scientific reports

OPEN

Combined nanopore and single-molecule real-time sequencing survey of human betaherpesvirus 5 transcriptome

Balázs Kakuk[1,5], Dóra Tombácz[1,2,3,5], Zsolt Balázs[1], Norbert Moldován[1], Zsolt Csabai[1], Gábor Torma[1], Klára Megyeri[4], Michael Snyder[3] & Zsolt Boldogkői[1]✉

Long-read sequencing (LRS), a powerful novel approach, is able to read full-length transcripts and confers a major advantage over the earlier gold standard short-read sequencing in the efficiency of identifying for example polycistronic transcripts and transcript isoforms, including transcript length- and splice variants. In this work, we profile the human cytomegalovirus transcriptome using two third-generation LRS platforms: the Sequel from Pacific BioSciences, and MinION from Oxford Nanopore Technologies. We carried out both cDNA and direct RNA sequencing, and applied the LoRTIA software, developed in our laboratory, for the transcript annotations. This study identified a large number of novel transcript variants, including splice isoforms and transcript start and end site isoforms, as well as putative mRNAs with truncated in-frame ORFs (located within the larger ORFs of the canonical mRNAs), which potentially encode N-terminally truncated polypeptides. Our work also disclosed a highly complex meshwork of transcriptional read-throughs and overlaps.

Next-generation short-read sequencing (SRS) platforms have revolutionized genomics and transcriptomics sciences, and while they are still invaluable in sequencing studies, the now state-of-the art long-read sequencing methods (LRS) are becoming more popular and represent an even more powerful approach in transcriptome research. Most genes encode multiple transcript isoforms[1] that are mRNAs or non-coding RNAs (ncRNAs) transcribed from the same locus, but have different transcriptional start sites (TSSs), or transcriptional end sites (TESs), or are the results of alternative splicing[2,3]. The reconstruction of all transcribed isoforms for each gene is challenging with the currently available bioinformatics tools since they have been developed for the analysis of SRS data[4,5].

Since LRS technologies are able to read full-length RNA molecules, they offer a solution to disclose the full spectrum of complex transcriptomes and offer an insight that is unachievable via SRS methods[6]. LRS platforms are currently commercially available by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), which provide read lengths of ~15 kb for PacBio and > 30 kb for ONT that surpass lengths of most transcripts. Both techniques were applied for the investigation of transcriptomic complexity of human cell lines[7] and various organisms, such as mammals[8], fish[9] and plants[10] and a number of viruses, such as poxviruses[11], baculoviruses[12], coronaviruses[13], circoviruses[14], adenoviruses[15]; and herpesviruses[16–19].

Since viral genomes are small and compact, they are ideal subjects for transcriptome analysis with the LRS techniques, as these methods still have a relatively low throughput compared to the SRS techniques[20]. These LRS-based studies repeatedly concluded that transcriptional complexity had previously been underestimated in all of the examined viruses[21]. In addition, ONT is capable of sequencing not only DNA[22] but also RNA in its native form[23]. Direct RNA sequencing (dRNA-Seq) does not require reverse transcription and PCR amplification therefore, it does not produce spurious transcripts, which are common artifacts of these techniques. While dRNA-Seq has its own limitations[24], it can be used to validate and to expand cDNA-based LRS studies[16].

[1]Department of Medical Biology, Faculty of Medicine, University of Szeged, Somogyi B. u. 4, 6720 Szeged, Hungary. [2]MTA-SZTE Momentum GeMiNI Research Group, University of Szeged, Somogyi B. u. 4, 6720 Szeged, Hungary. [3]Department of Genetics, School of Medicine, Stanford University, 300 Pasteur Dr, Stanford, CA, USA. [4]Department of Medical Microbiology and Immunobiology, Faculty of Medicine, University of Szeged, Szeged 6720, Hungary. [5]These authors contributed equally: Balázs Kakuk and Dóra Tombácz. ✉email: boldogkoi.zsolt@med.u-szeged.hu

Human cytomegalovirus (HCMV, also termed *Human betaherpesvirus 5*) infects 60–90% of the population worldwide[25] and can cause mononucleosis-like symptoms in adults[26], and severe life-threatening infections in newborns[27]. It can infect various human cells, including fibroblasts, epithelial cells, endothelial cells, smooth muscle cells, and monocytes[28]. HCMV has a linear double-stranded DNA genome (235 ± 1.9 kbps), which is the largest genome among human herpesviruses[29]. Its E-type genome structure consists of two large domains: the unique long (UL) and the unique short (US), each flanked by terminal (TRL and TRS) and internal (IRL and IRS) inverted repeats[30]. In addition, it encodes four major long non-coding RNAs (lncRNAs) (RNA1.2, RNA2.7, RNA4.9, and RNA5.0)[31], as well as at least 16 pre-miRNAs and 26 mature miRNAs[32–34]. Although the functions of most genes in infective stages have been identified, many remain uncharacterized[29]. The HCMV genome was shown to express more than 751 translated open reading frames (ORFs)[35,36], although most of them are very short and located upstream of the canonical ORFs. The compact genome with high gene density has many overlapping transcriptional units, which share common 5′ or 3′ ends, complex splicing patterns, antisense transcription, and transcription of lncRNAs and micro RNAs (miRNAs)[37]. Nested genes are special forms of the 3´-coterminal transcripts, since they have truncated in-frame open reading frames (ORFs), which possess different initiation but have common termination sites[38]. These add even more complexity to the genome regulation and expand coding potential of the virus. Short-read RNA sequencing studies have discovered splice junctions and ncRNAs[39] and have shown that the most abundant HCMV transcripts are similarly expressed in different cell types[10].

In our previous work[40], we used the Pacific Biosciences RSII sequencing platform to investigate the HCMV transcriptome and detected 291 previously undescribed or only partially annotated transcript isoforms, including polycistronic (PC) RNAs and also transcriptional overlaps. However, the RSII method is biased toward cDNA sizes between 1 and 2 kbp, therefore the short and the very long transcripts have not been detected by this analysis. As it was concluded by others[41], involving other sequencing technologies for the analysis could provide additional insights into the operation of transcriptional machineries of HCMV. Following this concept, in this work, we analyzed the HCMV transcriptome applying a multi-technique approach including ONT MinION and the PacBio Sequel platforms and using both cDNA and native RNA sequencings. Our primary objective was to construct the most comprehensive HCMV transcriptome atlas currently available using data provided by the state-of-the-art LRS methods, and thus to gain a deeper understanding of this important human pathogenic virus.

Results

Long-read sequencing of HCMV transcriptome using a multi-technique approach. In this work, we analyzed the HCMV transcripts with ONT MinION technique using cDNA (sample ID: *Non_Cap*), Cap-selected cDNA (*Cap*), and native RNA libraries (*dRNA)* and with PacBio Sequel platform using a cDNA library[42] (*Sequel)*. We also included the data obtained in our previous work using PacBio RSII method[40] (*RSII)*. The reads from all 7 libraries (from RSII, Sequel and ONT sequencers) were remapped with minimap2 and reanalyzed with the LoRTIA program developed in our laboratory[43,44]. Figure 1 shows the sequencing platforms, library preparation methods and data analysis steps that were carried out in this work.

The read statistics of the different sequencing approaches is shown in Table 1, and the read length distributions are illustrated in Fig. 2. As the various methods have distinct advantages and limitations, the use of multiplatform transcriptomics approaches have proven to be valuable[43]. For example, dRNA sequencing produces incomplete reads since a 15–30 nt long sequence always lacks from the 5’-termini, and also in many cases poly(A) tails are also missing. Nonetheless, as dRNA-Seq is free of RT- and PCR-biases, it can be used for the validation of introns. However, due to its lower coverage and shorter average read lengths (Table 1 and Fig. 1) compared to the cDNA libraries, and its other biases, dRNA-Seq is advised to use in conjunction with other methods. The Cap-selected cDNA library (*Cap*) produced the highest throughput, but shorter average read-length due to the applied size-selection (> 500 nt). On the other hand, the *Sequel* library produced the longest reads on average but with a relatively low throughput compared to the ONT cDNA libraries.

The LoRTIA software was used to detect the TESs, TSSs and splice sites (hereafter referred to as ‘features’). This program checks the quality of sequencing adapters and poly(A) sequences and then identifies and filters out false TESs, TSSs and splice sites generated by RT, PCR and sequencing as a result of false-priming, template switching, RNA degradation[44], etc. In order to have higher confidence for the validity of the annotated features, stringent filtering criteria were used. The features were only accepted if either one of the two following criteria were met: they were detected by at least two different methods, or they were detected in at least four different samples, of which no more than two could have been RSII samples. For intron annotation, an additional criterion was that every intron had to be supported by at least one read in the native RNA library. In each sample a feature was considered to exist if at least two reads supported it. Subsequently, the LoRTIA software was used to assemble the transcripts based on these features.

The LoRTIA toolkit is able to combine the results of different datasets, therefore this workflow can be regarded as a robust approach for the identification of TESs, TSSs and introns. As a result of the feature detection and subsequent filtering procedure, 83 novel TSSs and 8 TESs were identified. The stringent filtering criteria led to the detection of 103 introns (28 of them are novel), all of them complying with the GT/AG rule. The features that passed the stringent filtering criteria, are termed as ‘high-confidence’ hereafter.

