1079. doi: 10.3389/fmicb.2017.01079
- Tombácz, D., Sharon, D., Szucs, A., Moldován, N., Snyder, M., and Boldogoi, Z. (2018). Transcriptome-wide survey of pseudorabies virus using next- and third-generation sequencing platforms. *Sci. Data* 5:180119. doi: 10.1038/sdata.2018.119
- Tyler, S. D., Peters, G. A., Grose, C., Severini, A., Gray, M. J., Upton, C., et al. (2007). Genomic cartography of varicella-zoster virus: a complete genome-based analysis of strain variability with implications for attenuation and phenotypic differences. *Virology* 359, 447–458. doi: 10.1016/j.virol.2006.09.037
- Wu, T. D., and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21, 1859–1875. doi: 10.1093/bioinformatics/bti310

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Tombácz, Prazsák, Moldován, Szűcs and Boldogkői. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

RESEARCH ARTICLE

Open Access

Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus

István Prazsák^{1†}, Norbert Moldován^{1†}, Zsolt Balázs¹, Dóra Tombácz¹, Klára Megyeri², Attila Szűcs¹, Zsolt Csabai¹ and Zsolt Boldogkői^{1*}

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 the 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 25 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 ORF63 exhibits a complex structural variation encompassing the splice sites of VZV latency transcripts. 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 of the 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, nroRNA

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) 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

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

[†]István Prazsák and Norbert Moldován contributed equally to this work.

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

Full list of author information is available at the end of the article

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), however, a recent study reported the identification of VZV latency-associated transcript (VLT) overlapping the ORF61 gene [9]. It has also been thought that VZV lacks the α -TIF protein encoded by the HSV's 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. Exceptions to this rule were found in the Kaposi's sarcoma-associated herpesvirus [19, 20] and in HSV-1 [21]. 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 [22, 23]. 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, ORF50 the glycoprotein M (gM), and the newly discovered, multiple spliced VLT [6, 9, 23–26].

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 [27], 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 [28] with an upstream uracil (U) exerting a stabilizing effect on the RNA-ADAR complex [29]. Additionally, RNA hyper-editing has been shown to play a crucial role in the viral life cycle by dodging the host's immune system [30] and also in the control of DNA replication [31].

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 [32]. Mammalian mRNAs contain an essential sequence context for translation initiation, known as the Kozak consensus sequence [33]. 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) [34]. 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 [32, 35]. In short upstream (u)ORFs translation reinitiation shows a positive correlation with distance between stop codon of uORF and the AUG of downstream ORF [36].

The hybridization-based microarray and the second-generation short-read sequencing (SRS) techniques have revolutionized genome and transcriptome research, including herpesviruses [37–40]. 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 multi-intronic 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 [41, 42].

Similar 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 [43, 44] or false priming [45]. 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 [46–49], including herpesviruses [18, 22, 23, 50–52]. These reports have revealed a hidden transcriptome complexity, which 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, 46], HSV-1 [53], and human cytomegalovirus (HCMV) [22]. Additionally, a large extent of transcriptional read-through and overlaps has also been uncovered by these studies [18, 22, 23]. 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 [54].

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 (MRC-5) 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% CO₂ atmosphere. The live attenuated OKA/Merck strain of varicella zoster virus (VZV) was cultured at 37 °C in MRC-5 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 MRC-5 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 50 µl 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 50 min according to the kit's recommendations by using an oligo(d)T-containing primer (5' -> 3': TCTC AGGCGTTTTTTTTTTTTTTTTTTT). 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.

Targeted sequencing of putative ONT transcripts

Validation of the TSSs of three 'near the transcription origin' (NTO) transcripts, and the spliced isoforms of the region was carried out using targeted long-read sequencing. The cDNA library was prepared using the ONT 1D protocol, as described in the 'Oxford Nanopore MinION cDNA sequencing and barcoding' section, with the following modifications: we used ribodepletion instead of poly(A) selection, and target-specific primers (Integrated DNA Technologies) were applied instead of oligo(dT)-priming for the cDNA preparation. The target-specific primers used in this study are listed in Table 1.

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

Table 1 Target-specific primers used for the detection of NTO transcripts

Primer name	Sequence
pNTO2	ACTTGCCTGTCGCTCTATCTTCCGGTCAGGATCT
pNTO3	ACTTGCCTGTCGCTCTATCTTCGACTGTGGAACT
pNTO4	ACTTGCCTGTCGCTCTATCTTCATGTGAGACAACC
pNTO1_3'	ACTTGCCTGTCGCTCTATCTTCTCGTGTGACG
pNTO1-ex2	ACTTGCCTGTCGCTCTATCTTCCTCGACGCTG

The primers are composed of a sequence complementary to the MinION cDNA adapter and a target-specific sequence (in *italics*)

GMAP v2017-04-24 with the default settings [55]. In order to annotate the 5' ends of the transcripts, we used our custom script (<https://github.com/zsolt-balazs/LoR-TIA>). In short: the last 16 nt of the MinION 5' strand switching adapter or the last 16 nt 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 or -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 [$k_0; \lambda$]) 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 rare long reads were inspected individually using the Integrative Genome Viewer (IGV) [56]. 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 positions 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. The computational pipeline is outlined in Additional file 1.

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 10 nt 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. The relative abundance of the transcript isoforms was calculated by dividing the number of reads with the total number of reads of the dataset; both resulting from the computational pipeline described before. A ± 10 nt-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 [57], 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 [58] 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 NTO3-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 NTO3 as the other one. Reads were visualized using the Geneious [59] software suite and IGV. GC-, CCAAT- and TATA-boxes were annotated using the Smith-Waterman algorithm for canonical motifs [60].

