Exploring the Transcriptomic Profile of Human Monkeypox Virus via CAGE and Native RNA Sequencing Approaches

Gergely Ármin Nagy[1#], Dóra Tombácz[1#], István Prazsák[1], Zsolt Csabai[1], Ákos Dörmő[1], Gábor Gulyás[1], Gábor Kemenesi[2,3], Gábor E. Tóth[2,3,4], Jiří Holoubek[5,6,7]; Daniel Růžek[5,6,7], Balázs Kakuk[1], Zsolt Boldogkői[1]*

[#]These authors contributed equally to this work

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

[1] Department of Medical Biology, Albert Szent-Györgyi Medical School, University of Szeged, Somogyi B. u. 4., 6720 Szeged, Hungary

[2] National Laboratory of Virology, Szentágothai Research Centre, University of Pécs, Pécs, Hungary

[3] Institute of Biology, Faculty of Sciences, University of Pécs, Pécs, Hungary

[4] Bernhard Nocht Institute for Tropical Medicine, WHO Collaborating Centre for Arbovirus and Hemorrhagic Fever Reference and Research, Hamburg, Germany

[5] Veterinary Research Institute, Hudcova 70, CZ-62100 Brno, Czech Republic

[6] Institute of Parasitology, Biology Centre of the Czech Academy of Sciences, Branisovska 31, CZ-37005 Ceske Budejovice, Czech Republic

[7] Department of Experimental Biology, Faculty of Science, Masaryk University, Kamenice, 753/5, Brno CZ-62500, Czech Republic

Running title: Transcriptomic architecture of human Monkeypox virus

Abstract

In this study, we employed short- and long-read sequencing technologies to delineate the transcriptional architecture of the human monkeypox virus and to identify key regulatory elements that govern its gene expression. Specifically, we conducted a transcriptomic analysis to annotate the transcription start sites (TSSs) and transcription end sites (TESs) of the virus by utilizing cap analysis of gene expression sequencing on the Illumina platform and direct RNA sequencing on the Oxford Nanopore technology device. Our investigations uncovered significant complexity in the use of alternative TSSs and TESs in viral genes. In this research, we also detected the promoter elements and poly(A) signals associated with the viral genes. Additionally, we identified novel genes in both the left and right variable regions of the viral genome.

Importance

Generally, gaining insight into how the transcription of a virus is regulated offers insights into the key mechanisms that control its life cycle. The recent outbreak of the human monkeypox virus has underscored the necessity of understanding the basic biology of its causative agent. Our results are pivotal for constructing a comprehensive transcriptomic atlas of the human monkeypox virus, providing valuable resources for future studies.

INTRODUCTION

Orthopoxvirus, a genus in the Poxviridae family, encompasses several significant human and animal pathogens. Orthopoxviruses include several species, most notably the variola virus, which causes smallpox; the monkeypox virus (MPXV); the cowpox virus; and the vaccinia virus (VACV), which is known for its use in smallpox vaccination (1–3). Over the course of centuries, smallpox claimed millions of lives until its successful eradication, thanks to an extensive worldwide vaccination initiative (4). Monkeypox virus can cause human disease, although the symptoms are typically mild (5). The human monkeypox virus (hMPXV), responsible for the 2022 outbreak, originated from zoonotic transmission from animals to humans (6). Genetic analysis has shown that hMPXV corresponds to the less virulent West African clade of MPXV (7, 8). Despite a decline in monkeypox cases in recent years, the risk of a future outbreak should not be underestimated.

Orthopoxviruses have a large linear double-stranded DNA genome, approximately 200 kilobase pairs long (9). Unlike the majority of mammalian DNA viruses, including herpesviruses and adenoviruses, which replicate in the nucleus, poxviruses, along with the African swine fever virus, replicate in the cytoplasm. The replication and transcription processes of poxviruses are carried out within specialized structures known as "viral factories" (10). The regulation of viral gene expression is governed by transcription factors specific to different stages, which bind selectively to the promoters of early (E), intermediate (I), and late (L) genes (11). The full transcription machinery is pre-packaged within the poxvirus virion, which allows for immediate expression of E genes once the virus has entered the cell and while the viral genome is still encapsulated. This is then followed by DNA replication and the subsequent expression of I and L gene classes, collectively termed as post-replicative (PR) genes. E genes are responsible for encoding proteins that synthesize DNA and RNA molecules and those that play a part in the interactions between the virus and the host. Meanwhile, PR genes primarily encode the structural elements of the virus (12).

Unlike herpesviruses, which tend to produce 3'-co-terminal transcripts by the adjacent tandem genes, poxviruses generate a vast diversity of 3'-ends (13), especially during the late stages of infection (14). The lack of splicing in poxvirus transcripts is attributed to their replication in the cytoplasm (15). Poxviruses have the unique ability to produce their own enzymes for capping, decapping, and polyadenylation, and they employ strategies such as mRNA decapping to inhibit host translation (16). Though poxvirus mRNAs generally resemble host mRNAs in structure, one distinctive trait is the presence of 5'-poly(A) leaders in PR mRNAs (17). Recent studies have revealed that poly(A) leaders provide the capability to utilize either cap-dependent or cap-independent translation initiation(18).

Several studies have explored the transcriptional impact of hMPXV infection across various cell types, predominantly utilizing micro-array-based techniques (19–22). These pioneering works have laid the foundation for the understanding of the viral transcription landscape. A notable limitation of the micro-array-based techniques is their inability to resolve important aspects of the transcriptome, particularly to detect the transcript isoforms (23). Determining the exact genomic location of the transcription start sites (TSSs) and transcription end sites (TESs) of the mRNAs is crucial in annotating viral genomes. Methods, such as S1 nuclease treatment with labeled probe-hybridization (24, 25) have been developed as early attempts to determine both 5'-ends (26–30) and 3'-ends (31–33) of poxviral mRNAs. Rapid amplification of cDNA ends (RACE) (34) and its modified versions [New-RACE (35), Target-RACE (36) and circular-RACE (37) are widely used PCR-based methods to identify both ends of cDNA transcripts. RACE was used to determine transcript boundaries in Poxviruses (38). Detection of cap is utilized in transcriptome research for identifying transcription initiation (39–41). Although microarray and PCR-based techniques offer high precision, they are limited to

analyzing only those transcripts for which probes or gene-specific primers exist. In contrast, total RNA-sequencing methods allow for the examination of the entire transcriptome. With the advent of next-generation sequencers, the bulk analysis of whole transcriptome features, including TSSs and TESs became possible.

Poxviruses are unique among the viruses, because they have their own capping (42) and decapping enzymes (43, 44). Cap Analysis of Gene Expression (CAGE) (45) uses cloned tags at the 5'-end of mRNAs (46, 47) for selective detection of capped RNA ends. Originally, CAGE was developed for Sanger-sequencing, but later it was adapted for high-throughput sequencing methods (48, 49) to reproduce a global snapshot of transcriptional start points and promoter elements (50, 51). CAGE sequencing (CAGE-Seq) is optimal for short-read sequencing (SRS) (52). Many variations of this approach have been developed so far (53–55). CAGE-Seq has also been applied to explore the poxviral mRNA 5'-ends (56) and promoter elements (57). While polyA-selected SRS was used to identify 3'-transcript ends (56, 58). SRS provides a high-throughput, base-precision map of transcriptional activity. However, reverse transcription-dependent techniques are unable to circumvent the drawbacks occurring during cDNA-syntheses, such as e.g. template switching (59, 60), false priming (61), or spurious antisense transcription (62).

Long-read sequencing (LRS) methods, such as single-molecule real-time (SMRT) and nanopore sequencings (63, 64) are able to read entire mRNAs, making them indispensable in transcriptome research (65, 66). Oxford Nanopore technologies (ONT) allows for the direct sequencing of native RNA molecules (dRNA-Seq). This approach eliminates the generation of false products that may arise during the library preparation process, specifically during the reverse-transcription, second strand synthesis, and PCR steps. The limitation of this technique is its reduced precision in annotating the 5'-ends of mRNAs (67). This issue can be mitigated by integrating dRNA-Seq with 5'-end sensitive PCR-free direct cDNA (dcDNA) sequencing (dcDNA-Seq) or CAGE-Seq methods (56, 68–71).

LRS cDNA-Seq approach has been applied for the analysis of dynamic VACV transcriptome (14, 23, 72, 73). Host cell transcriptome was recently inferred upon hMPXV infection (74). However, the transcriptome of hMPXV itself has not been analyzed.

Our objective in this study was to identify transcription start sites (TSSs) and transcription end sites (TESs) of hMPXV which helps to annotate the complete viral transcriptome. Furthermore, we identified the promoter and poly(A) signal consensus elements of the hMPXV genes.

RESULTS

We employed two distinct sequencing approaches to identify the terminal regions of the hMPXV transcripts. TSSs were detected using CAGE-Seq on the Illumina MiSeq platform, whereas TESs were identified through dRNA-Seq on the ONT MinION device.

Transcription Start Sites

CAGE-Seq analysis identified a total of 3,676 TSS positions excluding the singletons (**Supplementary Table 1**). Although dRNA-Seq efficiently validates 5'-ends but encounters challenges due to incomplete sequencing of these termini (67). However, unlike cDNA-Seq techniques, dRNA-Seq is free from common artifacts. Therefore, we opted to utilize this method to validate the results obtained from CAGE-Seq (**Figure 1A, Supplementary Figure 1**). A total of 2,625 transcription start sites (TSSs) were confirmed by dRNA-Seq within a 25-nucleotide window, likely representing an underestimate of the overall TSSs. We analyzed the distribution of dRNA-Seq read ends in the proximity of CAGE-Seq signals and found that the 5'-ends detected by dRNA-Seq were most frequently positioned on average 11 nucleotides downstream from the TSSs identified by CAGE-Seq (**Figure 1B**). The missing nucleotides at

the 5′ end result from the premature release of the RNA molecules by the motor protein. We further filtered the 2,625 positions by eliminating those with fewer than 10 supporting reads, resulting in a total of 720 positions by excluding those supported by fewer than 10 reads (**Supplementary Figure 2**). Subsequently, we analyzed which of these positions were within 40 nt upstream of a predicted promoter. This latter analysis yielded a final count of 401 positions (**Figure 2**).

Furthermore, employing another novel TSS clustering algorithm within the TSSr package (peakclu), we identified 646 clusters of CAGE signals, each with a single dominant peak (**Supplementary Table 2**). Comparing these dominant peaks of the clusters with the dataset of 401 filtered TSS positions, we identified a set of CAGE signals comprising 133 positions that met all the filtering criteria (**Figure 1 C and D; Supplementary Table 2.**).

This shows that both clustering and unclustering of CAGE signals lead to robust TSS detection, demonstrating their consistency. Using the shape score index, peak analysis of CAGE-Seq data revealed two major types of TSS distributions: broad and narrow range. The analysis indicated that the majority of the clusters consist of single peaks, with the vast majority the clusters does not surpass 10 nt in width (**Figure 3**).

The distinguishing characteristic of poxviral mRNAs is the presence of poly(A)-leader sequences at the 5'-ends of late mRNAs (75, 76). Despite the absence of 11 nt on average at the 5'-end of dRNA-Seq reads, the presence of a 5'-poly(A) leader enables the sequencing of the entire molecule, as shown in **Figure 3C**. We estimated the number of 5'-poly(A) leaders and found that 10% of CAGE-Seq reads and 5% of dRNA-Seq reads contain at least 3 A bases **(Figure 4)**.

TSS positions were sorted according to their abundance. The top 5 TSSs surpass a read depth of 1,000 in CAGE-Seq (**Supplementary Table 3**). Among these, three TSSs stand out with exceptionally high CAGE-Seq signals, each showing count values exceeding 10,000. The highest CAGE-Seq signal represents 13% of the total and nearly 44% of the count for the top 5 TSSs. In dRNA-Seq, the most abundant 5'-end position belongs to the gene OPG110, which encodes the ankyrin-motif containing protein D8L. Out of the most abundant 5'-CAGE-Seq positions, three coincided with the most abundant dRNA-Seq positions belonging to the following genes: OPG065, OPGOPG110 and OPG022. **Supplementary Table 3** provides a summary of the orthologues and functions of genes associated with the most abundant TSS positions.

Promoter elements

Our understanding of promoter elements in Orthopoxviruses primarily stems from research on VACV (77, 78). Poxviruses use distinct promoter motifs in the early and late phase of infection (79). Given the close phylogenetic relationship between VACV and hMPXV (80, 81), the promoter motifs of the former virus were employed to identify corresponding elements in hMPXV (**Table 1**).

Type	Kinetics	Consensus	Reference
Promoter	Early	AAAANTGAAAANNA	Yang et al. 2010
	Late	TAAATG/NNNTNNNNNNNNNTAAATG	Rosel et al. 1986; Yang et al. 2011;
	Group I	NNNNNNNNNNYNWNWWWTWWWNNNNNWTAAATG	Yang et al. 2011;
	Group IIB	NATWWNWNNNHTAAAAANNDNNNNHNNDWWNTAAAYN	
	Group IIA	NRNNWNWTNWMWNWWWWTDNNNNH	

	Intermediate	NNNATNNNNNNNNTAAAAANNNNNNNNNNNNNTAAA	Yang et al. 2013
	Mixed	NNNA/TNNNNNNNNTNNNNNNNNNTAAATGGNNN	Yang et al 2011
	Mixed	NTAWAD	Tombácz and Prazsák et al. 2020
PAS	Early	UUUUUNU	Yang, Reynolds et al. 2011; Yang et al 2012

Table 1.

