

# Elucidating the Transcriptome of Turkey Hemorrhagic Enteritis Virus

3

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

<sup>5</sup> Abraham Quaye<sup>1\*</sup>, Bret Pickett<sup>\*</sup>, Joel S. Griffitts<sup>\*</sup>, Bradford K. Berges<sup>\*</sup>, Brian D. Poole<sup>†\*</sup>

<sup>6</sup>\*Department of Microbiology and Molecular Biology, Brigham Young University

7 1 First-author

<sup>8</sup> † Corresponding Author

## **9 Corresponding Author Information**

<sup>10</sup> brian\_poole@byu.edu

11 Department of Microbiology and Molecular Biology,

<sup>12</sup> 4007 Life Sciences Building (LSB),

<sup>13</sup> Brigham Young University,

**14 Provo, Utah**

15

16 **ABSTRACT**

17 **Background:** Hemorrhagic enteritis (HE) is a disease affecting 6-12-week-old turkeys characterized by *im-*  
18 *munosuppression (IS)* and bloody diarrhea. This disease is caused by *Turkey Hemorrhagic Enteritis Virus*  
19 (*THEV*) of which avirulent strains (*THEV-A*) that do not cause HE but retain the immunosuppressive ability  
20 have been isolated. The *THEV-A* Virginia Avirulent Strain (VAS) is still used as a live vaccine despite its  
21 immunosuppressive properties. *Our objective is to understand the genetic basis by which VAS induces*  
22 *IS.* The transcriptome of *THEV* was studied to set the stage for further experimentation with specific viral  
23 genes that may mediate IS.

24 **Methods:** After infecting a turkey B-cell line (MDTC-RP19) with the VAS vaccine strain, samples in tripli-  
25 cates were collected at 4-, 12-, 24-, and 72-hours post-infection. Total RNA was subsequently extracted,  
26 and poly-A-tailed mRNA sequencing done. After trimming the raw sequencing reads with the FastQC, reads  
27 were mapped to the *THEV* genome using Hisat2 and transcripts assembled with StringTie. An in-house  
28 script was used to consolidate transcripts from all time-points, generating the final transcriptome. PCR, gel  
29 electrophoresis, and Sanger sequencing were used to validate all identified splice junctions.

30 **INTRODUCTION**

31 Adenoviruses (AdVs) are non-enveloped icosahedral-shaped DNA viruses, causing infection in virtually all  
32 vertebrates. Their double-stranded linear DNA genomes range between 26 and 45kb in size, producing a  
33 broad repertoire of transcripts via a highly complex alternative splicing pattern (1, 2). The AdV genome is  
34 one of the most optimally economized; both the forward and reverse DNA strands harbor protein-coding  
35 genes, making it highly gene-dense. There are 16 genes termed “genus-common” that are homologous in  
36 all AdVs; these are thought to be inherited from a common ancestor. All other genes are termed “genus-  
37 specific”. “Genus-specific” genes tend to be located at the termini of the genome while “genus-common”  
38 genes are usually central (1). This pattern is observed in *Adenoviridae*, *Poxviridae*, and *Herpesviridae* (1,  
39 3, 4). The family *Adenoviridae* consists of five genera: *Mastadenovirus* (MAdV), *Aviadenovirus*, *Ataden-  
40 ovirus*, *Ictadenovirus*, and *Siadenovirus* (SiAdV) (5, 6). Currently, there are three recognized members  
41 of the genus SiAdV: frog adenovirus 1, raptor adenovirus 1, and turkey adenovirus 3 also called turkey  
42 hemorrhagic enteritis virus (THEV) (5, 7–10). Members of SiAdV have the smallest genome size (~26 kb)  
43 and gene content (~23 genes) of all known AdVs, and many “genus-specific” putative genes of unknown  
44 functions have been annotated (see **Figure 1**) (1, 2, 7).

45 Virulent THEV strains (THEV-V) and avirulent strains (THEV-A) of THEV are serologically indistinguishable,  
46 infecting turkeys, chickens, and pheasants and the THEV-V cause different clinical diseases in these birds  
47 (2, 11). In turkeys, the THEV-V cause hemorrhagic enteritis (HE), a debilitating acute disease affecting pre-  
48 dominantly 6-12-week-old turkeys characterized by immunosuppression (IS), weight loss, intestinal lesions  
49 leading to bloody diarrhea, splenomegaly, and up to 80% mortality (11–13). HE is the most economically  
50 significant disease caused by any strain of THEV (11). While the current vaccine strain (a THEV-A isolated  
51 from a pheasant, Virginia Avirulent Strain [VAS]) have proven effective at preventing HE in young turkey  
52 poulets, it still retains the immunosuppressive ability. Thus, vaccinated birds are rendered more susceptible  
53 to opportunistic infections and death than unvaccinated cohorts leading to substantial economic losses (11,  
54 14–16). The induced IS also interferes with vaccination schemes for other infections of turkeys (11, 14).  
55 To eliminate this immunosuppressive side-effect of the vaccine, a thorough investigation of the culprit viral  
56 factors (genes) mediating this phenomenon is essential. However, the transcriptome (splicing and gene ex-  
57 pression patterns) of THEV has not been characterized, making the investigation of specific viral genes for  
58 possible roles in causing IS impractical. A well-characterized transcriptome of THEV is required to enable  
59 the next leap forward in THEV research - experimentation with specific viral genes that may mediate IS.  
60 Myriads of studies have elucidated the AdV transcriptome in fine detail (17, 18). However, a large pre-

61 ponderance of studies focus on MAdVs - specifically human AdVs - thus, most of the current knowledge  
62 regarding AdV gene expression and replication is based on MAdV studies, which is generalized for all other  
63 AdVs (6, 19). MAdV genes are transcribed in a temporal manner; therefore, genes are categorized into five  
64 early transcription units (E1A, E1B, E2, E3, and E4), two intermediate (IM) units (pIX and IVa2), and one  
65 major late unit (MLTU or major late promoter [MLP]), which generates five families of late mRNAs (L1-L5).  
66 An additional gene (UXP or U exon) is located on the reverse strand. The early genes encode non-structural  
67 proteins such as enzymes or host cell modulating proteins, primarily involved in DNA replication or provid-  
68 ing the necessary intracellular niche for optimal replication while late genes encode structural proteins. The  
69 immediate early gene E1A is expressed first, followed by the the delayed early genes, E1B, E2, E3 and  
70 E4. Then the intermediate early genes, IVa2 and pIX are expressed followed by the late genes (6, 17, 18).  
71 MAdV makes an extensive use of alternative RNA splicing to produce a very complex array of mRNAs. All  
72 but the pIX mRNA undergo at least one splicing event. The MLTU produces over 20 distinct splice variants  
73 all of which contain three non-coding exons at the 5'-end (collectively known as the tripartite leader, TPL)  
74 (17, 18). There is also an alternate 5' three non-coding exons present in varying amounts on a subset of  
75 MLTU mRNAs (known as the x-, y- and z-leaders). Lastly, there is the i-leader exon, which is infrequently  
76 included between the second and third TPL exons, and codes for the i-leader protein (20). Thus, the MLTU  
77 produces a complex repertoire of mRNA with diverse 5' untranslated regions (UTRs) spliced onto differ-  
78 ent 3' coding exons which are grouped into five different 3'-end classes (L1-L5). Each transcription unit  
79 (TU) contains its own promoter driving the expression of all the array of mRNA transcripts produced via  
80 alternative splicing of the genes encoded in the unit(6, 17, 18). During translation of AdV mRNA, recent  
81 studies strongly suggest the potential usage of secondary start codons; adding to what was already a highly  
82 complex system for gene expression (17).

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

91 **RESULTS**

92 **Overview of sequencing data and analysis pipeline outputs**

93 A previous study by Zeinab *et al* showed that almost all THEV transcripts were detectable beginning at  
94 4 hours (22). Therefore, infected MDTC-RP19 cells were harvested at 4-, 12-, 24-, and 72-hours post-  
95 infection(h.p.i) to ensure an amply wide time window to sample all transcripts. Our paired-end RNA se-  
96 quencing (RNA-seq) experiment yielded an average of **107.1** million total reads of **149bp** in length per  
97 time-point, which were simultaneously mapped to both the virus (THEV) and host (*M.gallipavo*) genomes  
98 using the Hisat2 (23) alignment program. A total of **18.1** million reads from all time-points mapped to the  
99 virus genome; this provided good coverage/depth, leaving no regions unmapped. The mapped reads to  
100 the virus genome increased substantially from **432** reads at 4 h.p.i to **16.9** million reads at 72 h.p.i (**Table**  
101 **1**, **Figure 2a**). From the mapped reads, we identified a total of **2,457** unique THEV splice junctions from all  
102 time-points, with splice junctions from the later time-points being supported by significantly more sequence  
103 reads than earlier time-points. For example all the **13** unique junctions at 4 h.p.i had less than 10 reads  
104 supporting each one, averaging a mere **2.8** reads/junction. Conversely, the **2374** unique junctions at 72 h.p.i  
105 averaged **898.4** reads/junction, some junctions having coverage as high as **322,677** reads. The substantial  
106 increases in splice junctions and mapping reads to the THEV genome over time denotes an active infection,  
107 and correlates with our quantitative PCR (qPCR) assay quantifying the total number of viral genome copies  
108 over time (**Figure 2b**). Using StringTie (23), an assembler of RNA-seq alignments into potential tran-  
109 scripts, the mapped reads for each time-point were assembled into transcripts using the genomic location  
110 of the predicted THEV ORFs as a guide. In the consolidated transcriptome, a composite of all unredudant  
111 transcripts from all time points, we counted a total of **28** transcripts all of which are novel, and using 3'  
112 Rapid Amplification of cDNA Ends (3'RACE) and other methods, we further identified ##### unique splice  
113 variants. Although some exons in some transcripts match the predicted ORFs exactly, most of our identified  
114 exons are longer, spanning multiple predicted ORFs (**Figure 3**). The complete list of unique splice junctions  
115 mapped to THEV's genome has been submitted to the National Center for Biotechnology Information Gene  
116 Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo>) under **accession no. XXXXXX**.

