Characterizing the Transcriptome of Turkey Hemorrhagic Enteritis Virus

**Running Title:** Novel Insights into Turkey Hemorrhagic Enteritis Virus Transcriptome

Abraham Quaye1\*, Brett Pickett\*, Joel S. Griffitts\*, Bradford K. Berges\*, Brian D. Poole\*

\*Department of Microbiology and Molecular Biology, Brigham Young University  
1First-author  
 Corresponding Author

**Corresponding Author Information**  
[brian\_poole@byu.edu](mailto:brian_poole@byu.edu)  
Department of Microbiology and Molecular Biology,  
4007 Life Sciences Building (LSB),  
Brigham Young University,  
Provo, Utah

## ABSTRACT

**Background:** Hemorrhagic enteritis (HE) is a disease affecting 6-12-week-old turkeys characterized by *immunosuppression (IS)* and bloody diarrhea. This disease is caused by *Turkey Hemorrhagic Enteritis Virus (THEV)* of which avirulent strains (THEV-A) that do not cause HE but retain the immunosuppressive ability have been isolated. The THEV-A Virginia Avirulent Strain (VAS) is still used as a live vaccine despite its immunosuppressive properties. *Our objective is to understand the genetic basis by which VAS induces IS*. The *transcriptome of THEV* was studied to set the stage for further experimentation with specific viral genes that may mediate IS.  
**Methods:** After infecting a turkey B-cell line (MDTC-RP19) with the VAS vaccine strain, samples in triplicates were collected at 4-, 12-, 24-, and 72-hours post-infection. Total RNA was subsequently extracted, and poly-A-tailed mRNA sequencing done. After trimming the raw sequencing reads with the FastQC, reads were mapped to the THEV genome using Hisat2 and transcripts assembled with StringTie. An in-house script was used to consolidate transcripts from all time-points, generating the final transcriptome. PCR, gel electrophoresis, and Sanger sequencing were used to validate all identified splice junctions.

## INTRODUCTION

Adenoviruses (AdVs) are non-enveloped icosahedral-shaped DNA viruses, causing infection in virtually all vertebrates. Their double-stranded linear DNA genomes range between 26 and 45kb in size, producing a broad repertoire of transcripts via highly complex alternative splicing patterns (1, 2). The AdV genome is one of the most optimally economized; both the forward and reverse DNA strands harbor protein-coding genes, making it highly gene-dense. There are 16 genes termed “genus-common” that are homologous in all AdVs; these are thought to be inherited from a common ancestor. All other genes are termed “genus-specific”. “Genus-specific” genes tend to be located at the termini of the genome while “genus-common” genes are usually central (1). This pattern is observed in *Adenoviridae*, *Poxviridae*, and *Herpesviridae* (1, 3, 4). The family *Adenoviridae* consists of five genera: *Mastadenovirus* (MAdV), *Aviadenovirus*, *Atadenovirus*, *Ichtadenovirus*, and *Siadenovirus* (SiAdV) (5, 6). Currently, there are three recognized members of the genus SiAdV: frog adenovirus 1, raptor adenovirus 1, and turkey adenovirus 3 also called turkey hemorrhagic enteritis virus (THEV) (5, 7–10). Members of SiAdV have the smallest genome size (~26 kb) and gene content (~23 genes) of all known AdVs, and many “genus-specific” putative genes of unknown functions have been annotated (see **Figure 1**)(1, 2, 7).

Virulent THEV strains (THEV-V) and avirulent strains (THEV-A) of THEV are serologically indistinguishable, infecting turkeys, chickens, and pheasants, with the THEV-V causing different clinical diseases in these birds (2, 11). In turkeys, the THEV-V cause hemorrhagic enteritis (HE), a debilitating acute disease affecting predominantly 6-12-week-old turkeys characterized by immunosuppression (IS), weight loss, intestinal lesions leading to bloody diarrhea, splenomegaly, and up to 80% mortality (11–13). HE is the most economically significant disease caused by any strain of THEV (11). While the current vaccine strain (a THEV-A isolated from a pheasant, Virginia Avirulent Strain [VAS]) has proven effective at preventing HE in young turkey poults, it still retains the immunosuppressive ability. Thus, vaccinated birds are rendered more susceptible to opportunistic infections and death than unvaccinated cohorts leading to substantial economic losses (11, 14–16). The induced IS also interferes with vaccination schemes for other infections of turkeys (11, 14). To eliminate this immunossupressive side-effect of the vaccine, a thorough investigation of the culprit viral factors (genes) mediating this phenomenon is essential. However, the transcriptome (splicing and gene expression patterns) of THEV has not been characterized, making the investigation of specific viral genes for possible roles in causing IS impractical. A well-characterized transcriptome of THEV is required to enable experimentation with specific viral genes that may mediate IS.

Myriads of studies have elucidated the AdV transcriptome in fine detail (17, 18). However, a large preponderance of studies focus on MAdVs — specifically human AdVs. Thus, most of the current knowledge regarding AdV gene expression and replication is based on MAdV studies, which is generalized for all other AdVs (6, 19). MAdV genes are transcribed in a temporal manner; therefore, genes are categorized into five early transcription units (E1A, E1B, E2, E3, and E4), two intermediate (IM) units (pIX and IVa2), and one major late unit (MLTU or major late promoter [MLP] region), which generates five families of late mRNAs (L1-L5) based on the polyadenylation site. An additional gene (UXP or U exon) is located on the reverse strand. The early genes encode non-structural proteins such as enzymes or host cell modulating proteins, primarily involved in DNA replication or providing the necessary intracellular niche for optimal replication while late genes encode structural proteins. The immediate early gene E1A is expressed first, followed by the the delayed early genes, E1B, E2, E3 and E4. Then the intermediate early genes, IVa2 and pIX are expressed followed by the late genes (6, 17, 18). Noteworthily, the MLP shows basal transcriptional activity during early infection (before DNA replication), with a comparable efficiency to other early viral promoters, but reaches its maximal activity during late infection (after DNA replication). However, during early infection the repertoire of late transcripts from the MLP is restricted until late infection (6). MAdV makes an extensive use of alternative RNA splicing to produce a very complex array of mRNAs. All but the pIX mRNA undergo at least one splicing event. The MLTU produces over 20 distinct splice variants all of which contain three non-coding exons at the 5’-end (collectively known as the tripartite leader, TPL) (17, 18). There is also an alternate 5’ three non-coding exons present in varying amounts on a subset of MLTU mRNAs (known as the x-, y- and z-leaders). Lastly, there is the i-leader exon, which is infrequently included between the second and third TPL exons, and codes for the i-leader protein (20). Thus, the MLTU produces a complex repertoire of mRNA with diverse 5’ untranslated regions (UTRs) spliced onto different 3’ coding exons which are grouped into five different 3’-end classes (L1-L5) based on polyadenylation site. Each transcription unit (TU) contains its own promoter driving the expression of all the array of mRNA transcripts produced via alternative splicing in the unit (6, 17, 18). During translation of AdV mRNA, recent studies strongly suggest the potential usage of secondary start codons; adding to what was already a highly complex system for gene expression (17).

High throughput sequencing methods have facilitated the discovery of many novel transcribed regions and splicing isoforms. It is also a very powerful tool to study alternative splicing under different conditions at an unparalleled depth (18, 21). In this paper, a paired-end deep sequencing experiment was performed to characterize for the first time the transcriptome of THEV (VAS vaccine strain) during different phases of the infection, yielding the first THEV splicing map. Our paired-end sequencing allowed for reading **149** bp long high quality (mean Phred Score of 36) sequences from each end of cDNA fragments, which were mapped to the genome of THEV.

## RESULTS

**Overview of sequencing data and analysis pipeline outputs**  
A previous study by Zeinab *et al* showed that almost all THEV transcripts were detectable beginning at 4 hours (22). Therefore, infected MDTC-RP19 cells were harvested at 4-, 12-, 24-, and 72-hours post-infection(h.p.i) to ensure an amply wide time window to sample all transcripts. Our paired-end RNA sequencing (RNA-seq) experiment yielded an average of **107.1** million total reads of **149**bp in length per time-point, which were simultaneously mapped to both the virus (THEV) and host (*M. gallopavo*) genomes using the Hisat2 (23) alignment program. A total of **18.1** million reads from all time-points mapped to the virus genome; this provided good coverage/depth, leaving no regions unmapped. The mapped reads to the virus genome increased substantially from **432** reads at 4 h.p.i to **16.9** million reads at 72 h.p.i (**Table 1**, **Figure 2a**). From the mapped reads, we identified a total of **2,457** unique THEV splice junctions from all time-points, with splice junctions from the later time-points being supported by significantly more sequence reads than earlier time-points. For example all the **13** unique junctions at 4 h.p.i had less than 10 reads supporting each one, averaging a mere **2.8** reads/junction. Conversely, the **2,374** unique junctions at 72 h.p.i averaged **898.4** reads/junction, some junctions having coverage as high as **322,677** reads. The substantial increases in splice junction and mapping reads to the THEV genome over time denotes an active infection, and correlates with our quantitative PCR (qPCR) assay quantifying the total number of viral genome copies over time (**Figure 2b**).

Using StringTie (23), an assembler of RNA-seq alignments into potential transcripts, the mapped reads for each time point were assembled into transcripts using the genomic location of the predicted THEV ORFs as a guide. In the consolidated transcriptome, a composite of all unredundant transcripts from all time points, we counted a total of **29** novel transcripts. ~~and using 3’ Rapid Amplification of cDNA Ends (3’RACE) and other methods, we further identified #### unique splice variants~~. Although some exons in some transcripts match the predicted ORFs exactly, most of our identified exons are longer, spanning multiple predicted ORFs (**Figure 3**).

We validated the splice junctions in all transcripts by PCR amplification of viral cDNA, cloning, and Sanger sequencing (**Supplementary PCR methods**). The complete list of unique splice junctions mapped to THEV’s genome has been submitted to the National Center for Biotechnology Information Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo>) under **accession no. XXXXXX**.

**Changes in THEV splicing profile over time**  
AdV gene expression occurs under exquisite temporal control with each promoter typically producing one or few pre-mRNAs that undergo alternative splicing to yield the manifold repertoire of complex transcripts. To evaluate the activity of each promoter over time, StringTie and Ballgown (a program for statistical analysis of assembled transcriptomes) (24) were used to estimate and normalize expression levels of all transcripts for each time point in Fragments Per Kilobase of transcript per Million mapped reads (FPKM) units. Very few unique splice junctions, reads, and transcripts were counted at 4 h.p.i; hence, this time point was excluded in this analysis.