The sequencing libraries were downsampled to the library size of the smallest library (*dRNA*) in order to make it possible to compare the efficiency of the different sequencing methods and library preparations in terms of TES, TSS and intron detection. The extent to which the high-confidence features were detected in the downsampled sets show how efficient the respective method is in terms of detecting the TESs, TSSs and introns, regardless of read count and library size. Figure 3 shows the precision and accuracy (recall) of the different libraries in terms of detecting these high-confidence features and the effect of downsampling. In terms of intron detection, from among the downsampled libraries the highest recall (73%) was achieved in the *Sequel* sequencing library

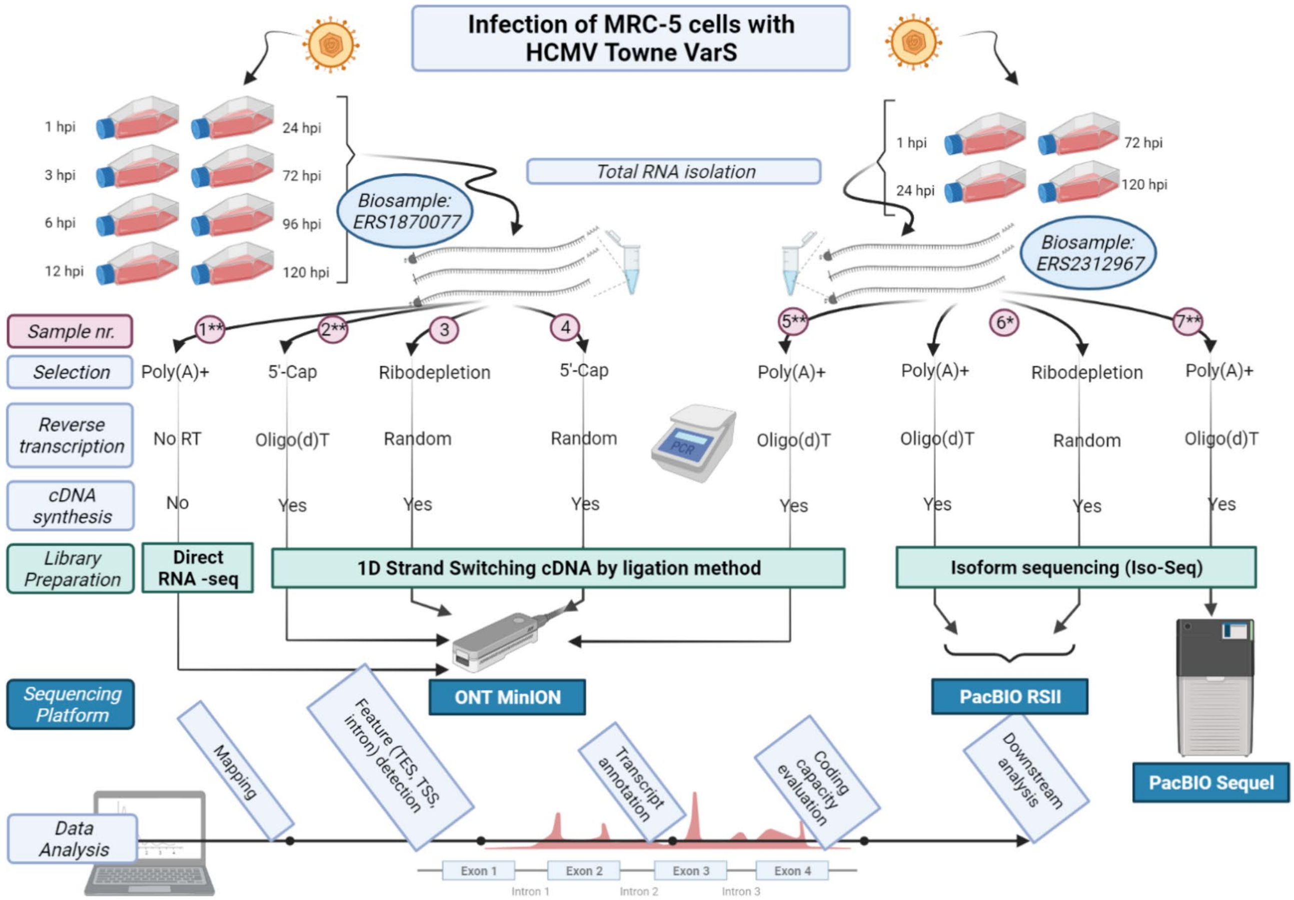

Figure 1. An overview of the utilized sequencing platforms, library preparation and sequencing methods and bioinformatic analyses of the resulting sequencing reads. The sample numbers in this figure correspond to the sample numbers in Table 1. The star in sample_6 (RSII) indicates that the data and analysis from this sample have been reported in our previous publication, while double stars indicate that only the sequencing data have been published in our previous publication. Created with BioRender.com.

Sample nr	Sample ID	Mapped read statistics					
		Sample size	Read count	Min. length	Mean length	Max. length	Mean coverage
1	*dRNA***	1	30,366	97	787	8,397	101,8
2	*Cap***	1	576,557	118	1,073	5,137	2,245.4
3	*Non-cap random*	1	151,234	190	996	8,246	645,9
4	*Cap random*	1	39,776	232	936	8,992	159,6
5	*Non_Cap***	1	352,485	135	1,570	10,731	2,271.4
6	*RSII**	8	61,375	83	1,250	5,833	329
7	*Sequel***	1	40,233	84	2,021	7,239	325,6

Table 1. Statistics of the reads, mapped to the Towne varS genome, according to each sample library used. The first column corresponds to the sample numbers in Fig. 1, whereas the second column corresponds to the identifier that is used throughout the study for referring to each sample. Green background corresponds to ONT, whereas blue to PacBio sequencing. The stars in the first column mean the same to as in Fig. 1.

(precision: 59%), which was comparable even to that of the *dRNA* library (68%). *RSII* showed a somewhat lower recall (54%), but its precision was somewhat higher (72%). With the exception of the *Cap* sample, which was similar to the PacBio samples, in the ONT samples the precision was generally higher (less false positives), while the recall was generally lower (less true positives). In the case of full libraries, the poly(A)-selected Cap sample (*Cap*) showed a very high recall (95%), but many false positives as well, as this library was the largest; while the *Cap random* sample showed a high precision, but there were many valid features that it could not detect (recall = 22%), again because of the library size, which in this case was small, comparable to that of the size of the *dRNA* sample. TES detection was more efficient in the PacBio samples, as both the precision and recall were higher in the downsampled libraries, (except for the *dRNA* and the *Cap* samples whose precision was similar,

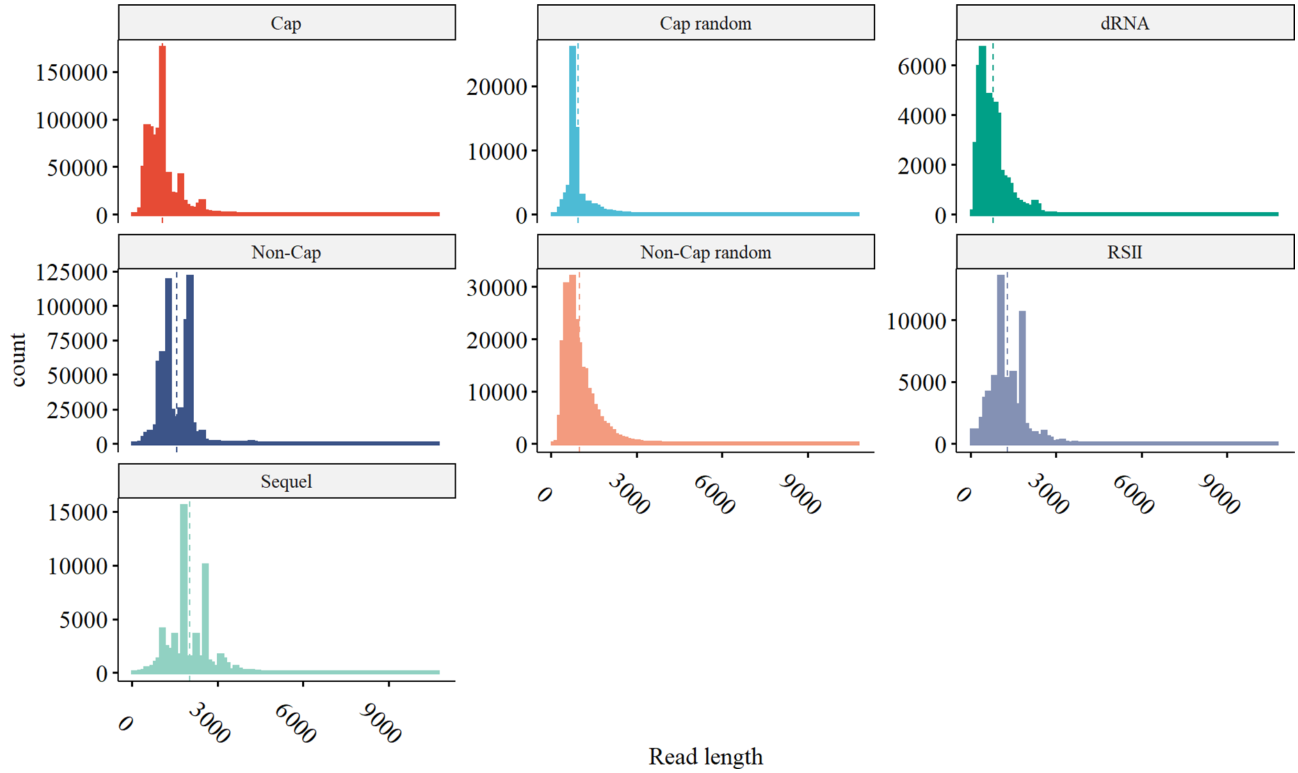

Figure 2. Mapped read densities per sample, with median lengths represented as dotted lines.

but not their recall). Moreover, the full PacBio libraries showed comparable values to the full ONT libraries, despite their much lower coverage. The Sequel sample performed the best (94% recall and 81% precision), even though its read count was an order of magnitude lower than that of the *Cap* and *Non_Cap* samples (Table 1). In the case of TSSs, the PacBio samples showed a similar performance to the *Cap* and better than the *Non_Cap* cDNA samples, both in the downsampled and full libraries. Thus, generally, the PacBio samples showed better efficiency in terms of TES and TSS detection than the ONT libraries, and similar performance in intron detection to the dRNA library; as after downsampling, both their precision and recall values were higher.