117 **Changes in THEV splicing profile over time**

118 AdV gene expression occurs under exquisite temporal control, supervised by designated promoters for  
119 each transcription unit (TU) or region. Each promoter typically produce one or few pre-mRNAs that undergo  
120 alternative splicing to yield the manifold repertoire of complex transcripts characteristic of AdVs (17, 18).  
121 To evaluate the activity of each promoter over time, Firstly, Ballgown, a program for statistical analysis of  
122 assembled transcriptomes (24) was used to estimate and normalize expression levels of all transcripts for

123 each time point in Fragments Per Kilobase of transcript per Million mapped reads (FPKM) units. Very few  
124 unique splice junctions, reads, and transcripts were counted at 4 h.p.i; hence, this time point was excluded  
125 in this analysis.

126 Individually, TRXPT\_21 (DBP) - from the E2 region - was the most significantly expressed at 12 h.p.i,  
127 comprising about **33.58%** of the total transcripts. Transcripts in the E3 and E4 regions also contributed  
128 significant proportions, and noticeably, some MLP region transcripts. The later time points were dominated  
129 by the MLP region; transcripts TRXPT\_10 and TRXPT\_14 were most abundantly expressed at 24 and 72  
130 h.p.i, respectively, as expected (**Figure 4a**). When we performed analysis of the FPKM values of transcripts  
131 per region we found a similar pattern: the E2 region was the most abundantly expressed at 12 h.p.i, after  
132 which the MLP region assumes predominance (**Figure 4b**). Secondly, we estimated relative abundances  
133 of all splice junctions for each time point using the raw reads. We counted as significantly expressed only  
134 junctions with coverage of at least 1% of the total splice junction reads counted at the given time point.  
135 At 12 h.p.i, **18** junctions meet the 1% threshold, and were comprised of predominantly early region (E1,  
136 E2, E3, and E4) junctions, albeit the MLP was the single most preponderant region overall, constituting  
137 **38.8%** of all the junctions reads. (**Table 2a**). The levels of the top most abundant junctions at 12 h.p.i were  
138 maintained also 24 h.p.i as the most significantly expressed. However, here, the MLP-derived junctions  
139 were unsurprisingly even more preponderant overall, accounting for **45.7%** of all the junction reads counted  
140 (**Table 2b**). At 72 h.p.i, the trend of increased activity of the MLP continued as expected; at this time, the  
141 MLP-derived junctions were not only the most abundant overall - accounting for **67.4%** of all junctions reads -  
142 but also contained the most significantly expressed individual junctions (**Table 2c**. Also see **Supplementary**  
143 **Tables 1a-c; Figure 4c**). When we limited this analysis to only junctions in the final transcriptome, the  
144 relative abundances of the junctions for each region over time was generally similar to the pattern seen with  
145 all the junctions included (**Figure 4d**).  
146 Furthermore, we analyzed splice donor and acceptor site nucleotide usage over time to investigate any  
147 peculiarities that THEV may show, generally or over the course of the infection. We found that most splice  
148 donor-acceptor sequences were unsurprisingly the canonical GU-AG nucleotides.

149 **Early Region 1 (E1) transcripts.** This region in MAdVs is the first transcribed after successful entry of  
150 the viral DNA into the host cell nucleus, albeit at low levels (18). The host transcription machinery solely  
151 mediates the transcription of this region. After their translation, the E1 proteins in concert with a myriad of  
152 host transcription factors activate the other viral promoters (6). Only two ORFs (ORF1 [sialidase] and Hyd)  
153 are predicted in this region; however, we discovered **four** novel transcripts in this region, which collectively  
154 contain **3** unique splice junctions (**Figure 5**). Most of the encoded proteins of the novel transcripts are

155 distinct from the predicted ORFs, although they all have the potential to encode the Hyd protein as the 3'-  
156 most coding sequence (CDS) if secondary start codon usage is considered. The 5'-most CDS of TRXPT\_1  
157 is multi-exonic, encoding a 17.9 kilodalton (kDa) protein of 160 residues [amino acids (aa)]. The CDS  
158 begins in the first exon, starting at position 211, spans the second exon, and terminates in the third exon  
159 at position 2312. From the 5'-most start codon (SSC), TRXPT\_2 encodes the largest protein in this region  
160 - a 64.3 kDa, 580 aa protein with the same SSC as TRXPT\_1 (position 211). This CDS spans almost  
161 the entire predicted ORF1 and Hyd, coming short in two regards: it is spliced from 1655 to 1964 (ORF1's  
162 C-terminus, including the stop codon), and its stop codon (STC; position 2312) is 13 bp short of the Hyd  
163 STC. However, it has an SSC 102 bp upstream and in-frame with ORF1's predicted SSC. TRXPT\_2 CDS  
164 therefore, shares substantial protein sequence similarity with ORF1 but not with Hyd, as the SSC of Hyd  
165 is not in-frame. TRXPT\_3 is almost identical to TRXPT\_1, except for the lack of TRXPT\_1's second exon.  
166 From our RNA-seq data, all E1 transcripts share the same transcription termination site (TTS; at position  
167 2325bp); however, TRXPT\_3 and TRXPT\_4 seem to have transcription start sites (TSS) downstream of  
168 the TSS of TRXPT\_1 and TRXPT\_2 (at position 54bp). Given that studies in MAdVs show that E1 mRNAs  
169 share not only a common TTS but also the TSS, and only differ from each other regarding the introns (18),  
170 it is likely that TRXPT\_3 and TRXPT\_4 are incomplete, and their actual TSS just like the TTS are identical  
171 for all E1 transcripts. Regardless of the TSS considered for TRXPT\_3, the coding potential (CP) remains  
172 unaffected. Its 5'-most CDS, beginning at 1965 and sharing the same STC as TRXPT\_1 and TRXPT\_2,  
173 produces a 13.1 kDa, 115 residue protein. This CDS (ORF4) was predicted in an earlier study (25) but was  
174 excluded in later studies (1, 12); however, our data suggests it is a bona fide ORF. The CP of TRXPT\_4 is  
175 affected by the TSS considered; if we consider its unmodified TSS, then its CP is the same as TRXPT\_3  
176 (ORF4 as the first CDS and Hyd as second CDS if the first SSC is skipped). However, if we assume that  
177 TRXPT\_4 shares the same TSS as TRXPT\_1, then the 5'-most CDS is a distinct multi-exonic 15.9 kDa,  
178 143 aa protein with the same SSC as TRXPT\_1 and TRXPT\_2 but with a unique STC. All splice junctions  
179 of the transcripts in this region (except the junction for TRXPT\_4) have been validated by cloning and  
180 Sanger sequencing of cDNA (**Figure 5b; supplementary PCR methods**). Finally, during our validation of  
181 TRXPT\_2, ORF1 was present on the agarose gel and Sanger sequencing results as a bona fide transcript  
182 (**supplementary PCR methods**). This is corroborated by 3'-RACE experiment, which shows transcripts  
183 spanning the entire ORF1 and Hyd ORFs without any splicing with a poly-A tail immediately after the TTS  
184 of transcripts in this region. The 5'-most CDS of this transcript would encode ORF1. However, the SSC  
185 of the predicted ORF1 is in-frame but downstream of TRXPT\_2 SSC, suggesting that the predicted ORF1  
186 CDS is a truncated ORF; it shares the same TSS, SSC, and TTS as TRXPT\_2, but has a unique STC.

187 **Early Region 2 (E2) and Intermediate Region (IM) transcripts.** The E2 TU expressed on the anti-sense  
188 strand, is subdivided into E2A and E2B and encodes three classical AdV proteins essential for genome  
189 replication: pTP and Ad-pol (E2B proteins), and DBP (E2A protein) (17, 18). Unlike MAdV where two  
190 promoters (E2-early and E2-late) were discovered (17), we discovered only a single promoter from which  
191 both E2A and E2B transcription is initiated. However, similar to MAdVs, E2A and E2B transcripts have  
192 distinct TTSs, and the E2B transcripts share the TTS of the IVa2 transcript of the IM region (17, 18) (**Figure**  
193 **6**).