Individually, TRXPT\_21 (DBP) — from the E2 region — was the most significantly expressed at 12 h.p.i, comprising about **33.58**% of the total transcripts. Transcripts in the E3 and E4 regions also contributed significant proportions, and noticeably, some MLP region transcripts. The later time points were dominated by the MLP region — TRXPT\_10 and TRXPT\_14 were the most abundantly expressed at 24 and 72 h.p.i, respectively, as expected (**Figure 4a**). When we performed analysis of the FPKM values of transcripts per region we found a similar pattern: the E2 region was the most abundantly expressed at 12 h.p.i, after which the MLP region assumes predominance (**Figure 4b**). Secondly, we estimated relative abundances of all splice junctions at each time point using the raw reads. We counted as significantly expressed only junctions with coverage of at least 1% of the total splice junction reads at the given time point. At 12 h.p.i, **18** junctions meet the 1% threshold, and were comprised of predominantly early region (E1, E2, E3, and E4) junctions, albeit the MLTU was the single most preponderant region overall, constituting **38.8**% of all the junction reads (**Table 2a**). The top most abundant junctions at 12 h.p.i remained the most significantly expressed at 24 h.p.i also. However, here, the MLP-derived junctions were unsurprisingly even more preponderant overall, accounting for **NA**% of all the junction reads counted (**Table 2b**). At 72 h.p.i, the trend of increased activity of the MLP continued as expected; at this time, the MLP-derived junctions were not only the most abundant overall — accounting for **67.5**% of all junctions reads — but also contained the most significantly expressed individual junctions (**Table 2c**. Also see **Supplementary Tables 1a-c**; **Figure 4c**). When we limited this analysis to only junctions in the final transcriptome, the relative abundances of the junctions for each region over time was generally similar to the pattern seen with all the junctions included (**Figure 4d**).  
We also analyzed splice donor and acceptor site nucleotide usage over time to investigate any peculiarities that THEV may show, generally or over the course of the infection. We found that most splice donor-acceptor sequences were unsurprisingly the canonical GU-AG nucleotides.

**Early Region 1 (E1) transcripts**  
This region in MAdVs is the first transcribed after successful entry of the viral DNA into the host cell nucleus, albeit at low levels (18). The host transcription machinery solely mediates the transcription of this region. After their translation, the E1 proteins in concert with a myriad of host transcription factors activate the other viral promoters (6). Only two ORFs (ORF1 [sialidase] and Hyd) are predicted in this region; however, we discovered **four** novel transcripts in this region, which collectively contain **3** unique splice junctions (**Figure 5**). Most of the encoded proteins of the novel transcripts are distinct from the predicted ORFs, although they all have the potential to encode the predicted Hyd protein as the 3’-most coding sequence (CDS) if secondary start codon usage is considered as reported for other AdVs (17, 18). The 5’-most CDS of TRXPT\_1 is multi-exonic, encoding a novel 17.9 kilodalton (kDa), 160 residue [amino acids (aa)] protein. From its 5’-most start codon (SSC), TRXPT\_2 encodes the largest protein in this region — a 64.3 kDa, 580 aa protein with the same SSC as TRXPT\_1 (position 211bp). This CDS spans almost the entire predicted ORF1 and Hyd, coming short in two regards: it is spliced from 1655 to 1964bp (ORF1’s C-terminus, including the stop codon), and it’s stop codon (STC; position 2312) is 13 bp short of the Hyd’s STC. However, it has an SSC 102 bp upstream and in-frame with ORF1’s predicted SSC. Thus, the CDS of TRXPT\_2 shares substantial protein sequence similarity with ORF1 but not with Hyd, as the SSC of Hyd is not in-frame. Without its splice site removing the ORF1 STC, TRXPT\_2 would encode a longer variant of ORF1, starting from an upstream SSC. TRXPT\_3 is almost identical to TRXPT\_1, except for the lack of TRXPT\_1’s second exon. Our RNA-seq data shows that all E1 transcripts share the same transcription termination site (TTS; at position 2325bp); however, TRXPT\_3 and TRXPT\_4 seem to have transcription start sites (TSS) downstream of the TSS of TRXPT\_1 and TRXPT\_2 (at position 54bp). Given that studies in MAdVs show that E1 mRNAs share not only a common TTS but also the TSS, and only differ from each other regarding the internal splicing (18), it is likely that TRXPT\_3 and TRXPT\_4 are incomplete, and their actual TSS just like the TTS are identical for all E1 transcripts. Regardless of the TSS considered for TRXPT\_3, the coding potential (CP) remains unaffected. Its 5’-most CDS, beginning at 1965bp and sharing the same STC as TRXPT\_1 and TRXPT\_2, produces a 13.1 kDa, 115 residue protein (ORF4). ORF4 was predicted in an earlier study (25) but was excluded in later studies (1, 12); however, our data suggests it is a bona fide ORF. Unlike TRXPT\_3, the CP of TRXPT\_4 is affected by the TSS considered; if we consider its unmodified TSS, then its CP is the same as TRXPT\_3 (ORF4 as the first CDS and Hyd as second CDS if the first SSC is skipped). However, if we assume that TRXPT\_4 shares the same TSS as TRXPT\_1, then the 5’-most CDS is a distinct, novel, multi-exonic 15.9 kDa, 143 aa protein with the same SSC as TRXPT\_1 and TRXPT\_2 but with a unique STC. The splice junctions of the transcripts in this region (except the junction for TRXPT\_4) were validated by cloning of viral cDNA and Sanger sequencing (**Figure 5**; **Supplementary PCR methods**). During the validation of TRXPT\_2, ORF1 was present on the agarose gel (an unspliced band size) and Sanger sequencing results as a bona fide transcript (**Supplementary PCR methods**). This was corroborated by our 3’RACE experiment, which showed a transcript (TRXPT\_2B) spanning the entire ORF1 and Hyd ORFs without any splicing, with a poly-A tail immediately after the TTS of transcripts in this region. The 5’-most CDS of this transcript (TRXPT\_2B) would encode ORF1. However, TRXPT\_2B has an upstream and in-frame SSC to the predicted SSC of ORF1, suggesting that the predicted ORF1 CDS is truncated; it shares the same TSS, SSC, and TTS as TRXPT\_2, but has a unique STC.

**Early Region 2 (E2) and Intermediate Region (IM) transcripts**  
The E2 TU expressed on the anti-sense strand, is subdivided into E2A and E2B and encodes three classical AdV proteins essential for genome replication: pTP and Ad-pol (E2B proteins), and DBP (E2A protein) (17, 18). Unlike MAdV where two promoters (E2-early and E2-late) are known (17), we discovered only a single TSS (E2 TSS; 18,751bp) from which both E2A and E2B transcription is initiated. However, similar to MAdVs, E2A and E2B transcripts have distinct TTSs, and the E2B transcripts share the TTS of the IVa2 transcript of the IM region (17, 18) (**Figure 6**).

The E2A ORF, DBP is one of three THEV ORFs predicted to be spliced from two exons. The corresponding transcript (TRXPT\_21) found in our data matches this predicted splicing pattern precisely but with a non-coding additional exon at the 5’-end (E2-5’UTR) at position 18,684-18,751 bp. Thus, TRXPT\_21 is a three-exon transcript encoding DBP (380 residues, 43.3 kDa) precisely as predicted. This transcript (TRXPT\_21) was also corroborated in a 3’-RACE experiment. Additionally, from the 3’-RACE, a splice variant of TRXPT\_21 which retains the second intron leading to a 2-exon transcript was found. This transcript (TRXPT\_21B), albeit longer due to retaining the second intron and possessing a short 3’ UTR, encodes a truncated isoform of DBP because the first SSC utilized by TRXPT\_21, is followed shortly by STCs in the retained intron, and does not yield any viable product. Utilizing the SSC 173 bp downstream of TRXPT\_21’s SSC yields a 346 residue, 39.3 kDa product, which is in-frame of DBP but entirely contained in the second exon. TRXPT\_21 and TRXPT\_21B share a common TTS but TRXPT\_21B as seen in our 3’-RACE data, extends 39 bp into an adenine-thymine (A-T) rich sequence before the poly-A tail sequence occur, suggesting this position (16,934bp) as the bona fide E2A TTS (**Figure 6**).

The E2B region transcripts also start with the E2-5’UTR but extend thousands of base pairs downstream to reach the TTS at 2334bp in the IM region, which is immediately followed by an A-T rich sequence (position 2323-2339bp) where polyadenylation probably occurs. Interestingly, the TTS of the E1 region (position 2,325bp) on the sense strand is also in the immediate vicinity of this A-T rich sequence, which is almost palindromic; hence it likely serves as the polyadenylation signal for both E1 and E2B/IM transcripts. The E2B transcripts, TRXPT\_6 and TRXPT\_7 are almost identical except for an extra splice junction at the 3’-end of TRXPT\_6, making TRXPT\_6 a five-exon transcript and TRXPT\_7, four exons (**Figure 6**). TRXPT\_7 has the CP for IVa2 and both classical proteins (pTP and Ad-pol) encoded in this region, of which the pTP ORF is predicted to be spliced from two exons just like in all other AdVs. The predicted splice junction of pTP is corroborated by our data; however, the full transcript is markedly longer than the predicted ORF: there are two novel non-coding 5’ exons, the third exon (containing the SSC of pTP) is significantly longer than predicted, and the last exon containing the bulk of the CDS is more than triple the predicted size of pTP. The first two exons are 5’-UTRs because the SSC here is immediately followed by STCs; hence, the 5’-most SSC (position 10,995bp) of the third exon which matches the predicted SSC of pTP is utilized. The encoded product is identical to the predicted pTP ORF (597 residues; 70.5 kDa). If secondary SSC (secSSC) usage is considered, with SSC at 6768bp and STC at 3430bp, the encoded product is identical to the predicted Ad-pol (polymerase) ORF (1112 residues; 129.2 kDa). TRXPT\_6 differs from TRXPT\_7 by containing an extra splice site at 3447-3515bp. However, the CP remains similar to that of TRXPT\_7 except the Ad-pol encoded from the secSSC is a truncated isoform with a new STC resulting from the splice site. We also found a novel short transcript (TRXPT\_15) entirely nested within the terminal exon of TRXPT\_7 but with a unique splice site. This transcript is an incomplete construction from the mapped reads as it contains an incomplete CDS. However, we validated the this splice junction to be genuine (**Supplementary PCR methods**).

The IM region is a single-transcript TU, encoding a single classical protein, IVa2. The promoter expressing this single transcript (TRXPT\_5) is embedded in E2B region and shares a TTS with E2B transcripts (17, 18). TRXPT\_5 is a two-exon transcript spliced at 3447-3615bp exactly as the last intron of TRXPT\_6. The first exon is an UTR, except the last 2 nucleotides, which connect with the first nucleotide of second exon to form the 5’-most SSC. This first SSC is 4 codons upstream and in-frame of the predicted IVa2 SSC. Regardless of the SSC considered, the encoded protein (IVa2) is largely unaffected. Except for the four extra residues at the N-terminus (considering the 5’-most SSC), the entire protein sequence is identical to IVa2.

**Early Region 3 (E3) transcripts**.  
The E3 region is wholly contained in the MLTU and encodes proteins involved in modulating and evading the host immune defenses. In MAdVs, this region contains seven ORFs expressed from several transcripts which share the same TSS (from the E3 promoter) but have different TTSs (6, 17, 18). However, some E3 transcripts use the TSS of the MLP. Due to sharing the same TSS, in MAdVs, secSSC usage is heavily relied on for gene expression in this region except for 12.5K and transcripts using the MLP’s TSS, as utilizing only the first SSC cannot produce all the other transcripts in this TU (17).