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 (Fig. 1). Both cap-selected and

non-cap-selected protocols were applied for the analyses of the whole transcriptome. Additionally, targeted sequencing was used with the non-cap-selected protocol for validating the TSS and splice junctions of the NTO transcripts. Sequencing of the non-cap-selected non-targeted sample yielded 619,687 reads of which 66,455 mapped to the VZV reference genome (GeneBank Accession: NC_001348.1) with an average mapped read length of 1349 bp, and an average coverage of 625, while 553,232 reads mapped to the host cell genome (*Homo sapiens* - GRCh37, BioProject number: PRJNA31257).

Targeted sequencing yielded 1,792,059 reads of which 328,112 mapped to the VZV genome, with an average mapped read length of 651 bp, and an average coverage of 1710, and 1,463,947 reads mapped to the host cell genome. The 2.41% of the targeted reads mapping to the VZV genome included the target-specific primers and mapped to the correct position, whereas, 24.112% of the reads included the primer but mapped off-target, while for 73.477% the target-specific primer was missing. These latter two categories could be a result of nonspecific priming or template switching.

Sequencing of the cap-selected samples yielded 6,918,054 reads, of which 614,192 mapped to the viral genome, with an average coverage of 1200, and 6,303,862 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) (Fig. 2 panel a). Intriguingly, we obtained the same poor result with other herpesviruses, including PRV [23] HSV-1 (our unpublished results), however, the cap-selection technique performed very well in 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 (Fig. 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.

In total, we detected 1124 5'-ends and 255 3'-ends in the non-cap-selected dataset, and 1428 5'-ends and 279 3'-ends in the cap-selected dataset (Additional file 2: Table S1-S4). 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 (Table 2). 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.

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 putative TSSs and 71 TESs listed in Additional file 2: Table S5. 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 3).

Twenty-two and five transcript isoforms were identified with the non-cap-selected, and cap-selected dataset,

respectively with a relative abundance higher than 1% (Additional file 2: Table S5).

In this work, we detected an enrichment of As and Us in an interval of up to -10 nt upstream of the putative TES, and an abundant GT-rich region downstream in the +10 nt interval (Fig. 3). We annotated altogether 240 novel transcripts of which 143 were

confirmed by at least three reads, additionally, we mapped the TSS and TES of 37 previously detected transcripts lacking former mapping of RNA end positions (Additional file 2: Table S5-S6).

Putative mRNAs

Transcripts embedded into 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 of 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 25 embedded viral transcripts containing tORFs. We could identify promoters for 10 of these transcripts, located at a mean distance of 93.85 ± 32.18 nt (mean $\pm$ SD) from their TSSs (Additional file 2: Table S7). 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. Eighteen of the *possibly protein coding* VZV transcripts contain a PAS at a 19.44 ± 8.91 (mean $\pm$ SD) distance upstream their TESs (Additional file 2: Table S7).

Table 2 The number of read ends following each step of filtering

Filter	Non-cap-selected		Cap-selected	
	5' ends	3' ends	5' ends	3' ends
None	1124	255	1428	279
Peak analysis and Poisson probability	151	92	202	116
Template switching/false priming	102	76	181	100

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 (Fig. 1, Additional file 2: Table S5). 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 $\pm$ SD) upstream

Table 3 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

their TSSs, and 19 have a canonical PAS upstream their TESs at a mean distance of 14.96 ± 11.59 (mean $\pm$ 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 18 embedded ncRNAs, of which eight were 5'-truncated, while ten were 3'-truncated transcripts (Additional file 2: Table S5). 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 $\pm$ SD) upstream their TSSs, while 21 have a PAS at a 16.35 ± 8.87 (mean $\pm$ 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 (Fig. 1). The LAT RNA has been described in every member of the alpha-herpesvirus subfamily [61, 62], 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 (Fig. 4).

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 18 novel 5'-UTR length variants, 7 being shorter and 11 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 $\pm$ SD). Additionally, we detected eight 3'-UTR length isoform, all with a canonical PAS 22 ± 7.65 (mean $\pm$ SD) upstream their TESs. An intriguing finding is the putative TSS at the genome position + 4 belonging to the rare transcripts 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

Fig. 3 The frequency of nucleotides in the vicinity of transcriptional end sites (TES). The ± 10 nt vicinity of $N=61$ unique TESs was analyzed using the online WebLogo v3.6.0 service. 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 TES

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 ($\leq 1\%$) 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 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.

ORF63, homologue to the HSV-1 and the PRV *us1* gene, is one of the main regulator 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 [50, 53] (Fig. 5). Ten splice junctions of ORF63 splice variants coincide with those of the VLT_{ly} splice variants [9], thus they were labeled as VLT-ORF63-C (Fig. 6). These were confirmed using targeted sequencing. It is noted that one of the splice donor sites, present in both NTO1v1 and NTO1v2 is GC, which differ from the canonical splice donor sequence.

We detected all of the three ORF50 splice isoforms previously described [24], 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 4).

In four isoforms, splicing affects the sizes of the ORFs, all producing hypothetical proteins truncated near the N-terminal. Roviš and colleagues [63] have identified the majority of VZV proteins using monoclonal antibodies. In some cases, lower-size bands appeared in their Western blots, which they explained to be the result of putative post-translational modification, or proteolytic cleavage. We propose splicing as possible option for these nonspecific protein isoforms, listed in Table 5.

