Dynamic transcriptome landscape of Vaccinia virus revealed by long-read sequencing approach

István Prazsák[1], Dóra Tombácz[1] , Zsolt Balázs[1], Zoltán Maróti[2], Tibor Kalmár[2], Norbert Moldován[1], Attila Szűcs[1], Zsolt Csabai[1], Michael Snyder[3] ,Béla Dénes[4] and Zsolt Boldogkői[1]

[1]Department of Medical Biology, Faculty of Medicine, University of Szeged, Hungary, [2]Department of Pediatrics and Pediatric Health Center, Faculty of Medicine, University of Szeged, Hungary, [3]Department of Genetics, School of Medicine, Stanford University, California, USA, [4]Veterinary Diagnostic Directorate of the National Food Chain Safety Office, Budapest, Hungary

Introduction

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. Vaccinia virus (VACV) is the prototypic member of the Orthopoxvirus genus. It is closely related to the *Variola virus* that was eradicated as a result of a successful global vaccination program using live VACV. VACV is extensively utilized as an expression and a gene delivery vector, e.g. in oncolytic treatments. It is also known as a model system for the analysis of virus-host interactions. Poxviruses are able to replicate independently in the cytoplasm of the host cell because they encode own DNA Polymerase. The virus has a relatively large (approximately 195 kbp) double-stranded DNA genome coding for about 220 proteins. The VACV genes are divided into two main temporal classes: pre-replicative and post-replicative, they are further grouped into early (E1 and E2), intermediate (I), and late (L) genes, according to Bernard Moss.

Aims

Several traditional methods including short-read sequencing techniques (SRS) has been already used to resolve transcriptional start and end sites (TSSs, TESs) of VACV, however variable read-trough or polycistronism has remained hidden in the VACV transcriptome. Despite the scientific relevance of VACV, no long-read sequencing techniques (LRS) have been adopted for VACV transcriptome to date. The aim of this study was the transcriptional reannotation of VACV, description of potential new TSSs and TESs and discovery of transcript length isoforms.

Keywords: full-length, long-read sequencing, Oxford Nanopore Technologies, Pacific Biosciences, transcript kinetics, viral life cycle, Vaccinia virus, Poxviridae

Methods

African green monkey kidneys fibroblast cell culture (*CV-1 cell line* derived from Cercopithecus aethiops) was infected with VACV strain WR until cytopathic effect has been appeared. Cells were harvested in different time points p.i. than lysed and total RNA was isolated according with the manufacturer's protocol (Macherey Nagel, NucleoSpin® RNA isolation kit). RNA isolation was controlled by RT real time PCR with gene specific primers. Before sequencing poly(A)-selection of mRNAs was done according with the library preparation instructions of the Pacific Biosciences (PacBio) Iso-Seq method using the Clontech SMARTer PCR cDNA Synthesis Kit and No Size Selection' or the "BluePippin size-selection" protocol. An other series of cDNA libraries were prepared in accordance to the Oxford Nanopore Technologies (ONT - 1D strand switching cDNA by ligation) kit and in a parallel sample Cap-selection of 5'-ends of mRNAs was prepared using the Cap-specific Lexogen –kit (Fig 1). Long read sequencing was performed on ONT MinIon, PacBio RSII and Sequel platforms. Minimap2 was used to align sequenced reads. Degraded viral mRNA, strand switching of reverse transcriptase, incorrect binding of sequencing primers and failed alignment of mapping program all can produce false TSS, TES and splice sites. We used biostatistics filtration of reads to determine the subset of reads bearing detectable barcodes, and primers at both read ends, or poly(A)-tail at the 3' ends and our in house developed python package the LoRTIA ((*https://github.com/zsolt-balazs/LoRTIA*) bioinformatics pipeline was used to analyze transcriptomic signals and annotate the transcripts of host and VACV.