In THEV, only one ORF (E3) was predicted in this region. However, we identified six novel transcripts here (**TRXPT\_22, TRXPT\_23, TRXPT\_24, TRXPT\_25, TRXPT\_26, TRXPT\_27**) (**Figure 7**). We identified two distinct TSSs — one similar to the classic MAdV E3 TSS (position 18,230bp) and the other about 500 bp downstream at 18,727bp. The E3 transcripts collectively have the CP for several predicted THEV ORFs: 100K, 22K, 33K, pVIII, E3, Fiber (IV), and ORF7 belonging to the MLTU; however, some CDSs are nonidentical due to unpredicted splicing or the use of an upstream, in-frame SSC. For instance, 33K is one of the few THEV ORFs predicted to be spliced from two exons; however, we discovered it to be a significantly longer four-exon ORF expressed from TRXPT\_24. The first two exons of L33K were not predicted but the last two match the predicted exons and the CDS is in-frame. However, the first 20bp of the predicted 33K (including the SSC at 20,142bp) is spliced out as as part of the second intron of TRXPT\_24. Thus, the bona fide 33K is a 19.8 kDa, 171 residue protein (L33K) spanning four exons instead of the predicted 120 aa protein. TRXPT\_24 also has the CP for the ORFs, pVIII and E3 (a longer variant starting from and upstream, in-frame SSC) if we consider downstream SSC usage. Also, 22K (89 residues) is a single-exon ORF predicted to use the same SSC as 33K (20,142bp). However, just like 33K, all the transcripts in this region exclude the predicted SSC as part of their introns; hence 22K as predicted is not identical to any expressed ORF. TRXPT\_29 has its SSC upstream of 22K’s predicted SSC but is spliced to overlap a portion of 22K’s CDS. However, the overlapping sequence is short and not in-frame of 22K. The 5’-most product of TRXPT\_29 is a novel 73 residue protein (8.3aK) entirely different from 22K. TRXPT\_23 being spliced identically as TRXPT\_29 also encodes this novel 73 aa protein (8.3aK) from its first SSC. Similarly, TRXPT\_22 also encodes a 73 aa novel protein (8.3bK) from its first SSC that shares over 80% similarity with 8.3aK, but it differs from 8.3aK at the C-terminus. Considering downstream SSC usage, both TRXPT\_22 and TRXPT\_23 can encode pVIII and E3 in that order, but TRXPT\_23 being longer, has the CP for the Fiber ORF also. As the splice junctions of TRXPT\_22, TRXPT\_23, TRXPT\_24, and TRXPT\_29 essentially share the same genomic space, their validation was done with a single primer pair and they were differentiated from each other by cloning and Sanger sequencing.

In addition to corroborating the splice junctions for the aforementioned transcripts, the Sanger sequencing results also showed another splice variant undetected in our RNA-seq transcriptome. This was a three-exon transcript (TRXPT\_30) with its first and last exons spliced identically as TRXPT\_23, but which also has the second exon of TRXPT\_24 (**Figure 7**). The first CDS on TRXPT\_30 spans all three exons, with the STC in the terminal (third) exon, producing a novel 140 residue, 15.7kDa protein (L22K). Interestingly, the last 81 C-terminus residues are identical to 22K (89 residues); only the first 7 residues are lacking due to splicing. Hence, we may consider L22K as a long variant of the predicted 22K ORF. Albeit the TSS and TTS of TRXPT\_30 was not seen, we presume that they are similar to TRXPT\_23, in which case it would also have the downstream CP of TRXPT\_23. TRXPT\_25 is the largest transcript in the TU. It also utilizes the classic E3 TSS but has distinct TTS. It is a two exon transcript, encoding a novel protein (t100K; 543 residues), which is a shorter isoform of the predicted 100K ORF. Considering secSSC usage on this transcript yields the predicted ORF, 22K. It also has the CP for pVIII and E3 in that order. Furthermore, during the validation of TRXPT\_25’s splice junction using primers that span its junction (18350-18717bp), we noticed a DNA band that corresponds to the full unspliced sequence (**Supplementary PCR methods**). As TRXPT\_25 only falls short of encoding the complete predicted 100K protein due to its splice junction, this band (which we cloned and validated by Sanger sequencing) suggests that the predicted 100K is indeed expressed. This transcript (TRXPT\_25B) although not seen in our RNA-seq data, likely shares the same TSS and TTS as TRXPT\_25. Lastly, TRXPT\_26 and TRXPT\_27 share the same TSS, unique from the other transcripts in this region but with distinct TTSs. TRXPT\_26 is a three-exon transcript but the first two are UTRs. It encodes pVIII as the 5’-most ORF and has the CP for E3 and Fiber in that order. TRXPT\_27 on the other hand, is only a two-exon transcript but similar to TRXPT\_26, only the terminal exon contains the CDSs. It encodes Fiber as the 5’-most ORF, and ORF7 downstream with secSSC usage. TRXPT\_13, which seems to be an E3 transcript that uses the MLP TSS is discussed under the MLTU transcripts.

**Early Region 4 (E4) transcripts**. This transcription unit (TU) is the found at the tail-end (3’-end) of the genome on the anti-sense strand. Based on nucleotide position, ORF7 and ORF8 were predicted in this region (1); however, as ORF7 is neither on the same strand as ORF8 nor transcribed from a promoter in the E4 region, only ORF8 can legitimately be classified as a transcript in this TU. This is corroborated by our RNA-seq data, as only one transcript was identified in this region on the anti-sense strand (**Figure 8**). The transcript (TRXPT\_28) spans 25192-26247 and is spliced at 25701-26055; hence, a two-exon transcript. The second exon fully matches the predicted ORF8 with 12 extra base pairs at the 3’-end; however, the encoded protein is an exact match. There is a SSC in the first exon at position 26246 (second nucleotide of the transcript). The encoded protein from this SSC is in-frame with the predicted SSC of ORF8 in the second exon; hence, the bulk of this longer protein (26.4 kDa, 229 aa) is identical to the predicted ORF8 protein.

**Major Late Transcription Unit (MLTU) or MLP Region transcripts**  
The MLTU transcripts dominate the late phase (i.e, after DNA replication) of the AdV infectious cycle. The MLP produces all late mRNAs by alternative splicing and alternative polyadenylation of a primary transcript, which are grouped into five transcript classes (L1-L5). These late proteins primarily act as capsid proteins, promote virion assembly, and direct genome packaging (6, 17, 18). Similar to other AdVs, most of THEV’s coding capacity falls within this TU. Specifically, about 13 out of the 23 predicted ORFs were assigned to this TU, some of which we have found to belong to the E3 TU instead. Our RNA-seq data revealed 12 transcripts (**TRXPT\_8, TRXPT\_9, TRXPT\_10, TRXPT\_11, TRXPT\_12, TRXPT\_13, TRXPT\_14, TRXPT\_16, TRXPT\_17, TRXPT\_18, TRXPT\_19, TRXPT\_20**) in this TU. We identified the untranslated TPL at the 5’ end of most transcripts in this region as expected. However, for three transcripts (**TRXPT\_16, TRXPT\_17, TRXPT\_18**), a different leader sequence (sTPL) is used: the first TPL exon is substituted for a unique first exon, found between the first and second TPL exons. Also, TRXPT\_20 seems to include only the third TPL exon (**Figure 9**).

We identified five TTSs (10,549bp, 12,709bp, 16,870bp, 17,891bp, 20,865bp) in this TU, corresponding to the five late mRNA classes (L1-L5), respectively, as found in all AdVs. L1 mRNAs include TRXPT\_8, which comprises the TPL (non-coding) and the CDS-containing terminal exon. This transcript encodes the 52K ORF exactly as predicted with the SSC beginning from the first nucleotide of the terminal exon. L2 mRNAs include TRXPT\_16, TRXPT\_17, and TRXPT\_18, all of which consist of the sTPL (also non-coding) followed by their respective terminal exons. TRXPT\_16 encodes pIIIa exactly as predicted as the 5’-most ORF, and also has the CP for the ORFs, III and pVII in that order. TRXPT\_17 encodes the ORF, III (penton), and TRXPT\_18 encodes the ORF pVII exactly as predicted. The L3 mRNAs include TRXPT\_14 and TRXPT\_20, of which TRXPT\_14 utilizes the full TPL whereas TRXPT\_20 uses only the third TPL exon (ex3TPL). Both transcripts have the CP for the ORF, hexon (II) but hexon is the only ORF encoded on TRXPT\_14, whereas the 5’-most ORF on TRXPT\_20 is pX (pre-Mu) followed by pVI and hexon in that order. L4 mRNAs include TRXPT\_9, TRXPT\_10, TRXPT\_11, and TRXPT\_13 all of which begin with the TPL followed by three (TRXPT\_9, TRXPT\_10, and TRXPT\_13) or four (TRXPT\_11) coding exons. These are the largest transcripts found in the transcriptome, each one possessing the CP for several similar late proteins. Normally, MLTU transcripts encoding particular ORFs splice the TPL onto a splice site just upstream of the ORF to be expressed (17). While this holds true for most MLTU ORFs, several late ORFs (pVI, protease, and ORF7) do not have such close proximity splicing but are contained in larger transcripts such as the L4 mRNAs, strongly suggesting the use of non-standard ribosomal initiation mechanisms such as secSSC utility and ribosome shunting found in other AdVs for their translation (17, 26). TRXPT\_9 and TRXPT\_10 are very similar but not identical. The last exon of TRXPT\_9 seems to be truncated and probably shares the same TTS as the other L4 mRNAs. They are both 6-exon transcripts encoding pVII as the 5’-most ORF (fourth exon) and also have the CP for pX, pVI, hexon, a longer variant of protease (Lprot) — uses an in-frame, upstream SSC than predicted, and 14K (a novel unpredicted 120 aa protein). TRXTP\_10 (and TRXPT\_9 with the L4 TTS) also has the CP for pVIII and E3. TRXPT\_11 is a seven-exon mRNA with hexon as it’s 5’-most ORF but it also has the CP for Lprot, 14K, L33K, and also pVIII and E3 in that order. TRXPT\_13 seems to be an E3 ORF utilizing the MLP TSS. It encodes pVIII and E3 in that order similar to TRXPT\_22 but lacks TRXPT\_22’s novel first ORF.

Lastly, the L5 class includes only TRXPT\_12 which contains the TPL and a coding terminal exon. Its 5’-most ORF is fiber (IV) but it also has the CP for the THEV specific gene, ORF7. TRXPT\_12’s CP is identical TRXPT\_27 of the the E3 TU albeit they differ in their 5’-UTRs.