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

In our earlier work, we reported [64] 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-RNA: [65]) and HCMV (OriLyt: [66]; RNA4.9: [67]). The VZV genome contains exclusively OriS, and lacks the replication origin at the UL segment. We identified nine 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 (Fig. 5 panel a). NTO2 starts at the same TSS as NTO1v1 but terminates 30 bp downstream of its splice donor site. The 5' end of the reads of NTO3 and NTO4 are positioned downstream of NTO2. These transcripts were present in our cap-selected but not in the non-cap-selected sample presumably because of their low abundance. In order to

augment the detection of these transcripts, we used target-specific primers (Fig. 5 panel a.) for sequencing library preparation, followed by MinION cDNA sequencing. The starting positions detected in the cap-selected sample were not confirmed by the targeted sequencing of these transcripts, thus we hypothesize that NTO3 and NTO4 starts at the closest upstream confirmed TSS, that of NTO1v1. The arrangement of the NTO transcripts is similar to that observed in PRV, where the TSS of the PTO-US1 (positionally similar to the NTO1v1) is located at the same genetic locus as PTO (locationally similar to the NTO2, NTO3 and NTO4, but not homologous) (Fig. 5 panels a and c). Our in silico analysis revealed that the poly(A) cleavage site of NTO2 is located within a canonical CA element, with a GU-rich region starting 13 nt downstream from the TES. The NTO3 has a CTAAA poly(A) signal [68] starting 20 nt upstream from its TES, with a cleavage site in a homopolymer C run and with a consensus GU-rich region starting 25 nt downstream from the TES. The NTO4 has an ATATAA poly(A) signal [69] starting 24 nt upstream the TES, the

cleavage site being marked by a consensus CA signal, followed by a GU-rich region 33 nt downstream. No enrichment of adenines was observed in the genomic region of the TES-s in any of the three NTO transcripts, which would be the source of strand switching or false priming. Based on these results, 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] (Fig. 6).

Polycistronic and complex transcripts

A major issue of SRS, as well 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

VLT_{ly}-ORF63-C4, VLT_{ly}-ORF63-C5, VLT_{ly}-ORF63-C6 and VLT_{ly}-ORF63-64-C1 encoding hypothetical protein 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* (Fig. 7). The N-terminal overhangs of the putative proteins show no homology with any other known proteins in online databases.

Transcripts can overlap each other in parallel (tandem or head-to-tail), convergent (tail-to-tail) or divergent (head-to-head) manner (Fig. 8). RNAs identified and

Table 4 Splice junction sites of the VZV transcriptome

Intron Start	Intron End	Strand	Intron Length	Donor Site	Acceptor Site	Accepted as novel	Intron depth	Transcript name	References
585	714	+	129	GT	AG		45,62%	ORF0-1-C-SP	[25]
5004	4155	–	849	GT	AG	✓	2,90%	ORF5-4-SP	
16,712	16,886	+	174	GT	AG	✓	5,04%	ORF12-13-SP	
43,826	43,505	–	321	GT	AG	✓	4,11%	ORF24-SP	
43,950	43,863	–	87	GT	AG	✓	1,40%	ORF24-SP-2	
61,790	63,564	+	1774	GT	AG	✓	2,40%	–	
77,870	78,148	+	278	GT	AG	✓	11,29%	ORF42-43-AS	
78,938	78,039	–	899	GT	AG	✓	38,88%	ORF42/ORF45-SP-2	
81,537	78,039	–	3498	GT	AG		100,00%	ORF42/45-SP-1	[6]
81,537	79,053	–	2484	GT	AG	✓	12,72%	ORF42/ORF45-SP-2	
87,759	86,782	–	977	GT	AG		4,15%	ORF50B	[24]
87,759	86,982	–	777	GT	AG		2,33%	ORF50A	[24]
87,436	86,782	–	654	GT	AG		0,62%	ORF50C	[24]
86,982	87,436	–	454	GT	AG		0,33%	ORF50D	
101,728	102,420	+	692	GT	AG		87,50%	VL _{T_{ly}}	[9]
102,484	102,853	+	369	GT	AG		100,00%	VL _{T_{ly}}	[9]
102,983	103,827	+	844	GT	AG		2,42%	VL _{T_{ly}}	[9]
103,924	104,293	+	369	GT	AG		3,99%	VL _{T_{ly}}	[9]
104,433	104,768	+	335	GT	AG		5,22%	VL _{T_{ly}}	[9]
104,824	110,509	+	5685	GT	AG	✓	2,18%	ORF63-SP2-C	
108,830	110,509	+	1679	GC	AG	✓	11,12%	NT01v1 (ORF63-SP-1)	
114,920	115,050	+	130	GT	AG	✓	1,04%	ORF67-SP	
121,067	119,388	–	1679	GC	AG	✓	10,65%	NT01v1 (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 Additional file 2: Table S6. The TSS and TES belonging to the transcript with splice sites in row 7 could not be determined with certainty

annotated with a certain TSS and TES in this work form a total of 481 overlaps, of which 15 are head-to-head, 450 are head-to-tail and 16 are tail-to-tail (Additional file 2: Table S8). The overlaps can be full or partial. Full overlaps can be formed between the RNA molecules encoded by polycistronic transcription units (Fig. 8 panel a.), 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

Table 5 The proteins and hypothetical proteins produced by spliced transcripts

Transcript	Host protein length (aa)	Spliced protein length (aa)	Intron position	Splice donor	Splice acceptor
ORF5-4-SP	ORF4: 452	452	5' UTR	4141	2783
ORF12-13-SP	ORF12: 661	233	in frame, 5'truncated	16,214	17,088
ORF24-SP	ORF24: 269	137	in frame	44,021	43,286
ORF42/45-SP-2	ORF42/45: 747	380	in frame, 5'truncated	82,593	78,969
ORF50D	ORF50: 435	–	–	–	–
NT01v1 (ORF63-SP-1)	ORF63: 278	278	5' UTR	110,581	111,417
ORF63-SP-2-C	ORF63: 278	278	5' UTR	110,581	111,417
ORF67-SP	354	311	in frame, 5'truncated	114,496	115,559

isoforms) or transcriptional read-through (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 (Fig. 8 panel c).