The promoter motifs used to scan viral promoter and PAS sequences are categorized by their kinetics, based on data from literature on experiments related to VACV gene expression. PAS stands for poly(A) signal.

We identified 1,369 putative promoters within a 100-nt interval upstream of TSSs using the FIMO (Find Individual Motif Occurrences) program. The resulting predicted promoters, along with their p- and q-values, are listed in **Supplementary Table 1c**. The best-matching motifs, associated with the names of ORFs, are organized according to their q-values and detailed in **Supplementary Table 1d**. The average distance between each TSS and its predicted promoter motif was determined to be 26 nucleotides, with the most frequent distance observed being 1 nucleotide (**Figure 5A**). This finding is consistent with results from studies conducted on VACV (56).

Transcription End Sites

Direct RNA sequencing, based on poly(A) selection, was employed to identify the 3'-ends of hMPXV RNAs, using the LoRTIA (82) tool for TES annotation. A total of 3,241 positions were identified (excluding singlets), with 496 of these positions validated by a minimum of 6 reads (**Supplementary Figure 3**). Among these, 135 positions were further validated by ePAS signals within 50-nucleotide distance **(Figure 6 and Supplementary Table 4**).

Poly(A) signals

Orthopoxviruses utilize their unique enzymatic machinery to recognize polyadenylation signals (PASs) and to synthesize the poly(A)-tail of viral mRNAs. VACV early mRNAs are characterized by a UUUUUNU early PAS (ePAS), leading to a premature and homogenous end of early mRNAs (56, 83). Using a motif scanning algorithm (FIMO), we identified 734 ePASs, as detailed in **Supplementary Table 4.** Of these, 313 ePASs were found 50 nt upstream of TESs, validating 135 of the previously mentioned 496 TESs, as reported in **Supplementary Table 4**. The average distance of ePAS from TESs is 24 nt, which is in concordance with VACV data (56, 57). One benefit of dRNA-Seq is its ability to directly analyze the native poly(A) tails of RNAs. In the analysis of 232,258 hMPXV mRNAs, the mean poly(A)-tail length was found to be 97.91 nt (with an SD of 51.07 nt) according to Nanopolish, and 82.21 nt (with an SD of 43.48 nt) as measured by Dorado. The most frequent poly(A)-tail lengths were 86 nt and 71 nt (**Figure 7, Supplementary Table 5**).

UTRs of hMPXV genes

The hMPXV genome displays the densely packed and sequentially arranged gene structure common to Orthopoxviruses. This layout creates many short intergenic regions, with an average distance of 129 nucleotides between genes, which often causes the untranslated regions (UTRs) of neighboring genes to overlap. Following the annotation of TSSs and TESs, we identified the canonical UTR for each ORF in hMPXV. To determine the 5'-UTRs, we initially aligned the filtered TSS positions with the coordinates of a given ORF and selected the most abundant closest TSS as canonical.

We found that 118 out of 190 ORFs had an associated TSS, while the remainder either failed to meet our strict criteria or shared a common TSS with other ORFs. The length of the 5'-UTRs ranged from 0 to 763 nt, with an average of 44 nt (see **Supplementary Table 6a**). This excludes cases where the TSS was located within the host ORF. The 5'-UTRs can also be distinguished by their TSS distribution. We discovered that 63 ORFs have a single, highly abundant TSS, while 55 ORFs could be associated with non-single peak type of TSSs. Additionally, 20 TSSs were found in the upstream neighboring ORF, and 7 TSSs were detected within the host ORF, as detailed in **Supplementary Table 6b**.

It is known that VACV produces heterogeneous 3'-ends (23, 58), therefore determining the length of 3'-UTRs is challenging. We examined the 3'-UTRs of hMPXV based on the closest TES to a given ORF and found that out of 190 ORFs 113 are assigned to TESs. The mean length of 3'-UTRs was found to be 176 nt. According to our data, almost half of the canonical 3'-UTRs are terminated in the downstream ORFs (**Supplementary Table 7)**.

Putative novel genes

An in-depth analysis of TSS positions showed CAGE-Seq signals within intergenic spaces located at the variable ends of the genome. These signals, identified in both the right and left terminal regions, were validated by the ends of dRNA-Seq reads (**Table 2**).

The new genes were further corroborated by the prediction of promoter elements and by dRNA-Seq identifying their TESs (**Supplementary Table 8**). Three of the most abundant novel TSSs are demonstrated on **Figure 8.**

TSS (CAGE)	TES (dRNA)	strand	localization	promoter start	adjacent downstream ORF	adjacent upstream ORF
6,936	6,230	-	LTR	6,930; 6,949	MA001-005 (D2L)	MA001-004 (Ankyrin CPXV-017)
9,501	9,203	-	LTR	9,504	MA001-007 (D4L)	MA001-008 (D5L)
152,144	152,157	+	RTR	152,117	MA001-158 (A49R)	MA001-159 (A50R)
157,160	157,506	+	RTR	157,126	MA001-160 (A51R)	MA001-161 (B1R)
168,981	169,692	+	RTR	168,951	MA001-170 (B10R)	MA001-171 (B11R)
187,189	187,794	+	RTR	187,160	MA001-182 (B21R)	MA001-183 (R1R)

Table 2. List of novel TSS and TES positions in intergenic region of hMPXV

Novel TSSs and TESs have been identified in both the left and right variable regions of the hMPXV genome. Their positions were determined based on sequence alignment against the first public hMPXV reference sequence (ON563414.3) from the 2022 outbreak (84). The locations of the TSSs are indicated as follows: left terminal region (LTR) and right terminal region (RTR). The possible lengths of ORFs are calculated by taking the coordinates from the first ATG to the following STOP codon, along with the dRNA-seq reads.

Despite their short predicted ORFs, a pBLAST search revealed homology with poxviral sequences for three entities: a hypothetical ankyrin-repeat containing protein (located between B21R and R1R), a kelch-like motif containing a possible protein-coding sequence (located between A49R and A50R), and another unknown protein-coding gene situated in the intergenic area of ORF004 and D2L (**Figure 8**, **Supplementary Table 8**).

Discussion

Long-read RNA sequencing (lrRNA-Seq) enables the capture of entire transcripts, facilitating the examination of isoform diversity. Although there have been advancements in decreasing

sequencing errors and improving base calling, challenges in identifying transcript sequences from lrRNA-Seq data persist due to biases associated with RNA degradation, library preparation, and read mapping. Unlike traditional RNA-Seq, which relies on conversion of RNA to cDNA, dRNA-Seq directly sequences RNA molecules. This approach circumvents the errors such as false priming during reverse transcription introduced during library preparation (85–87). In this study, we employed dRNA-Seq on ONT MinION platform to identify the precise TESs of hMPXV, known for their considerable diversity in poxviruses (72). Detection of poly(A) signals was used for the validation of dRNA-Seq results. The lrRNA-Seq methods, particularly of ONT approach, have been found to produce a pervasive 5'-truncation of transcripts, potentially leading to incorrect identification of false TSSs (88). Our previous investigations (14, 89, 90) have also uncovered a diverse range of 5'- and 3'-transcript ends in various viruses, many of which, particularly the transcription start sites (TSSs), are likely non-functional or could even be of non-biological origin.

To address this issue, we employed CAGE sequencing on Illumina MiSeq platform, a well-established method for detecting the 5'-ends of capped RNA molecules. While CAGE-Seq is highly reliable, we cannot exclude the possibility that a certain fraction of degraded RNA molecules is also detected by this technique, since it has been shown that mammalian cells contain enzymes in the cytoplasm capable of generating caps onto uncapped RNAs (91). A key issue is the absence of software capable of unequivocally differentiating genuine RNA molecules from technical artifacts. In light of this, our study focused on the annotation of main transcript ends, but also provided data on the low-abundance putative TSSs and TESs.

We compared the 5′-ends of mRNAs from CAGE-Seq libraries, to those generated by dRNA-Seq and detected that a significant portion of dRNA-read ends are accumulated on average 11 nt downstream of a TSSs (**Figure 1B**). This discrepancy is mainly due to poor-quality ends of dRNA-Seq reads, which fail to align when local alignment methods are used. To overcome this phenomenon SRS and LRS methods are combined (92, 93), or adapter ligation is carried out (94).

VACV is the best-studied representative of Orthopoxviruses. Since VACV and hMPXV are phylogenetically closely related (95), their promoter motifs are expected to be very similar. Therefore, we scanned the hMPXV genome using a set of VACV promoter modules. The validation of TSSs and TESs was carried out by identifying nearby consensus sequences and poly(A) signals, respectively. We also compiled a list of high-abundance putative transcript ends where cis-regulatory sequences could not be identified nearby. Integration of short- and long-read sequencing data provided a high-resolution map of the viral transcript ends. Extremely high levels of transcriptional activity were detected in both the core and terminal regions of the viral genome. Additionally, we observed mRNA readthrough at the peak of the circularized genome. The positions of the most abundant TSSs, along with their corresponding host ORF, and their VACV orthologues are listed in the **Supplementary Table 3**. The most abundant TSS belongs to the gene gp011, which codes for a short, non-essential protein termed D8L containing an ankyrin-like peptide domain. This domain plays a role in host immune evasion by blocking IL-1 receptors (96) and modulating the NF-κB pathway (97). The second most abundant TSS belongs to the gene gp052, which might have evolved via episodic positive selection in response to immune selection (81) and host antiviral response (101). In the dRNA-Seq analysis, the most abundant TSS is associated with the hMPXVgp095 gene, which plays a critical role in replication and for virion morphogenesis (98, 99).

Our findings on TSS-pattern align with previous studies, confirming the existence of two major TSS types: single-peak and broad-range CAGE-Seq signal distributions. Similar patterns have been observed in Orthopoxviruses (56), Herpesviruses (93) and other organisms (100, 101). More precise mapping of the TSSs and additional mutagenesis studies are needed to further explore the transcriptomic structure of Poxviruses.

Termination of poxvirus transcription requires the interaction between a U(5)NU consensus sequence and the assembly of a ternary complex, which includes the viral termination factor (VTF) and the RAP94 protein, causing strict 3'-termination of transcripts (102, 103). Unlike early mRNAs, PR RNAs exhibit high heterogeneity in length because the ePAS is unrecognized by the poxvirus transcription termination complex (56, 58). The transcription of Orthopoxvirus genes often terminates within the downstream ORFs (56, 58).

Using oligodT selection-based library screening, canonical TES positions were assigned to the annotated ORFs. However, our dRNA-Seq analysis shows that not all ORFs can be assigned canonical TESs due to the presence of TESs likely used by more than one gene in hMPXV. A similar pattern of TES distribution was revealed in VACV using LRS (23, 72), suggesting the formation of co-terminal transcription units. Our LRS method also enabled the annotation of 73 ePAS, confirming the existence of early canonical TESs (**Supplementary Table 4**). We detected a 3'-UTR architecture similar to VACV in the hMPXV transcriptome.

We found that the average length of 5'-UTRs in hMPXV is short, consistent with findings reported by others for other Orthopoxviruses (56, 72). In some rare cases (**Supplementary Table 6b**), anomalous TSSs were located downstream to the annotated start codon, suggesting alternative ATG usage by the virus (23, 104). The presence of 5'-poly(A) leader is a characteristic feature of the poxviral of mRNAs (58). Furthermore, VACV is a cytoplasmic virus, possessing two enzymes (D9, D10) functioning as decapping enzymes in mRNA degradation and translation regulation. In our study, we also detected the poly(A) leaders in both the dRNA and CAGE samples. Although literature suggests an average length of 35 nt for these sequences (105), we observed shorter lengths in hMPXV. However, it is important to consider that these shorter lengths may be underestimations due to the possible incomplete sequencing of the 5'-end.

Direct RNA sequencing confirmed the presence of polyadenylated novel mRNAs in the intergenic region. This region of Poxviruses is thought to be responsible for host-virus interactions therefore, a similar function is expected for the novel genes. Farlow and colleagues (106) reported mutations in a cidofovir-resistant MPXV strain in the same genomic region. They speculated about the presence of a hypothetical yet unknown ankyrin-like protein-coding gene which we can confirm here. On the other hand, this virus is classified within the European Clade II B.1 lineage of hMPXV. Phylogenetic studies show a relatively high mutation rate within this lineage (107, 108). This accelerated evolution is suggested to be driven by the action of the cellular APOBEC3 nucleic acid editing enzyme in the terminal genomic region (109–111). Genotyping hMPXV via gene or genome sequencing and identifying point mutations are frequently employed to track the pandemic's progression (8, 112). Several studies have aimed to elucidate the pathogenicity and virulence of hMPXV by examining variations in the terminal region, which encodes proteins involved in immune modulation (113–115). Nonetheless, transcriptomic studies provide the benefit of describing the functional units of the viral genome, rather than merely analyzing gene variants.