194 The E2A ORF, DBP is one of three THEV ORFs predicted to be spliced from two exons. The corresponding  
195 transcript (TRXPT\_21) found in our data matches this predicted splicing pattern exactly but with a non-  
196 coding additional exon at the 5'-end (E2-5'UTR) at position 18,684-18,751 bp. Thus a three-exon transcript  
197 with an identical encoded protein (DBP; 380 residues, 43.3 kDa). This transcript (TRXPT\_21) was also  
198 corroborated in a 3'-RACE experiment. Additionally, from the 3'-RACE, a splice variant of TRXPT\_21 which  
199 retains the second intron leading to a 2-exon transcript was found. This transcript (TRXPT\_21B), albeit  
200 longer due to retaining the second intron and possessing a short 3' UTR, encodes a truncated isoform of  
201 DBP because the first SSC utilized by TRXPT\_21, is followed shortly by STCs in the retained intron, and  
202 does not yield any viable product. Utilizing the SSC 173 bp downstream of TRXPT\_21's SSC yields a 346  
203 residue, 39.3 kDa product, which is in-frame of DBP but entirely contained in the second exon. TRXPT\_21  
204 and TRXPT\_21B share a common TSS but TRXPT\_21B as seen in our 3'-RACE data, extends 39 bp into  
205 an adenine-thymine (A-T) rich sequence before the poly-A tail sequence occur, suggesting this position  
206 (16,934bp) as the bona fide E2A TTS (**Figure 6**).

207 The E2B region transcripts share the TSS of E2A - begins with E2-5'UTR - but extend thousands of base  
208 pairs downstream to reach the TTS at 2334bp in the IM region, which is immediately followed by an A-T  
209 rich sequence (position 2323-2339bp) where polyadenylation probably occurs. Interestingly, the TTS of  
210 the E1 region on the sense strand is in the immediate vicinity of this A-T rich sequence, and the A-T rich  
211 sequence is almost palindromic; hence it likely serves as the polyadenylation signal for both E1 and E2B/IM  
212 transcripts. The E2B transcripts, TRXPT\_6 and TRXPT\_7 are almost identical except for an extra splice  
213 junction at the 3'-end of TRXPT\_6, making TRXPT\_6 a five-exon transcript and TRXPT\_7, four (**Figure**  
214 **6**). TRXPT\_7 has the CP for IVa2 and both classical proteins (pTP and Ad-pol) encoded in this region, of  
215 which the pTP ORF is predicted to be spliced from two exons just like in all other AdVs. The predicted splice  
216 junction of pTP is corroborated by our data; however, the full transcript is markedly longer than the predicted  
217 ORF: there are two novel 5' exons, the third exon (containing the SSC of pTP) is significantly longer than  
218 predicted, and the last exon containing the bulk of the CDS is more than triple the predicted size of pTP.

219 The first two exons are 5'-UTRs because the SSC here is immediately followed by STCs; hence, the 5'-  
220 most SSC (position 10,995bp) of the third exon which matches the predicted SSC is utilized. The encoded  
221 product is identical to the predicted pTP ORF (597 residues; 70.5 kDa). If secondary SSC (secSSC) usage  
222 is considered, with SSC at 6768bp and STC at 3430bp, the encoded product is identical to the predicted  
223 Ad-pol ORF (1112 residues; 129.2 kDa). ~~finish trxpt\_7 and then discuss trxpt\_6.~~

224 The IM region is a single-transcript TU, encoding a single classical protein, IVa2. The promoter expressing  
225 this single transcript (TRXPT\_5) is embedded in E2B region and shares a TTS with E2B transcripts (17,  
226 18). TRXPT\_5 is a two-exon transcript located on the reverse strand spliced at 3447-3615. The first exon  
227 is an UTR, except the last 2 nucleotides, which connect with the first nucleotide of second exon to form the  
228 5'-most SSC. This first SSC is 4 codons upstream and in-frame of the predicted IVa2 SSC. Regardless of  
229 the SSC considered, the encoded protein (IVa2) is largely unaffected. Except for the four extra residues at  
230 the N-terminus (considering the 5'-most SSC), the entire protein sequence is identical.

231 The splice junction of ~~TRXPT\_5 add trxpts from E2~~ were confirmed by cloning and Sanger sequencing of  
232 cDNA (**supplementary PCR methods**).

233 **Early Region 3 (E3) transcripts.**

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

240 In THEV, only one ORF (E3) was predicted in this region. However, we identified six transcripts here  
241 (**TRXPT\_22, TRXPT\_23, TRXPT\_24, TRXPT\_25, TRXPT\_26, TRXPT\_27**), all of which are novel (**Figure**  
242 **7**). We identified two distinct TSSs - one similar to the classic MAdV E3 TSS (position 18,230bp) and the  
243 other about 500 bp downstream at 18,727bp. The E3 transcripts collectively have the CP for several THEV  
244 ORFs: 100K, 22K, 33K, pVIII, E3, Fiber (IV), and ORF7 belonging to the MLTU; however, some CDSs are  
245 nonidentical due to unpredicted splicing. For instance, 33K is one of the few THEV ORFs predicted to be  
246 spliced from two exons; however, we discovered it to be a four-exon ORF expressed from TRXPT\_24. The  
247 first two exons were not predicted but the last two, match the predicted exons precisely and the CDS is in-  
248 frame. However, the first 20bp of the predicted 33K (including the SSC) is spliced as part of the second  
249 intron of TRXPT\_24. Thus, the bona fide 33K is a 19.8 kDa, 171 residue protein spanning four exons.  
250 TRXPT\_24 can also encode the ORFs, pVIII and E3 if we consider downstream SSC usage. TRXPT\_22

and TRXPT\_23 share the TSS of TRXPT\_24 (E3 promoter; 18,230bp) but have distinct TTSs. Considering the first SSCs, they both encode novel 73 residue proteins that is over 80% similar but differ at the C-termini. Considering downstream SSC usage, both transcripts can encode pVIII and E3 in that order, but TRXPT\_23 being longer, has the CP for the Fiber ORF also. 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 the junction (18350-18717bp), we noticed a DNA band that corresponds to the full unspliced sequence (**supplementary PCR methods** **Table 1**). As TRXPT\_25 only falls short of encoding the complete 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 have distinct TTSs. TRXPT\_26 is a three-exon transcript but the first two are UTRs - only the terminal exon encodes proteins. 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 similarly, 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 as 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 reverse 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 reverse 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 SSC of ORF8 in the second exon; hence, the C-terminus of this longer protein (26.4 kDa, 229 aa) would be the identical to the predicted ORF8 protein. ~~The splice junction of TRXPT\_28 was validated by cloning and Sanger sequencing of cDNA (supplementary PCR methods).~~

**Major Late Transcription Unit (MLTU) or MLP Region transcripts.**

283 **DISCUSSION/CONCLUSIONS**

284 In the original study where the ORFs of THEV were predicted, ORF4 was predicted in the E1 region span-  
285 ning the Hyd gene. However, later studies predicted and preferred Hyd instead of ORF4; hence, the current  
286 prediction map. However, this study shows that while both Hyd and ORF4 may be both expressed, ORF4 is  
287 most likely the bona fide gene. For fig2a: There is a dramatic increase of mean coverage/depth from **2.42**  
288 at 4 h.p.i to **95,042** at 72 h.p.i, strongly demonstrating an active infection. Unexpectedly, the pileup of reads  
289 seems consistently skewed over similar regions of the genome. We could speculate that the temporal gene  
290 expression regulation of THEV is different from MAdVs or this could simply mean that the infection was not  
291 well synchronized. However, the relative proportions over these similar regions shows some variation over  
292 time. For fig2b: titer reaching a plateau at 120 h.p.i, probably due to high cell death TRXPT\_2 and ORF1  
293 are isoforms Presumably, if the junction reads were normalized, MLTU would not be predominant at 12hpi.  
294 The TTSs were all in the context of A-T rich sequences; which presumably serve as polyA signals.