Upstream ORFs

Using *in silico* methods, we detected 44 potential (u)ORFs on the 5'-UTRs of 81 VZV transcripts (Additional file 2: Table S9). Five of the longer TSS variants contain uORFs, while their shorter isoform does not. We have previously described this phenomenon in HCMV [52] and PRV [23]. The average size of an uORF was 54.9 ± 44.64 nt (mean $\pm$ 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 $\pm$ SD, median = 110 nt). This space between the two uORF and the canonical ORF is enough for a potential reinitiation event. We identified Kozak consensus sequences in five uORFs (Additional file 2: Table S9).

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 $\rightarrow$ G (Fig. 9 panel a and 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) (Fig. 9 panel c), making 22.07% of all As of the transcript edited. The targeted sequencing dataset revealed a similar A $\rightarrow$ G substitution pattern in the NTO3 overlapping reads as the cap-selected data. This suggest a hyper-editing event in NTO3. Using the MEME software suite [70] we could detect a slight enrichment of the GU dinucleotide preceding

the editing site resulting in a GUAG motif (the editing site is underlined) (Fig. 9 panel d).

Using the 'RNAstructure' software suite [58], we predicted the secondary structure with the lowest free energy of the NTO3 ssRNA (Additional file 3 a and b) and of the NTO3-ORF62-5' dsRNA hybrid, using both the unedited and edited forms of the asRNA (Additional file 3 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, 71–78]. 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 increased the number of full length transcripts in VZV by one order of magnitude. We identified 108 novel transcripts, including novel mRNA molecules, monocistronic transcripts, UTR isoforms, as well as multigenic transcripts. Novel splice sites and splice variants were also detected. Additionally, we discovered five sncRNAs and 21 lncRNAs and

Fig. 8 Types of overlaps between VZV transcripts. Reads were visualized using IGV. Blue arrow-rectangles represent transcript annotations. The overlaps are framed in orange. **a.** Parallel overlaps of the ORF5–4-3 cluster. **b.** Divergent overlaps of the ORF8 and ORF9 and ORF9A-9 transcripts. **c.** Convergent overlaps of the ORF17 and ORF18-AT transcripts

annotated their TSSs and TESs with base pair precision. The precise annotation of the viral transcriptome is essential for understanding the genetic regulation of VZV. Different TSS isoforms can be regulated by alternative promoters, like for the *ul44* gene of HCMV, where two transcript isoforms initiate downstream of a canonical TATA promoter and are expressed at early time points of the viral infection, while the third promoter is non-canonical, and

the mRNA isoform starting downstream from it is transcribed at a later phase [79]. Furthermore, alternative TESs may influence RNA metabolism by allowing or forbidding the binding of micro (mi)RNAs to the viral RNA [80] or by facilitating or preventing deadenylation [81]. The enrichment of As and Us upstream of PASs in mammalian systems is well-established [82–84]. 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 [85].

Similar to other herpesviruses [22, 53], this study also revealed a complex meshwork of transcriptional read-throughs and overlaps. It has been known earlier that the polycistronic transcription units include co-terminal multigenic RNA molecules, which represent a large extent of parallel overlaps along the entire viral genome. We can raise the question as to whether the downstream sequences of multicistronic transcripts have any other functions besides their possible coding potential, 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 [54].

Understanding the factors influencing the expression of *orf63* is of particular interest as its transcript is translated predominantly during latent infections of human ganglia [9, 86]. It has been previously shown that *orf63* has no trans-regulatory effect on the expression of *orf62*, which encodes the major immediate-early transactivator protein of VZV [87]. Nevertheless, deletion of *orf63* has demonstrated that it is indispensable for VZV replication, and its product has a protein interaction site with the translated *orf62* [86]. Other experiments have shown that deletion of *orf63* increases the expression of *orf62* [88], suggesting that there is a link between the regulation of the two genes. A possible explanation may be the transcriptional interference caused by the head-to-head overlaps between ORF62 and the isoforms of ORF63 (NTO1v1 and NTO1v2) in the late phase of the viral life cycle.

Polycistronic (especially bicistronic) transcripts are also common in eukaryotic organisms, but their generation is explained by trans-splicing and not by transcriptional read-through [89]. 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 read-through 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 [90]. Very long complex transcripts form a distinct category among multigenic transcripts because of their oppositely oriented ORFs and increased size. Similarly to other herpesviruses [22, 53] 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 [91].

It has been previously thought that the OriS of VZV lacks any overlapping transcripts [6]. Although Davison and colleagues (1986) hypothesized the existence of non-coding RNA in the intergenic region between *orf62* and *orf63* [6]. 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 [92] through the interplay between the transcriptional and replication apparatuses [54]. 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 [64]. Furthermore, the intergenic region between *orf62* and *orf63* contains binding sites for T cell proteins including heat shock factor 1 (HSF-1), nuclear factor 1 (NF-1), NF-κB/Rel, and octamer binding proteins (Oct-1), as well as cyclic AMP-responsive elements (CRE) [93]. The presence of NTO and ORF63 isoforms overlapping this region could interfere with the binding of these host cell proteins. Herpesviruses can evade host cell interferon production and T cell response by producing miRNAs derived from lncRNAs, or antisense RNAs [94–96]. Short noncoding RNAs have been recently documented in the *orf61–62–63* region of VZV [78]. The NTO and ORF63 isoforms detected in our study could be a possible source of these sncRNAs.