Materials and Methods

Virus propagation and RNA isolation

The methods for cell culture, virus propagation, and RNA isolation are detailed in the **Supplemental Text**. Briefly, the hMPXV isolate was propagated in CV-1 cell lines at a multiplicity of infection (MOI) of 5, with three replicates, in 75 cm² flasks. The infected cells were then incubated at 37°C for 2, 6, 12, and 24 hours. RNA was isolated using the Nucleospin RNA Mini Kit (Macherey Nagel) according to the manufacturer's protocol at each time point, followed by DNase treatment to remove residual DNA. Thereafter, polyadenylated RNA enrichment was carried out using Lexogen's Poly(A) RNA Selection Kit

V1.5. RNA samples were bound to beads, washed, and hybridized. After incubation and washing, the polyadenylated RNA was eluted in nuclease-free water and stored at -80°C for subsequent analysis.

Native RNA sequencing

The Oxford Nanopore Technologies SQK-RNA002 kit was utilized to sequence the RNA molecules. For library preparation, we used fifty nanograms (in 9 µl) of a pooled sample of poly(A) $^{(+)}$ RNAs. The initial step involved the ligation of a 1 µl RT Adapter (110nM; part of the ONT Kit) to the RNA sample using a mix of 3µl NEBNext Quick Ligation Reaction Buffer (New England BioLabs), 0.5µl RNA CS (ONT Kit), and 1.5µl T4 DNA Ligase (2M U/ml New England BioLabs). This process was conducted at RT for 10 minutes. Subsequently, the cDNA strand was synthesized using SuperScript III Reverse Transcriptase (Life Technologies), with the reaction taking place at 50°C for 50 minutes, followed by a 10-minute inactivation phase at 70°C. After this, the sequencing adapters from ONT's DRS kit were ligated to the cDNA at RT for 10 minutes using the T4 DNA ligase enzyme and NEBNext Quick Ligation Reaction Buffer. The final direct RNA library was sequenced on an R9.4 SpotON Flow Cell. To wash the direct RNA-seq and direct cDNA-seq libraries after each enzymatic reaction, RNAClean XP beads and AMPure XP beads (both sourced from Beckman Coulter) were employed.

Cap Analysis of Gene Expression

The detailed protocol is described in the **Supplemental Methods.** Briefly, to investigate TSS patterns in hMPXV, we used CAGE-Seq. Total RNA (5 µg) was prepared into CAGE-Seq libraries, starting with RNA denaturation and first-strand cDNA synthesis using the CAGE™ Preparation Kit. Post synthesis, the RNA was oxidized, and biotin was attached to the 5′-Cap. Biotinylated RNA underwent Cap-trapping on Streptavidin beads, followed by sequential washing and cDNA release. The capped cDNAs were isolated and treated with RNase to remove residual RNA. Streptavidin beads were prepared and washed, and linkers were attached to the cDNAs. After ligation, samples were treated with Shrimp Alkaline Phosphatase and USER enzyme to prepare for second-strand cDNA synthesis. Following synthesis, the samples underwent multiple purification steps and were sequenced on an Illumina MiSeq instrument. The sample concentration and library quality were assessed using Qubit 4.0 and TapeStation, ensuring accurate transcription start site profiling. The CAGE sequencing was performed on the MiSeq platform with v2 (using 150 cycles) and v3 (using 300 cycles) reagent kit.

Bioinformatics

CAGE sequencing analysis:

The reads derived from CAGE-Seq were mapped using STAR to the reference genome with the following parameters: *STAR --runThreadN 8 --outSAMunmapped Within --alignIntronMax 1000*. The bam files were merged after mapping into one dataset (**Supplementary Figure 4**). The downstream analysis was conducted within an R environment (version: 4.2). Due to technical artifacts and stochastic transcriptional activities, TSSs inferred from CAGE-Seq may not represent *bona fide* TSSs. Therefore, we applied the TSSr program (https://github.com/Linlab-slu/TSSr) for CAGE-Seq signal analysis, which effectively handles this problem (116). As one function of TSSr did not work properly, we removed the soft-clips from the alignments using the script at GitHub (https://github.com/gabor-gulyas/softclipremover). The *getTSSs* function was used with two sets of parameters: one for the core region and one for the repeat regions. In the core region, default parameters were used, however the threshold for the mapping quality in the terminal repeats needed to be decreased (*mapq* >= 3) to include the secondary alignments that have lower values. The distribution of CAGE-signals has been calculated by the Shape Index (SI)

score of TSSr's *shapeCluster* function. TSS clusters were identified by the '*peakclu*' algorithm in TSSr. The *clusterTSS* function calculates the inter-quantile width of TSS clusters based on the cumulative distribution of CAGE signals. At least 80% CAGE signals within a cluster, were defined as the 5'-and 3'-boundaries of the TSS clusters ((116).

Long-read direct RNA sequencing analysis:

During sequencing, the reads generated were basecalled using the fast model of the Guppy program (https://community.nanoporetech.com). We performed the mapping using Minimap2 (version: 2.17-r941) with the following parameters: minimap2 -ax splice -Y -C5 -t4 --cs. The reference genome was downloaded from NCBI GenBank (accession: ON563414.3) (84). Furthermore, we used the LoRTIA pipeline, developed in our laboratory, for assessing sequencing adapter quality and poly(A) sequences. It also helps eliminate false TESs that could arise from several sources, as described earlier (61). To ensure the alignments were not results of internal priming events, we applied the talon_label_reads submodule of the TALON software package(117)

The LoRTIA program (https://github.com/zsolt-balazs/LoRTIA) was used with the following parameters: *LoRTIA five_score=16.0, three_score=16.0 three_adapter='AAAAAAAAAAAAAAA', five_adapter='GCTGATATTGCTGGG'* to identify 5'- and 3'-adapters on the sequencing reads and to determine the TES positions. To estimate the length of polyA tails of viral native RNAs two methods were used: Nanopolish 1.) (https://github.com/jts/nanopolish) using the polyA command with default parameters and Dorado 2.) (https://github.com/nanoporetech/dorado) using the following parameters: *--estimate-poly-a --min-qscore 6.*

Identifying the promoter elements and poly(A) signals of hMPXV:

These sequence elements were identified using FIMO (Find Individual Motif Occurrences) (118). For promoter identification the following command was used: *fimo --oc . --verbosity 1 --bgfile --nrdb-- --thresh 1.0E-4 motifs.meme ON563414.3.fasta*, while for PAS identification the same command was used with the exception of lowering the threshold to 10^{-3} *(--thresh 1.0E-3).*

Poly(A)-tail length estimation:

We implemented poly(A) tail length estimator packages from Nanopolish (119) and Dorado (v0.5.3) to retrieve the length of poly(A)-tails of viral mRNAs. While the 5'- poly(A) leader sequences were counted at the 5'-soft-clipped region of mapped mRNAs allowing 1 mismatch after 3 bases of As/Ts.

Data availability

Bam files from CAGE-Seq have been deposited in the European Nucleotide Archive and are available under the Project Accession: PRJEB60061. dRNA-Seq data are available from the PRJEB56841 study.

Acknowledgements: The research was funded by the National Research, Development and Innovation Office (NRDIO), through the Researcher-initiated research projects (Grant number: K 142674) awarded to ZB. GÁN was supported by the New National Excellence Program (ÚNKP-23-3-SZTE-306) of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund. GET and GK were supported by the National Research, Development and Innovation Office, Hungary under grant RRF-2.3.1-21-2022–00010. The Article Processing Charges (APC) were paid by the Open Access Fund of the University of Szeged: 6683. We would like to thank Ferenc Jakab for his intellectual contributions to the preparation of this article and Fanni V. Földes, Brigitta Zana, and Zsófia Lanszki for their assistance in propagating the virus.

Abbreviations

CAGE: cap analysis of gene expression

cDNA-Seq. cDNA sequencing

dRNA-Seq: direct RNA sequencing

ePAS: early poly(A)-signal

LRS: long-read sequencing

hMPXV: human monkeypox virus

PAS: poly(A)-signal

PR: post-replicative

SRS: short-read sequencing

TES: transcription end site

TSS: transcription start site

VACV: Vaccinia virus

Additional footnote:

The arrangement of the co-first authors' names was determined based on alphabetical order.

Author contributions

GÁN: Performed bioinformatics, analyzed and interpreted data, produced figures.

BK: Performed bioinformatics, analyzed and interpreted data.

ZC: Isolated total RNA and took part in dRNA sequencing.

ÁD: Contributed to RNA isolation, executed poly(A)-selection, and direct cDNA sequencing.

GET: Conducted viral infection.

GG: Contributed to data analysis and bioinformatics.

GK: Conducted viral infection.

JH: Cultivated the virus.

DR: Cultivated the virus.

IP: Engaged in data analysis and writing the manuscript.

DT: Participated in experiment design, data analysis.

ZB: Conceived and designed experiments, supervised the project.

All authors reviewed and approved the final paper.

Ethics declarations

Not applicable

Competing interests

The authors declare that there are no competing interests.

FIGURE LEGEND

Figure 1: Distribution and characterization of 5'-ends of hMPXV mRNAs

A. The figure shows the raw read coverage of all superimposed CAGE-Seq and dRNA-Seq reads around all annotated TSSs (dashed black line represents the position of TSSs). The x-axis represents the distance from the TSS, while the y-axis indicates the coverage. The CAGE-Seq is a composite of 150 bp and 300 bp libraries. The figure demonstrates that the coverage of dRNA-Seq and CAGE-Seq reads generally agrees, providing a clear signal for detecting transcriptional start positions.

B. The histogram illustrates the distribution of 5`-ends of dRNA-Seq reads around all CAGE-Seq TSSs in a +/- 25 nt window. The x-axis represents the distance from the TSS, while the y-axis indicates the sum of read counts. The dRNA-Seq 5`-ends most frequently accumulate 11 nt downstream from the TSSs, which is seen as two dominant peaks on the histogram.

C. Venn diagram shows the initial number of putative TSSs in CAGE-Seq and dRNA-Seq and their intersection before applying the filtering criteria.

D. Onion diagram, showing the number of CAGE-Seq signals according to the filtering method implemented in this study.

a: All detected CAGE-Seq peak signals, except singletons, corroborated by dRNA 5'-ends located within a 25 nucleotide window downstream from the TSS.

b: Number of CAGE-Seq peaks with at least 10 read counts corroborated by dRNA 5'-ends located within a 25 nucleotide window downstream from the TSS.

c: CAGE-Seq signals with at least 10 read counts, corroborated by a promoter motif detected within a 40 nt interval and co-terminating with dRNA-Seq reads within a 25 nt window downstream from the TSS.

d: Number of dominant TSS signals within the clusters of CAGE-Seq signals that match the filtered TSS data.

Figure 2. Distribution of filtered TSSs along the viral genome

The figure illustrates the annotated genome of hMPXV (ON563414.3), depicting the positions of TSSs determined by CAGE-Seq. We applied specific filtering criteria to identify these TSSs: a minimum of 10 CAGE-Seq signals at a position, a predicted promoter motif within a 40-nucleotide window upstream of the TSS, and at least one dRNA-seq 5'-end with a minimum read count of 2 within a 25-nucleotide window downstream from the TSS. This resulted in a total of 401 TSSs. TSSs on the positive strand are illustrated in red, and those on the negative strand in blue. The x-axis denotes the values of CAGE-Seq peaks at each genomic position on a logarithmic scale, and the y-axis denotes the genomic positions.

Figure 3. Cluster analysis of CAGE-tags by TSSr

A: Histogram of Shape Index (SI) scores of TSSs. Higher SI values indicate sharper core promoters, with an SI value of 2 corresponds to a single peak per cluster.

B: The histogram displays the distribution of inter-quantile widths of TSS clusters in TSSr. The majority of peaks occurred within a 50 nt distance around a given TSS.

C: Histogram of 5'-ends around TSSs, according to the two types of TSS clusters within a 50 nt distance in the two libraries (dRNA-Seq and CAGE-Seq). Broad-range clusters feature a wider distribution of TSSs, whereas single-peak clusters exhibit a more concentrated distribution of TSSs. TSSs are grouped together based on their shape values. The dRNA-Seq reveals an 11-nucleotide shift in the accumulation of 5' ends, accompanied by a distinct single peak indicating that a portion of the reads has been completely sequenced.

Figure 4. Distribution of the length of the 5'-poly(A) leader sequence in hMPXV

The distribution of the length of the 5'-poly(A) leader sequence in CAGE-Seq and dRNA-Seq samples from both the + and - strands. The x-axis denotes the length of the poly(A) leader (excluding values of 0), while the y-axis represents the number of reads.

Figure 5. Promoter elements in hMPXV genome

A: Distribution of promoter motifs within a 40 nt interval upstream of TSSs.

B: The consensus motifs of early promoters are illustrated by WebLogo.

C: Base composition probability near TSSs associated with post-replicative promoters. The TSS within the conserved TAAAT-motif is indicated by dashed-line.

Figure 6. Distribution of filtered TESs

The figure displays the global distribution of TES positions with a minimum count of 6 in the dRNA-Seq data. The x-axis represents the count on a log10 scale, while the y-axis indicates the genomic position.

Figure 7. Characterization of TESs and poly(A)-tails of hMPXV mRNAs

A: The PASs of the early ORFs are located within 50 nt upstream of the TESs, which are represented by a dashed red line.

B: The TES is characterized by a dominant A/T nucleotide composition.

C: The poly(A)-tail length distribution of viral dRNA-Seq reads estimated by Nanopolish.