Transcript annotation and abundance estimation. To annotate transcripts, we used the *transcript_annotator* function of LoRTIA pipeline, that identify transcripts by searching for reads with correct 5'- and 3'-adapters that span from a high-confidence TSS to a high-confidence TES (within a 10-nt window). This method is highly accurate for annotating transcript isoforms, as the read that supports each isoform must have a correct poly(A) tail and 5'-Cap, and eliminating spurious transcripts and fragments. This, however means that the reads that do not carry the correct adapters are omitted, thus they are not included when calculating transcript abundance. The resulting transcripts were further filtered to have at least 2 such reads in different libraries (or at least 3 reads in the same library).

Then, we compared the identified transcripts with the previous dataset[40] and with other literature sources (Supplementary File 1). We termed a transcript identical to another if their termini were within a 10-nt window, and if their intron composition matched. If the difference was larger than 10-nt or the intron composition differed, then the transcript was termed either a length isoform or a splice-variant, respectively. Categorizing and naming of transcripts were carried out in-line with our previous convention[45].

After the stringent filtering procedure, a total of 576 transcripts were annotated (Fig. 4, Supplementary File 2). This does not include 22 5'-truncated or short transcripts that were filtered subsequently, because they were not supported by dRNA sequencing (please see *Feature detection and transcript annotations* in Materials and Methods section). This is a significant increase compared to the previously described 291 transcripts using RSII sequencing, underscoring the advantages of using multiple sequencing approaches. Although, 312 transcripts have already been described: 244 in our previous dataset[40] and 68 in other sources, the remaining 264 transcripts are novel. Note that the lack of confirmation of certain earlier described TSSs, TESs, introns and transcripts due to using more stringent criteria for the annotation in this study, does not necessarily mean that they do not exist.

In order to detect transcripts that would otherwise elude identification, both samples (Biosample ERS1870077 and ERS2312967) consisted of a mixture of cell cultures of different hpi (1–120 hpi) and were sequenced on different platforms and with different library preparation methods. These were selected to complement each other (Fig. 1, also refer to Materials and Method section), and not necessarily to precisely estimate relative transcript abundance. For example, the non-cap selected cDNA sequencing library was size-selected (> 500 nt). The read length distribution shows (Fig. 1) that the Sequel platform has a similar molecule-size preference to the RSII platform, while the MinION platform produces shorter reads. Also, some of the samples (e.g. the native RNA)

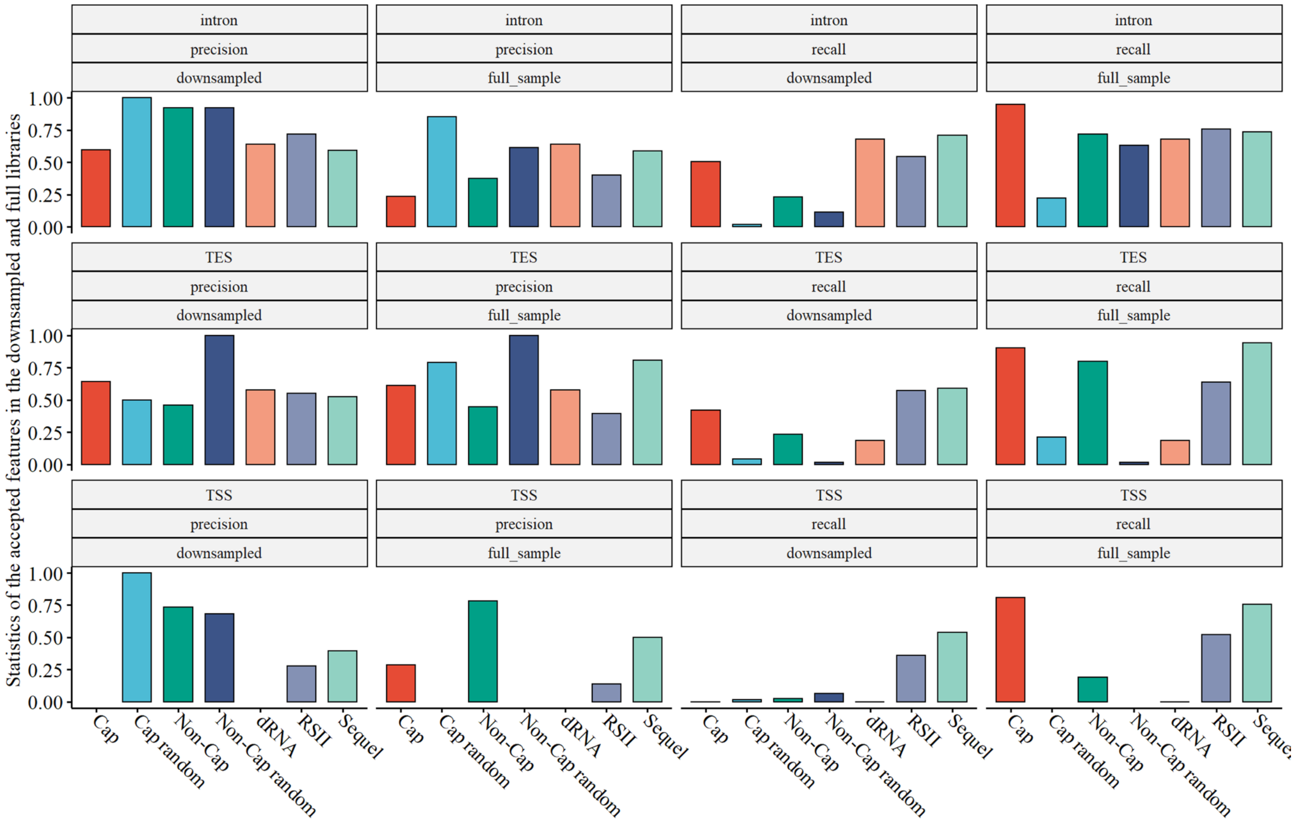

Figure 3. Effect of downsampling on the recall and precision of the detection of the high-confidence features (introns, TESs and TSSs) in the different sequencing methods and library preparations. The features were termed high-confidence if they passed the stringent filtering criteria. Recall was calculated as the ratio of high-confidence features that was found in the respective samples, while precision was calculated as the ratio of true positives in all the hits of respective samples. The downsampling of reads were carried out to match the sizes of the libraries to the size of the smallest library (dRNA).

yielded low coverage. For these reasons, the comparison of transcript abundances between these different datasets is unfeasible. The sum of the read counts (as produced with the LoRTIA program), from all samples for each transcript was thus used to categorize the transcripts into 5 abundance groups (Fig. 4). The read counts of the transcripts (on a log2 scale) are shown in Supplementary Figure S1. The transcripts are grouped according to which gene they originate from and also according to which dataset they were described in (source). The number of transcripts in each abundance category is summarized in Supplementary Fig. S2.

According to this abundance grouping, many of the newly described transcripts have a high read count (13 'very highly' and 23 'highly' abundant). Some of them are the products of the highly abundant *rl2-rl4* and *rl7-rl13* gene clusters[39]. Hence these are genes not only expressed in high quantity but also with high isoform polymorphism as well. The medium abundance category is distributed relatively evenly among the novel and previously described transcripts: 79 and 92. Many newly described transcripts are expressed in 'very low' (71) or 'low' abundance (98). These results show that the reason for why many of the newly described transcripts were undetected as of now is not that they are expressed in such a low abundance, but the previously used techniques were not able to differentiate between the transcript variants and isoforms.