We detected a hyper-editing process in the novel NTO3 transcript. This phenomenon was previously observed in members of the herpesvirus family including alphaherpesviruses [97–99]. This process plays a crucial role in the cell's innate immunity [100], while it can be

hijacked by some viruses to evade inactivation [30]. Additionally, in hepatitis delta virus hyper-editing is indispensable for replication [31]. 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 [101]. 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 inhibiting RNA interference [102]. 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) [103, 104]. 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 [105]. 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. The scarcity of the novel splice isoforms implies further validation, using techniques capable of detecting rare transcripts like targeted direct RNA sequencing.

At least six VZV-transcripts have been shown to be expressed in latency [106, 107], 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 life-cycle, 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 distinct 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 splicing patterns during the life cycle of virus. Similar alternative splicing has been described in VLT, assuming distinct functions for it in lytic and latent phase [9, 105].

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 [108]. 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 stages of the viral lifecycle [22]. TSSs and TESs detected by former S1 nuclease and primer extension experiments agree with our results with base pair precision, demonstrating that our comprehensive LRS transcriptomic survey of VZV complemented with the proposed bioinformatics pipeline can provide a useful background to design other viral transcriptome reannotation analysis by ONT platform. It is also important to know which transcripts are affected by a genetic modification of the virus. The organization principles of the viral genome and the viral replication can only be understood with a detailed knowledge of the transcriptome.

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 discovery of novel TSS and TES isoforms can lead to a better understanding of the regulation of the viral gene expression through alternative promoters or through RNA turnover, whereas uORFs discovered in this work can modulate translation. 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, especially the data on novel putative mRNAs and ORFs, may also provide useful information to target transcript candidates for controlling this virus.

Additional files

Additional file 1: The procedure of finding novel transcriptional isoforms. The flowchart describes the pipeline used to detect novel TSSs and TES-s. (PDF 505 kb)

Additional file 2: The list of 5', 3' ends, TSSs, TESs, promoter motifs, polyadenylation signals, transcriptional overlaps and uORFs identified by this study. (XLSX 383 kb)

Additional file 3: 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 NTO3 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 IU base pairs is marked with black arrows. (TIFF 1530 kb)

Abbreviations

ADAR1: Adenosine deaminase acting on RNA type 1; AS: Antisense (transcript); AT: Alternative termination; C: Complex (transcript); HCMV: Human cytomegalovirus; HSV-1: Herpes simplex virus type 1; lncRNA: Long non-coding RNA; LRS: Long-read sequencing; miRNA: MicroRNA; 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; tORF: Truncated open reading frame; -TR: 5'-truncated (transcript); 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

Acknowledgements

Not applicable.

Author contributions

IP analyzed the data, participated in RNA-isolation 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. ZBa carried out the analysis of the targeted sequencing. 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 and participated in the MinION sequencing. KM carried out the virus infection and propagated the cells. ZB conceived, designed and co-ordinated the study and wrote the manuscript. All authors have read and approved the final version of the manuscript.

Funding

This work was supported by the Swiss-Hungarian Cooperation Programme [SH/7/2/8] and NKFIH OTKA [K128247] to ZB. The work was also supported by the Bolyai János Scholarship of the Hungarian Academy of Sciences to DT.

Availability of data and materials

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

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Department of Medical Biology, Faculty of Medicine, University of Szeged, Szeged 6720, Hungary. ²Department of Medical Microbiology and Immunobiology, Faculty of Medicine, University of Szeged, Szeged 6720, Hungary.