D: The poly(A)-tail length distribution of viral reads estimated by Dorado.

Figure 8. Novel hMPXV genes

The figure shows the localization of the three most abundant novel genes in the hMPXV genome. These putative novel genes are located within intergenic positions. ORFs are indicated with blue boxes in both the right and left terminal regions of the genome. Novel TSSs are indicated by asterisks. The dRNA reads visualized in IGV reveal a novel gene located between D2L and ORF004 at the left terminal region. A novel TSS is located between the ORFs A49R and A50R, and downstream of the B21R gene in the right terminal region of hMPXV. The letters above the genome indicate the HindIII fragments of hMPXV (source: ViralZone).

TABLE LEGEND

Table 1. Cis-regulatory sequences used for promoter and PAS annotation

The promoter motifs used to scan viral promoter and PAS sequences are categorized by their kinetics, based on data from literature on experiments related to VACV gene expression. PAS stands for poly(A) signal.

Table 2. List of novel TSS and TES positions in intergenic region of hMPXV

Novel TSSs and TESs have been identified in both the left and right variable regions of the hMPXV genome. Their positions were determined based on sequence alignment against the

first public hMPXV reference sequence (ON563414.3) from the 2022 outbreak (84). The locations of the TSSs are indicated as follows: left terminal region (LTR) and right terminal region (RTR). The possible lengths of ORFs are calculated by taking the coordinates from the first ATG to the following STOP codon, along with the dRNA-seq reads.

SUPPLEMENTARY MATERIAL

Supplementary Figure 1. Read coverage of CAGE-Seq and dRNA-Seq around the TSSs

This figure illustrates the coverage of CAGE and dRNA-Seq reads in the regions surrounding the TSS positions within a 500-nucleotide window on both sides, separated by strands

Supplementary Figure 2. Putative TSSs detected by TSSr

This figure displays the distribution of putative TSS positions following various filtering steps, shown on a logarithmic scale.

A: All putative TSS positions before any filtering (altogether 9,599 TSSs are shown).

B: Putative TSS positions with a CAGE signal of 10 or more (altogether 720 TSSs are shown).

C: Putative TSS positions requiring a minimum CAGE signal of 10, validated by a promoter within a 40-nucleotide window, and by dRNA-Seq 5'-ends within a 25-nucleotide window (altogether 401 TSSs are shown).

Supplementary Figure 3. Putative TESs detected by LoRTIA

This figure shows the distribution of putative TES positions after various filtering steps, presented on a logarithmic scale.

A: All putative TES positions before any filtering (altogether 3,241 TESs are shown).

B: Putative TES positions confirmed by 6 or more dRNA-Seq reads (altogether 496 TESs are shown).

C: Putative TES positions requiring a minimum dRNA-Seq reads of 6, validated by a poly(A) signal within a 50-nucleotide window (altogether 135 TESs are shown).

Supplementary Figure 4. Correlation matrix of the three sequenced samples

CAGE-Seq was conducted with three replicates for each of three samples (A, B, C). The plots demonstrate consistency in CAGE-Seq signal positions across all compared bam files. The bam files were all merged into one file.

Supplementary Table 1. CAGE-Seq peaks detected by TSSr

The table contains the list of the CAGE-Seq signals. (a) This table summarizes all by TSSr detected CAGE-Seq positions along with their p-score and q-score values. (b) dRNA-Seq read's 5'-positions are listed here with their count values.

(c) The genomic positions of the predicted promoter motifs are listed in a separate column, highlighting the most significant position within a 100 nt binned fraction upstream of a given TSS predicted by FIMO

(d) The table contains the best matching promoter elements upstream of the known ORFs according to their predicted highest q-values. All positions are aligned to the genome: ON563414.3.

Supplementary Table 2. Clusters of TSSs and list of shape scores

The table contains clusters of TSSs. The dominant CAGE-Seq signal with count values are listed with the shape-score given by TSSr.Shape-index.

Supplementary Table 3. List of the most abundant TSSs

The table shows the top five most abundant CAGE- and dRNA-Seq signals and the adjacent ORF names with their function.

Supplementary Table 4. List of TESs

This table presents dRNA-Seq 3'-end positions determined by the LoRTIA toolkit. The U5NU ePAS motif, scanned within a 50 nt binned fraction upstream of a given TES, is listed in a separate column according to their genomic positions.

Supplementary Table 5. Estimated poly(A)-tail lengths

The table contains the estimated length of poly(A)-tails of dRNA-Seq reads. (a) The table contains the output data of Nanopolish. (b) The table contains dorado output for poly(A)-tail length estimation.

Supplementary Table 6. List of 5'-UTRs

The table categorizes hMPXV ORFs and associated TSSs based on their count values, and their proximity to the ORFs. The 5'-UTRs were determined based on the distance of the canonical TSS belonging to the ORFs**.**

Supplementary Table 7. List of 3'-UTRs

The table categorizes ORFs and associated TESs based on their count values, and their distance to the ORFs. The 3'-UTRs were determined based on the distance of the canonical TES belonging to the ORFs**.**

Supplementary Table 8. Novel TSSs found in intergenic regions

This table contains a detailed description of the novel genes found in intergenic regions of hMPXV. Novel TSSs and TESs have been identified in the intergenic regions of the hMPXV genome, following the reference sequence ON563414.3. The TSS locations are indicated as LTR (left terminal region) and RTR (right terminal region). The potential lengths of open reading frames (ORFs) are calculated based on the coordinates from the first ATG to the subsequent STOP codon, in conjunction with dRNA-Seq reads.

Supplementary Methods. This file contains a detailed description of the virus propagation and CAGE-Seq protocol applied.

References

1. Diven DG. 2001. An overview of poxviruses. J Am Acad Dermatol 44:1–16.

2. Moss B. 2013. Poxvirus DNA Replication. Cold Spring Harb Perspect Biol 5:a010199.

3. Moss B, Smith GL. 2021. Poxviridae: The Viruses and Their Replication, p. 573–613. *In* Peter M. Howley MD, David M. Knipe PhD, Jeffrey L. Cohen, Blossom A. Damania Ph.D (eds.), Field's Virology, 7th ed. Wolters Kluwer.

4. Elwood JM. 1989. Smallpox and its eradication. J Epidemiol Community Health 43:92–92.

5. Adler H, Gould S, Hine P, Snell LB, Wong W, Houlihan CF, Osborne JC, Rampling T, Beadsworth MB, Duncan CJ, Dunning J, Fletcher TE, Hunter ER, Jacobs M, Khoo SH,

Newsholme W, Porter D, Porter RJ, Ratcliffe L, Schmid ML, Semple MG, Tunbridge AJ, Wingfield T, Price NM, Abouyannis M, Al-Balushi A, Aston S, Ball R, Beeching NJ, Blanchard TJ, Carlin F, Davies G, Gillespie A, Hicks SR, Hoyle M-C, Ilozue C, Mair L, Marshall S, Neary A, Nsutebu E, Parker S, Ryan H, Turtle L, Smith C, van Aartsen J, Walker NF, Woolley S, Chawla A, Hart I, Smielewska A, Joekes E, Benson C, Brindley C, Das U, Eyton-Chong CK, Gnanalingham C, Halfhide C, Larru B, Mayell S, McBride J, Oliver C, Paul P, Riordan A, Sridhar L, Storey M, Abdul A, Abrahamsen J, Athan B, Bhagani S, Brown CS, Carpenter O, Cropley I, Frost K, Hopkins S, Joyce J, Lamb L, Lyons A, Mahungu T, Mepham S, Mukwaira E, Rodger A, Taylor C, Warren S, Williams A, Levitt D, Allen D, Dixon J, Evans A, McNicholas P, Payne B, Price DA, Schwab U, Sykes A, Taha Y, Ward M, Emonts M, Owens S, Botgros A, Douthwaite ST, Goodman A, Luintel A, MacMahon E, Nebbia G, O'Hara G, Parsons J, Sen A, Stevenson D, Sullivan T, Taj U, van Nipsen tot Pannerden C, Winslow H, Zatyka E, Alozie-Otuka E, Beviz C, Ceesay Y, Gargee L, Kabia M, Mitchell H, Perkins S, Sasson M, Sehmbey K, Tabios F, Wigglesworth N, Aarons EJ, Brooks T, Dryden M, Furneaux J, Gibney B, Small J, Truelove E, Warrell CE, Firth R, Hobson G, Johnson C, Dewynter A, Nixon S, Spence O, Bugert JJ, Hruby DE. 2022. Clinical features and management of human monkeypox: a retrospective observational study in the UK. Lancet Infect Dis 22:1153–1162.

6. Damon IK. 2011. Status of human monkeypox: clinical disease, epidemiology and research. Vaccine 29:D54–D59.

7. Luna N, Ramírez AL, Muñoz M, Ballesteros N, Patiño LH, Castañeda SA, Bonilla-Aldana DK, Paniz-Mondolfi A, Ramírez JD. 2022. Phylogenomic analysis of the monkeypox virus (MPXV) 2022 outbreak: Emergence of a novel viral lineage? Travel Med Infect Dis 49:102402.

8. Isidro J, Borges V, Pinto M, Sobral D, Santos JD, Nunes A, Mixão V, Ferreira R, Santos D, Duarte S, Vieira L, Borrego MJ, Núncio S, de Carvalho IL, Pelerito A, Cordeiro R, Gomes JP. 2022. Phylogenomic characterization and signs of microevolution in the 2022 multi-country outbreak of monkeypox virus. Nat Med 28:1569–1572.

9. Hendrickson RC, Wang C, Hatcher EL, Lefkowitz EJ. 2010. Orthopoxvirus Genome Evolution: The Role of Gene Loss. Viruses 2:1933–1967.

10. Walsh D. 2017. Poxviruses: Slipping and sliding through transcription and translation. PLoS Pathog 13:e1006634.

11. Baldick CJ, Keck JG, Moss B. 1992. Mutational analysis of the core, spacer, and initiator regions of vaccinia virus intermediate-class promoters. J Virol 66:4710–4719.

12. Yang Z, Martens CA, Bruno DP, Porcella SF, Moss B. 2012. Pervasive Initiation and 3′-End Formation of Poxvirus Postreplicative RNAs. Journal of Biological Chemistry 287:31050–31060.

13. Condit RC, Moussatche N, Traktman P. 2006. In A Nutshell: Structure and Assembly of the Vaccinia Virion. Adv Virus Res 66:31–124.

14. Tombácz D, Prazsák I, Szűcs A, Dénes B, Snyder M, Boldogkői Z. 2018. Dynamic transcriptome profiling dataset of vaccinia virus obtained from long-read sequencing techniques. Gigascience 7.

15. McFadden G. 2005. Poxvirus tropism. Nat Rev Microbiol 3:201–213.

16. Moss B. 2007. Field's Virology, 5th ed. Lippincott Williams & Wilkins, Wolters Kluwer, Philadelphia.

17. Davison AJ, Moss B. 1989. Structure of vaccinia virus late promoters. J Mol Biol 210:771–784.

18. Dhungel P, Cao S, Yang Z. 2017. The 5'-poly(A) leader of poxvirus mRNA confers a translational advantage that can be achieved in cells with impaired cap-dependent translation. PLoS Pathog 13.

19. Alkhalil A, Hammamieh R, Hardick J, Ichou MA, Jett M, Ibrahim S. 2010. Gene expression profiling of monkeypox virus-infected cells reveals novel interfaces for host-virus interactions. Virol J 7:173.

20. Rubins KH, Hensley LE, Bell GW, Wang C, Lefkowitz EJ, Brown PO, Relman DA. 2008. Comparative Analysis of Viral Gene Expression Programs during Poxvirus Infection: A Transcriptional Map of the Vaccinia and Monkeypox Genomes. PLoS One 3:1–12.

21. Rubins KH, Hensley LE, Relman DA, Brown PO. 2011. Stunned Silence: Gene Expression Programs in Human Cells Infected with Monkeypox or Vaccinia Virus. PLoS One 6:e15615.

22. Bourquain D, Dabrowski PW, Nitsche A. 2013. Comparison of host cell gene expression in cowpox, monkeypox or vaccinia virus-infected cells reveals virus-specific regulation of immune response genes. Virol J 10:61–74.

23. Tombácz D, Prazsák I, Torma G, Csabai Z, Balázs Z, Moldován N, Dénes B, Snyder M, Boldogkői Z. 2021. Time-Course Transcriptome Profiling of a Poxvirus Using Long-Read Full-Length Assay. Pathogens 10:919.

24. Berk AJ, Sharp PA. 1977. Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. Cell 12:721–732.

25. Weaver RF, Weissmann C. 1979. Mapping of RNA by a modification of the Berk-Sharp procedure: the 5′ termini of 15 S β-globin mRNA precursor and mature 10 S β-globin mRNA have identical map coordinates. Nucleic Acids Res 7:1175–1194.

26. Lee-Chen G-J, Niles EG. 1988. Map positions of the 5′ ends of eight mRNAs synthesized from the late genes in the vaccinia virus HindIII D fragment. Virology 163:80–92.

27. Traktman P, Sridhar P, Condit RC, Roberts BE. 1984. Transcriptional mapping of the DNA polymerase gene of vaccinia virus. J Virol 49:125–131.