295 **MATERIALS AND METHODS**

296 **Cell culture and THEV Infection**

297 The Turkey B-cell line (MDTC-RP19, ATCC CRL-8135) was grown as suspension cultures in 1:1 complete  
298 Leibovitz's L-15/McCoy's 5A medium with 10% fetal bovine serum (FBS), 20% chicken serum (ChS), 5%  
299 tryptose phosphate broth (TPB), and 1% antibiotics solution (100 U/mL Penicillin and 100ug/mL Strepto-  
300 mycin), at 41°C in a humidified atmosphere with 5% CO<sub>2</sub>. Infected cells were maintained in 1:1 serum-  
301 reduced Leibovitz's L15/McCoy's 5A media (SRLM) with 2.5% FBS, 5% ChS, 1.2% TPB, and 1% antibiotics  
302 solution (100 U/mL Penicillin and 100ug/mL Streptomycin). A commercially available HE vaccine was pur-  
303 chased from Hygieia Biological Labs as a source of THEV-A (VAS strain). The stock virus was titrated using  
304 an in-house qPCR assay with titer expressed as genome copy number(GCN)/mL, similar to Mahshoub *et*  
305 *al* (26) with modifications. Cells were infected at a multiplicity of infection (MOI) of 100 GCN/cell and sam-  
306 ples in triplicates were harvested at 4-, 12-, 24-, and 72-h.p.i for RNA-seq. The infection was repeated but  
307 samples in triplicates were harvested at 12-, 24-, 36-, 48-, and 72-h.p.i for PCR validation of novel splice  
308 sites. Still one more independent infection was done at time points ranging from 12 to 168-h.p.i for qPCR  
309 quantification of virus titers.

310 **RNA extraction and Sequencing**

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

319 **Validation of Novel Splice Junctions**

320 All splice junctions identified in this work are novel except one predicted splice site each for pTP and DBP,  
321 which were corroborated in our work. However, these predicted splice junctions had not been experimen-  
322 tally validated hitherto, and we identified additional novel exons, giving the complete picture of these tran-

scripts. 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 Table 1. 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 (~supplementary PCR methods). 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* (27) 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

Analysis of our sequence reads were analyzed following a well established protocol described by Pertea *et al* (23), using SNAKE MAKE 7.24.0 to drive the pipeline. Briefly, sequencing reads were trimmed with the FastQC – version 0.11.9 (28) program to achieve an overall Mean Sequence Quality (Phred Score) of 36. Trimmed reads were mapped to the complete 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 without relying on known splice sites. The generated 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 gff3 file from NCBI cont m. aining the predicted ORFs of THEV as a guide. A custom script was used

<sup>352</sup> to consolidate all transcripts from all time-points without redundancy, generating the final transcriptome of  
<sup>353</sup> THEV.

354 **SCRIPTS AND SUPPLEMENTARY MATERIALS**

355 **DATA AVAILABILITY**

356 **CODE AVAILABILITY**

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

359 **ACKNOWLEDGMENTS**

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

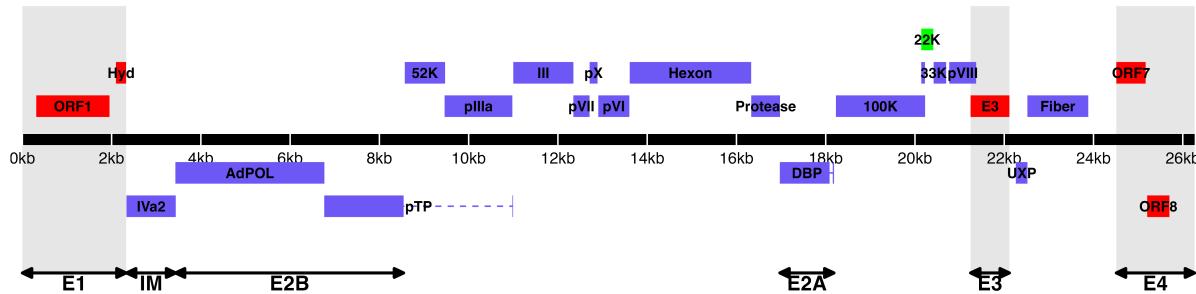
362 REFERENCES

- 363 1. Davison A, Benko M, Harrach B. 2003. Genetic content and evolution of adenoviruses. *The Journal*  
364 of general virology
- 365 2. Harrach B. 2008. Adenoviruses: General features, p. 1–9. *In* Mahy, BWJ, Van Regenmortel, MHV  
366 (eds.), *Encyclopedia of virology* (third edition). Book Section. Academic Press, Oxford.
- 367 3. Upton C, Slack S, Hunter AL, Ehlers A, Roper RL. 2003. Poxvirus orthologous clusters: Toward  
368 defining the minimum essential poxvirus genome. *Journal of virology* 77:7590–7600.
- 369 4. McGeoch D, Davison AJ. 1999. Chapter 17 - the molecular evolutionary history of the herpesviruses,  
370 p. 441–465. *In* Domingo, E, Webster, R, Holland, J (eds.), *Origin and evolution of viruses*. Book  
Section. Academic Press, London.
- 371 5. Harrach B, Benko M, Both GW, Brown M, Davison AJ, Echavarría M, Hess M, Jones M, Kajon A,  
372 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.
- 373 6. Guimet D, Hearing P. 2016. 3 - adenovirus replication, p. 59–84. *In* Curiel, DT (ed.), *Adenoviral*  
374 vectors for gene therapy (second edition). Book Section. Academic Press, San Diego.
- 375 7. Kovács ER, Benkő M. 2011. Complete sequence of raptor adenovirus 1 confirms the characteristic  
376 genome organization of siadenoviruses. *Infection, Genetics and Evolution* 11:1058–1065.
- 377 8. Davison AJ, Wright KM, Harrach B. 2000. DNA sequence of frog adenovirus. *J Gen Virol* 81:2431–  
378 2439.
- 379 9. Kovács ER, Jánoska M, Dán Á, Harrach B, Benkő M. 2010. Recognition and partial genome char-  
380 acterization by non-specific DNA amplification and PCR of a new siadenovirus species in a sample  
originating from parus major, a great tit. *Journal of Virological Methods* 163:262–268.
- 381 10. Katoh H, Ohya K, Kubo M, Murata K, Yanai T, Fukushi H. 2009. A novel budgerigar-adenovirus  
382 belonging to group II avian adenovirus of siadenovirus. *Virus Research* 144:294–297.
- 383 11. Beach NM. 2006. Characterization of avirulent turkey hemorrhagic enteritis virus: A study of the  
384 molecular basis for variation in virulence and the occurrence of persistent infection. Thesis.

- 385 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.  
386 J Gen Virol 90:1978–85.
- 387 13. Gross WB, Moore WE. 1967. Hemorrhagic enteritis of turkeys. Avian Dis 11:296–307.
- 388
- 389 14. Rautenschlein S, Sharma JM. 2000. Immunopathogenesis of haemorrhagic enteritis virus (HEV) in  
390 turkeys. Dev Comp Immunol 24:237–46.
- 391 15. Larsen CT, Domermuth CH, Sponenberg DP, Gross WB. 1985. Colibacillosis of turkeys exacerbated  
392 by hemorrhagic enteritis virus. Laboratory studies. Avian Dis 29:729–32.
- 393 16. Dhama K, Gowthaman V, Karthik K, Tiwari R, Sachan S, Kumar MA, Palanivelu M, Malik YS, Singh  
394 RK, Munir M. 2017. Haemorrhagic enteritis of turkeys – current knowledge. Veterinary Quarterly  
37:31–42.
- 395 17. Donovan-Banfield I, Turnell AS, Hiscox JA, Leppard KN, Matthews DA. 2020. Deep splicing plasticity  
396 of the human adenovirus type 5 transcriptome drives virus evolution. Communications Biology 3:124.
- 397 18. Zhao H, Chen M, Pettersson U. 2014. A new look at adenovirus splicing. Virology 456-457:329–341.
- 398
- 399 19. Wolfrum N, Greber UF. 2013. Adenovirus signalling in entry. Cell Microbiol 15:53–62.
- 400
- 401 20. Falvey E, Ziff E. 1983. Sequence arrangement and protein coding capacity of the adenovirus type 2  
402 "i" leader. Journal of Virology 45:185–191.

- 403 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  
404 cells. *Nature* 489:101–108.
- 405 22. Aboeza Z, Mabsoub H, El-Bagoury G, Pierson F. 2019. In vitro growth kinetics and gene expression  
406 analysis of the turkey adenovirus 3, a siadenovirus. *Virus Research* 263:47–54.
- 407 23. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of  
408 RNA-seq experiments with HISAT, StringTie and ballgown. *Nature Protocols* 11:1650–1667.
- 409 24. Jack Fu [Aut], Alyssa C. Frazee [Aut, Cre], LeonardoCollado-Torres [Aut], Andrew E. Jaffe [Aut],  
410 Jeffrey T. Leek[Aut, Ths]. 2017. Ballgown. Bioconductor.
- 411 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,  
412 hemorrhagic enteritis virus. *Virology* 249:307–315.
- 413 26. Mabsoub HM, Evans NP, Beach NM, Yuan L, Zimmerman K, Pierson FW. 2017. Real-time PCR-  
414 based infectivity assay for the titration of turkey hemorrhagic enteritis virus, an adenovirus, in live  
vaccines. *Journal of Virological Methods* 239:42–49.
- 415 27. Green MR, Sambrook J. 2019. Rapid amplification of sequences from the 3' ends of mRNAs: 3'-  
416 RACE. *Cold Spring Harbor Protocols* 2019:pdb.prot095216.
- 417 28. 2015. FastQC.  
418