Received: 4 September 2018 Accepted: 19 November 2018

Published online: 04 December 2018

References

- Kennedy PGE. Varicella-zoster virus latency in human ganglia. *Rev Med Virol.* 2002;12:327–34.
- Cohrs RJ, Gilden DH, Mahalingam R. Varicella zoster virus latency, neurological disease and experimental models: an update. *Front Biosci.* 2004;9:751–62.
- Arvin AM. Investigations of the pathogenesis of varicella zoster virus infection in the SCIDhu mouse model. *Herpesviridae.* 2006;13:75–80.
- Zerboni L, Sen N, Oliver SL, Arvin AM. Molecular mechanisms of varicella zoster virus pathogenesis. *Nat Rev Micro.* 2014;12:197–210.
- Maresova L, Pasioka 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.
- Davison AJ, Scott JE. The complete DNA sequence of varicella-zoster virus. *J Gen Virol.* 1986;67:1759–816.
- 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.
- 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.
- 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.
- 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.
- Kinchington PR, Houghland 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.
- 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.
- Moriuchi H, Moriuchi M, Straus SE, Cohen JL. 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.
- 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.
- 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.
- 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. American Society for Microbiology*. 2003;77:11718–32.
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 Public Libr Sci*. 2016;11: 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.
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. Griffiths A, Coen DM. An unusual internal ribosome entry site in the herpes simplex virus thymidine kinase gene. *Proc Natl Acad Sci U S A*. 2005;102: 9667–72.
22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. Walkley CR, Li JB. Rewriting the transcriptome: adenosine-to-inosine RNA editing by ADARs. *Genome Biol BioMed Central*. 2017;18:205.
28. Eggington JM, Greene T, Bass BL. Predicting sites of ADAR editing in double-stranded RNA. *Nat Commun*. 2011;2:319.
29. 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.
30. Zahn RC, Schelp I, Utermöhlen O, von Laer D. A-to-G hypermutation in the genome of lymphocytic choriomeningitis virus. *J Virol. American Society for Microbiology*. 2007;81:457–64.
31. 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.
32. Kozak M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene Elsevier*. 2005;361:13–37.
33. Kozak M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell*. 1986;44: 283–92.
34. Pisarev AV, Kolupaeva VG, Pisareva VP, Merrick WC, Hellen CUT, Pestova TV. 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 Cold Spring Harbor Laboratory Press*. 2006;20:624–36.
35. Luukkainen 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 American Society for Microbiology (ASM)*. 1995;69:4086–94.
36. Kozak M. Constraints on reinitiation of translation in mammals. *Nucleic Acids Res Oxford University Press*. 2001;29:5226–32.
37. 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.
38. 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.
39. 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.
40. 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.
41. Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, et al. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat Commun Nat Publ Group*. 2017;8:16027.
42. Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, Jaeger E, Chen F, et al. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat Biotechnol*. 2015;33:736–42.
43. Luo GX, Taylor J. Template switching by reverse transcriptase during DNA synthesis. *J Virol*. 1990;64:4321–8.
44. Cocquet J, Chong A, Zhang G, Veitia RA. Reverse transcriptase template switching and false alternative transcripts. *Genomics*. 2006;88:127–31.
45. 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.
46. 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:fny013.
47. 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.
48. 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;5:01040–17.
49. 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 Nat Publ Group*. 2018;8:8604.
50. Tombacz D, Sharon D, Olah P, Csabai Z, Snyder M, Boldogkoi Z. Strain Kaplan of Pseudorabies Virus Genome Sequenced by PacBio Single-Molecule Real-Time Sequencing Technology. *Genome Announc*. 2014;2: e00628–14 e00628–14.
51. 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.
52. 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 Nat Publ Group*. 2017;4:170194.
53. 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.
54. Boldogkői Z. Transcriptional interference networks coordinate the expression of functionally related genes clustered in the same genomic loci. *Front Genet Frontiers*. 2012;3:122.
55. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics Oxford University Press*. 2005; 21:1859–75.
56. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform Oxford University Press*. 2013;14:178–92.
57. 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.
58. Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*. 2010;11:129.
59. 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.
60. Narang V, Sung W-K, Mittal A. Computational modeling of oligonucleotide positional densities for human promoter prediction. *Artif Intell Med Elsevier*. 2005;35:107–19.
61. 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 American Society for Microbiology (ASM)*. 1987;61:3820–6.
62. Jin L, Scherba G. Expression of the pseudorabies virus latency-associated transcript gene during productive infection of cultured cells. *J Virol American Society for Microbiology*. 1999;73:9781–8.
63. Lenac Roviš T, Bailer SM, Pothineni VR, Ouwendijk WJD, Šimić H, Babić M, et al. Comprehensive analysis of varicella-zoster virus proteins using a new monoclonal antibody collection. *J Virol*. 2013;87:6943–54.