28. Hruby DE, Maki RA, Miller DB, Ball LA. 1983. Fine structure analysis and nucleotide sequence of the vaccinia virus thymidine kinase gene. Proceedings of the National Academy of Sciences 80:3411–3415.

29. Jones E V, Moss B. 1985. Transcriptional mapping of the vaccinia virus DNA polymerase gene. J Virol 53:312–315.

30. Luttge BG, Moyer RW. 2005. Suppressors of a Host Range Mutation in the Rabbitpox Virus Serpin SPI-1 Map to Proteins Essential for Viral DNA Replication. J Virol 79:9168–9179.

31. Lee-Chen G-J, Bourgeois N, Davidson K, Condit RC, Niles EG. 1988. Structure of the transcription initiation and termination sequences of seven early genes in the vaccinia virus HindIII D fragment. Virology 163:64–79.

32. Antczak JB, Patel DD, Ray CA, Ink BS, Pickup DJ. 1992. Site-specific RNA cleavage generates the 3' end of a poxvirus late mRNA. Proceedings of the National Academy of Sciences 89:12033–12037.

33. D'Costa SM, Antczak JB, Pickup DJ, Condit RC. 2004. Post-transcription cleavage generates the 3′ end of F17R transcripts in vaccinia virus. Virology 319:1–11.

34. Frohman MA, Dush MK, Martin GR. 1988. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proceedings of the National Academy of Sciences 85:8998–9002.

35. Scotto-Lavino E, Du G, Frohman MA. 2006. Amplification of 5′ end cDNA with "new RACE." Nat Protoc 1:3056–3061.

36. Bower NI, Johnston IA. 2010. Targeted rapid amplification of cDNA ends (T-RACE)--an improved RACE reaction through degradation of non-target sequences. Nucleic Acids Res 38:e194.

37. McGrath PT. 2011. Characterizing cDNA ends by circular RACE. Methods in Molecular Biology 772:257–265.

38. Wennier ST, Brinkmann K, Steinhäußer C, Mayländer N, Mnich C, Wielert U, Dirmeier U, Hausmann J, Chaplin P, Steigerwald R. 2013. A Novel Naturally Occurring Tandem Promoter in Modified Vaccinia Virus Ankara Drives Very Early Gene Expression and Potent Immune Responses. PLoS One 8:e73511.

39. Adams JM, Cory S. 1975. Modified nucleosides and bizarre 5′-termini in mouse myeloma mRNA. Nature 255:28–33.

40. Furuichi Y. 2015. Discovery of m(7)G-cap in eukaryotic mRNAs. Proceedings of the Japan Academy, Series B 91:394–409.

41. Wei CM, Moss B. 1975. Methylated nucleotides block 5'-terminus of vaccinia virus messenger RNA. Proceedings of the National Academy of Sciences 72:318–322.

42. Ensinger MJ, Martin SA, Paoletti E, Moss B. 1975. Modification of the 5'-terminus of mRNA by soluble guanylyl and methyl transferases from vaccinia virus. Proceedings of the National Academy of Sciences 72:2525–2529.

43. Parrish S, Moss B. 2007. Characterization of a Second Vaccinia Virus mRNA-Decapping Enzyme Conserved in Poxviruses. J Virol 81:12973–12978.

44. Parrish S, Resch W, Moss B. 2007. Vaccinia virus D10 protein has mRNA decapping activity, providing a mechanism for control of host and viral gene expression. Proceedings of the National Academy of Sciences 104:2139–2144.

45. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proceedings of the National Academy of Sciences 100:15776–15781.

46. Carninci P, Kvam C, Kitamura A, Ohsumi T, Okazaki Y, Itoh M, Kamiya M, Shibata K, Sasaki N, Izawa M, Muramatsu M, Hayashizaki Y, Schneider C. 1996. High-Efficiency Full-Length cDNA Cloning by Biotinylated CAP Trapper. Genomics 37:327–336.

47. Suzuki Y, Yoshitomo-Nakagawa K, Maruyama K, Suyama A, Sugano S. 1997. Construction and characterization of a full length-enriched and a 5′-end-enriched cDNA library. Gene 200:149–156.

48. Maeda N, Nishiyori H, Nakamura M, Kawazu C, Murata M, Sano H, Hayashida K, Fukuda S, Tagami M, Hasegawa A, Murakami K, Schroder K, Irvine K, Hume DA, Hayashizaki Y, Carninci P, Suzuki H. 2008. Development of a DNA barcode tagging method for monitoring dynamic changes in gene expression by using an ultra-high-throughput sequencer. Biotechniques 45:95–97.

49. Valen E, Pascarella G, Chalk A, Maeda N, Kojima M, Kawazu C, Murata M, Nishiyori H, Lazarevic D, Motti D, Marstrand TT, Tang M-HE, Zhao X, Krogh A, Winther O, Arakawa T, Kawai J, Wells C, Daub C, Harbers M, Hayashizaki Y, Gustincich S, Sandelin A, Carninci P. 2009. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. Genome Res 19:255–265.

50. Takahashi H, Kato S, Murata M, Carninci P. 2012. CAGE (Cap Analysis of Gene Expression): A Protocol for the Detection of Promoter and Transcriptional Networks. Methods in Molecular Biology 786:181–200.

51. Batut P, Dobin A, Plessy C, Carninci P, Gingeras TR. 2013. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. Genome Res 23:169–180.

52. de Hoon M, Hayashizaki Y. 2008. Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. Biotechniques 44:627–628, 630, 632.

53. Kanamori-Katayama M, Itoh M, Kawaji H, Lassmann T, Katayama S, Kojima M, Bertin N, Kaiho A, Ninomiya N, Daub CO, Carninci P, Forrest ARR, Hayashizaki Y. 2011. Unamplified cap analysis of gene expression on a single-molecule sequencer. Genome Res 21:1150–1159.

54. Leenen FAD, Vernocchi S, Hunewald OE, Schmitz S, Molitor AM, Muller CP, Turner JD. 2016. Where does transcription start? 5′-RACE adapted to next-generation sequencing. Nucleic Acids Res 44:2628–2645.

55. Policastro RA, Zentner GE. 2021. Global approaches for profiling transcription initiation. Cell Reports Methods 1:100081.

56. Yang Z, Bruno DP, Martens CA, Porcella SF, Moss B. 2011. 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.

827 57. Yang Z, Reynolds SE, Martens CA, Bruno DP, Porcella SF, Moss B. 2011. Expression
828 Profiling of the Intermediate and Late Stages of Poxvirus Replication. J Virol 85:9899–
829 9908.

830 58. Yang Z, Martens CA, Bruno DP, Porcella SF, Moss B. 2012. Pervasive Initiation and
831 3′-End Formation of Poxvirus Postreplicative RNAs. Journal of Biological Chemistry
832 287:31050–31060.

833 59. Kanagawa T. 2003. Bias and artifacts in multitemplate polymerase chain reactions
834 (PCR). J Biosci Bioeng 96:317–323.

835 60. Cocquet J, Chong A, Zhang G, Veitia RA. 2006. Reverse transcriptase template
836 switching and false alternative transcripts. Genomics 88:127–131.

837 61. Balázs Z, Tombácz D, Csabai Z, Moldován N, Snyder M, Boldogkői Z. 2019.
838 Template-switching artifacts resemble alternative polyadenylation. BMC Genomics
839 20:824.

840 62. Mourão K, Schurch NJ, Lucoszek R, Froussios K, MacKinnon K, Duc C, Simpson G,
841 Barton GJ. 2019. Detection and mitigation of spurious antisense expression with
842 RoSA. F1000Res 8:819.

843 63. McCarthy A. 2010. Third Generation DNA Sequencing: Pacific Biosciences' Single
844 Molecule Real Time Technology. Chem Biol 17:675–676.

845 64. Thompson JF, Milos PM. 2011. The properties and applications of single-molecule
846 DNA sequencing. Genome Biol 12:217.

847 65. Bayega A, Wang YC, Oikonomopoulos S, Djambazian H, Fahiminiya S, Ragoussis J.
848 2018. Transcript Profiling Using Long-Read Sequencing Technologies. Methods in
849 Molecular Biology 1783:121–147.

850 66. Boldogkői Z, Moldován N, Balázs Z, Snyder M, Tombácz D. 2019. Long-Read
851 Sequencing – A Powerful Tool in Viral Transcriptome Research. Trends Microbiol
852 27:578–592.

853 67. Soneson C, Yao Y, Bratus-Neuenschwander A, Patrignani A, Robinson MD, Hussain
854 S. 2019. A comprehensive examination of Nanopore native RNA sequencing for
855 characterization of complex transcriptomes. Nat Commun 10:3359.

856 68. Depledge DP, Srinivas KP, Sadaoka T, Bready D, Mori Y, Placantonakis DG, Mohr I,
857 Wilson AC. 2019. Direct RNA sequencing on nanopore arrays redefines the
858 transcriptional complexity of a viral pathogen. Nat Commun 10:754.

859 69. Olasz F, Tombácz D, Torma G, Csabai Z, Moldován N, Dörmő Á, Prazsák I, Mészáros
860 I, Magyar T, Tamás V, Zádori Z, Boldogkői Z. 2020. Short and Long-Read Sequencing
861 Survey of the Dynamic Transcriptomes of African Swine Fever Virus and the Host
862 Cells. Front Genet 11.

863 70. Fülöp Á, Torma G, Moldován N, Szenthe K, Bánáti F, Almsarrhad IAA, Csabai Z,
864 Tombácz D, Minárovits J, Boldogkői Z. 2022. Integrative profiling of Epstein–Barr
865 virus transcriptome using a multiplatform approach. Virol J 19:7.

71. Tombácz D, Kakuk B, Torma G, Csabai Z, Gulyás G, Tamás V, Zádori Z, Jefferson VA, Meyer F, Boldogkői Z. 2022. In-Depth Temporal Transcriptome Profiling of an Alphaherpesvirus Using Nanopore Sequencing. Viruses 14:1289.

72. Tombácz D, Prazsák I, Csabai Z, Moldován N, Dénes B, Snyder M, Boldogkői Z. 2020. Long-read assays shed new light on the transcriptome complexity of a viral pathogen. Sci Rep 10:13822.

73. Maróti Z, Tombácz D, Prazsák I, Moldován N, Csabai Z, Torma G, Balázs Z, Kalmár T, Dénes B, Snyder M, Boldogkői Z. 2021. Time-course transcriptome analysis of host cell response to poxvirus infection using a dual long-read sequencing approach. BMC Res Notes 14:239.

74. Wang Y, Zhang J, Li M, Jia M, Yang L, Wang T, Wang Y, Kang L, Li M, Kong L. 2024. Transcriptome and proteomic analysis of mpox virus F3L-expressing cells. Front Cell Infect Microbiol 14.

75. Patel DD, Pickup DJ. 1987. Messenger RNAs of a strongly-expressed late gene of cowpox virus contain 5′-terminal poly(A) sequences. EMBO J 6:3787–3794.

76. Ahn BY, Moss B. 1989. Capped poly(A) leaders of variable lengths at the 5’ ends of vaccinia virus late mRNAs. J Virol 63:226–232.

77. Miner JN, Weinrich SL, Hruby DE. 1988. Molecular dissection of cis-acting regulatory elements from 5'-proximal regions of a vaccinia virus late gene cluster. J Virol 62:297–304.

78. Cowley R, Greenaway PJ. 1990. Nucleotide sequence comparison of homologous genomic regions from variola, monkeypox, and vaccinia viruses. J Med Virol 31:267–271.

79. Ahn B-Y, Gershon PD, Jones E V., Moss B. 1990. Identification of *rpo30*, a Vaccinia Virus RNA Polymerase Gene with Structural Similarity to a Eucaryotic Transcription Elongation Factor. Mol Cell Biol 10:5433–5441.

80. Ferrareze PAG, Pereira e Costa RA, Thompson CE. 2023. Genomic characterization and molecular evolution of human monkeypox viruses. Arch Virol 168:278.

81. Molteni C, Forni D, Cagliani R, Mozzi A, Clerici M, Sironi M. 2023. Evolution of the orthopoxvirus core genome. Virus Res 323:198975.

82. Balázs Z, Tombácz D, Szűcs A, Csabai Z, Megyeri K, Petrov AN, Snyder M, Boldogkői Z. 2017. Long-Read Sequencing of Human Cytomegalovirus Transcriptome Reveals RNA Isoforms Carrying Distinct Coding Potentials. Sci Rep 7:15989.

83. Mohamed MR, Niles EG. 2003. UUUUUNU Stimulation of Vaccinia Virus Early Gene Transcription Termination. Journal of Biological Chemistry 278:39534–39541.

84. Ma Y, Chen M, Bao Y, Song S. 2022. MPoxVR: A comprehensive genomic resource for monkeypox virus variant surveillance. The Innovation 3:100296.

85. Fuller CW, Kumar S, Porel M, Chien M, Bibillo A, Stranges PB, Dorwart M, Tao C, Li Z, Guo W, Shi S, Korenblum D, Trans A, Aguirre A, Liu E, Harada ET, Pollard J, Bhat A, Cech C, Yang A, Arnold C, Palla M, Hovis J, Chen R, Morozova I, Kalachikov S, Russo JJ, Kasianowicz JJ, Davis R, Roever S, Church GM, Ju J. 2016. Real-time

907 single-molecule electronic DNA sequencing by synthesis using polymer-tagged
908 nucleotides on a nanopore array. Proceedings of the National Academy of Sciences
909 113:5233–5238.