## DISCUSSION/CONCLUSIONS

For fig2a: There is a dramatic increase of mean coverage/depth from **2.42** at 4 h.p.i to **95,042** at 72 h.p.i, strongly demonstrating an active infection. Unexpectedly, the pileup of reads seems consistently skewed over similar regions of the genome. We could speculate that the temporal gene expression regulation of THEV is different from MAdVs or this could simply mean that the infection was not well synchronized. However, the relative proportions over these similar regions shows some variation over time. For fig2b: titer reaching a plateau at 120 h.p.i, probably due to high cell death TRXPT\_2 and ORF1 are isoforms Presumably, if the junction reads were normalized, MLTU would not be predominant at 12hpi. The TTSs were all in the context of A-T rich sequences; which presumably serve as polyA signals. All splice junctions were confirmed by cloning and Sanger sequencing of cDNA (**Supplementary PCR methods**). We did not find the x,y,z or i-leaders for MLP transcripts probably because THEV doesn’t use it due to its smaller size The E3 ORF has an upstream, in-frame SSC.

## MATERIALS AND METHODS

### Cell culture and THEV Infection

The Turkey B-cell line (MDTC-RP19, ATCC CRL-8135) was grown as suspension cultures in 1:1 complete Leibovitz’s L-15/McCoy’s 5A medium with 10% fetal bovine serum (FBS), 20% chicken serum (ChS), 5% tryptose phosphate broth (TPB), and 1% antibiotics solution (100 U/mL Penicillin and 100ug/mL Streptomycin), at 41oC in a humidified atmosphere with 5% CO2. Infected cells were maintained in 1:1 serum-reduced Leibovitz’s L15/McCoy’s 5A media (SRLM) with 2.5% FBS, 5% ChS, 1.2% TPB, and 1% antibiotics solution (100 U/mL Penicillin and 100ug/mL Streptomycin). A commercially available HE vaccine was purchased from Hygieia Biological Labs as a source of THEV-A (VAS strain). The stock virus was titrated using an in-house qPCR assay with titer expressed as genome copy number(GCN)/mL, similar to Mahshoub *et al* (27) with modifications. Cells were infected at a multiplicity of infection (MOI) of 100 GCN/cell and samples in triplicates were harvested at 4-, 12-, 24-, and 72-h.p.i for RNA-seq. The infection was repeated but samples in triplicates were harvested at 12-, 24-, 36-, 48-, and 72-h.p.i for PCR validation of novel splice sites. Still one more independent infection was done at time points ranging from 12 to 168-h.p.i for qPCR quantification of virus titers.

### RNA extraction and Sequencing

Total RNA was extracted from infected cells using Thermofishers’ RNAqueous™-4PCR Total RNA Isolation Kit (#AM1914) per manufacturer’s instructions. An agarose gel electrophoresis was performed to check RNA integrity. The RNA quantity and purity was initially assessed using nanodrop, and RNA was used only if the A260/A280 ratio was 2.0 ± 0.05 and the A260/A230 ratio was >2 and <2.2. Extracted total RNA samples were sent to LC Sciences, Houston TX for poly-A-tailed mRNA sequencing where RNA integrity was checked with Agilent Technologies 2100 Bioanalyzer High Sensitivity DNA Chip and poly(A) RNA-seq library was prepared following Illumina's TruSeq-stranded-mRNA sample preparation protocol. Paired-end sequencing was performed on Illumina's NovaSeq 6000 sequencing system.

### Validation of Novel Splice Junctions

All splice junctions identified in this work are novel except one predicted splice site each for pTP and DBP, which were corroborated in our work. However, these predicted splice junctions had not been experimentally validated hitherto, and we identified additional novel exons, giving the complete picture of these transcripts. The novel splice junctions in this work discovered in the assembled transcripts using the StringTie transcript assembler which we validated by PCR and Sanger Sequencing are shown in **Supplementary PCR methods**. Briefly, we designed primers that crossed a range of novel exon-exon boundaries for each specific transcript in a transcription unit (TU) paired with their respective universal primers for the TU. Each forward primer contained a KpnI restriction site and reverse primers, an XbaI site. After first-strand cDNA synthesis with SuperScript™ III First-Strand Synthesis System, these primers were used in a targeted PCR amplification, the products analyzed with agarose gel electrophoresis to confirm expected band sizes, cloned by traditional restriction enzyme method, and Sanger sequenced to validate these splice junctions at the sequence level.

### 3’ Rapid Amplification of cDNA Ends (3’-RACE)

We performed a rapid amplification of sequences from the 3’ ends of mRNAs (3’-RACE) experiment using a portion of the extracted total RNA of infected MDTC-RP19 cells used for the RNA-seq experiment as explained above. We followed the protocol described by Green *et al* (28) with modifications. Briefly, 1ug of total RNA was reverse transcribed to cDNA using SuperScript™ IV First-Strand Synthesis System following the manufacturing instructions using an adapter-primer with a 3’-end poly(T) and a 5’-end BamHI restriction site. A gene-specific sense primer with a 5’-end KpnI restriction site paired with an anti-sense adapter-primer with a 5’-end BamHI site were used to amplify target sections of the cDNA using Invitrogen’s Platinum™ Taq DNA polymerase High Fidelity, following manufacturer’s instructions. The PCR amplicons were restriction digested, cloned, and Sanger sequenced.

### Computational Analysis of RNA Sequencing Data: Mapping and Transcript characterization

Our sequence reads were analyzed following a well established protocol described by Pertea *et al* (23), using Snakemake - version 7.24.0 (29), a popular workflow management system to drive the pipeline. Briefly, sequencing reads were trimmed with the FastQC - version 0.11.9 (30) program to achieve an overall Mean Sequence Quality (Phred Score) of 36. Trimmed reads were mapped simultaneously to the complete genomic sequence of avirulent turkey hemorrhagic enteritis virus strain Virginia (<https://www.ncbi.nlm.nih.gov/nuccore/AY849321.1/>) and *Meleagris gallopavo* (<https://www.ncbi.nlm.nih.gov/genome/?term=Meleagris+gallopavo>) using Hisat2 - version 2.2.1 (23) with default settings. The generated alignment (BAM) files from each infection time point were filtered for reads mapping to the THEV genome and fed into StringTie - version 2.2.1 (23) using a GTF file derived from a gff3 file obtained from NCBI, which contains the predicted ORFs of THEV as a guide. A custom script was used to consolidate all transcripts from all time-points without redundancy, generating the transcriptome of THEV. See **Supplementary Computational Analysis** for the details of transcript expression level estimations and splice junction read counts.

## SCRIPTS AND SUPPLEMENTARY MATERIALS

### DATA AVAILABILITY

### CODE AVAILABILITY

All the code/scripts written for analysis of the data are available on github (<https://github.com/Abraham-Quaye/thev_transcriptome>)

## ACKNOWLEDGMENTS

LC Sciences - RNA sequencing was done here  
Eton Bioscience, Inc, San Diego, CA - All Sanger sequencing validations was done here

## REFERENCES

1. Davison A, Benko M, Harrach B. 2003. [Genetic content and evolution of adenoviruses](https://doi.org/10.1099/vir.0.19497-0). The Journal of general virology 84:2895–908.

2. Harrach B. 2008. [Adenoviruses: General features](https://doi.org/10.1016/B978-012374410-4.00680-4), p. 1–9. *In* Mahy, BWJ, Van Regenmortel, MHV (eds.), Encyclopedia of virology (third edition). Book Section. Academic Press, Oxford.

3. Upton C, Slack S, Hunter AL, Ehlers A, Roper RL. 2003. [Poxvirus orthologous clusters: Toward defining the minimum essential poxvirus genome](https://doi.org/10.1128/jvi.77.13.7590-7600.2003). Journal of virology 77:7590–7600.

4. McGeoch D, Davison AJ. 1999. [Chapter 17 - the molecular evolutionary history of the herpesviruses](https://doi.org/10.1016/B978-012220360-2/50018-0), p. 441–465. *In* Domingo, E, Webster, R, Holland, J (eds.), Origin and evolution of viruses. Book Section. Academic Press, London.

5. Harrach B, Benko M, Both GW, Brown M, Davison AJ, Echavarría M, Hess M, Jones M, Kajon A, Lehmkuhl HD, Mautner V, Mittal S, Wadell G. 2011. Family adenoviridae. Virus Taxonomy: 9th Report of the International Committee on Taxonomy of Viruses 125–141.

6. Guimet D, Hearing P. 2016. [3 - adenovirus replication](https://doi.org/10.1016/B978-0-12-800276-6.00003-6), p. 59–84. *In* Curiel, DT (ed.), Adenoviral vectors for gene therapy (second edition). Book Section. Academic Press, San Diego.

7. Kovács ER, Benkő M. 2011. [Complete sequence of raptor adenovirus 1 confirms the characteristic genome organization of siadenoviruses](https://doi.org/10.1016/j.meegid.2011.03.021). Infection, Genetics and Evolution 11:1058–1065.

8. Davison AJ, Wright KM, Harrach B. 2000. [DNA sequence of frog adenovirus](https://doi.org/10.1099/0022-1317-81-10-2431). J Gen Virol 81:2431–2439.

9. Kovács ER, Jánoska M, Dán Á, Harrach B, Benkő M. 2010. [Recognition and partial genome characterization by non-specific DNA amplification and PCR of a new siadenovirus species in a sample originating from parus major, a great tit](https://doi.org/10.1016/j.jviromet.2009.10.007). Journal of Virological Methods 163:262–268.

10. Katoh H, Ohya K, Kubo M, Murata K, Yanai T, Fukushi H. 2009. [A novel budgerigar-adenovirus belonging to group II avian adenovirus of siadenovirus](https://doi.org/10.1016/j.virusres.2009.04.012). Virus Research 144:294–297.

11. Beach NM. 2006. [Characterization of avirulent turkey hemorrhagic enteritis virus: A study of the molecular basis for variation in virulence and the occurrence of persistent infection](http://scholar.lib.vt.edu/theses/available/etd-08142006-145339/). Thesis.

12. Beach NM, Duncan RB, Larsen CT, Meng XJ, Sriranganathan N, Pierson FW. 2009. [Comparison of 12 turkey hemorrhagic enteritis virus isolates allows prediction of genetic factors affecting virulence](https://doi.org/10.1099/vir.0.010090-0). J Gen Virol 90:1978–85.

13. Gross WB, Moore WE. 1967. Hemorrhagic enteritis of turkeys. Avian Dis 11:296–307.

14. Rautenschlein S, Sharma JM. 2000. [Immunopathogenesis of haemorrhagic enteritis virus (HEV) in turkeys](https://doi.org/10.1016/s0145-305x(99)00075-0). Dev Comp Immunol 24:237–46.

15. Larsen CT, Domermuth CH, Sponenberg DP, Gross WB. 1985. Colibacillosis of turkeys exacerbated by hemorrhagic enteritis virus. Laboratory studies. Avian Dis 29:729–32.

16. Dhama K, Gowthaman V, Karthik K, Tiwari R, Sachan S, Kumar MA, Palanivelu M, Malik YS, Singh RK, Munir M. 2017. [Haemorrhagic enteritis of turkeys – current knowledge](https://doi.org/10.1080/01652176.2016.1277281). Veterinary Quarterly 37:31–42.