Transcript isoforms. The alternative use of transcription start sites (TSS isoforms) results in shorter (labeled here: 'S') or longer ('L') 5'- untranslated regions (5'-UTRs) compared to the canonical transcript, but containing the same main ORFs, although short upstream ORF (uORF) compositions may differ. In this dataset, we detected 15 novel short TSS isoforms, which include short isoforms of the RL4, RL2 and RL3 transcripts (8, 2 and 1 isoforms, respectively), as well as one isoform of UL99, US34A and UL25 transcripts (Fig. 4). We used the reads obtained by dRNA sequencing to validate the TSSs of the short length isoforms. Since dRNA sequencing reads lack 10–25 nucleotides from their 5'-termini (Supplementary Fig. S3), we accepted those transcripts, where at least one (LoRTIA filtered) dRNA read mapped (in the correct orientation) near their 5'-ends, and the read was shorter than the transcript but with no more than 25 nucleotides. We also detected 55 novel long TSS isoforms; the majority of these were expressed from the *rl2-rl8* genomic region similarly to the S isoforms. Here the transcripts use both alternative TESs and TSSs (Fig. 4): we describe 22 L isoforms from the *rl2* gene, 4 from *rl7*, 2 from *rl13*, *rl3* and *rl5* as well. Besides this region, we found that novel L transcript variants expressed from genes: *ul132, rs2* and *us6* expressed 6, 4 and 2, respectively, while the *ul26 ul34 ul48 ul29a, ul78, ul73 ul123 ul5*

Figure 4. Genes (upper panel), ORFs (middle panel) and the annotated transcripts (bottom panel) of HCVM LT907985.2. Genes are colored according their function (as reviewed by Van Damme and Van Loock[29]), while transcripts are colored according to their read counts: Very High: read count higher than 500; High: between 100 and 500; Medium: between 100 and 25; Low: between 25 and 10; and Very Low: between 2 and 10.

us34a genes expressed 1 novel L isoform; as well as the *ul100, ul105, ul115 and ul82* genes, however in these cases, no L isoform was described previously.

In this study, 5 novel alternative TES variants (labelled with –AT in their names) were also identified: one TES isoform of the UL25 and UL54, and two of the UL95 mRNA.

In addition, we detected 46 novel splice isoforms of which one is an intron-retention variant (NSP, non-splice variant) of *ul123*. Three SP isoforms were found from the ul71 gene, while the rest were expressed from the *ul106-129* region: *ul112, ul123* and also and *ul129* genes were found to express novel splice variants (2, 2 and 1, respectively).

Non-coding transcripts and transcriptional overlaps. In this work, 71 RNA molecules were identified that did not fully encompass any of the 385 long ORFs (>=10 AA), described via ribosome-profiling by Stern-Ginossar et al[35] (we refer to these ORFs as 'validated' hereafter). However, many of these transcripts did contain experimentally validated short ORFs from the same source and in silico predicted in-frame co-terminal ORFs: some transcripts of *rl2-rl13, ul29a, ul71-ul74a, ul80, ul111a, ul123, ul129* genes; or only predicted ORFs, these are some transcripts of the following genes or genomic regions: *rl7, ul40, ul54, ul82, ul100-102, ul132* and *us30-34*. UL54.5-AT-L is an exception, as the in silico predicted ORF that it carries is not co-terminal with the canonical UL54 ORF, but it is in-frame with it. These transcripts are considered *embedded* transcripts rather than non-coding and will be discussed in the following section. The rest of the non-coding transcripts include 5 novel lncRNAs found to be expressed from the non-coding gene *rna4.9,* three from the *rs2* gene, two from *ul111* (*rna5.0*) and 1 from *ul102* (UL102-NC).

Another important type of lncRNAs are the antisense RNAs (asRNAs) molecules that are controlled by their own promoters[20,46]. We identified 3 such asRNAs from the *us30* gene region (and several complex transcripts as well, described in the following section), and detected an additional 7 that were described previously. The only gene in the vicinity that is in the same orientation as these transcripts is the *us33*, which starts 764 nts downstream to the start of these transcripts. It is unlikely however, that its promoter is involved in the transcription of these asRNAs, as no detected transcripts initiated from that gene. The annotation of two novel AS transcripts were also possible where no full-length AS transcripts were found before: UL45AS and UL29AS-S, as both are partially antisense to a canonical gene (ul29 and ul45, respectively). The UL45A transcript carries the experimentally validated ORFL114W.

On the other hand, the in silico predicted ORF carried by UL29AS-S transcript is not from experimentally validated list of ORFs, however it is co-terminal (and in-frame) with the validated ORFL82W. Thus, it is possible that these antisense transcripts have a coding capacity after all. Upon closer inspection of the LoRTIA results of this region, we found that two other TSSs were also detected, and the transcripts were annotated, only during the filtering procedure they were not considered as high-confidence. These two putative transcripts do carry the validated ORFL82W, and the longer TSS isoform carries an additional short ORF as well. We included these two transcripts in the analysis (UL29AS and UL29AS-L), but note that they need further confirmation.

In addition, many transcripts contain antisense segments, but they cannot be considered as true AS transcripts, rather they are the products of either transcriptional read-through between convergent genes (as is the case in TES isoforms encoded by *us1, ul48, ul53, ul54, ul8, ul89, ul103, ul115* and *us33*), or transcriptional overlaps between divergently oriented genes (*ul20, ul123* and *us34*).

We carried out a genome-wide analysis of transcriptional overlaps: the number of overlapping transcripts were calculated using a 10-nt sliding window. The transcriptional orientation of genes relative to the adjacent genes can be either convergence (→←), divergence (←→), or co-orientation (→→)[20]. Parallel overlaps between co-oriented genes are common throughout the entire viral genome[38], whereas divergent and convergent overlaps are restricted into distinct genomic locations. We found a novel convergent transcriptional overlap between the transcripts encoded by *us32* and *us33* genes and we also detected a novel divergent transcriptional overlap in the *ul100-102* region.

Multigenic transcripts. Long read sequencing techniques are especially suitable for distinguishing the polycistronic mRNAs from monocistronic transcripts[18]. We detected 55 novel polycistronic (PC) and 19 complex transcripts (CX) (Supplementary Fig. S4). CX transcripts are multigenic RNA molecules that contain at least one oppositely oriented gene[49]. The PC transcripts contain either additional introns or genes which were earlier considered to produce exclusively monocistronic transcripts or contain novel introns. Novel polycistronism was found in the following genes (genomic regions): *us7 ul74a ul95-96, ul114-116, rl7-9* and *us7-9*. We also identified 12 novel CX transcripts which were generated by transcriptional readthroughs or the use of long 5'-UTRs. The rest of the multigenic transcripts are TSS or TES variants of already described PC or CX RNA molecules. The identified PC transcripts include 34 bicistronic, 14 tricistronic, 4 tetracistronic, 2 pentacistronic and 1 hexacistronic (Supplementary Fig. S4). Polycistronic transcripts of nested genes were also found in *ul34* (1) and *ul49* (1).

In the *ul89-99* region, until now the polycistronic transcription units were shown to produce two families of nested 3´-coterminal transcripts, encompassing ORFs UL92–UL94 and UL93–UL99, respectively that are differentially regulated[48]. Here we identified two novel TESs that produce five novel TES transcripts, two in UL95 and three in UL96, respectively (Fig. 4, Supplementary File 1). The transcripts associated with the former TES are partially antisense to UL89 and carry three distinct in-frame ORFs (two short and one long), whereas transcripts associated with the latter TES are either monocistronic transcripts of UL96 or are polycistronic UL95-UL96. The bicistronic UL95-UL96 transcripts carry the same ORFs. Until now, transcripts produced from this region always described to end in UL99 and no data was found on the monocistronic, or bicistronic transcripts. UL95 protein

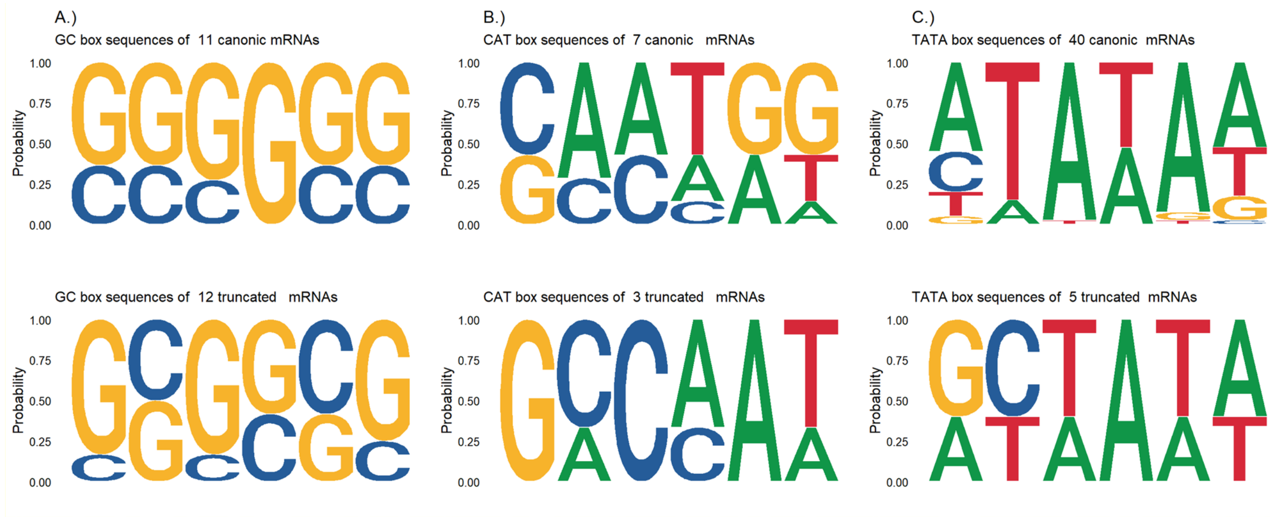

Figure 5. Weblogo of promoter elements: GC-boxes (**A**), CAT-boxes (**B**) and TATA-boxes (**C**) of truncated and canonical transcripts. Only those host genes were selected for the comparison that contained embedded genes. The number of each transcript that was found to contain promoter elements are shown above the respective weblogos.

is required for late viral gene expression and, consequently, for viral growth[46] and is associated with latency[47], thus their monocistronic variant may be important in regulating late gene expression as well.