910 86. Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, DuBois RM, Forsberg
911 EC, Akeson M, Vollmers C. 2017. Nanopore long-read RNAseq reveals widespread
912 transcriptional variation among the surface receptors of individual B cells. Nat
913 Commun 8:16027.

914 87. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N,
915 Admassu T, James P, Warland A, Jordan M, Ciccone J, Serra S, Keenan J, Martin S,
916 McNeill L, Wallace EJ, Jayasinghe L, Wright C, Blasco J, Young S, Brocklebank D,
917 Juul S, Clarke J, Heron AJ, Turner DJ. 2018. Highly parallel direct RNA sequencing on
918 an array of nanopores. Nat Methods 15:201–206.

919 88. Calvo-Roitberg E, Daniels RF, Pai AA. 2023. Challenges in identifying mRNA
920 transcript starts and ends from long-read sequencing data. bioRxiv
921 https://doi.org/10.1101/2023.07.26.550536.

922 89. Moldován N, Torma G, Gulyás G, Hornyák Á, Zádori Z, Jefferson VA, Csabai Z,
923 Boldogkői M, Tombácz D, Meyer F, Boldogkői Z. 2020. Time-course profiling of
924 bovine alphaherpesvirus 1.1 transcriptome using multiplatform sequencing. Sci Rep
925 10:20496.

926 90. Torma G, Tombácz D, Csabai Z, Moldován N, Mészáros I, Zádori Z, Boldogkői Z.
927 2021. Combined Short and Long-Read Sequencing Reveals a Complex Transcriptomic
928 Architecture of African Swine Fever Virus. Viruses 13:579.

929 91. Trotman JB, Schoenberg DR. 2018. A recap of RNA recapping. WIREs RNA 10.

930 92. O'Grady T, Wang X, Höner zu Bentrup K, Baddoo M, Concha M, Flemington EK.
931 2016. Global transcript structure resolution of high gene density genomes through
932 multi-platform data integration. Nucleic Acids Res 44: e145.

933 93. Braspenning SE, Sadaoka T, Breuer J, Verjans GMGM, Ouwendijk WJD, Depledge
934 DP. 2020. Decoding the Architecture of the Varicella-Zoster Virus Transcriptome.
935 mBio 11:e01568.

936 94. Ibrahim F, Oppelt J, Maragkakis M, Mourelatos Z. 2021. TERA-Seq: true end-to-end
937 sequencing of native RNA molecules for transcriptome characterization. Nucleic Acids
938 Res 49:e115–e115.

939 95. Gubser C, Hué S, Kellam P, Smith GL. 2004. Poxvirus genomes: a phylogenetic
940 analysis. Journal of General Virology 85:105–117.

941 96. Kluczyk A, Siemion IZ, Szewczuk Z, Wieczorek Z. 2002. The immunosuppressive
942 activity of peptide fragments of vaccinia virus C10L protein and a hypothesis on the
943 role of this protein in the viral invasion. Peptides (NY) 23:823–834.

944 97. Herbert M, Squire C, Mercer A. 2015. Poxviral Ankyrin Proteins. Viruses 7:709–738.

945 98. Boyle KA, Greseth MD, Traktman P. 2015. Genetic Confirmation that the H5 Protein
946 Is Required for Vaccinia Virus DNA Replication. J Virol 89:6312–6327.

947 99. Rabaan AA, Abas AH, Tallei TE, Al-Zaher MA, Al-Sheef NM, Fatimawali, Al-Nass
948 EZ, Al-Ebrahim EA, Effendi Y, Idroes R, Alhabib MF, Al-Fheid HA, Adam AA, Bin
949 Emran T. 2023. Monkeypox outbreak 2022: What we know so far and its potential
950 drug targets and management strategies. J Med Virol 95:e28306.

951 100. Rach EA, Yuan H-Y, Majoros WH, Tomancak P, Ohler U. 2009. Motif composition,
952 conservation and condition-specificity of single and alternative transcription start sites
953 in the Drosophila genome. Genome Biol 10:R73.

954 101. Haberle V, Forrest ARR, Hayashizaki Y, Carninci P, Lenhard B. 2015. CAGEr: precise
955 TSS data retrieval and high-resolution promoterome mining for integrative analyses.
956 Nucleic Acids Res 43:e51.

957 102. Christen LA, Piacente S, Mohamed MR, Niles EG. 2008. Vaccinia virus early gene
958 transcription termination factors VTF and Rap94 interact with the U9 termination motif
959 in the nascent RNA in a transcription ternary complex. Virology 376:225–235.

960 103. Yang Z, Moss B. 2009. Interaction of the Vaccinia Virus RNA Polymerase-Associated
961 94-Kilodalton Protein with the Early Transcription Factor. J Virol 83:12018–12026.

962 104. Yang Z, Cao S, Martens CA, Porcella SF, Xie Z, Ma M, Shen B, Moss B. 2015.
963 Deciphering Poxvirus Gene Expression by RNA Sequencing and Ribosome Profiling. J
964 Virol 89:6874–6886.

965 105. Václav Vopálenský, Michal Sýkora, Zora Mělková, Tomáš Mašek, Martin Pospíšek.
966 2020. Transcripts of vaccinia virus postreplicative genes do not contain a 5'
967 methylguanosine cap. bioRxiv.

968 106. Farlow J, Ichou MA, Huggins J, Ibrahim S. 2010. Comparative whole genome
969 sequence analysis of wild-type and cidofovir-resistant monkeypoxvirus. Virol J 7.

970 107. Scarpa F, Sanna D, Azzena I, Cossu P, Locci C, Angeletti S, Maruotti A, Ceccarelli G,
971 Casu M, Fiori PL, Petrosillo N, Ciccozzi M. 2022. Genetic Variability of the
972 Monkeypox Virus Clade IIb B.1. J Clin Med 11:6388.

973 108. Ndodo N, Ashcroft J, Lewandowski K, Yinka-Ogunleye A, Chukwu C, Ahmad A,
974 King D, Akinpelu A, Maluquer de Motes C, Ribeca P, Sumner RP, Rambaut A,
975 Chester M, Maishman T, Bamidele O, Mba N, Babatunde O, Aruna O, Pullan ST,
976 Gannon B, Brown CS, Ihekweazu C, Adetifa I, Ulaeto DO. 2023. Distinct monkeypox
977 virus lineages co-circulating in humans before 2022. Nat Med 29:2317–2324.

978 109. Suspène R, Raymond KA, Boutin L, Guillier S, Lemoine F, Ferraris O, Tournier J-N,
979 Iseni F, Simon-Lorière E, Vartanian J-P. 2023. APOBEC3F Is a Mutational Driver of
980 the Human Monkeypox Virus Identified in the 2022 Outbreak. J Infect Dis 228:1421–
981 1429.

982 110. Delamonica B, Davalos L, Larijani M, Anthony SJ, Liu J, MacCarthy T. 2023.
983 Evolutionary potential of the monkeypox genome arising from interactions with human
984 APOBEC3 enzymes. Virus Evol 9.

985 111. Jahankir MJB, Sounderrajan V, Rao SS, Thangam T, Kamariah N, Kurumbati A,
986 Harshavardhan S, Ashwin A, Jeyaraj S, Parthasarathy K. 2024. Accelerated mutation
987 by host protein APOBEC in Monkeypox virus. Gene Rep 34.

112. Pfaff F, Hoffmann D, Beer M. 2022. Monkeypox genomic surveillance will challenge lessons learned from SARS-CoV-2. The Lancet 400:22–23.

113. Gigante CM, Korber B, Seabolt MH, Wilkins K, Davidson W, Rao AK, Zhao H, Smith TG, Hughes CM, Minhaj F, Waltenburg MA, Theiler J, Smole S, Gallagher GR, Blythe D, Myers R, Schulte J, Stringer J, Lee P, Mendoza RM, Griffin-Thomas LA, Crain J, Murray J, Atkinson A, Gonzalez AH, Nash J, Batra D, Damon I, McQuiston J, Hutson CL, McCollum AM, Li Y. 2022. Multiple lineages of monkeypox virus detected in the United States, 2021–2022. Science (1979) 378:560–565.

114. Lopera JG, Falendysz EA, Rocke TE, Osorio JE. 2015. Attenuation of monkeypox virus by deletion of genomic regions. Virology 475:129–138.

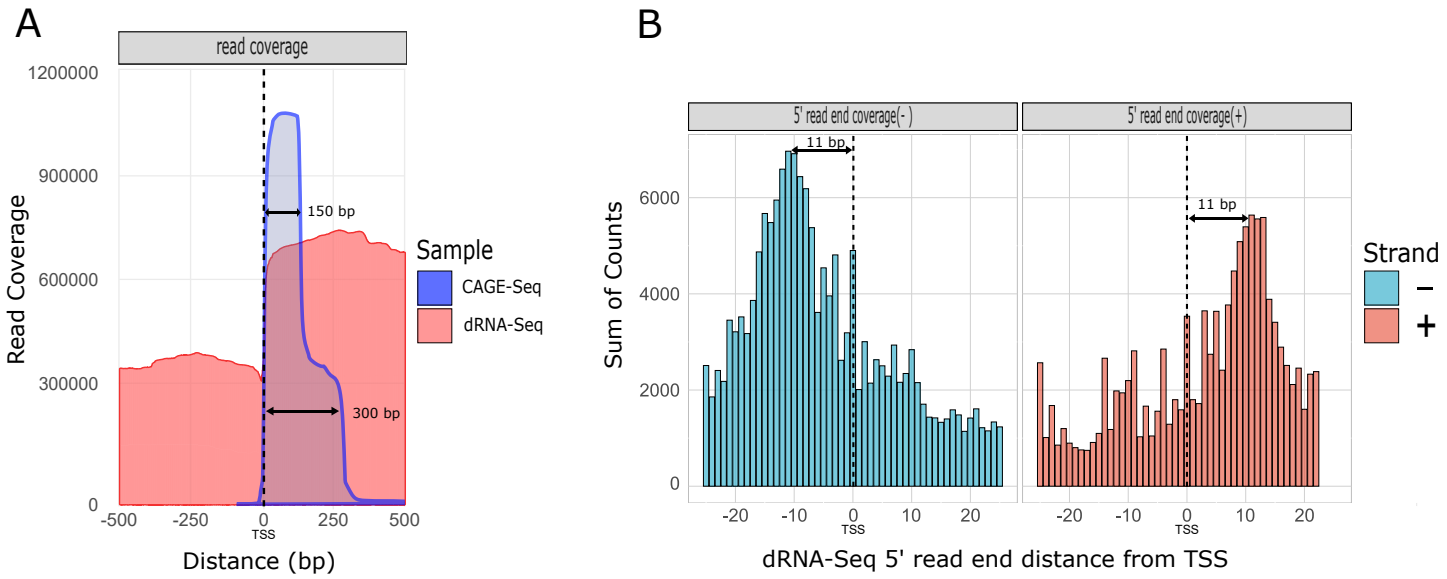
115. Li H, Zhang H, Ding K, Wang X-H, Sun G-Y, Liu Z-X, Luo Y. 2022. The evolving epidemiology of monkeypox virus. Cytokine Growth Factor Rev 68:1–12.

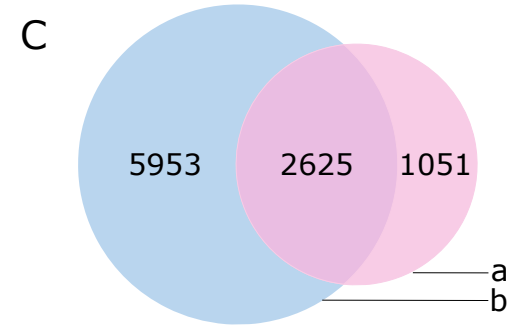
116. Lu Z, Berry K, Hu Z, Zhan Y, Ahn T-H, Lin Z. 2021. TSSr: An R package for comprehensive analyses of TSS sequencing data. NAR Genom Bioinform 3.

117. Dana Wyman, Gabriela Balderrama-Gutierrez, Fairlie Reese, Shan Jiang, Sorena Rahmanian, Stefania Forner, Dina Matheos, Weihua Zeng, Brian Williams, Diane Trout, Whitney England, Shu-Hui Chu, Robert C. Spitale, Andrea J. Tenner, Barbara J. Wold, Ali Mortazavi. 2020. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. bioRxiv https://doi.org/https://doi.org/10.1101/672931.

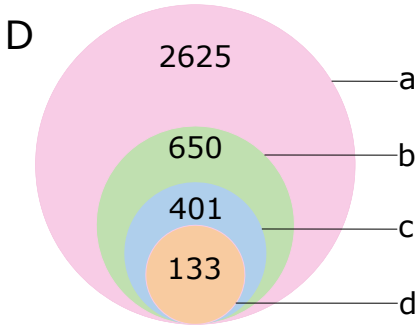
118. Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics 27:1017–1018.

119. Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J, Sadowski N, Holmes N, de Jesus JG, Jones KL, Soulette CM, Snutch TP, Loman N, Paten B, Loose M, Simpson JT, Olsen HE, Brooks AN, Akeson M, Timp W. 2019. Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat Methods 16:1297–1305.





a: CAGE-Seq signals (≥2 reads)
b: dRNA-Seq signals (≥2 reads)



a: CAGE-Seq signals corroborated by dRNA-Seq reads.

b: CAGE-Seq signals (≥10 reads) corroborated by dRNA-Seq reads.

c: CAGE-Seq signals (≥10 reads) corroborated by dRNA-Seq reads and promoter motif.

d: CAGE-Seq signals (≥10 reads) corroborated by dRNA-Seq and promoter motif, matching the position of a dominant peak within a cluster.

Figure 1: Distribution and characterization of 5'-ends of hMPXV mRNAs

A. The figure shows the raw read coverage of all superimposed CAGE-Seq and dRNA-Seq reads around all annotated TSSs (dashed black line represents the position of TSSs). The x-axis represents the distance from the TSS, while the y-axis indicates the coverage. The CAGE-Seq is a composite of 150 bp and 300 bp libraries. The figure shows that the coverage of dRNA- and CAGE-Seq reads are generally in agreement and the coverage of reads provides a clear signal regarding the transcriptional start position detection.
B. The histogram illustrates the distribution of 5`-ends of dRNA-Seq reads around all CAGE-Seq TSSs in a +/- 25 nt window. The x-axis represents the distance from the TSS, while the y-axis indicates the sum of read counts. The dRNA-Seq 5`-ends most frequently accumulate 11 nt downstream from the TSSs, which is seen as two dominant peaks on the histogram.
C. Venn diagram shows the initial number of putative TSSs in CAGE-Seq and dRNA-Seq and their intersection before applying the filtering criteria.
D. Onion diagram, showing the number of CAGE-Seq signals according to the filtering method implemented in this study.
a: All detected CAGE-Seq peak signals, except singletons, corroborated by dRNA 5'-ends located within a 25 nucleotide window downstream from the TSS.
b: Number of CAGE-Seq peaks with at least 10 read counts corroborated by dRNA 5'-ends located within a 25 nucleotide window downstream from the TSS.
c: CAGE-Seq signals with at least 10 read counts, corroborated by a promoter motif detected within a 40 nt interval and co-terminating with dRNA-Seq reads within a 25 nt window downstream from the TSS.
d: Number of dominant TSS signals within the clusters of CAGE-Seq signals that match the filtered TSS data.

Figure 2. Distribution of filtered TSSs along the viral genome

The figure illustrates the annotated genome of hMPXV (ON563414.3), depicting the positions of TSSs determined by CAGE -Seq. We applied specific filtering criteria to identify these TSSs: a minimum of 10 CAGE -Seq signals at a position, a predicted promoter motif within a 40 -nucleotide window upstream of the TSS, and at lea st one dRNA -seq 5'-end with a minimum read count of 2 within a 25 -nucleotide window downstream from the TSS. This resulted in a total of 401 TSSs. TSSs on the positive strand are illustrated in red, and those on the negative strand in blue. The x -axis denotes the values of CAGE-Seq peaks at each genomic position on a logarithmic scale, and the y-axis denotes the genomic positions.

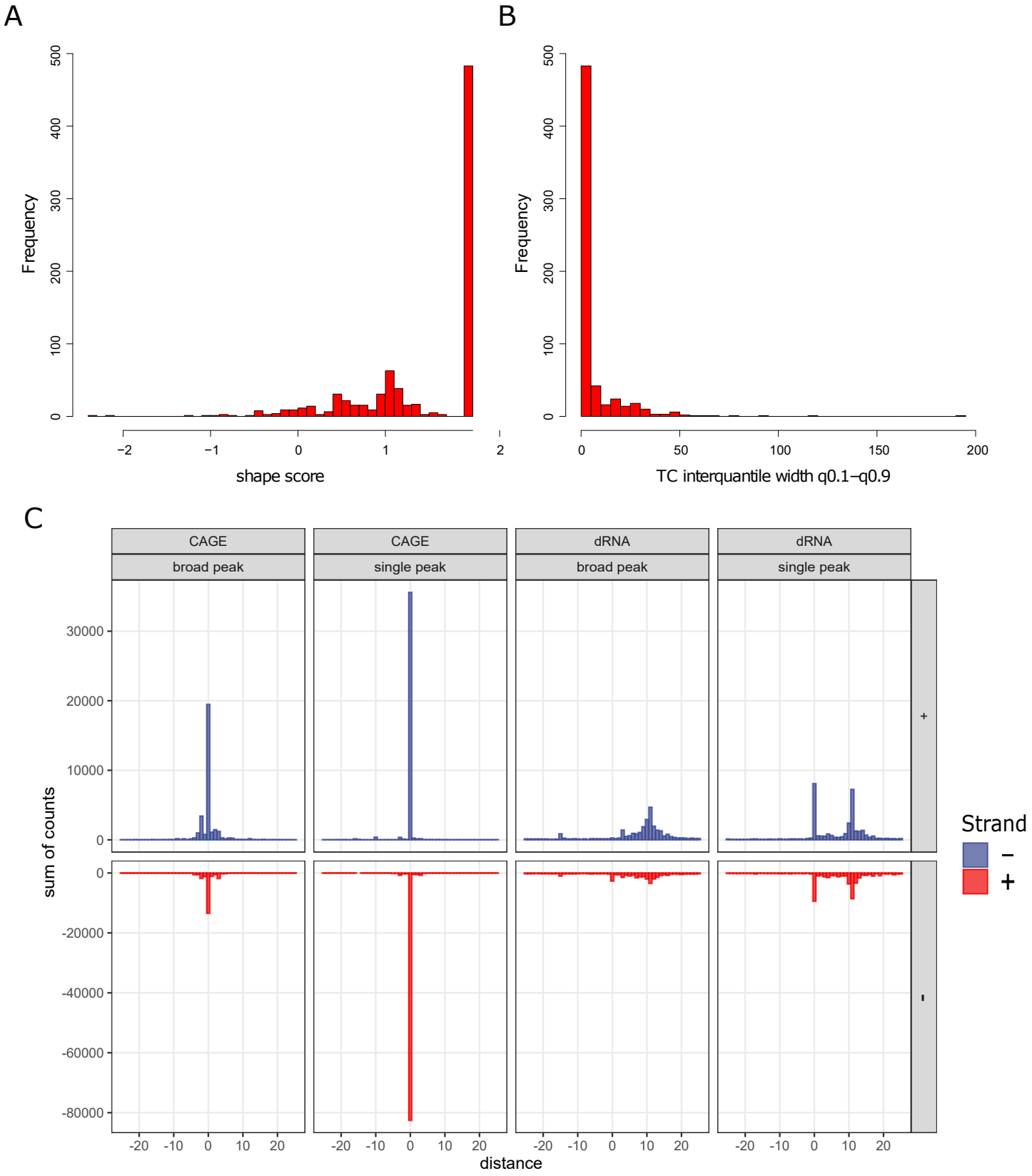


Figure 3. Cluster analysis of CAGE-tags by TSSr

A: Histogram of Shape Index (SI) scores of TSSs. Higher SI values indicate sharper core promoters, with an SI value of 2 corresponds to a single peak per cluster.

B: The histogram displays the distribution of inter -quantile widths of TSS clusters (TC) in TSSr. Th e majority of peaks occurred within a 50 nt distance around a given TSS.

C: Histogram of 5'-ends around TSSs, according to the two types of TSS clusters within a 50 nt distance in the two libraries (dRNA-Seq and CAGE-Seq). Broad-range clusters feature a w ider distribution of TSSs, whereas single -peak clusters exhibit a more concentrated distribution of TSSs. TSSs are grouped together based on their shape values. dRNA-Seq reveals a shift of 11 nucleotides in the accumulation of 5' ends, with the presence of a distinct single peak indicating that a portion of the reads has been sequenced completely.

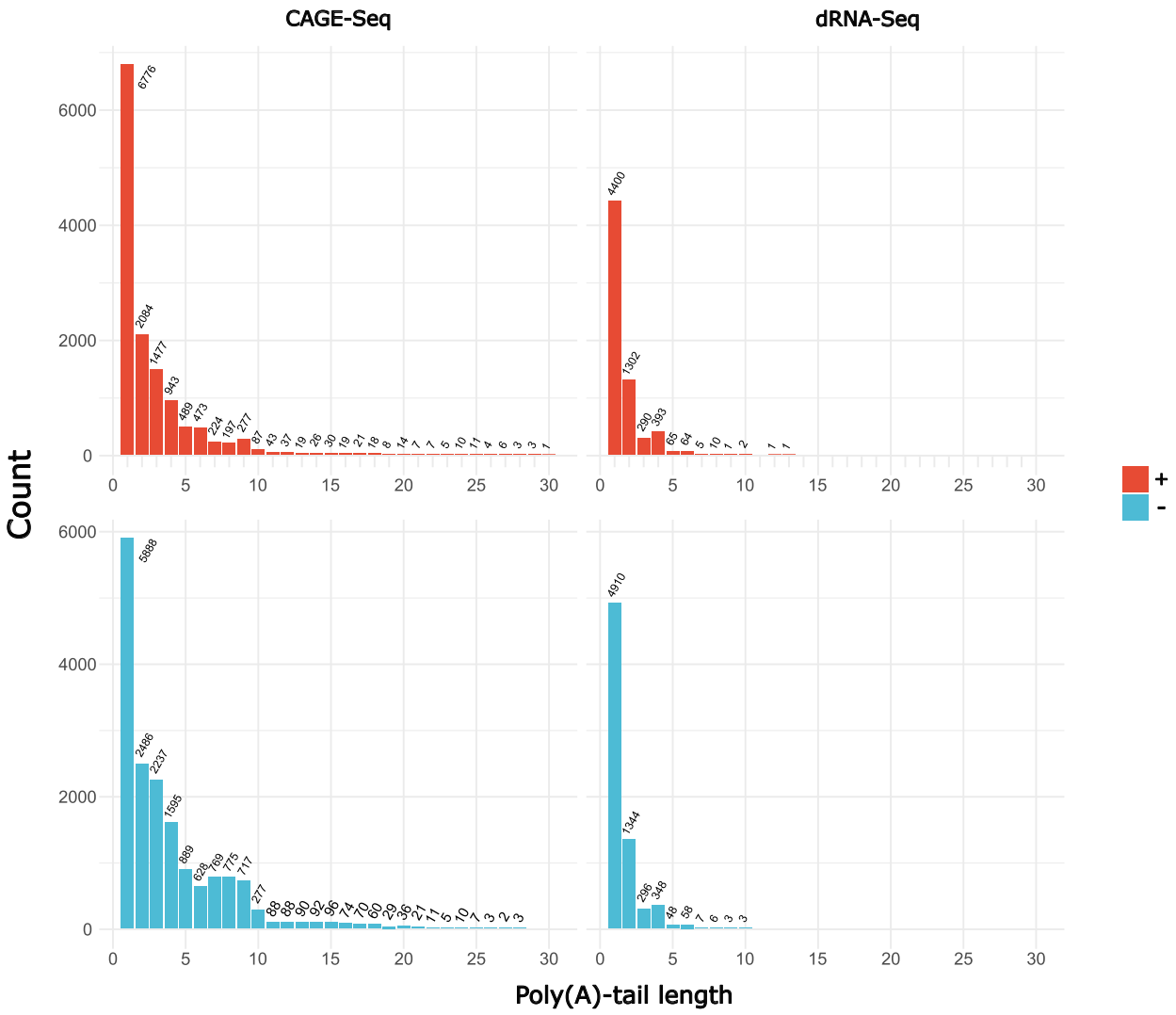


Figure 4. Distribution of the length of the 5'-poly(A) leader sequence in hMPXV

The distribution of the length of the 5' -poly(A) leader sequence in CAGE -Seq and dRNA-Seq samples from both the + and - strands. The x-axis denotes the length of the poly(A) leader (excluding values of 0), while the y -axis represents the number of reads.