17. Donovan-Banfield I, Turnell AS, Hiscox JA, Leppard KN, Matthews DA. 2020. [Deep splicing plasticity of the human adenovirus type 5 transcriptome drives virus evolution](https://doi.org/10.1038/s42003-020-0849-9). Communications Biology 3:124.

18. Zhao H, Chen M, Pettersson U. 2014. [A new look at adenovirus splicing](https://doi.org/10.1016/j.virol.2014.04.006). Virology 456-457:329–341.

19. Wolfrum N, Greber UF. 2013. [Adenovirus signalling in entry](https://doi.org/10.1111/cmi.12053). Cell Microbiol 15:53–62.

20. Falvey E, Ziff E. 1983. [Sequence arrangement and protein coding capacity of the adenovirus type 2 "i" leader](https://doi.org/10.1128/jvi.45.1.185-191.1983). Journal of Virology 45:185–191.

21. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See L-H, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR. 2012. [Landscape of transcription in human cells](https://doi.org/10.1038/nature11233). Nature 489:101–108.

22. Aboezz Z, Mahsoub H, El-Bagoury G, Pierson F. 2019. [In vitro growth kinetics and gene expression analysis of the turkey adenovirus 3, a siadenovirus](https://www.sciencedirect.com/science/article/abs/pii/S016817021830529X?via%3Dihub). Virus Research 263:47–54.

23. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. [Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and ballgown](https://doi.org/10.1038/nprot.2016.095). Nature Protocols 11:1650–1667.

24. Jack Fu [Aut], Alyssa C. Frazee [Aut, Cre], LeonardoCollado-Torres [Aut], Andrew E. Jaffe [Aut], Jeffrey T. Leek[Aut, Ths]. 2017. [Ballgown](https://doi.org/10.18129/B9.BIOC.BALLGOWN). Bioconductor.

25. Pitcovski J, Mualem M, Rei-Koren Z, Krispel S, Shmueli E, Peretz Y, Gutter B, Gallili GE, Michael A, Goldberg D. 1998. [The complete DNA sequence and genome organization of the avian adenovirus, hemorrhagic enteritis virus](https://doi.org/10.1006/viro.1998.9336). Virology 249:307–315.

26. Yueh A, Schneider RJ. 1996. [Selective translation initiation by ribosome jumping in adenovirus-infected and heat-shocked cells.](https://doi.org/10.1101/gad.10.12.1557) Genes &amp; Development 10:1557–1567.

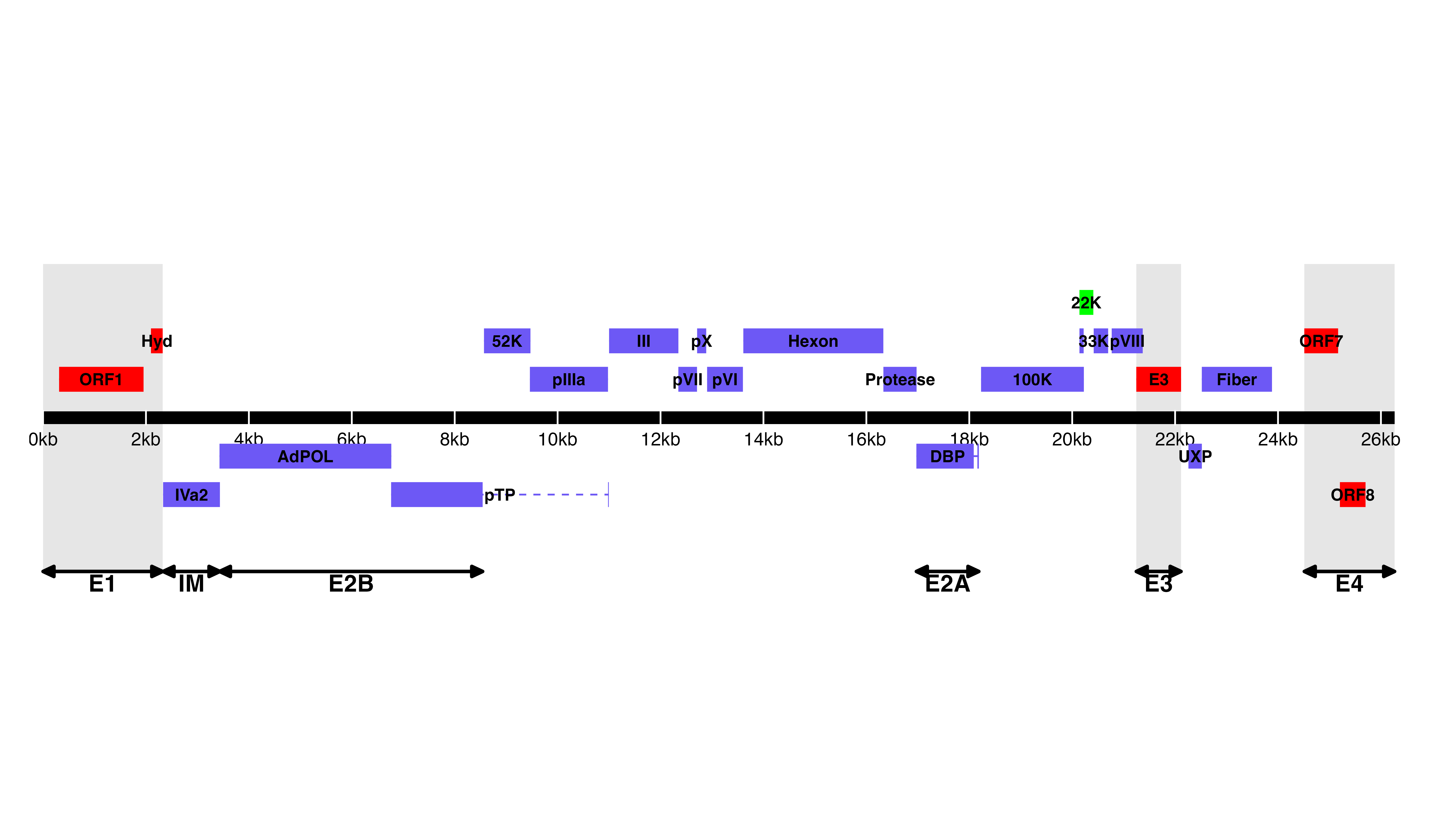
27. Mahsoub HM, Evans NP, Beach NM, Yuan L, Zimmerman K, Pierson FW. 2017. [Real-time PCR-based infectivity assay for the titration of turkey hemorrhagic enteritis virus, an adenovirus, in live vaccines](https://doi.org/10.1016/j.jviromet.2016.11.002). Journal of Virological Methods 239:42–49.