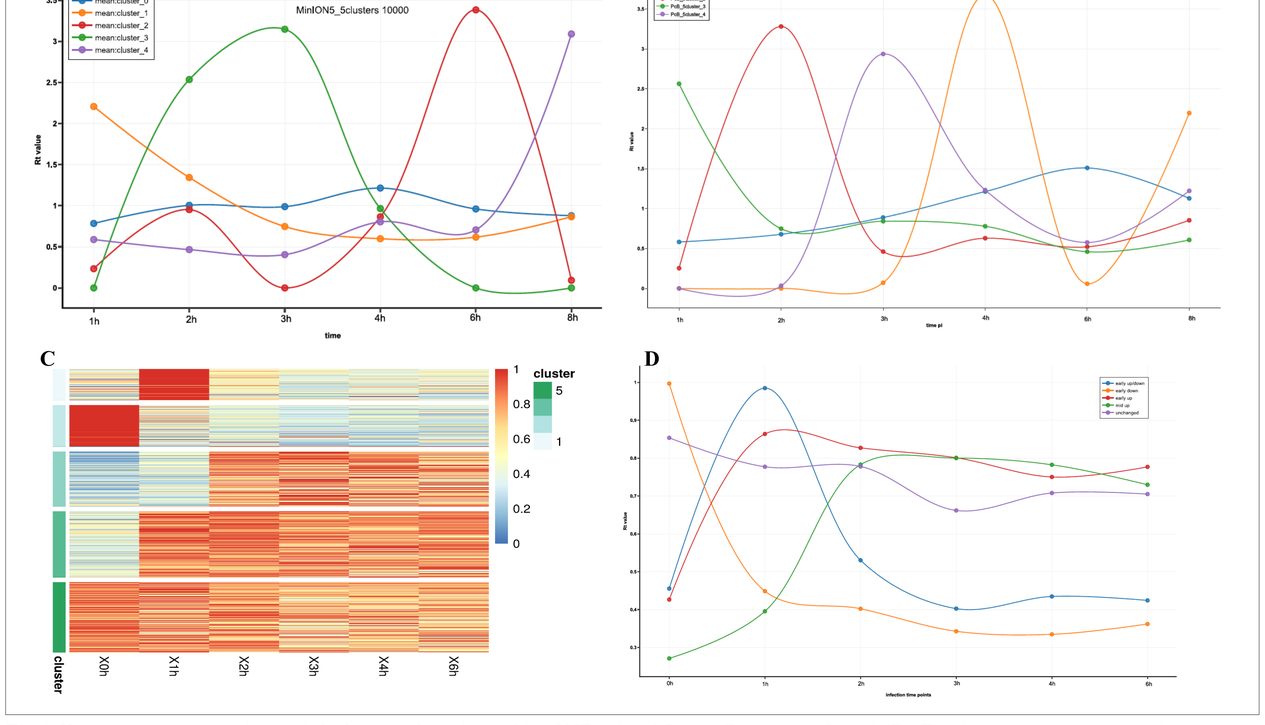

Fig. 3 Hierarchical clustering of transcripts obtained by Oxford Nanopore Sequencing (ONT; A) and Pacific Biosciences Sequel (B). The dataset contains annotated transcripts with a minimum read count of 5 reads. Five distinct clusters characterize the viral transcriptome. Based on the number of clusters, k-means clustering was generated, the mean of clusters is represented on **Fig. 4**.

Fig 4. K-means clustering analysis of **viral transcripts** obtained by ONT (A) and Pacific Biosciences Sequel (B). The dataset contains annotated transcripts with a minimum read count of 5 reads. We classified VACV transcripts into five kinetic groups: *Early group* with a continuously inclined Rt values from the beginning of the infection, the *second cluster* of viral transcripts shows a local maximum at 2h p.i., confirming previous findings of VACV transcriptomic analysis, after the replication a *third cluster* appears reaching a maximum of gene expression at 3h and 4h p.i. , *the fourth cluster* of transcripts has a clear *late* activity, while many of the viral transcripts show slight variability, containing transcripts with *constant* gene activity. We tried to render functions to the transcripts, some of them are well characterized gene products of VACV, although many transcripts has been classified kinetically first time with an unknown function (**Table 1**). Heatmap and k-means clustering of **host transcripts** based on the ONT dataset (C), (D). Results of ONT sequencing show no size selection bias of kinetic classes, compared to PacBio results (data not shown).