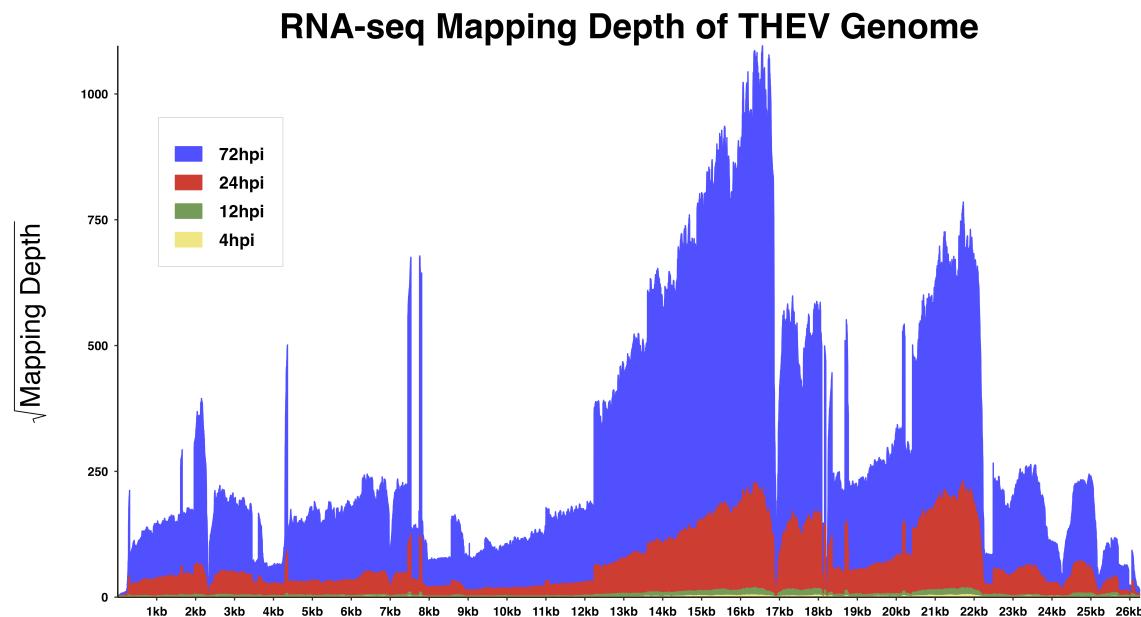
419 **TABLES AND FIGURES**



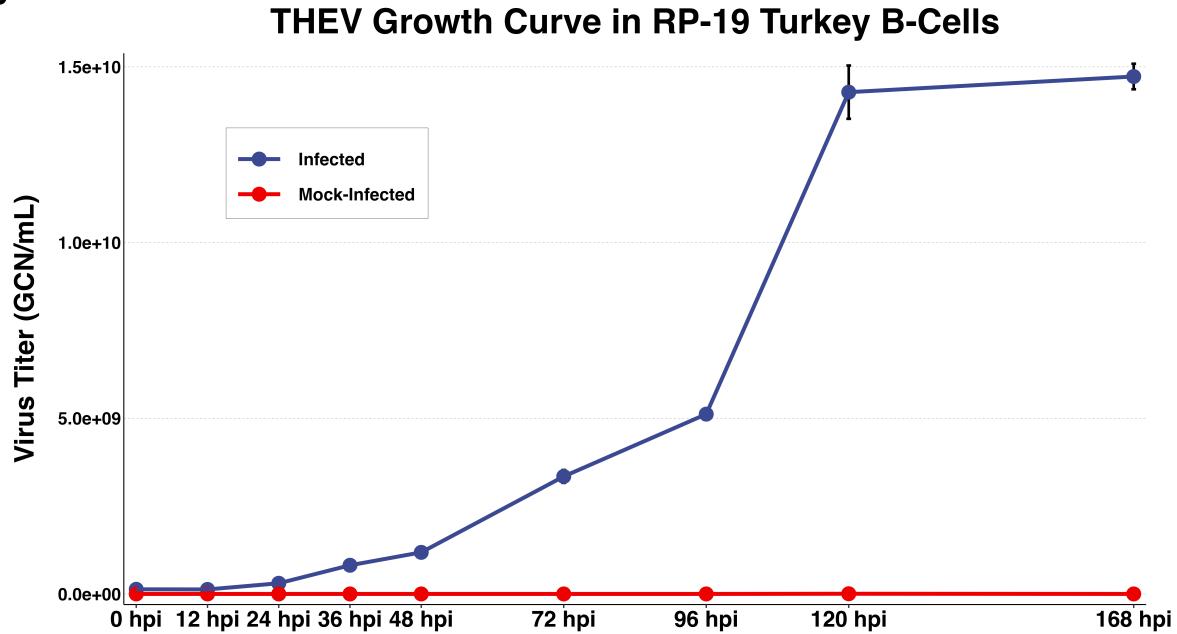
420

421 **Figure 1. Genomic map of THEV virulent strain annotated ORFs.** The central horizontal line repre-  
 422 sents the double-stranded DNA marked at 5kb intervals as white line breaks. Blocks represent viral genes.  
 423 Blocks above the DNA line are transcribed rightward, those below are transcribed leftward. pTP, DBP and  
 424 33K predicted to be spliced are shown as having tails. Shaded regions indicate regions containing “genus-  
 425 specific” genes (colored red). Genes colored in blue are “genus-common”. Gene colored in light green is  
 426 conserved in all but Atadenoviruses. The UXP (light blue) is an incomplete gene present in almost all AdVs.  
 427 Regions comprising the different transcription units are labelled at the bottom (E1, E2A, E2B, E3, E4, and  
 428 IM); the unlabeled regions comprise the MLTU.

A



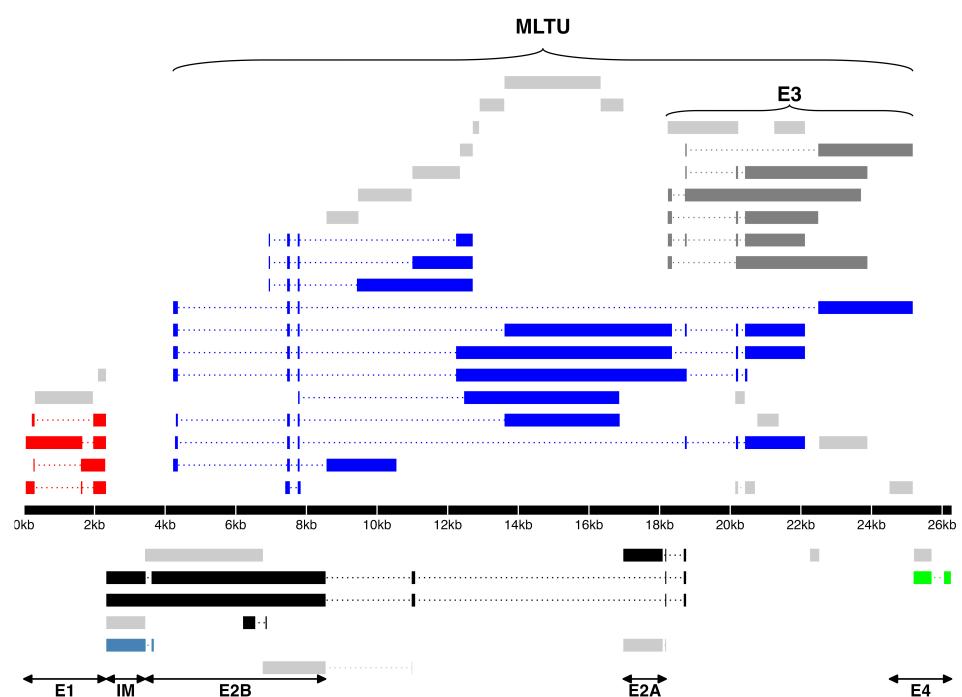
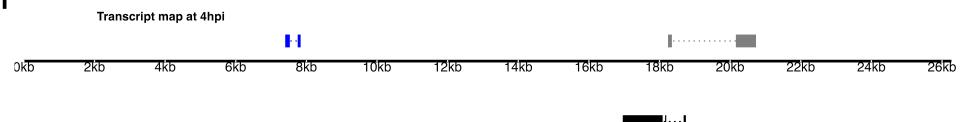
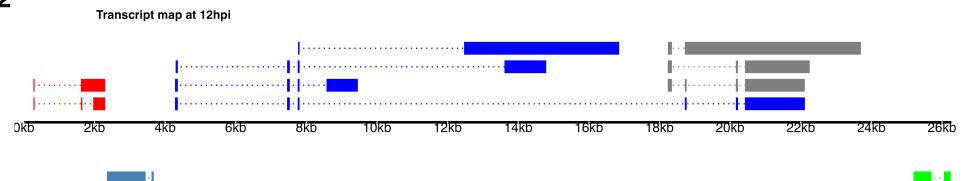
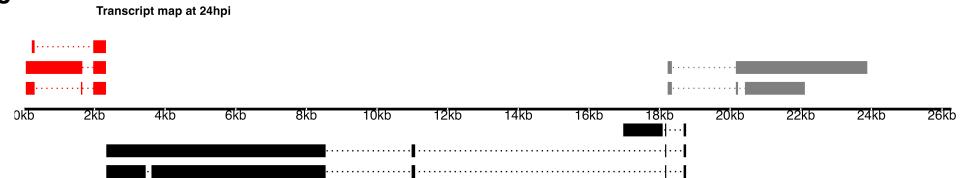
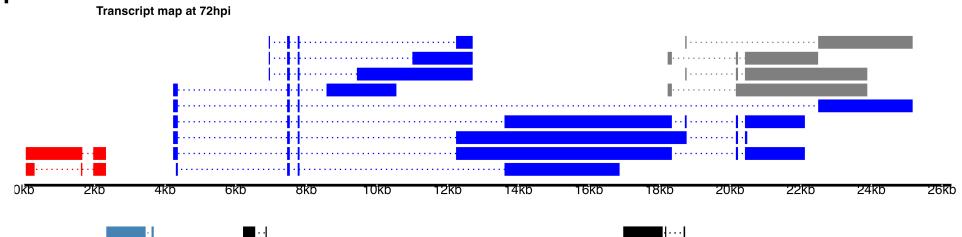
B