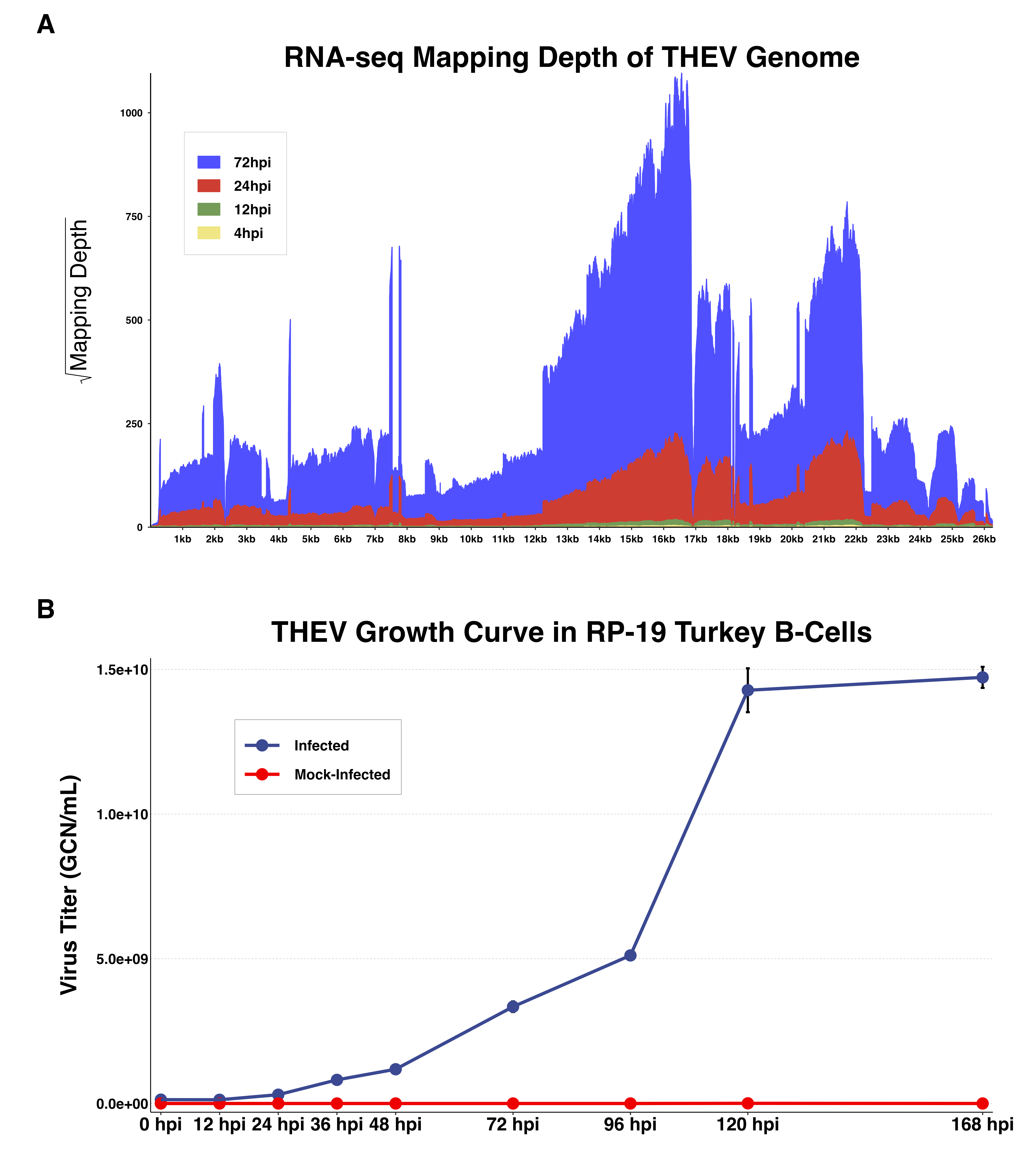
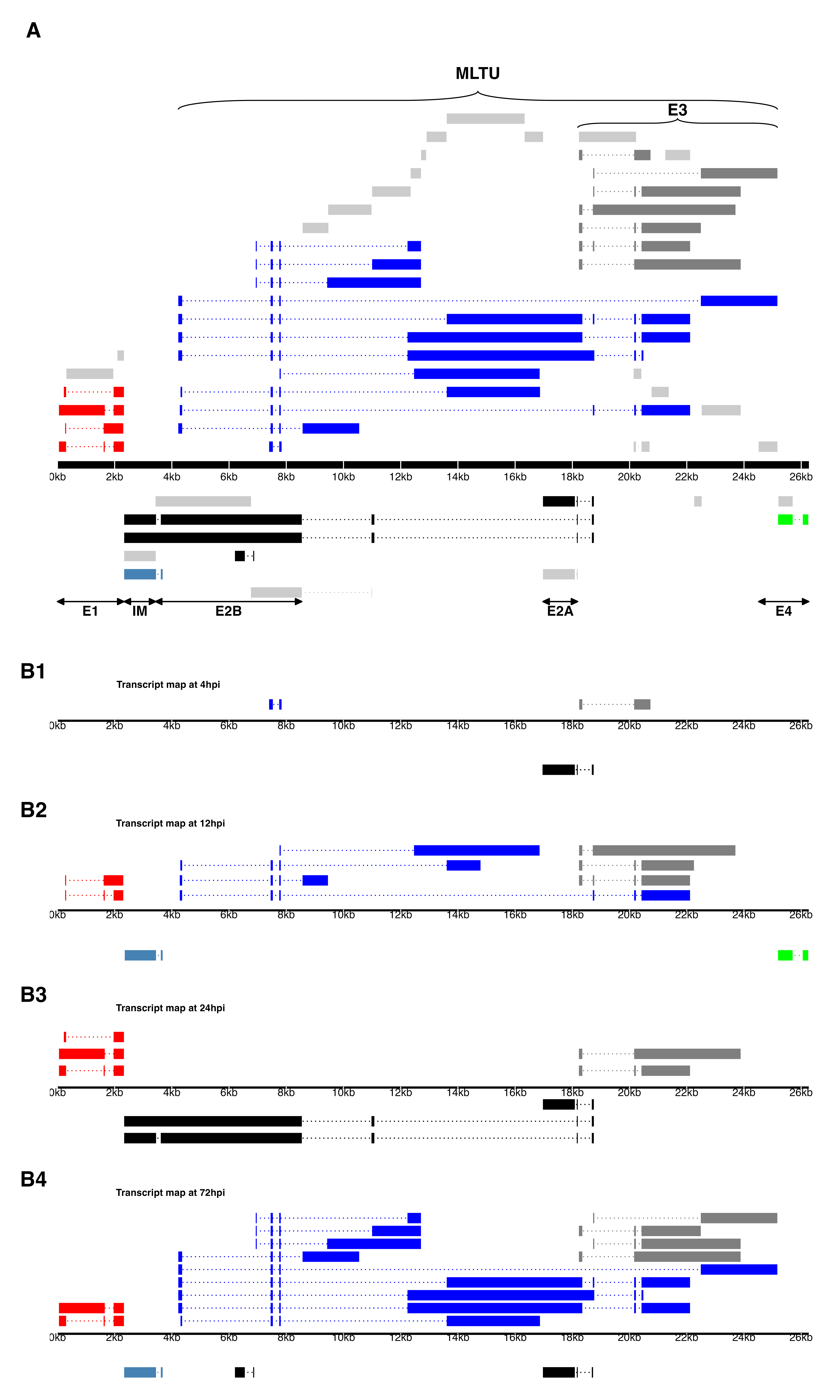
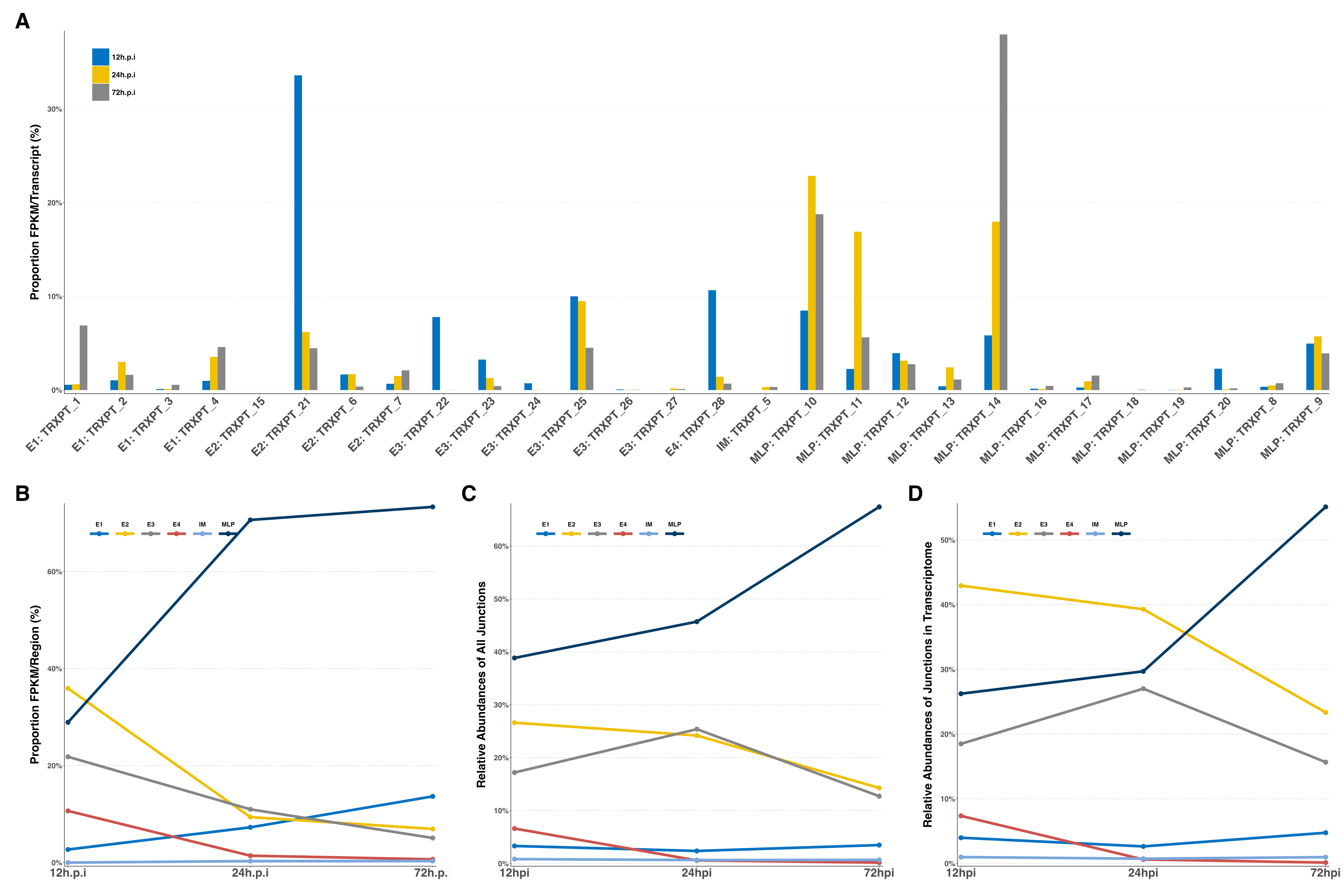
28. Green MR, Sambrook J. 2019. [Rapid amplification of sequences from the 3’ ends of mRNAs: 3’-RACE](https://doi.org/10.1101/pdb.prot095216). Cold Spring Harbor Protocols 2019:pdb.prot095216.

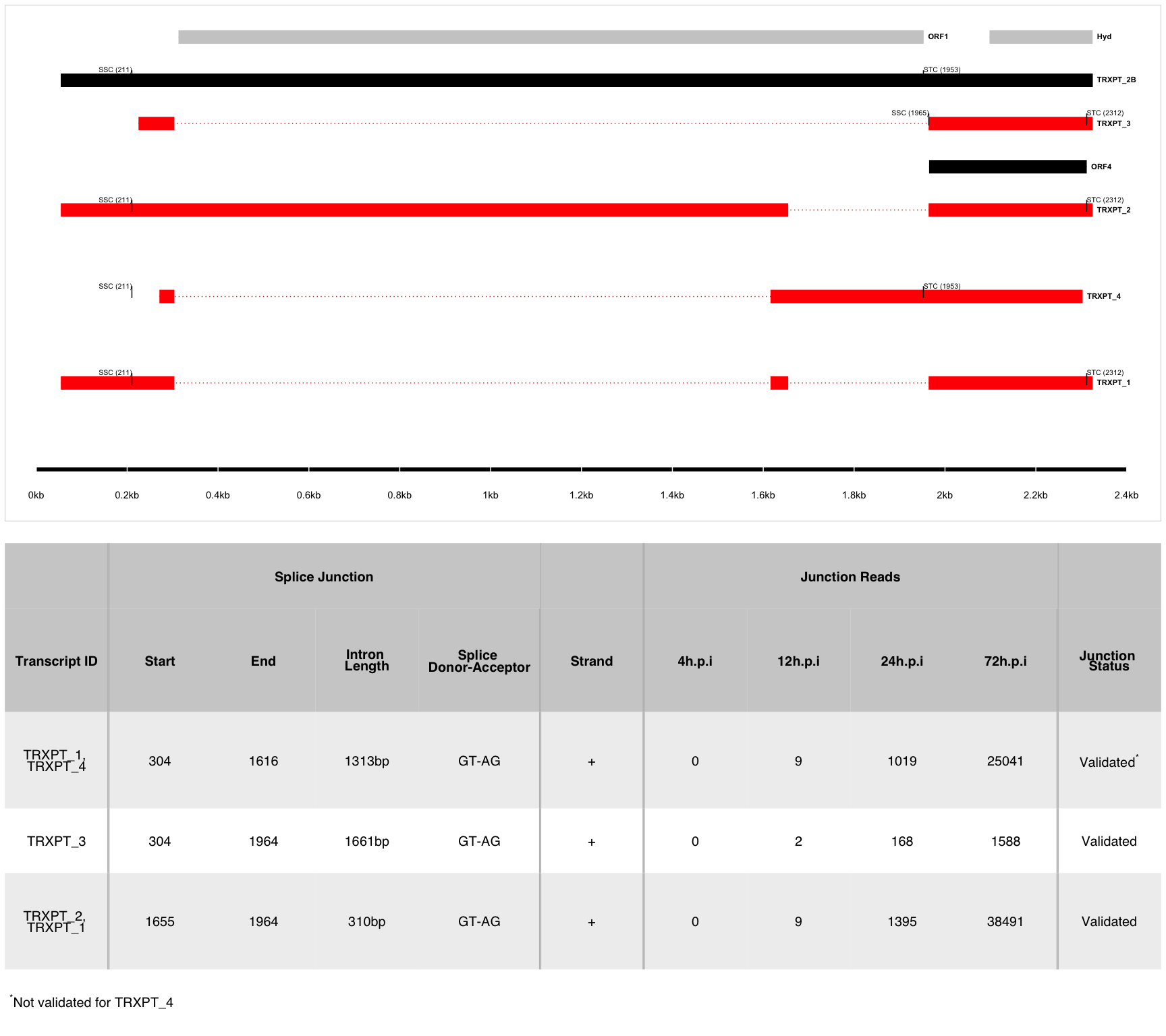
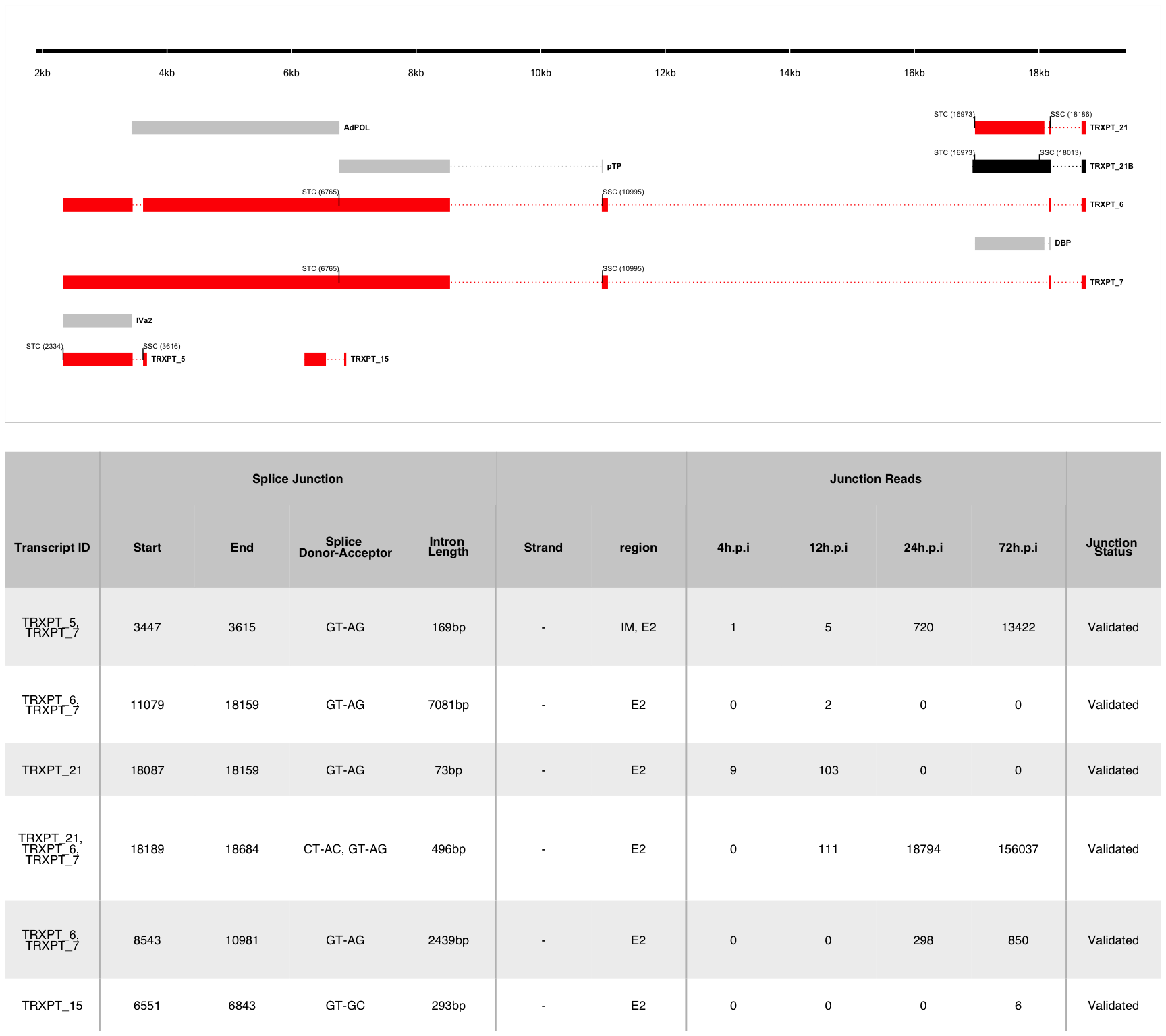
29. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, Wilm A, Holtgrewe M, Rahmann S, Nahnsen S, Köster J. 2021. [Sustainable data analysis with snakemake](https://doi.org/10.12688/f1000research.29032.2). F1000Research 10:33.