We identified eight PC transcripts at the *ul112-116* genomic region (Fig. 4); polycistronic transcripts have already been described in the *us6-us11* region[50], but the US9-US6 transcript is novel. In addition, five PC transcripts were expressed from the *ul42-44* region, and we also identified several polycistronic transcripts in the RL7-RL3 and in the RL11-RL13 region. The highest number of novel PC transcripts (25) were found in the *ul71-ul73* region; however, these are length variants of already described PC transcripts.

We detected three complex transcripts of *ul101* gene, which are antisense to *ul101* and partially antisense to *ul102* but sense to *ul100*. These are probably the products of the downstream convergent gene *ul103*. Antisense transcripts were described already from this gene[51], however none of them encodes the ORFL234C ORF (132 AA), which is located in the intron of the two spliced mRNAs encoded by this gene[51], nor the relatively shorter ORFL233C (43 AA).

Putative nested genes. The use of alternative TSSs may result in the expression of 5'-truncated transcripts that lack their canonical start codons. However, in some such cases the canonical ORF may contain downstream ATG(s) that gives rise to one or more 5'-truncated ORF(s). In most cases, these embedded ORFs are in-frame and 3' co-terminal with the host ORF, and if functional, can be considered as (a) nested protein-coding gene(s)[50]. The transcripts carrying them are considered *embedded* or *truncated* transcripts. Some of these transcripts are likely to be fragments and not functional mRNAs, thus in order to gain higher confidence in their existence, we only considered those transcripts as true RNAs of which 5'-ends overlapped with reads from the native RNA sequencing library (dRNA). Overall, 65 such transcripts were detected in this sequencing dataset of which 36 were described previously and 29 is novel. Each of these was validated by dRNA sequencing, 20 embedded transcripts were considered as low-confidence and filtered out. Twenty-six transcripts carry nested ORFs that were described previously using ribosome-profiling by Stern-Ginossar and coworkers[35]. However, 3 contain only *in-silico* predicted embedded ORFs (Supplementary File 1). If translated, the truncated mRNAs lead to an N-terminally truncated version of the protein encoded by the canonical ORF. We found truncated transcripts expressed by genes from which no such activity was previously described. These are as follows: *rl3, rl7, rl8, ul43, ul49, ul94, ul105, us28, us33a,* and *us34*. The names of the transcripts of these novel putative nested have a '*0.5*' suffix. Additional novel truncated transcripts mapped to genes, from which such transcripts have already been reported to be transcribed: *ul5, ul13, ul34, ul35, ul38, ul71, ul75, ul80, ul85, ul132, us6, us18, us22* and also *rl4* and *rl13*.

We carried out a promoter analysis of the transcripts: the presence and sequence composition of CAT, GC and TATA-boxes and their distances from the TSS of the transcripts were examined. In addition, the Kozak consensus sequences of the transcripts were analyzed as well. We then compared these features of the 5' truncated RNAs (with truncated ORFs) to the transcripts of their host genes, those that carry canonical ORFs. Differences in the sequence compositions of their promoter elements were detected (Fig. 5), however the number of transcripts where CAT-boxes were found was only 3 in the truncated and 7 in the canonical group. GC-boxes and TATA-boxes were found in more cases (Fig. 5) and an apparent difference was seen between the two transcripts groups, indicating that the transcriptional regulation of the embedded genes differ from their hosts. TSSs showed clear differences as well: the bases upstream of the C/G start site contain more A-s in the truncated transcripts (Supplementary Fig. S5). The Kozak sequence composition showed a modification mainly in the important −3 Kozak site from the consensus G/A to C/G, which weakens the translation initiation signal somewhat, however

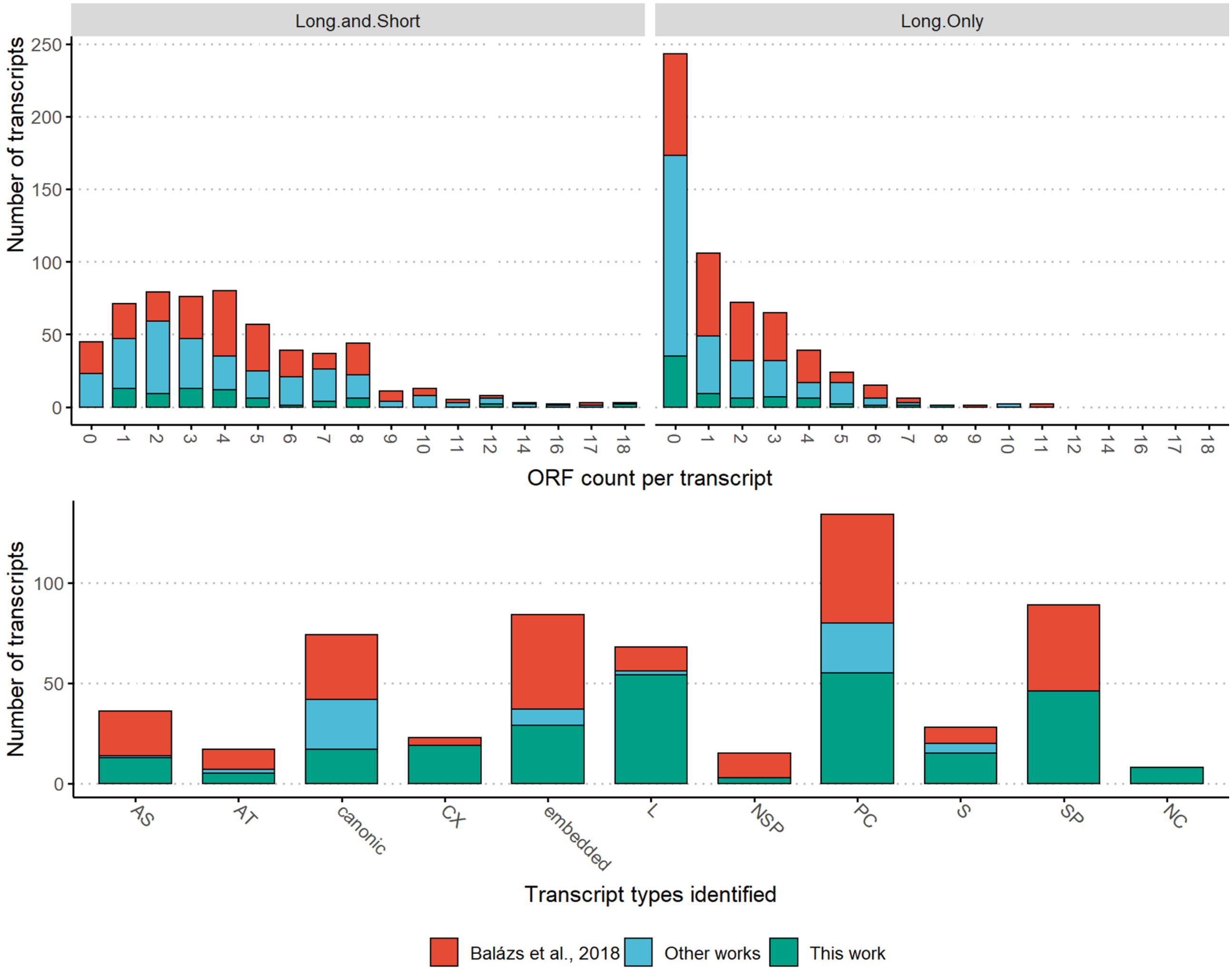

Figure 6. Upper panels: The frequency of transcripts, according to the number of ORFs they encompass (Long, or Long + Short, according to[35]). Lower panel: number of transcripts identified according to transcript categories. Transcripts described in this work are colored with blue, those that we described previously[40] are colored red, and those that we detected but were described in other works previously are colored green.

this did not cause an overall significant decrease of the mean Kozak sequence score in these transcripts (Supplementary Fig. S5). The differences altogether suggest that besides coding for different protein products, the embedded genes are differentially regulated both on the translation and on the transcription level, compared to their canonical counterparts.

Upstream ORFs and coding capacity. Upstream ORFs (uORFs) are a class of cis-regulatory elements within the 5' UTR of the respective mRNAs, and represent an alternative mode of translational regulation[52]. The uORFs are short (< 30 codons) and may initiate at near-cognate start codons; they generally repress the translation of the downstream main ORF, however they can stimulate translation of the downstream coding region as well[53]. The uORF composition in transcript variants (e.g. in short or long TSS variants) encoded by the same gene modulates their translational regulation[54]. In order to assess the coding capacity of the HCMV transcriptome, we transferred the experimentally validated ORF list (both short and long) published by Stern-Ginossar and collegues[35] to the Towne varS genome (from the Merlin genome) using BLAST. We included those ORFs from the in silico predicted list that were co-terminal and in-frame with the canonical ORFs (these are some of the embedded genes of the truncated transcripts) to the resulting ORF list. Subsequently, we mapped them to the detected transcripts and compared these ORF compositions to what was previously described[40]. Figure 6 upper panel shows the distribution of these ORFs in the transcripts and which dataset that are derived from. The analysis revealed 299 combinations of the validated ORFs (considering both short and long) carried by the transcripts. In an addition to the 195 combinations carried by transcripts that we detected, but were described earlier, we found 103 novel ones. Previously a total of 247 such combinations were described; thus, the results of this work greatly expand our earlier estimate of the coding capacity of the HCMV genome.