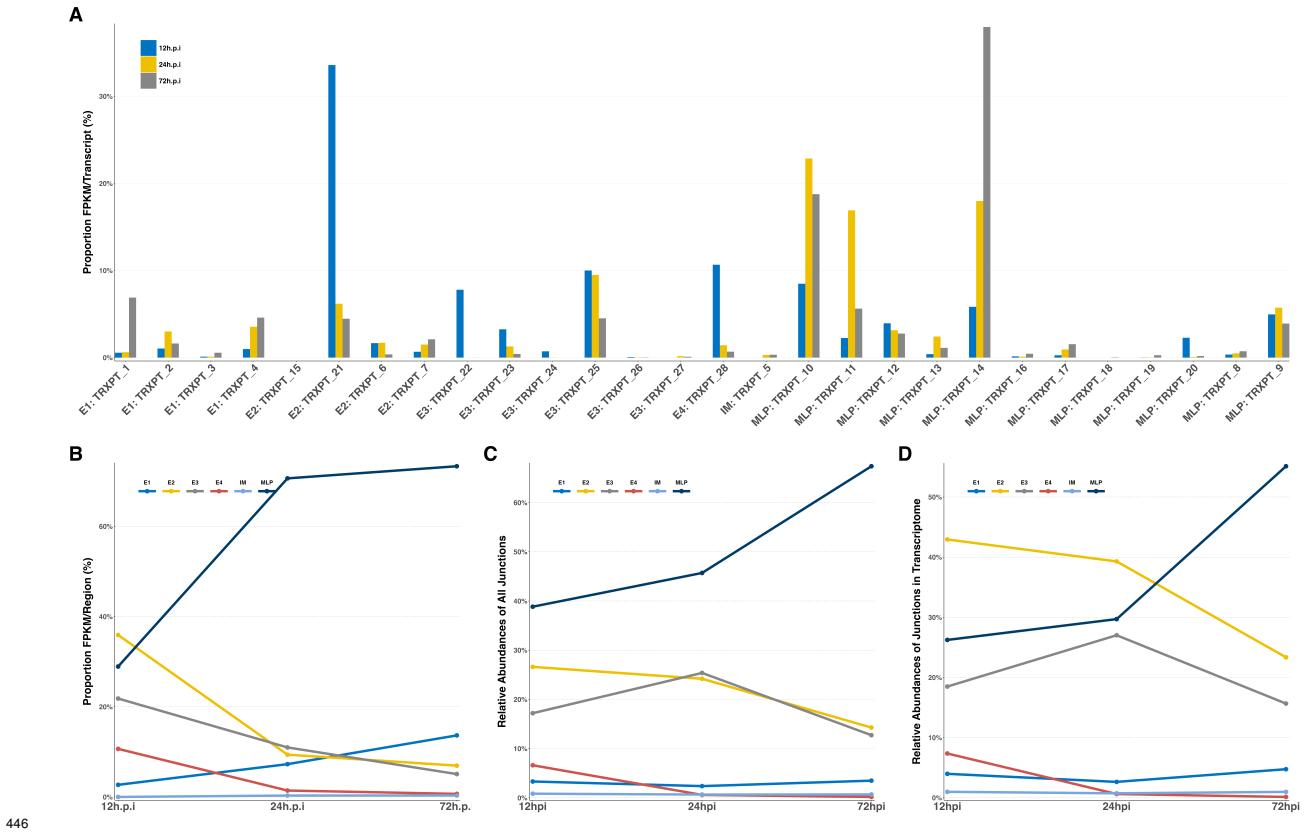
429

430 **Figure 2: Increasing levels of THEV over time. a) Per base coverage of sequence reads mapping**  
 431 **to THEV genome by time point.** The pileup of mRNA reads mapping to THEV genome at the base-pair  
 432 level for each indicated time point. b) **Growth curve of THEV (VAS vaccine strain) in MDTC-RP19 cell**  
 433 **line.** Virus titers were quantified with a qPCR assay. There is no discernible increase in virus titer up 12  
 434 h.p.i, after which there is a steady increase in virus titer is measured. The virus titer expands exponentially

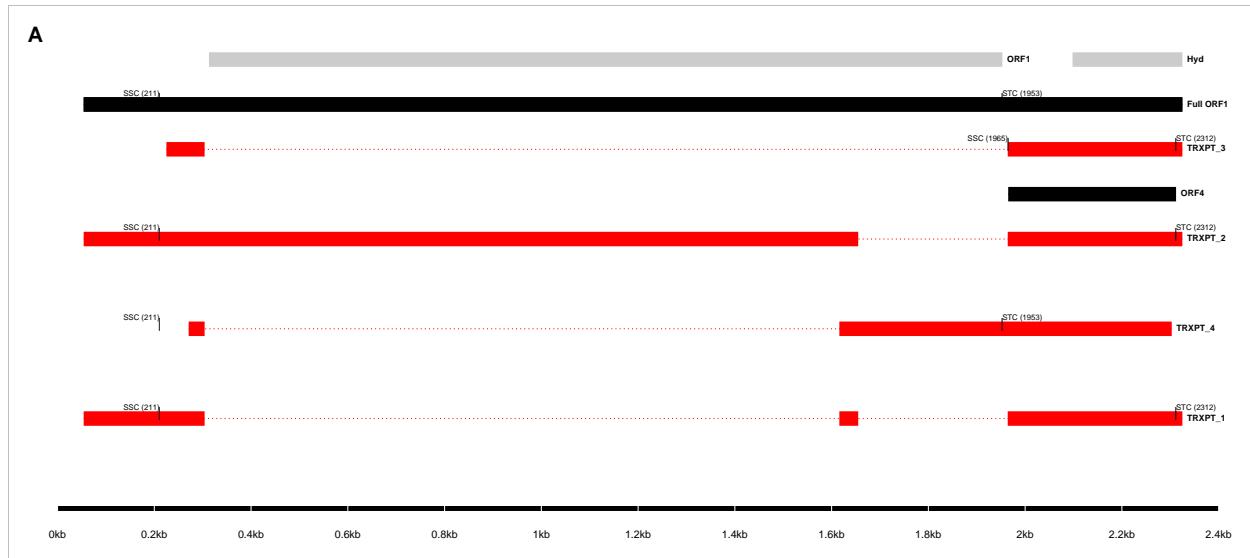
<sup>435</sup> beginning from 48 h.p.i, increasing by orders of magnitude before reaching a plateau at 120 h.p.i. GCN:  
<sup>436</sup> genome copy number.

**A****B1****B2****B3****B4**

438 **Figure 3. a) Composite transcriptome of THEV from RNA-seq.** THEV transcripts assembled from all  
439 time points by StringTie are unified forming this final transcriptome (splicing map). Transcripts belonging  
440 to the same transcription unit (TU) are located in close proximity on the genome and are color coded and  
441 labeled in this figure as such. The organization of TUs in the THEV genome is unsurprisingly similar to  
442 MAdVs; however, the MAdV genome shows significantly more transcripts. The TUs are color coded: E1  
443 transcripts - red, E2 - black, E3 - dark grey, E4 - green, MLTU - blue. Predicted ORFs are also indicated  
444 here, colored light grey. **b) THEV transcripts identified at given time points.** Transcripts are color coded  
445 as explained in **a**.