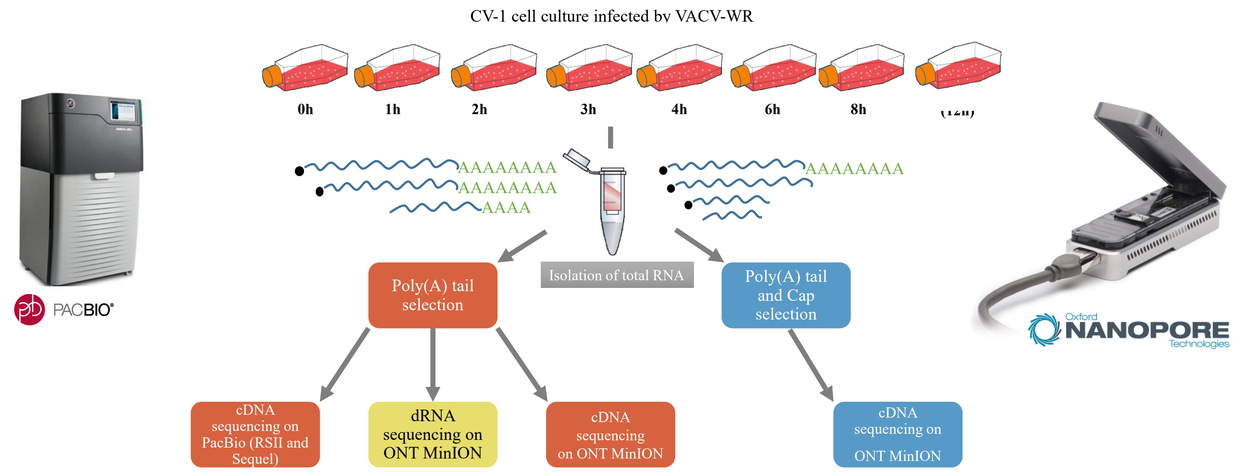

Fig 1. Schematic representation of experimental design. RNA was isolated in 8 different time points p.i. with poly(A) tail selection combined with Cap-selection to reveal specific 5' and 3' ends of transcripts during the course of infection. Long-read sequencing was conducted on different platforms to exclude technical bias of size selection.

Kinetic class	Maximum of activity	Viral genes	Function (according to UniProtKB)
cluster_0	intermediate	C6L	Prevents cellular antiviral state by blocking of interferon regulatory factors
		F4L	necessary for viral DNA synthesis
		H5R	generates homogeneous 3' ends of transcripts during transcription
		C10L	IL-1 receptor antagonist
		C11R	Epidermal growth factor-like protein (EGF-like protein)
cluster_1	early	E3L	inhibition of multiple cellular antiviral responses activated by dsRNA
		F11L	Stimulates microtubule dynamics and the motility of the host cells
		A37	unknown
		A46R	BCL2-like protein, disrupts the host immune response
		B15R	reduce the host inflammatory response by interacting with inteleukin-1 beta
		K3L	Viral mimic of eIF-2-alpha, prevents protein synthesis shutoff
cluster_2	late	O2L	Displays thioltransferase and dehydroascorbate reductase activities
		VACVWR_0015	unknown
		C1L	Poly(A) polymerase catalytic subunit
		H2R	Envelope protein part of the entry-fusion complex
		A33	plays an essential role in efficient cell-to-cell spread of viral particles

Table 1. The most typical genes in the identified clusters of viral transcriptome confirmed by both of the sequencing techniques