30. 2015. [FastQC](https://qubeshub.org/resources/fastqc).

## TABLES AND FIGURES

 **Figure 1. *Genomic map of THEV avirulent strain annotated ORFs***. The central horizontal line represents the double-stranded DNA marked at 5kb intervals as white line breaks. Blocks represent viral genes. Blocks above the DNA line are transcribed rightward, those below are transcribed leftward. pTP, DBP and 33K predicted to be spliced are shown as having tails. Shaded regions indicate regions containing “genus-specific” genes (colored red). Genes colored in blue are “genus-common”. Gene colored in light green is conserved in all but Atadenoviruses. The UXP (light blue) is an incomplete gene present in almost all AdVs. Regions comprising the different transcription units are labelled at the bottom (E1, E2A, E2B, E3, E4, and IM); the unlabeled regions comprise the MLTU.

 **Figure 2: Increasing levels of THEV over time. a) *Per base coverage of sequence reads mapping to THEV genome by time point***. The pileup of mRNA reads mapping to THEV genome at the base-pair level for each indicated time point. **b) *Growth curve of THEV (VAS vaccine strain) in MDTC-RP19 cell line***. Virus titers were a quantified with a qPCR assay. There is no discernible increase in virus titer up 12 h.p.i, after which a steady increase in virus titer is measured. The virus titer expands exponentially beginning from 48 h.p.i, increasing by orders of magnitude before reaching a plateau at 120 h.p.i. GCN: genome copy number.  
 **Figure 3. a) Transcriptome of THEV from RNA-seq**. THEV transcripts assembled from all time points by StringTie are unified forming this final transcriptome (splicing map). Transcripts belonging to the same transcription unit (TU) are located in close proximity on the genome and are color coded and labeled in this figure as such. The organization of TUs in the THEV genome is unsurprisingly similar to MAdVs; however, the MAdV genome shows significantly more transcripts. The TUs are color coded: E1 transcripts - red, E2 - black, E3 - dark grey, E4 - green, MLTU - blue. Predicted ORFs are also indicated here, colored light grey. **b) THEV transcripts identified at given time points**. Transcripts are color coded as explained in **a**.  
 **Figure 4: Changes in splicing and expression profile of THEV over time.** **a)** *Normalized (FPKM) expression levels of transcripts over time*. The expression levels (FPKM) of individual transcripts as a percentage of the total expression of all transcripts at each time point are indicated. Only transcripts from our RNA-seq data are included here. **b)** *Normalized (FPKM) expression levels of transcripts by region over time*. The expression levels of each region/TU as a percentage of the total expression of all transcripts at each time point are indicated. Region expression levels were calculated by summing up the FPKMs of all transcripts categorized in that region. **c)** *Relative abundances of all splice junctions grouped by region/TU over time*. After assigning all **2,457** unique junctions to a TU and the total junction reads counted at each time point for each region, the total junction reads for each TU plotted as percentage of all junction reads at each time point is indicated. Note that the junction read counts are not normalized. **d)** *Relative abundances of junctions in transcriptome grouped by region/TU over time*. This is identical to **(c)**, except that only the junctions found in the full transcriptome obtained from the RNA-seq data were included.

 **Figure 5: The splice map of the E1 transcription unit (TU).** Exons are depicted as boxes connected by introns (dotted lines). Transcripts from RNA-seq data are colored red, predicted ORFs are colored grey, and transcripts or ORFs discovered by other means are colored black. Each transcript or ORF is labelled with its name to the right. The start codon (SSC) and stop codon (STC) of the 5’-most CDS of each transcript is indicated with the nucleotide position in brackets. The region of the virus is depicted at the bottom as a black line with labels of the nucleotide positions for reference. The table shows sequence reads covering the splice junctions with information about their validation status using cloning and Sanger sequencing.  **Figure 6: The splice map of the E2 and IM TUs**. Exons are depicted as boxes connected by introns (dotted lines). Red transcripts are generated from RNA-seq data and predicted ORFs are colored grey. TRXPT\_21B discovered by 3’RACE is colored black. Each transcript or ORF is labelled with its name to the right. The SSC and STC of the 5’-most CDS of each transcript is indicated with the nucleotide position in brackets. The region of the virus is depicted at the bottom as a black line with labels of the nucleotide positions for reference. The table shows sequence reads covering the splice junctions with information about their validation status using cloning and Sanger sequencing.

 **Figure 7: The splice map of the E3 TU.** Exons are depicted as boxes connected by introns (dotted lines). Red transcripts are generated from RNA-seq data and predicted ORFs are colored grey. Transcripts discovered by other means are colored black. Each transcript or ORF is labelled with its name to the right. The start codon (SSC) and stop codon (STC) of the 5’-most CDS of each transcript is indicated with the nucleotide position in brackets. Similarly, the secondary SSC (secSSC) and secondary STC (secSTC) are shown. The region of the virus is depicted at the bottom as a black line with labels of the nucleotide positions for reference. The table shows sequence reads covering the splice junctions with information about their validation status using cloning and Sanger sequencing.

 **Figure 8: The splice map of the E4 TU.** Exons are depicted as boxes connected by introns (dotted lines). The transcript from RNA-seq data is colored red and the predicted ORF, grey. The transcript and ORF are labelled with their names to the right. The start codon (SSC) and stop codon (STC) of the 5’-most CDS is indicated with the nucleotide position in brackets. The region of the virus is depicted at the bottom as a black line with labels of the nucleotide positions for reference. The table shows sequence reads covering the splice junction with its validation status using cloning and Sanger sequencing.