446  
**Figure 4: Changes in splicing and expression profile of THEV over time.** **a)** Expression levels of  
447 transcripts over time. **b)** Expression levels of transcripts by region over time. **c)** Relative abundances of all  
448 splice junctions over time. **d)** Relative abundances of junctions in transcriptome.  
449

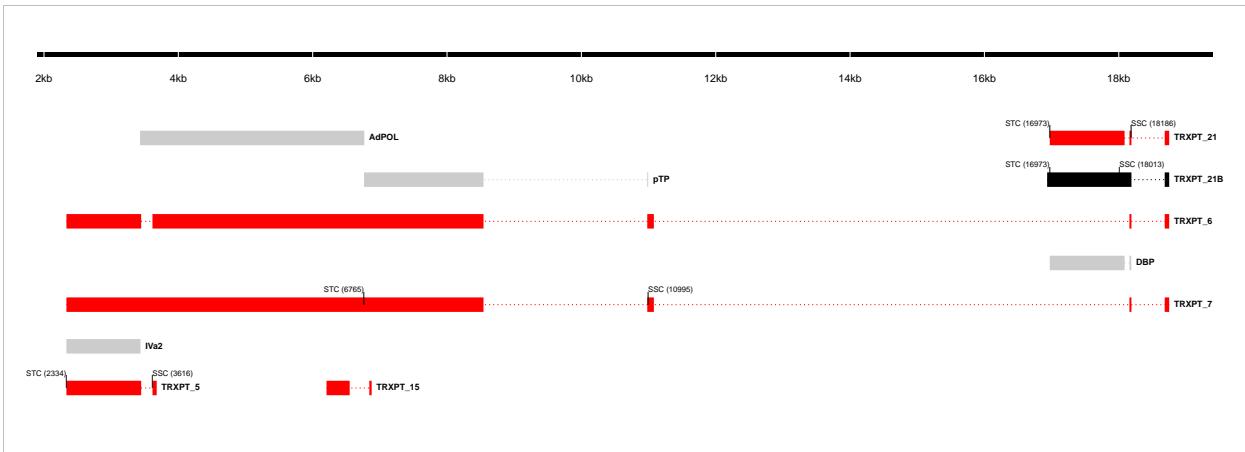


**B**

Transcript ID	Splice Junction					Strand	Junction Reads				Coding Potential	
	Start	End	Intron Length	Splice Donor-Acceptor			4h.p.i	12h.p.i	24h.p.i	72h.p.i		
TRXPT_1 TRXPT_4	304	1616	1313bp	GT-AG		+	0	9	1019	25041	Validated*	Hyd_iso_1, ORF4_novel
TRXPT_3	304	1964	1661bp	GT-AG		+	0	2	168	1588	Validated	Hyd_iso_2
TRXPT_2 TRXPT_1	1655	1964	310bp	GT-AG		+	0	9	1395	38491	Validated	ORF1_novel_iso, Hyd_iso_1

450 \* Not validated for TRXPT\_4

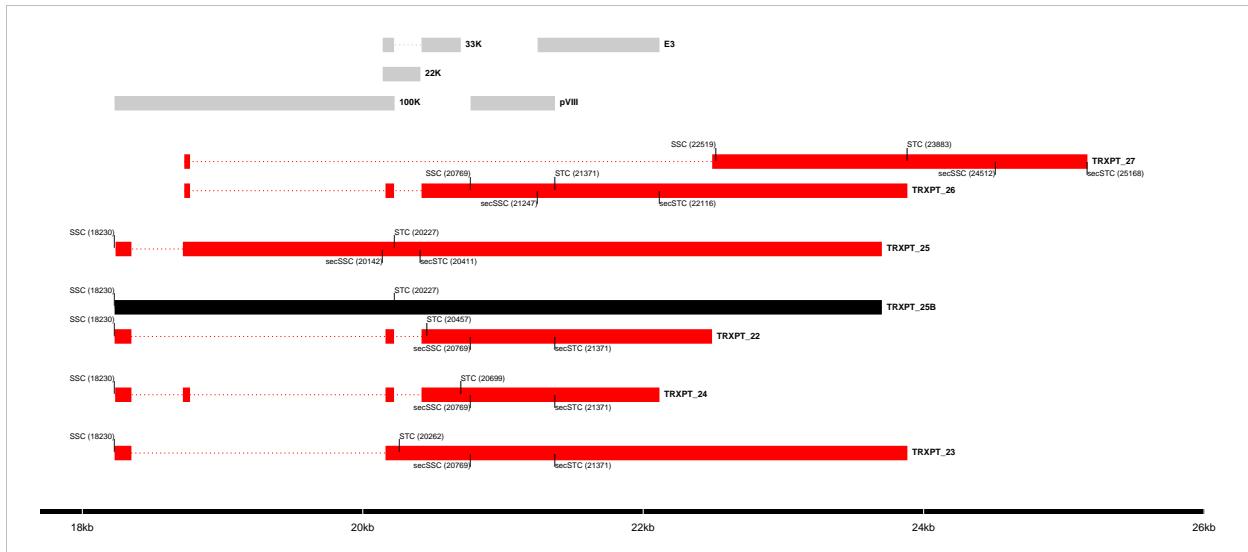
451 **Figure 5: E1 region transcripts. a)** The splice map of the E1 transcription unit. Exons are depicted as  
452 boxes connected by introns (dotted lines). Transcripts from RNA-seq data are colored red, predicted ORFs  
453 are colored grey, and the full ORF1 transcript and previously annotated ORF4 are colored black. Each  
454 transcript or ORF is labelled with its name to the right. The start codon (SSC) and stop codon (STC) of  
455 the 5'-most CDS of each transcript is indicated with the nucleotide position in brackets. The region of the  
456 virus is depicted at the bottom as a black line with labels of the nucleotide positions for reference **b)** The  
457 sequence reads covering the splice junctions are indicated with information about their validation status  
458 using cloning and Sanger sequencing.



Transcript ID	Splice Junction				Strand	region	Junction Reads				Junction Status
	Start	End	Splice Donor-Acceptor	Intron Length			4h.p.i	12h.p.i	24h.p.i	72h.p.i	
TRXPT_5, TRXPT_7	3447	3615	GT-AG	169bp	-	IM, E2	1	5	720	13422	Validated
TRXPT_6, TRXPT_7	11079	18159	GT-AG	7081bp	-	E2	0	2	0	0	Validated
TRXPT_21	18087	18159	GT-AG	73bp	-	E2	9	103	0	0	Validated
TRXPT_21, TRXPT_6, TRXPT_7	18189	18684	CT-AC, GT-AG	496bp	-	E2	0	111	18794	156037	Validated
TRXPT_6, TRXPT_7	8543	10981	GT-AG	2439bp	-	E2	0	0	298	850	Validated
TRXPT_15	6551	6843	GT-GC	293bp	-	E2	0	0	0	6	Unvalidated*