Cluster	Best hits	FDR
early up (65 genes)	no hit with <1 FDR	NA
early down (38 genes)	GO biological process: mesenchymal differentiation (5/9); STAT1, NRP1, TRIM28, TGFB2, SEMA3A	0.863
mid up (48 genes)	GO biological process: mitotic spindle assembly (4/4); SMC3, TPR, HNRNPU, CHMP2A	0.193
	GO biological process: positive regulation of viral life cycle (6/11); CHMP2A, TOP2A, FMR1, NUCKS1, LGALS1, SMC3	0.251
early up/down (11 genes)	GO biological process: regulation of signaling receptor activity (5/19); PHLDA2, HBEGF, CXCL3, CXCL8, CTGF	0.041 4*
	Panther protein class: signaling molecule	0.04*
	Panther Pathways: CCKR signaling map	0.007*

Table 2. Summary of the over representation analysis of the most typical genes in the identified clusters of host transcriptome reveled by ONT platform.

In the 768 highly expressed genes of the host organism we identified 4 distinct cluster of genes (early up – no or very low expression before virus infection, and constantly high expression at all later sampling points, early down – high expression before virus expression and constantly 0 or low expression at all later sampling points, early up/down no or low expression before virus infection, high expression at 1 hour sampling and no or low expression at later sampling points and mid up – no expression before virus infection that maximizes till 2 or 3 hours of sampling times). Over expression analysis identified significant cluster of genes only in the early-up/down gene cluster , however we also identified a few plausible set of genes in the mid-up cluster (**Table 2**).