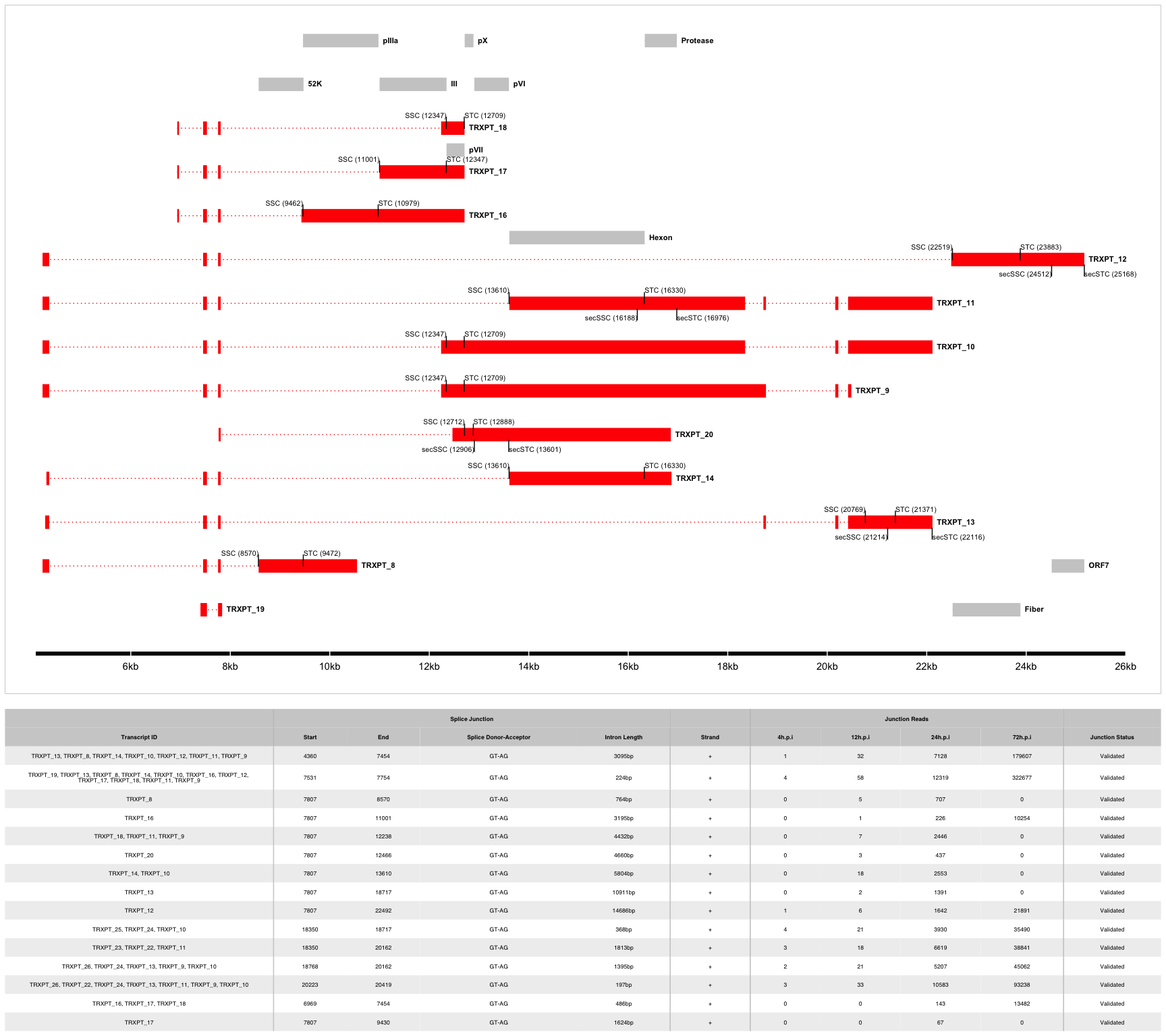
 **Figure 9: The splice map of the MLTU.** Exons are depicted as boxes connected by introns (dotted lines). The transcripts from our RNA-seq data are colored red and the predicted ORFs, grey. The transcripts and ORFs are labelled with their names to the right. The start codon (SSC) and stop codon (STC) of the 5’-most CDS of each transcript is indicated with the nucleotide position in brackets. Similarly, the secondary SSC (secSSC) and secondary STC (secSTC) are shown. The region of the virus is depicted at the bottom as a black line with labels of the nucleotide positions for reference. The table shows sequence reads covering the splice junctions with information about their validation status using cloning and Sanger sequencing.

Table 1: Overview of sequencing results

| **Metric** | **4h.p.i** | **12h.p.i** | **24h.p.i** | **72h.p.i** | **Total** |
| --- | --- | --- | --- | --- | --- |
| **Total reads** | 1.17e+08 | 7.63e+07 | 1.20e+08 | 1.15e+08 | 4.28e+08 |
| **Mapped   (Host)** | 1.04e+08  (89.06%) | 6.79e+07  (89.0393%) | 1.06e+08  (88.2719%) | 8.38e+07  (72.9802%) | 3.62e+08 |
| **Mapped   (THEV)** | 4.32e+02  ( 0.0004%) | 6.70e+03  ( 0.0088%) | 1.18e+06  ( 0.9841%) | 1.69e+07  (14.6904%) | 1.81e+07 |
| **Mean Per Base   Coverage/Depth** | 2.42 | 37.71 | 6,666.96 | 95,041.7 | 101,749 |
| **Total unique   splice junctions** | 13 | 37 | 236 | 2,374 | 2,457 |
| **Junction coverage   Total (at least 1 read)** | 37 | 605 | 115,075 | 2.13e+06 | 2.25e+06 |
| **Junction coverage   Mean reads** | 2.8 | 16.4 | 487.6 | 898.4 | 351.3 |
| **Junction coverage   (at least 10 reads)** | 0 | 13 | 132 | 1,791 | 1,936 |
| **Junction coverage   (at least 100 reads)** | 0 | 1 | 53 | 805 | 859 |
| **Junction coverage   (at least 1000 reads)** | 0 | 0 | 18 | 168 | 186 |

Table 2a: Most abundant splice junctions at 12h.p.i

| **Timepoint** | **Strand** | **Start** | **End** | **Splice\_Site** | **Region** | **Intron Length** | **Reads (Percentage)** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 12hpi | - | 18,087 | 18,159 | GT-AG | E2 | 72 bp | 103 (17%) |
| 12hpi | + | 18,189 | 18,684 | CT-AC | MLP | 495 bp | 97 (16%) |
| 12hpi | + | 7,531 | 7,754 | GT-AG | MLP | 223 bp | 58 (9.6%) |
| 12hpi | - | 25,701 | 26,055 | GT-AG | E4 | 354 bp | 37 (6.1%) |
| 12hpi | + | 20,223 | 20,419 | GT-AG | E3 | 196 bp | 33 (5.5%) |
| 12hpi | + | 4,360 | 7,454 | GT-AG | MLP | 3,094 bp | 32 (5.3%) |
| 12hpi | - | 18,751 | 20,668 | GT-AG | E2 | 1,917 bp | 22 (3.6%) |
| 12hpi | + | 18,350 | 18,717 | GT-AG | E3 | 367 bp | 21 (3.5%) |
| 12hpi | + | 18,768 | 20,162 | GT-AG | E3 | 1,394 bp | 21 (3.5%) |
| 12hpi | + | 7,807 | 13,610 | GT-AG | MLP | 5,803 bp | 18 (3%) |
| 12hpi | + | 18,350 | 20,162 | GT-AG | E3 | 1,812 bp | 18 (3%) |
| 12hpi | - | 18,189 | 18,684 | GT-AG | E2 | 495 bp | 14 (2.3%) |
| 12hpi | - | 18,751 | 21,682 | GT-AG | E2 | 2,931 bp | 10 (1.7%) |
| 12hpi | + | 304 | 1,616 | GT-AG | E1 | 1,312 bp | 9 (1.5%) |
| 12hpi | + | 1,655 | 1,964 | GT-AG | E1 | 309 bp | 9 (1.5%) |
| 12hpi | - | 18,087 | 18,163 | GT-AG | E2 | 76 bp | 8 (1.3%) |
| 12hpi | + | 7,807 | 12,238 | GT-AG | MLP | 4,431 bp | 7 (1.2%) |
| 12hpi | + | 7,807 | 22,492 | GT-AG | MLP | 14,685 bp | 6 (1%) |

Table 2b: Most abundant splice junctions at 24h.p.i

| **Timepoint** | **Strand** | **Start** | **End** | **Splice\_Site** | **Region** | **Intron Length** | **Reads (Percentage)** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 24hpi | - | 18,087 | 18,159 | GT-AG | E2 | 72 bp | 18,825 (16.4%) |
| 24hpi | + | 18,189 | 18,684 | CT-AC | MLP | 495 bp | 17,670 (15.4%) |
| 24hpi | + | 7,531 | 7,754 | GT-AG | MLP | 223 bp | 12,319 (10.7%) |
| 24hpi | + | 20,223 | 20,419 | GT-AG | E3 | 196 bp | 10,583 (9.2%) |
| 24hpi | + | 4,360 | 7,454 | GT-AG | MLP | 3,094 bp | 7,128 (6.2%) |
| 24hpi | + | 18,350 | 20,162 | GT-AG | E3 | 1,812 bp | 6,619 (5.8%) |
| 24hpi | + | 18,768 | 20,162 | GT-AG | E3 | 1,394 bp | 5,207 (4.5%) |
| 24hpi | + | 18,350 | 18,717 | GT-AG | E3 | 367 bp | 3,930 (3.4%) |
| 24hpi | - | 18,751 | 20,668 | GT-AG | E2 | 1,917 bp | 3,870 (3.4%) |
| 24hpi | + | 7,807 | 13,610 | GT-AG | MLP | 5,803 bp | 2,553 (2.2%) |
| 24hpi | + | 7,807 | 12,238 | GT-AG | MLP | 4,431 bp | 2,446 (2.1%) |
| 24hpi | + | 7,807 | 22,492 | GT-AG | MLP | 14,685 bp | 1,642 (1.4%) |
| 24hpi | + | 1,655 | 1,964 | GT-AG | E1 | 309 bp | 1,395 (1.2%) |
| 24hpi | + | 7,807 | 18,717 | GT-AG | MLP | 10,910 bp | 1,391 (1.2%) |
| 24hpi | - | 18,189 | 18,684 | GT-AG | E2 | 495 bp | 1,124 (1%) |
| 24hpi | - | 18,751 | 21,128 | GT-AG | E2 | 2,377 bp | 1,124 (1%) |
| 24hpi | + | 20,223 | 20,894 | GT-AG | E3 | 671 bp | 1,208 (1%) |

Table 2c: Most abundant splice junctions at 72h.p.i

| **Timepoint** | **Strand** | **Start** | **End** | **Splice\_Site** | **Region** | **Intron Length** | **Reads (Percentage)** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 72hpi | + | 7,531 | 7,754 | GT-AG | MLP | 223 bp | 322,677 (15.1%) |
| 72hpi | + | 4,360 | 7,454 | GT-AG | MLP | 3,094 bp | 179,607 (8.4%) |
| 72hpi | - | 18,087 | 18,159 | GT-AG | E2 | 72 bp | 161,336 (7.6%) |
| 72hpi | + | 18,189 | 18,684 | CT-AC | MLP | 495 bp | 146,425 (6.9%) |
| 72hpi | + | 20,223 | 20,419 | GT-AG | E3 | 196 bp | 93,238 (4.4%) |
| 72hpi | + | 7,807 | 13,610 | GT-AG | MLP | 5,803 bp | 81,420 (3.8%) |
| 72hpi | + | 7,807 | 12,238 | GT-AG | MLP | 4,431 bp | 77,616 (3.6%) |
| 72hpi | + | 18,768 | 20,162 | GT-AG | E3 | 1,394 bp | 45,062 (2.1%) |
| 72hpi | + | 1,655 | 1,964 | GT-AG | E1 | 309 bp | 38,491 (1.8%) |
| 72hpi | + | 18,350 | 20,162 | GT-AG | E3 | 1,812 bp | 38,841 (1.8%) |
| 72hpi | + | 18,350 | 18,717 | GT-AG | E3 | 367 bp | 35,490 (1.7%) |
| 72hpi | + | 304 | 1,616 | GT-AG | E1 | 1,312 bp | 25,041 (1.2%) |
| 72hpi | - | 18,751 | 20,668 | GT-AG | E2 | 1,917 bp | 26,338 (1.2%) |
| 72hpi | + | 7,807 | 12,904 | GT-AG | MLP | 5,097 bp | 21,946 (1%) |
| 72hpi | + | 7,807 | 22,492 | GT-AG | MLP | 14,685 bp | 21,891 (1%) |