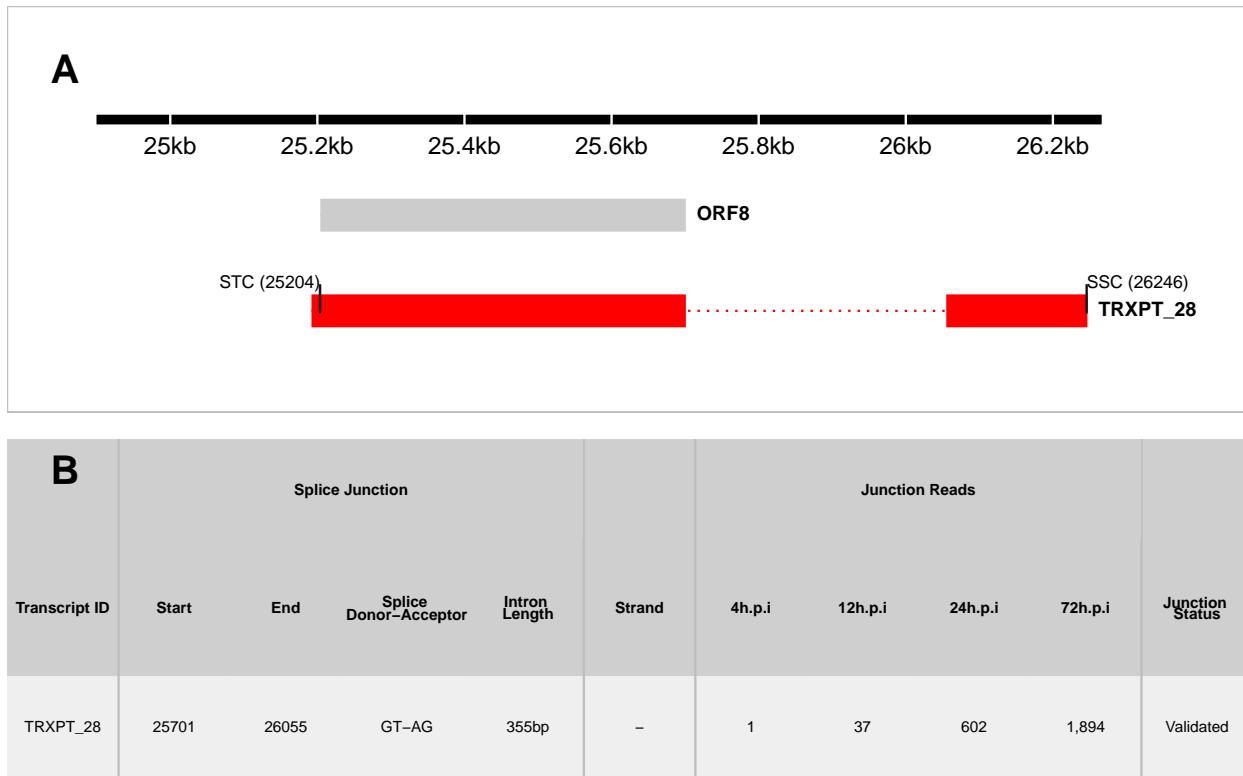
459 \*Incomplete transcript

460 **Figure 6: E2 and IM region transcripts.** **a)** The splice map of the E2 and IM transcription units. Exons  
 461 are depicted as boxes connected by introns (dotted lines). Red transcripts are generated from RNA-seq  
 462 data and predicted ORFs are colored grey. Each transcript or ORF is labelled with its name to the right.  
 463 The start codon (SSC) and stop codon (STC) of the 5'-most CDS of each transcript is indicated with the  
 464 nucleotide position in brackets. The region of the virus is depicted at the bottom as a black line with labels  
 465 of the nucleotide positions for reference **b)** The sequence reads covering the splice junctions are indicated  
 466 with information about their validation status using cloning and Sanger sequencing.

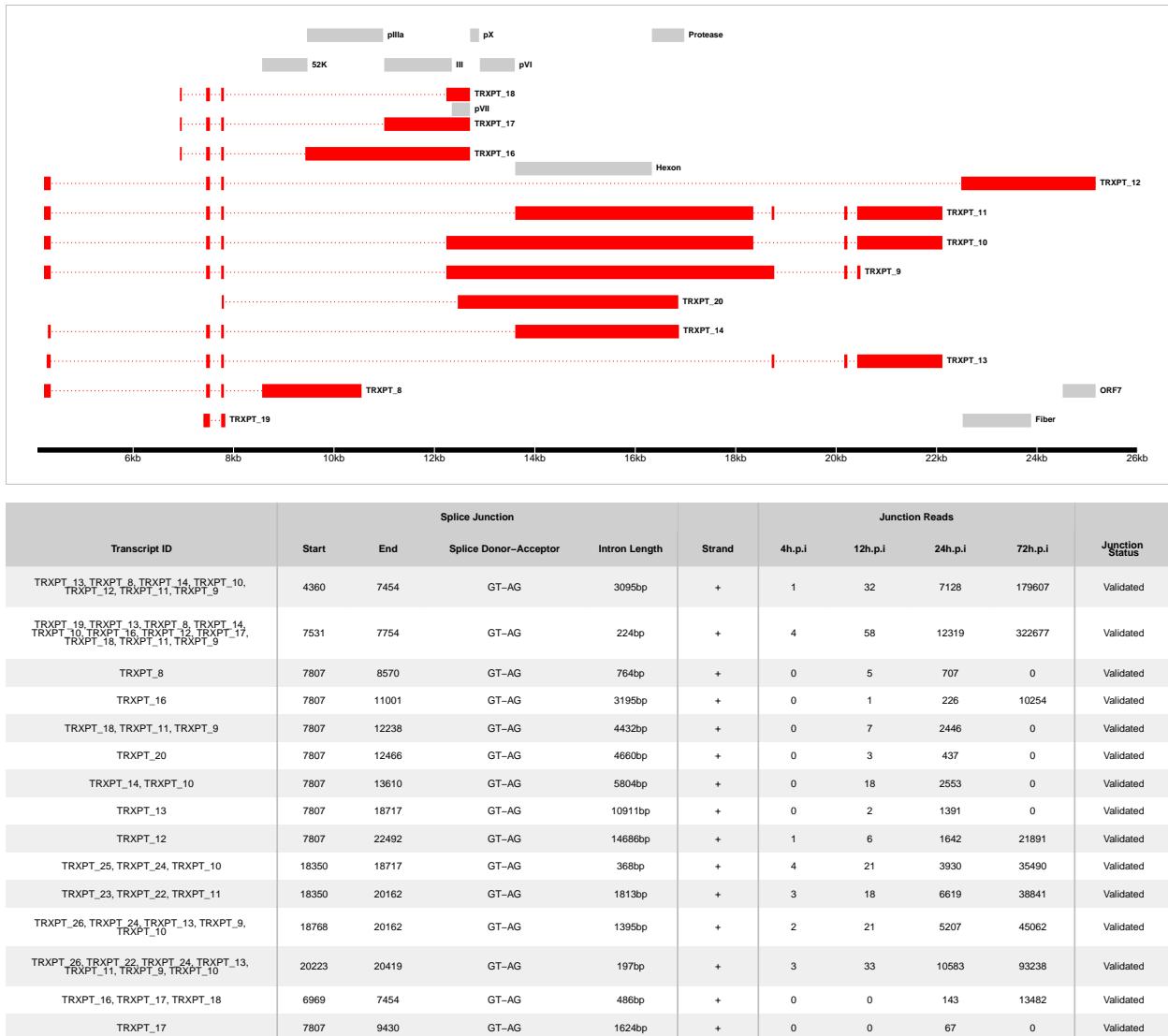


Transcript ID	Splice Junction				Strand	region	Junction Reads				Junction Status
	Start	End	Splice Donor-Acceptor	Intron Length			4h.p.i	12h.p.i	24h.p.i	72h.p.i	
TRXPT_25, TRXPT_24, TRXPT_10	18350	18717	GT-AG	368bp	+	E3, MLP	4	21	3930	35490	Validated
TRXPT_23, TRXPT_22, TRXPT_11	18350	20162	GT-AG	1813bp	+	E3, MLP	3	18	6619	38841	Validated
TRXPT_26, TRXPT_24, TRXPT_13, TRXPT_9, TRXPT_10	18768	20162	GT-AG	1395bp	+	E3, MLP	2	21	5207	45062	Validated
TRXPT_26, TRXPT_22, TRXPT_24, TRXPT_13, TRXPT_11, TRXPT_9, TRXPT_10	20223	20419	GT-AG	197bp	+	E3, MLP	3	33	10583	93238	Validated
467 TRXPT_27	18768	22492	GT-AG	3725bp	+	E3	0	0	101	1950	Validated

468 **Figure 7: E3 transcripts.** a) The splice map of the E3 transcription unit. Exons are depicted as boxes  
469 connected by introns (dotted lines). Red transcripts are generated from RNA-seq data and predicted ORFs  
470 are colored grey. Each transcript or ORF is labelled with its name to the right. The start codon (SSC) and  
471 stop codon (STC) of the 5'-most CDS of each transcript is indicated with the nucleotide position in brackets.  
472 The region of the virus is depicted at the bottom as a black line with labels of the nucleotide positions for  
473 reference b) The sequence reads covering the splice junctions are indicated with information about their  
474 validation status using cloning and Sanger sequencing.



476 **Figure 8: E4 region transcripts.** a) The splice map of the E4 transcription unit. Exons are depicted  
 477 as boxes connected by introns (dotted lines). The transcript from RNA-seq data is colored red and the  
 478 predicted ORF, grey. The transcript and ORF are labelled with their names to the right. The start codon  
 479 (SSC) and stop codon (STC) of the 5'-most CDS is indicated with the nucleotide position in brackets. The  
 480 region of the virus is depicted at the bottom as a black line with labels of the nucleotide positions for  
 481 reference b) The sequence reads covering the splice junction are indicated.



482

483 **Figure 9: MLTU transcripts. a)**

Table 1: Table 1: Overview of sequencing results

Metric	4h.p.i	12h.p.i	24h.p.i	72h.p.i	Total
<b>Total reads</b>	1.17e+08	7.63e+07	1.20e+08	1.15e+08	4.28e+08
<b>Mapped (Host)</b>	1.04e+08	6.79e+07	1.06e+08	8.38e+07	3.62e+08
<b>Mapped (THEV)</b>	4.32e+02	6.70e+03	1.18e+06	1.69e+07	1.81e+07
<b>Mean Per Base Coverage/Depth</b>	2.42	37.71	6,666.96	95,041.7	101,749
<b>Total unique splice junctions</b>	13	37	236	2374	2,457
<b>Junction coverage Total (at least 1 read)</b>	37	605	115075	2132806	2.25e+06
<b>Junction coverage Mean reads</b>	2.8	16.4	487.6	898.4	351.3
<b>Junction coverage (at least 10 reads)</b>	0	13	132	1791	1,936
<b>Junction coverage (at least 100 reads)</b>	0	1	53	805	859
<b>Junction coverage (at least 1000 reads)</b>	0	0	18	168	186

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

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

Table 3: Table 2b: Most abundant splice junctions at 24h.p.i

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

Table 4: Table 2c: Most abundant splice junctions at 72h.p.i

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