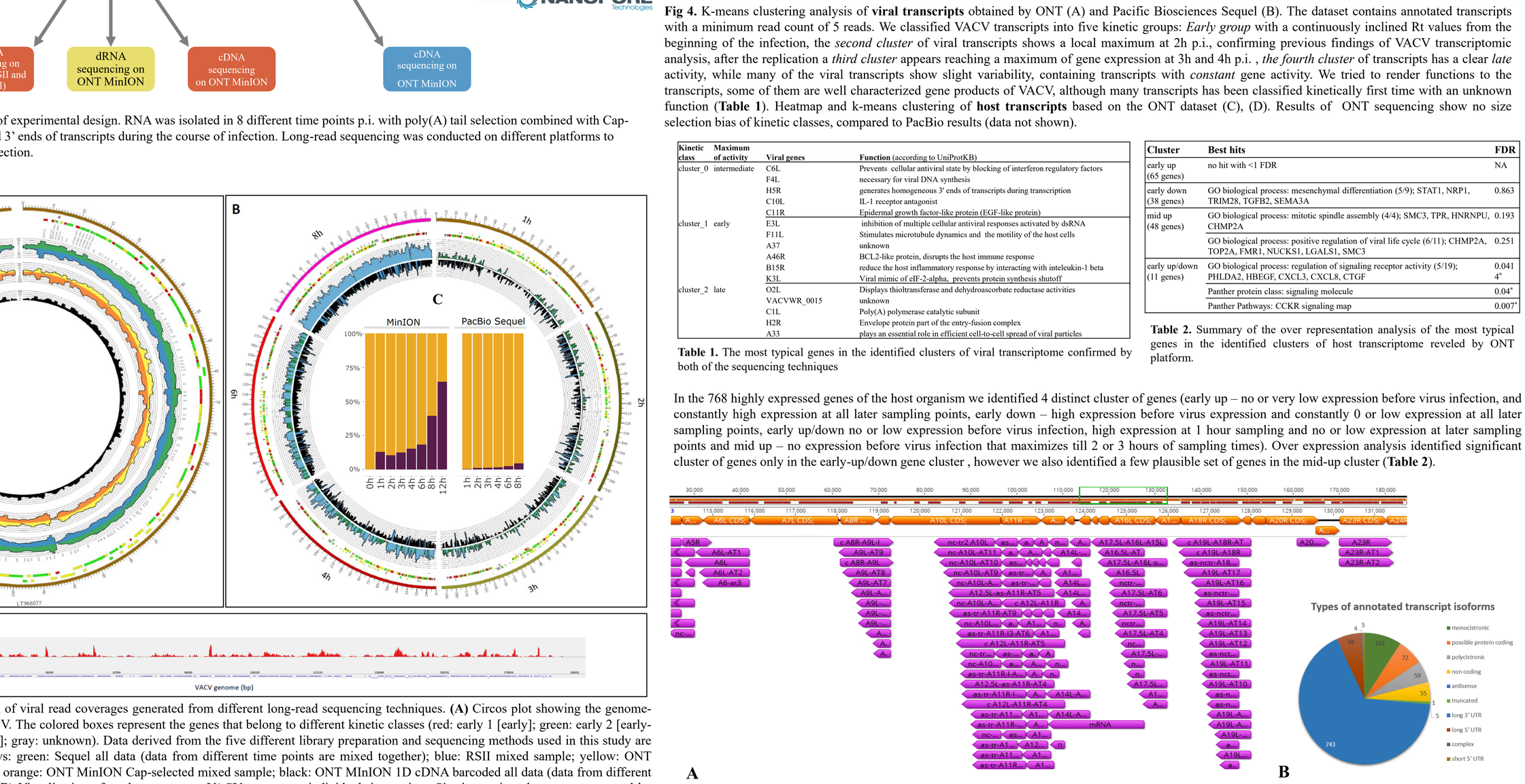

Fig. 5 (A) Excerption of VACV transcriptome at the irregular region of genome visualized in Geneious. The antisense RNA of A11R CDS and its 3' and 5' UTR isoforms are intensively expressed. It is almost impossible to distinguish the different transcripts of the given CDSs. Compared to other genomic regions this antisense hyper activity is a yet not known phenomenon in VACV. **(B)** Groups of novel VACV transcripts annotated in this study.

Fig 2. 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: post replicative [late]; gray: unknown). 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: ONT MinION 1D cDNA mixed sample; orange: ONT MinION Cap-selected mixed sample; black: ONT 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 (outer radius) have been visualized in a segmented circos plot (every segment represents an individual time point). **(C)** The fraction of reads mapping to the host (C. sabaeus) and the VACV genome. A reduction of read counts of the host and increase of read counts of the virus is observable in the ONT MinION data as the viral life cycle progresses. **(D)** Sashimi plot presentation of the ONT MinION dRNA-seq data across the VACV genome.

Summary:

Vaccinia virus lytic transcriptome and green monkey host transcriptome was sequenced by Oxford Nanopore Technologies and Pacific Biosciences long read sequencing platform based on Cap-selected, poly(A)-selected and direct mRNA isolation protocol in different time points of infection. Transcriptomic noise, false priming, failed mapping, template switching was filtered by *LoRTIA* program package. By obtaining full length transcripts we determined already known and new TSS and TES positions with base pair precision and first time annotated transcripts throughout the entire VACV genome. We detected an extensive read-trough in closely situated CDSs at the 3' ends forming bi-and tricistronic or nested gene clusters in especially in late phase of infection, and found novel possibly short protein coding RNA molecules called .5 variants of ORFs. A peculiar outcome of long read sequencing of Vaccinia transcriptome was finding of the so called *irregular or chaotic region*, which can be characterized by a hyper-intensive transcript initiation and termination forming hundreds of transcript isoforms in the O1L-O2L-I1L and A11R-A12R-A13L genomic region. We classified the viral and host transcripts according to their temporally expression and found 5 clusters. Early transcripts can be differentiated clear from the transcripts occurring late in infection, many of the transcripts remaining unchanged, however some transcript isoforms can be expressed by a time dependent manner.

Conclusion: Our data suggests a highly complex transcriptomic pattern in VACV than it was previously known, with an unexpected stochasticity in the post-replicative phase. These characteristics of transcription are uncommon among large DNA viruses, they might show a new level of viral gene expression regulation.

To see our publications in the topic of VACV transcriptomics scan the QR-code and click on the URLs:

The experiments were supported by the National Research, Development and Innovation Fund, Hungary- grant no: K 128247 and FK 128252