64. Tombácz D, Csabai Z, Oláh P, Havelda Z, Sharon D, Snyder M, et al. Characterization of novel transcripts in pseudorabies virus. *Viruses*. Multidisciplinary Digital Publishing Institute (MDPI). 2015;7:2727–44.
65. 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 National Academy of Sciences*. 1988;85:8454–8.
66. 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 American Society for Microbiology Journals*. 1996;70:5272–81.
67. Kitenko SV, Sacconi S, Izotova LS, Mirochnitchenko OV, Pestka S, Baluchova K, et al. Human cytomegalovirus harbors its own unique IL-10 homolog (cmvIL-10). *Proc Natl Acad Sci National Academy of Sciences*. 2000;97:1695–700.
68. Kainov YA, Aushev VN, Naumenko SA, Tchekina EM, Bazykin GA. Complex selection on human polyadenylation signals revealed by polymorphism and divergence data. *Genome Biol Evol Oxford University Press*. 2016;8:1971–9.
69. Pinto PAB, Henriques T, Freitas MO, Martins T, Domingues RG, Wyrzykowska PS, et al. RNA polymerase II kinetics in polo polyadenylation signal selection. *EMBO J European Molecular Biology Organization*. 2011;30:2431–44.
70. 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:W202–8.
71. 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.
72. 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;68:3571–81.
73. Kato T, Kitamura K, Hayakawa Y, Takahashi M, Kojima A, Sato S, et al. Transcription mapping of glycoprotein I (gpl) and gplV of varicella-zoster virus and immunological analysis of the gpl produced in cells infected with the recombinant vaccinia virus. *Microbiol Immunol*. 1989;33:299–312.
74. Meier JL, Straus SE. Varicella-zoster virus DNA polymerase and major DNA-binding protein genes have overlapping divergent promoters. *J Virol*. 1993;67:5753–81.
75. Kennedy PGE, Grinfeld E, Craigon M, Vierlinger K, Roy D, Forster T, et al. Transcriptomic analysis of varicella-zoster virus infection using long oligonucleotide-based microarrays. *J Gen Virol*. 2005;86:2673–84.
76. Grinfeld E, Ross A, Forster T, Ghazal P, Kennedy PGE. Genome-wide reduction in transcriptomic profiles of varicella-zoster virus vaccine strains compared with parental Oka strain using long oligonucleotide microarrays. *Virus Genes*. 2009;38:19–29.
77. 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:e01141–17.
78. Markus A, Golani L, Ojha NK, Borodiansky-Shteinberg T, Kinchington PR, Goldstein RS. Varicella-zoster virus expresses multiple small noncoding RNAs. *J Virol. American Society for Microbiology*. 2017;91:e01710–7.
79. Isomura H, Stinski MF, Kudoh A, Murata T, Nakayama S, Sato Y, et al. Noncanonical TATA sequence in the UL44 late promoter of human cytomegalovirus is required for the accumulation of late viral transcripts. *J Virol. American Society for Microbiology*. 2008;82:1638–46.
80. Barth S, Pfuhl T, Mamiani A, Eshes C, Roemer K, Kremmer E, et al. Epstein-Barr virus-encoded microRNA miR-BART2 down-regulates the viral DNA polymerase BALF5. *Nucleic Acids Res*. 2007;36:666–75.
81. Dickson AM, Anderson JR, Barnhart MD, Sokoloski KJ, Oko L, Opyrchal M, et al. Dephosphorylation of HuR protein during alphavirus infection is associated with HuR relocalization to the cytoplasm. *J Biol Chem. American Society for Biochemistry and Molecular Biology*. 2012;287:36229–38.
82. Legendre M, Gautheret D. Sequence determinants in human polyadenylation site selection. *BMC Genomics BioMed Central*. 2003;4:7.
83. 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.
84. Tian B, Graber JH. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip Rev RNA NIH Public Access*. 2012;3:385–96.
85. Hu J, Lutz CS, Wilusz J, Tian B. Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *RNA Cold Spring Harbor Laboratory Press*. 2005;11:1485–93.
86. Sommer MH, Zagha E, Serrano OK, Ku CC, Zerboni L, Baiker A, et al. Mutational analysis of the repeated open reading frames, ORFs 63 and 70 and ORFs 64 and 69, of varicella-zoster virus. *J Virol*. 2001;75:8224–39.
87. Kost RG, Kupinsky H, Straus SE. Varicella-zoster virus gene 63: transcript mapping and regulatory activity. *Virology*. 1995;209:218–24.
88. Hoover SE, Cohrs RJ, Rangel ZG, Gilden DH, Munson P, Cohen JL. Downregulation of varicella-zoster virus (VZV) immediate-early ORF62 transcription by VZV ORF63 correlates with virus replication in vitro and with latency. *J Virol American Society for Microbiology (ASM)*. 2006;80:3459–68.
89. 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.
90. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, et al. Antisense transcription in the mammalian transcriptome. *Science*. 2005;309:1564–6.
91. 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.
92. Huvel 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.
93. Jones JO, Sommer M, Stamatis S, Arvin AM. Mutational analysis of the varicella-zoster virus ORF62/63 intergenic region. *J Virol. American Society for Microbiology (ASM)*. 2006;80:3116–21.
94. Tycowski KT, Guo YE, Lee N, Moss WN, Vallery TK, Xie M, et al. Viral noncoding RNAs: More surprises. *Genes Dev*. 2015;29:567–84.
95. Qiu L, Wang T, Tang Q, Li G, Wu P, Chen K. Long non-coding RNAs: regulators of viral infection and the interferon antiviral response. *Front Microbiol*. 2018;9:1621.
96. Meng XY, Luo Y, Anwar MN, Sun Y, Gao Y, Zhang H, et al. Long non-coding RNAs: emerging and versatile regulators in host-virus interactions. *Front Immunol*. 2017;8:1663.
97. 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.
98. 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 American Society for Microbiology (ASM)*. 2007;81:13544–51.
99. 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.
100. 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.
101. Nishikura K. Editor meets silencer: crosstalk between RNA editing and RNA interference. *Nat Rev Mol Cell Biol*. 2006;7:919–31.
102. Scadden AD, Smith CW. RNAi is antagonized by A-to-G hyper-editing. *EMBO Rep European Molecular Biology Organization*. 2001;2:1107–11.
103. Scadden ADJ, Smith CW. Specific cleavage of hyper-edited dsRNAs. *EMBO J*. 2001;20:4243–52.
104. 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.
105. Depledge DP, Sadaoka T, Ouwendijk WJD. Molecular aspects of varicella-zoster virus latency. *Viruses*. 2018;7:349.
106. Cohen JL, 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.
107. 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.
108. 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.

Full Genome Sequence of the Western Reserve Strain of Vaccinia Virus Determined by Third-Generation Sequencing

István Prazsák,^a Dóra Tombácz,^a Attila Szűcs,^a Béla Dénes,^b Michael Snyder,^c Zsolt Boldogkői^a

^aDepartment of Medical Biology, Faculty of Medicine, University of Szeged, Szeged, Hungary

^bVeterinary Diagnostic Directorate of the National Food Chain Safety Office, Budapest, Hungary

^cDepartment of Genetics, School of Medicine, Stanford University, Stanford, California, USA

ABSTRACT The vaccinia virus is a large, complex virus belonging to the *Poxviridae* family. Here, we report the complete, annotated genome sequence of the neurovirulent Western Reserve laboratory strain of this virus, which was sequenced on the Pacific Biosciences RS II and Oxford Nanopore MinION platforms.

Poxviruses are generally brick-shaped (1), enveloped viruses that have a complex internal structure, including a relatively large double-stranded DNA genome and associated enzymes (2). In contrast to many other DNA viruses, poxviruses replicate and express their genomes within the cytoplasm rather than in the nucleus of the infected cell (3).

The vaccinia virus (VV) is a historically interesting and significant virus; it has been successfully used as a vaccine for immunization against human smallpox (4–6), which was declared eradicated in 1980 by the WHO thanks to global vaccination efforts (7). The linear genome of VV is approximately 190 kb in length; it is flanked by inverted terminal repeat sequences and encodes around 220 protein-coding genes (8). The most virulent strain of VV, the Western Reserve (WR) (9), was used for our study. The currently available third-generation (long-read) sequencing platforms, namely, the Pacific Biosciences (PacBio) RS II system and the Oxford Nanopore Technologies (ONT) MinION system, were used for the cDNA sequencing.

The CV-1 cell line (ATCC CCL-70) was infected with the laboratory WR strain of VV. The poly(A) fraction of the purified RNA was converted to cDNA following the Isoform Sequencing (Iso-Seq) protocol for PacBio sequencing and the cDNA-seq protocol (1D strand switching cDNA by ligation) for the MinION process. Seventeen single-molecule real-time (SMRT) cells with P6-C4 chemistry, as well as two MinION flow cells, were run on the PacBio and ONT systems, respectively. The PacBio sequencing resulted in 59,154 reads, while 16,175 reads were obtained from the MinION runs. Our data revealed extreme transcriptional activity across the entire VV genome. Both of the sequencing platforms that we used in this study cover the full-length viral DNA. The average coverages were 289.3× (PacBio) and 43.6× (ONT), and the mean sizes of the aligned read lengths were 953 bp (PacBio) and 525 bp (ONT).