Discussion

In this work, we applied two LRS platforms (ONT and PacBio) with various library preparations and sequencing methods: native RNA and cDNA with and without Cap-selection and using oligo(d)T, or random primers to re-annotate the HCMV transcriptome. We used minimap2 for mapping and the LoRTIA software, with a stringent filtering procedure, to identify transcripts, which lead to the identification of 81 novel TSSs, 8 novel TESs and 28 novel introns. Comparison of the downsampled libraries showed differences between the library preparation protocols and sequencing methods: generally, the PacBio samples showed better performance than the cDNA ONT libraries in terms of TES and TSS detection, and similar performance to that of the dRNA library with respect of intron detection.

We then used the identified features to assemble HCMV transcripts, which enabled the confirmation of 312 previously described, and the identification of 262 novel transcripts. The novel transcripts include 8 lncRNAs, 19 complex, 13 antisense, 29 putative protein coding (truncated), 55 polycistronic, 8 splice variant, 3 intron-retention variant, 54 5'-long 15 5'-short variants and 5 alternatively terminated transcripts. The novel multigenic transcripts include several PC transcripts that are expressed from genes where polycistronism has not been described before or contain novel introns.

In the *ul89-99* genomic region, we identified novel bicistronic and monocistronic transcripts that are likely regulated differentially as is the case with their the previously identified variants[48]. Although UL95p is required for viral growth, probably these variants were downregulated in previous studies to such an extent that they have gone undetected. In the *us30-34* region three PC, five CX, and 3 AS transcripts were detected (besides two embedded and two TSS variants). The complex transcripts are generated as a result of convergent transcriptional readthrough between *us32* and *us33*. We also identified two complex transcripts of UL102, which are antisense to UL101 and partially antisense to UL102, but sense to UL100. Previous studies confirmed antisense transcripts from this gene[51], however none of them contains the ORFL234C. At this point we can only speculate about the function of the complex transcripts. It is possible that the overlap causes transcriptional interference, as proposed earlier[18,55]. Several transcript isoforms were produced from *rl2-rl7* genes. A previous report detected a high transcriptional activity from the RL genes, mainly *rl4*[39]. The *ul71-ul73* region expressed many length variants of previously described PC transcripts. The novel antisense transcripts (UL45AS and UL29AS) were found to carry experimentally validated and/or predicted ORFs, thus their role is likely more complex than merely regulating their host genes.

Alternative TSS-usage can cause the 5'-truncation of the ORFs, which may result in the formation nested genes, wherein one or more ORFs that are in-frame and co-terminal with the canonical ORF. These putative embedded genes might encode N-terminally truncated polypeptides. The identification of novel TSSs enabled us to confirm the existence of many such embedded genes. To gain high confidence in the existence of their expressed truncated mRNAs, an additional validation step (using dRNA reads) was employed to the already stringent transcript filtering criteria. These truncated mRNAs represent potentially novel functionalities of the viral protein repertoire, if translated. The promoters of nested genes and their sequence composition around the TSSs differ substantially, compared to the host genes, which may refer differences in the transcriptional regulation. Their Kozak sequences also show dissimilarities, thus they are presumably regulated differentially on a translational level as well.

The truncated transcripts, along with multigenic RNAs, transcript isoforms (many times attributed to different uORF compositions) significantly expand the coding capacity of HCMV. In this work, we used the HCMV strain Towne, a lab-adapted HCMV strain harboring numerous mutations compared to clinical strains. While these mutations mainly affect structural and immunomodulatory proteins, which are not expected to significantly alter viral transcription, surely not the transcript isoforms. Other HCMV strains might have different transcriptional architectures in some genomic regions[56].

Human Cytomegalovirus is a highly prevalent infectious agent, which partly due to its complex genome and transcriptional architecture causes a lifelong infection. By using novel RNA sequencing methods, a deeper insight into its intricate transcription was achieved in this study. The herein reported results complemented with the previous ones represent the most detailed transcriptome of this important virus.

Materials and methods

In this study, we used the RSII sequencing datasets described in our earlier publications by Balázs et al.[57] (Biosample ERS1870077), and also by Balázs et al.[42] for the other sequencing datasets (both for ERS1870077 and for ERS2312967). The materials and methods used to generate those sequencing data are described here as well.

Cells and viruses. Human lung fibroblast cells [MRC-5; American Type Culture Collection (ATCC)] were cultured at 37 °C and 5% CO2-concentration in DMEM supplemented with 10% fetal bovine serum (Gibco Invitrogen), 100 units of potassium penicillin and 100 µg of streptomycin sulfate per 1 ml (Lonza). Four and eight T75 cell culture flasks were used for the ERS2312967 and ERS1870077 samples, respectively. Semi-confluent cells were infected with the Towne VarS (ATCC) strain of HCMV (a multiplicity of infection of 0.5 plaque-forming units (pfu) per cell). The viral-infected cells were then incubated for 1 h, which was followed by the removal of the culture medium and washing of the cells with phosphate-buffered saline. Subsequently, fresh culture medium was added, and cells were kept in the CO2 incubator for 24, 72, or 120 h, in case of ERS2312967, and for 1, 3, 6, 12, 24, 72, 96 or 120 h in case of ERS1870077.

RNA purification. Total RNA was extracted from each time point samples using the NucleoSpin RNA kit (Macherey–Nagel) as described in our earlier publications (Oláh et al., 2015 BMC Microbiology, Tombácz et al. 2018 Sci Data PRV).

PolyA(+) RNA purification and ribosomal RNA removal. 20 µl of the isolated RNA sample from each time point were pooled. The Oligotex mRNA Mini Kit (Qiagen) was used to select polyadenylated RNAs from both samples. Two different, poly(A)-selected libraries were prepared. For the analysis of the non-polyadenylated RNA fraction of the viral transcriptome, the ribosomal RNAs were removed using the RiboMinus™ Eukaryote System v2 (Thermo Fisher Scientific) according to the kit's instructions.

Library preparation and sequencing. *Biosample ERS2312967.* ***ONT MinION sequencing—direct RNA*** 500 ng poly(A)-selected RNA was used for direct RNA sequencing. First-strand cDNA was generated using SuperScript IV (Thermo Fischer Scientific) and the adapter primers (supplied by the ONT's Direct RNA Sequencing kit; SQK-RNA001). The library preparation was carried out with the ONT Ligation Sequencing 1D kit (SQK-LSK108) following the recommendations of the manufacturer.

ONT MinION sequencing—oligo(dT)-primed, Cap-selected cDNA Two micrograms from the total RNA sample was used to generate first strand cDNAs using the Lexogen TeloPrime Full-Length cDNA Amplification Kit. Oligo(dT) or random primers were used for the reverse transcription (RT). The ligation of the 5′ adapter to the samples was carried out overnight at 25 °C. The samples were amplified with PCR (30 cycles) using the reagents supplied by the TeloPrime kit. The libraries for nanopore sequencing were generated using the Ligation Sequencing 1D kit (SQK-LSK108, ONT) and the NEBNext End repair / dA-tailing Module NEB Blunt/TA Ligase Master Mix (New England Biolabs) according to the manufacturers' instructions.

ONT MinION sequencing—random-primed, non-Cap-selected cDNA RNA mixture from the rRNA-depleted sample was used to produce cDNA library for MinION sequencing. The RT reaction was carried out according to the ONT's Ligation Sequencing 1D kit (SQK-LSK108) using random primers instead of oligo(dT) primers and SuperScript IV (Thermo Fischer Scientific). The second-cDNA strand was primed with the strand-switching (5') adapter. The amplification of the samples was carried out with KapaHiFi DNA polymerase (Kapa Biosystems) enzyme applying 16 PCR cycles. For the ligation reaction, the NEBNext End repair / dA-tailing Module NEB Blunt/TA Ligase Master Mix (New England Biolabs) was used.

ONT MinION sequencing—random-primed, Cap-selected cDNA The Lexogen TeloPrime Kit was used to generate libraries from 5'-Capped RNAs with random primers to enrich the non-poly(A)-tailed RNAs or to capture the 5'-end of rare, very-long complex transcripts. The random-primed RT was followed by the ligation of the 5' adapter from the TeloPrime Kit at 25 °C, overnight. The sample was amplified through 30 PCR cycles with the KapaHiFi DNA polymerase enzyme (Kapa Biosystems). The library for nanopore sequencing was generated using the ONT's Ligation Sequencing 1D kit (SQK-LSK108) according to the manufacturer's recommendations.

All the ONT libraries were run on R9.4 SpotON Flow Cells with a MinION sequencing device.

Biosample ERS1870077. ***PacBio RSII sequencing*** A total RNA mixture containing samples in equal volume from all post infection (p.i.) time points was used for library preparation. For the enrichment of poly(A)-tailed RNAs, the Oligotex mRNA Mini Kit (Qiagen) was used. The Eukaryote System v2 (Ambion) kit was utilized to produce ribosomal RNA-free samples for random primer-based sequencing. Adapter-linked anchored oligo(dT) primers or random primers were used for the RT. The cDNA samples were prepared using the Clontech SMARTer PCR cDNA Synthesis Kit. The reactions were carried out according to the PacBio Isoform Sequencing (Iso-Seq) protocol. PCR reaction was performed on 16 cycles and 500 ng from the amplified samples were used to prepare the PacBio SMRTbell libraries with the PacBio DNA Template Prep Kit 2.0 and they were bound to MagBeads (MagBead Kit v2). The P6-C4 chemistry was used for sequencing. RSII SMRT Cells (v3) were applied with the RSII platform. Seven SMRT cells were utilized for sequencing the poly(A) + SMRTbell templates and one for the random-primed library.