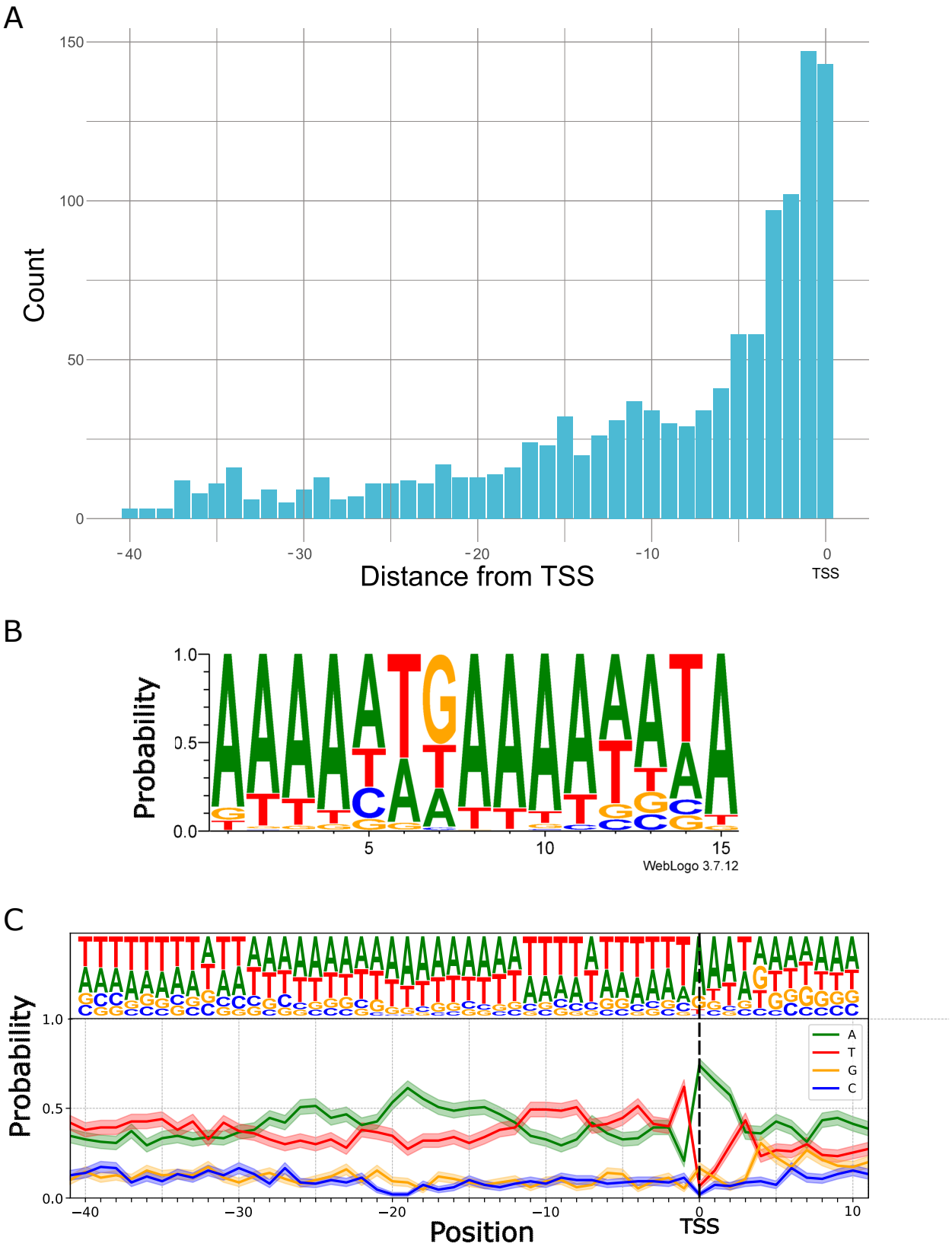


Figure 5. Promoter elements in hMPXV genome

A: Distribution of promoter motifs within a 40 nt interval upstream of TSSs.

B: The consensus motifs of early promoters are illustrated by WebLogo.

C: Base composition probability near TSSs associated with post -replicative promoters. The TSS within the conserved TAAAT -motif is indicated by dashed-line.

Figure 6. Distribution of filtered TESs

The figure displays the global distribution of TES positions with a minimum count of 6 in the dRNA-Seq data. The x-axis represents the count on a log10 scale, while the y-axis indicates the genomic position.

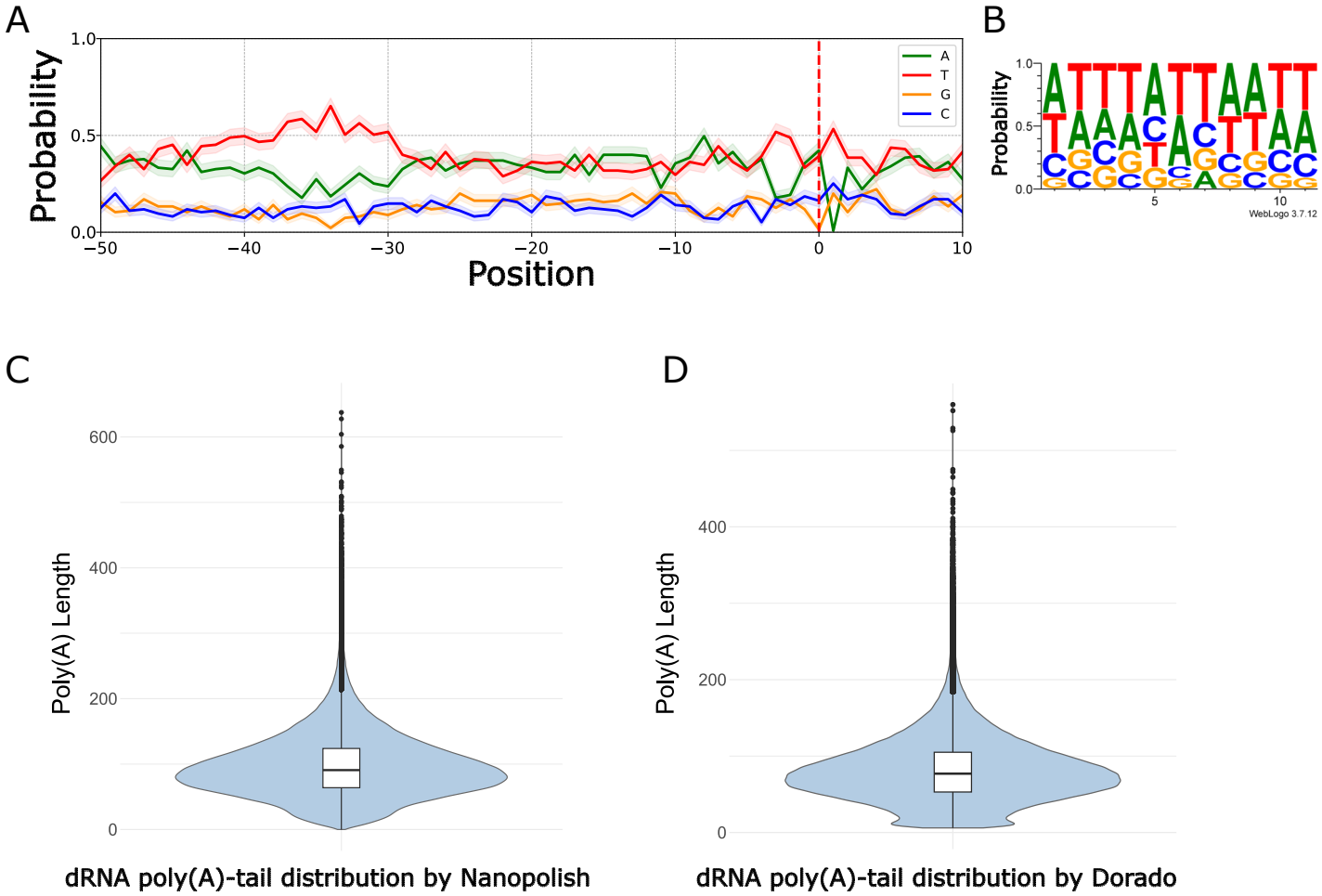


Figure 7. Characterization of TESs and poly(A)-tails of hMPXV mRNAs

A: The PASs of the early ORFs are located within 50 nt upstream of the TESs, which are represented by a dashed red line.

B: The TES is characterized by a dominant A/T nucleotide composition.

C: The poly(A)-tail length distribution of viral dRNA-Seq reads estimated by Nanopolish.

D: The poly(A)-tail length distribution of viral reads estimated by Dorado.

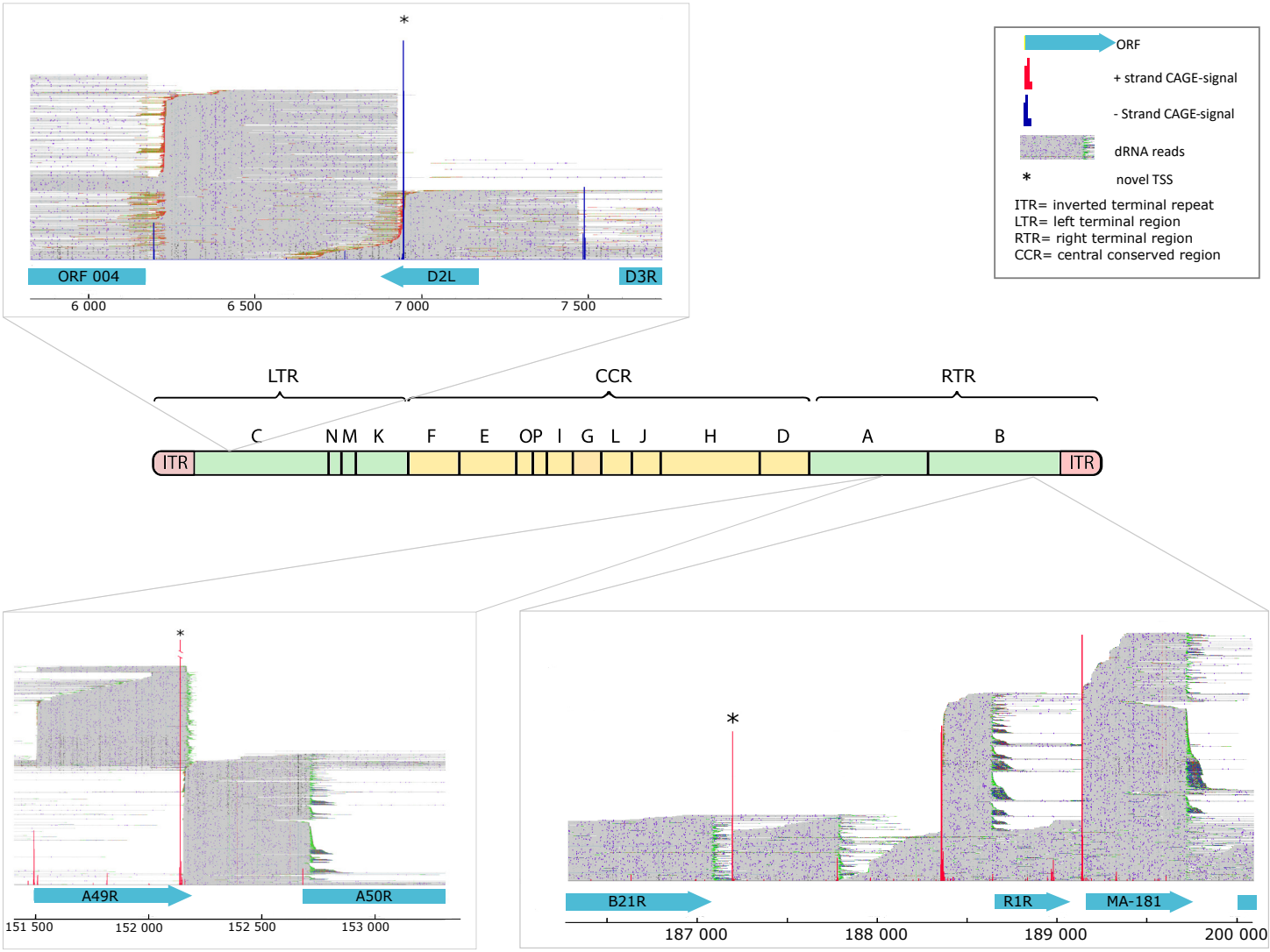


Figure 8. Novel hMPXV genes

The figure shows t he localization of the three most abundant novel g enes in the hMPXV genome. These putative novel genes are located within intergenic positions. ORFs are indicated with blue boxes in both the right and left terminal regions of the genome. Novel TSSs are indicated by asterisks. The dRNA reads visualized in IGV reveal a novel gene located between D2L and ORF004 at the left terminal region. A novel TSS is located between the ORFs A49R and A50R, and downstream of the B21R gene in the right terminal region of hMP XV. The letters above the genome indicate the HindIII fragments of hMPXV (source: ViralZone).

Type	Kinetics	Consensus	Reference
Promoter	Early	AAAANTGAAAANNA	Yang et al. 2010
	Late	TAAATG/NNNTNNNNNNNNNTAAATG	Rosel et al. 1986; Yang et al. 2011;
	Group I	NNNNNNNNNNYNWNWWWTWWWNNNNNWTAAATG	Yang et al. 2011;
	Group IIB	NATWWNWNNNHTAAAAANNDNNNNHNNDWWNTAAAYN	
	Group IIA	NRNNWNWTNWMWNWWWWTDNNNNH	
	Intermediate	NNNATNNNNNNNNTAAAAANNNNNNNNNNNNNNTAAA	
			Yang et al. 2013
	Mixed	NNNA/TNNNNNNNNTNNNNNNNNNTAAATGGNNN	Yang et al 2011
	mixed	NTAWAD	Tombácz and Prazsák et al. 2020
PAS		UUUUUNU	Yang, Reinolds et al. 2011; Yang et al 2012

TSS	TES	strand	localization	promoter start	adjacent downstream ORFs	adjacent upstream ORFs
6936	6230	-	LTR	6930; 6949	MA001-005 (D2L)	MA001-004 (Ankyrin CPXV-017)
9501	9203	-	LTR	9504	MA001-007 (D4L)	MA001-008 (D5L)
152144	152157	+	RTR	152117	MA001-158 (A49R)	MA001-159 (A50R)
157160	157506	+	RTR	157126	MA001-160 (A51R)	MA001-161 (B1R)
168981	169692	+	RTR	168951	MA001-170 (B10R)	MA001-171 (B11R)
187189	187794	+	RTR	187160	MA001-182 (B21R)	MA001-183 (R1R)

LTR= left terminal region

RTR= right terminal region