The PacBio raw reads were processed and mapped to the respective reference genome (GenBank accession no. NC_006998) with the BLASR long-read mapper (<https://github.com/PacificBiosciences/blasr>). GMAP (10) was also used for aligning the reads. The MinION sequencing reads were mapped to the reference genome by using Albacore version 1.2.6 software. The genome was assembled and the open reading frames were predicted by using Geneious software (11).

The complete genome sequence of the WR strain of VV is composed of 194,888 bp. The average G+C content of the genome is 33.3%. The virus contains 218 protein-coding genes. Our data show that the sequenced WR strain differs in 163 point

Received 18 December 2017 **Accepted** 23 February 2018 **Published** 15 March 2018

Citation Prazsák I, Tombácz D, Szűcs A, Dénes B, Snyder M, Boldogkői Z. 2018. Full genome sequence of the Western Reserve strain of vaccinia virus determined by third-generation sequencing. *Genome Announc* 6:e01570-17. <https://doi.org/10.1128/genomeA.01570-17>.

Copyright © 2018 Prazsák et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](#).

Address correspondence to Zsolt Boldogkői, boldogkoi.zsolt@med.u-szeged.hu. I.P. and D.T. contributed equally to this work.

mutations—among which were 58 nonsilent mutations—from the GenBank reference sequence. Eighteen insertions and four deletions were also detected.

The most common point mutation is a substitution of cytosine for thymidine. Most of the amino acid substitutions occurred in the C9L gene. A new coding sequence (A54L) was annotated. The majority of nucleic acid changes (21 events) were clustered in the F region of the viral genome. Mutations mainly occurred in viral genes playing a role in host defense modulation, transcription regulation, and viral replication.

Accession number(s). The complete and annotated genome sequence of the WR strain of VV has been deposited in the European Nucleotide Archive under accession no. [LT966077](#).

ACKNOWLEDGMENTS

This research was supported by the European Union and the State of Hungary through the European Social Fund under the framework of grant TÁMOP-4.2.6-14/1 to Z.B., by NIH Centers of Excellence in Genomic Science (CEGS) Center for Personal Dynamic Regulomes grant 5P50HG00773502 to M.S., and by a Bolyai János Scholarship of the Hungarian Academy of Sciences to D.T.

REFERENCES

1. Baxby D. 1996. Poxviruses, chapter 69. In Baron S (ed), Medical microbiology, 4th ed. University of Texas Medical Branch at Galveston, Galveston, TX. <https://www.ncbi.nlm.nih.gov/books/NBK8364>.
2. Condit RC, Moussatche N, Traktman P. 2006. In a nutshell: structure and assembly of the vaccinia virion. *Adv Virus Res* 66:31–124. [https://doi.org/10.1016/S0065-3527\(06\)66002-8](https://doi.org/10.1016/S0065-3527(06)66002-8).
3. Reeves PM, Smith SK, Olson VA, Thorne SH, Bornmann W, Damon IK, Kalman D. 2011. Variola and monkeypox viruses utilize conserved mechanisms of virion motility and release that depend on Abl and SRC family tyrosine kinases. *J Virol* 85:21–31. <https://doi.org/10.1128/JVI.01814-10>.
4. Moss B. 1991. Vaccinia virus: a tool for research and vaccine development. *Science* 252:1662–1667. <https://doi.org/10.1126/science.2047875>.
5. Paran N, Suez Y, Lustig S, Israely T, Schwantes A, Melamed S, Katz L, Preuss T, Hanschmann KM, Kalinke U, Erez N, Levin R, Velan B, Löwer J, Shafferman A, Sutter G. 2009. Postexposure immunization with modified vaccinia virus Ankara or conventional lister vaccine provides solid protection in a murine model of human smallpox. *J Infect Dis* 199:39–48. <https://doi.org/10.1086/595565>.
6. Jacobs BL, Langland JO, Kibler KV, Denzler KL, White SD, Holechek SA, Wong S, Huynh T, Baskin CR. 2009. Vaccinia virus vaccines: past, present and future. *Antiviral Res* 84:1–13. <https://doi.org/10.1016/j.antiviral.2009.06.006>.
7. Davidson T. 2017. Vaccines: history, science, and issues. Greenwood, Santa Barbara, CA.
8. Chung CS, Chen CH, Ho MY, Huang CY, Liao CL, Chang W. 2006. Vaccinia virus proteome: identification of proteins in vaccinia virus intracellular mature virion particles. *J Virol* 80:2127–2140. <https://doi.org/10.1128/JVI.80.5.2127-2140.2006>.
9. Zeh HJ, Downs-Canner S, McCart JA, Guo ZS, Rao UNM, Ramalingam L, Thorne SH, Jones HL, Kalinski P, Wieckowski E, O'Malley ME, Daneshmand M, Hu K, Bell JC, Hwang TH, Moon A, Breitbach CJ, Kirn DH, Bartlett DL. 2015. First-in-man study of Western Reserve strain oncolytic vaccinia virus: safety, systemic spread, and antitumor activity. *Mol Ther* 23: 202–214. <https://doi.org/10.1038/mt.2014.194>.
10. Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21:1859–1875. <https://doi.org/10.1093/bioinformatics/bti310>.
11. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28:1647–1649. <https://doi.org/10.1093/bioinformatics/bts199>.