PacBio Sequel sequencing RNA mixture from the polyA(+) samples was used to generate cDNA library for single-molecular real-time sequencing on PacBio's Sequel platform. The Clontech SMARTer PCR kit and the PacBio Iso-Seq protocol were used. The SMRTbell DNA Template Prep Kit 2.0 was used for the generation of libraries, which then were bound to Sequel DNA Polymerase 2.0. The PacBio's MagBead-loading protocol was used with MagBead Kit 2.0, and the sequencing was carried out on the Sequel instrument using a single Sequel SMRT Cell (v2) 1 M with Sequel Sequencing chemistry 2.1. The movie length was 10 h. Consensus sequences were generated with the SMRT-Link v5.0.1 software (Potter, 2016).

The optimal conditions for primer annealing and polymerase binding were determined with the PacBio's Binding Calculator in RS Remote.

ONT MinION sequencing—oligo(dT)-primed, non-Cap-selected PolyA(+) RNA fraction was used to generate cDNAs using SuperScript IV (Thermo Fischer Scientific) and adapter-linked oligod(T) primers. The 5′ adapter primers were ligated to allow for second-strand synthesis. The sample was amplified through 16 cycles using KapaHiFi enzyme. The PCR products were run on an UltraPure Agarose (Thermo Fischer Scientific) gel and cDNA fragments larger than 500 nt were purified using the Zymoclean Large Fragment DNA Recovery Kit (Zymo Research). The library was prepared using the ONT 1D kit (SQK-LSK108) and the NEBNext End repair / dA-tailing Module NEB Blunt/TA Ligase Master Mix (New England Biolabs) according to the kit's manual. The library was sequences on a R9.4 SpotON Flow Cells using a MinION devices.

Data validation. The measurement of the samples was carried out with Qubit 2.0 fluorometer (Life Technologies) while their quality was checked by Agilent 2100 Bioanalyzer (Agilent High Sensitivity DNA Kit). Samples with RNA Integrity Numbers higher than 9.5 were used for this study.

Read processing. All sequencing reads were aligned to the HCMV strain Towne VarS genome (LT907985.2) using minimap2[58], using options *-ax splice -Y -C5*. The mapped reads were not trimmed and may therefore contain terminal poly(A) sequences, 3` adapters or 5′ adapter sequences (AGAGTACATGGG in case of the Sequel, TGGATTGATATGTAATACGACTCACTATAG in the case of the CapSeq and TGCCATTACGGCCGGG in case of the not cap-selected cDNA sequencing). These sequences are soft clipped and can be used to determine read strandedness. Direct RNA sequencing reads do not contain 5′ adapters; read directions are determined by the sequencer as RNA molecules enter the nanopores with the poly(A)-tail first.

Read statistics were calculated using custom R scripts (available upon request).

Feature detection and transcript annotation. TheLoRTIA toolkit (https://github.com/zsolt-balazs/LoRTIA) was used with default parameters on the PacBio (*LoRTIA -5 AGAGTACATGGG --five_score 16 --check_in_soft 15 -3 AAAAAAAAAAAAAAA --three_score 18 -s poisson -f True*), the Capselected MinION (*LoRTIA -5 TGGATTGATATGTAATACGACTCACTATAG --five_score 16 --check_in_soft 15 -3 AAAAAAAAAAAAAAA --three_score 16 -s poisson -f True*) and the non-Cap selected MinION (*LoRTIA -5 TGCCATTACGGCCGGG --five_score 16 --check_in_soft 15 -3 AAAAAAAAAAAAAAA --three_score 16 -s poisson -f True*) mapping outputs to annotate TESs, TSSs, and introns. This algorithm first analyzes the reads, whether they contain correct adapters. Template switching artifacts are filtered out. From these filtered reads, TSSs and TESs are annotated if the ends of at least two of these reads (5' or 3', respectively) match. These potential features are then analyzed in a ± 50-nt window and a feature is only considered as valid if it is a local maximum (based on a Poisson distribution determined from the surrounding coverage), if at least two reads support and passes the minimum ratio of coverage filter (0.1%). The feature that has the highest number of supporting reads in a smaller (10-nt) window is then selected. The significance of all qualified features was tested against the Poisson or the negative binomial distributions and the p-value is corrected using the Bonferroni method (https://github.com/zsolt-balazs/LoRTIA/wiki/Stats). Introns are detected based on the cigar 'N' from the alignments (https://github.com/zsolt-balazs/LoRTIA/wiki/Samprocessor). In order to make sure that these features are consequently valid, they were only accepted if either one of the two following criteria were met: a given feature was detected by at least two different methods, or it was detected in at least four different samples, of which no more than two could have been RSII samples. For intron annotation, we set an additional criterion: one of the samples that supported it had to be the native RNA library. The results for each sequencing run were combined with the sum_gffs.py command. The transcript annotation was carried out by the *transcript_annotator.py* function of LoRTIA on the analyzed *.sam* files and the filtered feature sets (TSS, TES and intron). This algorithm searches for reads with correct adapters that were aligned from a high confidence TSS to a high confidence TES; and considers only these as transcripts (introns must match exactly). The resulting transcripts were further filtered to be supported by at least three reads, or two reads but in two different samples. To exclude fragmented transcripts, we used the reads obtained by dRNA sequencing to validate the TSSs of the short length isoforms and 5'-trincated (embedded) transcripts. Since dRNA sequencing reads lack 10–25 nucleotides from their 5'-termini, we accepted those transcripts, where at least one (LoRTIA filtered) dRNA read mapped (in the correct orientation) near their 5'-ends, and the read was shorter than the transcript but with no more than 25 nucleotides (SupplementaryFigure S3).

Downstream analysis. Transcript naming scheme: we used the same transcript naming scheme as in our previous studies[45]. Genome annotations (including ORFs, according to Stern-Ginossar et al., 2012[35]) were transferred to the Towne var-S genome with metablastr[59] and Liftoff [60]. Data evaluation, comparison of the transcripts, assessing their coding potentials, calculating their Kozak sequence score, carrying out ORF predictions and BLAST comparisons and generating visualizations were carried out with the following R packages: ORFik[61], Gviz[62] and tidygenomics[57,63] using custom R scripts. Promoter elements and PASs were searched with https://github.com/moldovannorbert/seqtools.

Data availability

Raw and mapped data files have been uploaded to the European Nucleotide Archive under the accession number PRJEB25680 (https://www.ebi.ac.uk/ena/data/view/PRJEB25680) and PRJEB22072 (https://www.ebi.ac.uk/ena/data/view/PRJEB22072). A '.gff' file of the identified transcripts is available on figshare (https://figshare.com/articles/dataset/Kakuk_et_al_2021_HCMV_gff/14762430). This file can be opened with any appropriate program (e.g. Genious), to visualize the transcripts, or imported into R (for example via: rtracklayer::import.gff). All data can be used without restrictions.

Received: 25 March 2021; Accepted: 28 June 2021
Published online: 14 July 2021

References

1. Gustincich, S. *et al.* The complexity of the mammalian transcriptome. *J. Physiol.* https://doi.org/10.1113/jphysiol.2006.115568 (2006).
2. Black, D. L. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* https://doi.org/10.1146/annurev.biochem.72.121801.161720 (2003).
3. Matlin, A. J., Clark, F. & Smith, C. W. J. Understanding alternative splicing: Towards a cellular code. *Nat. Rev. Mol. Cell Biol.* https://doi.org/10.1038/nrm1645 (2005).
4. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* https://doi.org/10.1038/nmeth.2714 (2013).
5. Engström, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* https://doi.org/10.1038/nmeth.2722 (2013).
6. Marx, V. Method of the year: Spatially resolved transcriptomics. *Nat. Methods* https://doi.org/10.1038/s41592-020-01033-y (2021).
7. Byrne, A. *et al.* Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* https://doi.org/10.1038/ncomms16027 (2017).

8. Chen, S. Y., Deng, F., Jia, X., Li, C. & Lai, S. J. A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. *Sci. Rep.* https://doi.org/10.1038/s41598-017-08138-z (2017).
9. Nudelman, G. *et al.* High resolution annotation of zebrafish transcriptome using long-read sequencing. *Genome Res.* **28**, 1415–1425 (2018).
10. Cheng, B., Furtado, A. & Henry, R. J. Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. *Gigascience* https://doi.org/10.1093/gigascience/gix086 (2017).
11. Tombácz, D. *et al.* Dynamic transcriptome profiling dataset of vaccinia virus obtained from long-read sequencing techniques. *Gigascience* **7**, 2 (2018).
12. Moldován, N. *et al.* Third-generation sequencing reveals extensive polycistronism and transcriptional overlapping in a baculovirus. *Sci. Rep.* https://doi.org/10.1038/s41598-018-26955-8 (2018).
13. Viehweger, A. *et al.* Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. *Genome Res.* https://doi.org/10.1101/gr.247064.118 (2019).
14. Moldován, N. *et al.* Multi-platform analysis reveals a complex transcriptome architecture of a circovirus. *Virus Res.* https://doi.org/10.1016/j.virusres.2017.05.010 (2017).
15. Price, A., Hayer, K., Depledge, D., Wilson, A. & Weitzman, M. Novel splicing and open reading frames revealed by long-read direct RNA sequencing of adenovirus transcripts. *bioRxiv* https://doi.org/10.1101/2019.12.13.876037 (2019).
16. Tombácz, D. *et al.* Long-read isoform sequencing reveals a hidden complexity of the transcriptional landscape of herpes simplex virus type 1. *Front. Microbiol.* **8**, 1079 (2017).
17. Depledge, D. P. *et al.* Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nat. Commun.* **10**, 1–13 (2019).
18. Prazsák, I. *et al.* Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus. *BMC Genom.* **19**, 1–20 (2018).
19. O'Grady, T. *et al.* Global transcript structure resolution of high gene density genomes through multi-platform data integration. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkw629 (2016).
20. Boldogkői, Z., Moldován, N., Balázs, Z., Snyder, M. & Tombácz, D. Long-read sequencing—a powerful tool in viral transcriptome research. *Trends Microbiol.* **27**, 578–592 (2019).
21. Depledge, D. P., Mohr, I. & Wilson, A. C. Going the distance: Optimizing RNA-seq strategies for transcriptomic analysis of complex viral genomes. *J. Virol.* **93**, 2 (2018).
22. Batista, F. M. *et al.* Whole genome sequencing of hepatitis A virus using a PCR-free single-molecule nanopore sequencing approach. *Front. Microbiol.* **11**, 1–9 (2020).
23. Keller, M. W. *et al.* Direct RNA sequencing of the coding complete influenza A virus genome. *Sci. Rep.* **8**, 1–8 (2018).
24. Soneson, C. *et al.* A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat. Commun.* **10**, 1–14 (2019).
25. Stevenson, E. V. *et al.* HCMV reprogramming of infected monocyte survival and differentiation: A goldilocks phenomenon. *Viruses* https://doi.org/10.3390/v6020782 (2014).
26. Wattré, P., Dewilde, A. & Lobert, P. E. Human cytomegalovirus pathogenesis: an overview. *La Rev. Med. interne* https://doi.org/10.1016/0248-8663(96)80723-5 (1995).
27. Wen, L. Z. *et al.* Cytomegalovirus infection in pregnancy. *Int. J. Gynecol. Obstet.* https://doi.org/10.1016/S0020-7292(02)00239-4 (2002).
28. Gerna, G. & Lilleri, D. Human cytomegalovirus (HCMV) infection/re-infection: Development of a protective HCMV vaccine. *New Microbiol.* **2**, 2 (2019).
29. Van Damme, E. & Van Loock, M. Functional annotation of human cytomegalovirus gene products: An update. *Front. Microbiol.* **5**, 218 (2014).
30. Murphy, E., Rigoutsos, I., Shibuya, T. & Shenk, T. E. Reevaluation of human cytomegalovirus coding potential. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 13585–13590 (2003).
31. Tai-Schmiedel, J. *et al.* Human cytomegalovirus long noncoding RNA4.9 regulates viral DNA replication. *PLOS Pathog.* **16**, e1008390 (2020).
32. Pavelin, J. *et al.* Systematic microRNA analysis identifies ATP6V0C as an essential host factor for human cytomegalovirus replication. *PLoS Pathog.* https://doi.org/10.1371/journal.ppat.1003820 (2013).
33. Ramette, A. *et al.* Biogeography: An emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. *Microb. Ecol.* **53**, 197–207 (2007).
34. Zhang, L., Yu, J. & Liu, Z. MicroRNAs expressed by human cytomegalovirus. *Virol. J.* **17**, 1–12 (2020).
35. Stern-Ginossar, N. *et al.* Decoding human cytomegalovirus. *Science* **338**, 1088–1093 (2012).
36. Patro, A. R. K. Subversion of immune response by human cytomegalovirus. *Front. Immunol.* **10**, 1155 (2019).
37. Sijmons, S., Van Ranst, M. & Maes, P. Genomic and functional characteristics of human cytomegalovirus revealed by next-generation sequencing. *Viruses* **6**, 1049–1072 (2014).
38. Ma, Y. *et al.* Human CMV transcripts: An overview. *Future Microbiol.* **7**, 577–593 (2012).
39. Gatherer, D. *et al.* High-resolution human cytomegalovirus transcriptome. *Proc. Natl. Acad. Sci.* **108**, 19755–19760 (2011).
40. Balázs, Z. *et al.* Long-read sequencing of human cytomegalovirus transcriptome reveals RNA isoforms carrying distinct coding potentials. *Sci. Rep.* **7**, 15989 (2017).
41. Martí-Carreras, J. & Maes, P. Human cytomegalovirus genomics and transcriptomics through the lens of next-generation sequencing: Revision and future challenges. *Virus Genes* **55**, 138–164 (2019).
42. Balázs, Z., Tombácz, D., Szűcs, A., Snyder, M. & Boldogkői, Z. Dual platform long-read RNA-sequencing dataset of the human cytomegalovirus lytic transcriptome. *Front. Genet.* **9**, 432 (2018).
43. Moldován, N. *et al.* Multi-platform sequencing approach reveals a novel transcriptome profile in pseudorabies virus. *Front. Microbiol.* **8**, 2708 (2018).
44. Balázs, Z. *et al.* Template-switching artifacts resemble alternative polyadenylation. *BMC Genom.* **20**, 824 (2019).
45. Tombácz, D. *et al.* Long-read assays shed new light on the transcriptome complexity of a viral pathogen. *Sci. Rep.* **10**, 1–13 (2020).
46. Lin, S., Zhang, L., Luo, W. & Zhang, X. Characteristics of antisense transcript promoters and the regulation of their activity. *Int. J. Mol. Sci.* **17**, 1–17 (2015).
47. Tombacz, D. *et al.* Characterization of novel transcripts in pseudorabies virus. *Viruses* https://doi.org/10.3390/v7052727 (2015).
48. Towler, J. C., Ebrahimi, B., Lane, B., Davison, A. J. & Dargan, D. J. Human cytomegalovirus transcriptome activity differs during replication in human fibroblast, epithelial and astrocyte cell lines. *J. Gen. Virol.* **93**, 1046–1058 (2012).
49. Isomura, H. *et al.* The human cytomegalovirus gene products essential for late viral gene expression assemble into prereplication complexes before viral DNA replication. *J. Virol.* **85**, 6629–6644 (2011).
50. Jones, T. R. & Muzithras, V. P. Fine mapping of transcripts expressed from the US6 gene family of human cytomegalovirus strain AD169. *J. Virol.* **65**, 2024–2036 (1991).
51. Zhang, G. *et al.* Antisense transcription in the human cytomegalovirus transcriptome. *J. Virol.* **81**, 11267–11281 (2007).
52. Akirtava, C. & McManus, C. J. Control of translation by eukaryotic mRNA transcript leaders—Insights from high-throughput assays and computational modeling. *WIREs RNA* https://doi.org/10.1002/wrna.1623 (2020).

53. Vincent, H., Ziehr, B. & Moorman, N. Human cytomegalovirus strategies to maintain and promote mRNA translation. *Viruses* **8**, 97 (2016).
54. Wethmar, K. The regulatory potential of upstream open reading frames in eukaryotic gene expression. *Wiley Interdiscip. Rev. RNA* https://doi.org/10.1002/wrna.1245 (2014).
55. Tombácz, D. *et al.* Multiple long-read sequencing survey of herpes simplex virus dynamic transcriptome. *Front. Genet.* **10**, 834 (2019).
56. Cha, T. A. *et al.* Human cytomegalovirus clinical isolates carry at least 19 genes not found in laboratory strains. *J. Virol.* https://doi.org/10.1128/jvi.70.1.78-83.1996 (1996).
57. Balázs, Z., Tombácz, D., Szucs, A., Snyder, M. & Boldogkoi, Z. Data Descriptor: Long-read sequencing of the human cytomegalovirus transcriptome with the Pacific Biosciences RSII platform. *Sci. Data* **4**, 170194 (2017).
58. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* https://doi.org/10.1093/bioinformatics/bty191 (2018).
59. Drost, H.-G. & Gogleva, A. metablastr: Perform Massive Local BLAST Searches. (2020).
60. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btaa1016 (2020).
61. Tjeldnes, H. *et al.* ORFik: a comprehensive R toolkit for the analysis of translation. *BioRXiv* (2021).
62. Hahne, F. & Ivanek, R. Visualizing genomic data using Gviz and bioconductor. *Methods Mol. Biol.* https://doi.org/10.1007/978-1-4939-3578-9_16 (2016).
63. Ahlmann-Eltze, C. tidygenomics: Tidy verbs for dealing with genomic data frames. (2019).

Author contributions

D.T. and Z.C. performed the sequencing work and the library preparations. K.M. cultured the cells and carried out the infections. B.K. wrote the scripts and performed the bioinformatic analyses, with help from G.T., Z.Ba. and N.M. B.K. and D.T. drafted the article. M.S. and Z.Bo. critically evaluated the paper. All named authors read and approved the final article.

Funding

The study was supported by National Research, Development and Innovation Office (Grant numbers: FK 128252 to DT, K 128247 to ZBo), Magyar Tudományos Akadémia (Grant number: Lendület (Momentum) I Program LP-2020/8 to DT), NIH Centers of Excellence in Genomic Science Center for Personal Dynamic Regulomes (Grant number: 5P50HG00773502 to MS), University of Szeged Open Access Fund (Grant number: 4675).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-93593-y.

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

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

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