JOURNAL OF VIROLOGY, Sept. 1993, p. 5502–5513
0022-538X/93/095502-12$02.00/0

Vol. 67, No. 9

Spliced Transcripts of Human Cytomegalovirus

WILLIAM D. RAWLINSON* AND BARCLAY G. BARRELL

Laboratory of Molecular Biology, Medical Research Council, Hills Road, Cambridge, CB2 2QH England

Received 8 March 1993/Accepted 2 June 1993

The availability of the human cytomegalovirus (HCMV) genomic sequence has resulted in more extensive knowledge of the overall coding capacity of the virus. Using polymerase chain reaction and rapid sequencing techniques, we have studied the splicing of mRNAs from a number of the predicted open reading frames (ORFs). Splicing was found between the UL122(IE2) ORF present within major immediate-early (MIE) region 2 and the downstream ORF (UL118) predicted to encode an incomplete glycoprotein. This locates the IE2 3′ donor site and provides evidence of a link between the MIE region and downstream ORFs. The downstream UL119-UL118-UL115 ORFs also undergo differential splicing, further increasing the known complexity of this region of the genome. A detailed map of the differential splicing within the region encoding the MIE ORF is presented. Also described are several previously unidentified spliced ORFs found in the long repeats and long unique regions, including one encoding a transcript with a large (4-kb) intron. The results show that spliced transcripts are encoded from throughout the genome at immediate-early, early, and late times postinfection.

Human cytomegalovirus (HCMV) has a double-stranded DNA genome consisting of 229,354 bp. Analysis of the protein-coding content of the complete DNA sequence was published from this laboratory in 1990, and at that time the genome was predicted to encode 208 open reading frames (ORFs) (8). These were predicted on the basis of their length, the overlap between them, and codon usage (3, 8). Although transcriptional data for an increasing number of these ORFs are available, function and gene products have been characterized for fewer than half of them (46). Of the total number of predicted ORFs, 47 belong to nine distinct gene families and 44 have significant homology with other herpesvirus ORFs with known or imputed functions (8).

Transcription from the HCMV genome is divided into three major temporal stages. Immediate-early (IE) genes are those first transcribed after infection in the presence of inhibitors of protein synthesis, predominantly from the ORFs UL123(IE1), UL122(IE2) (53), UL36-UL38 (29), and US3 (58). The UL123(IE1)-UL122(IE2) ORFs make up part of the major IE (MIE) region, which comprises approximately 16 kb of the long unique (UL) part of the genome and hybridizes to 88% of IE RNA (16, 49, 51, 54). Early genes are transcribed in the absence of viral DNA synthesis and are distributed throughout the genome, although the repeat regions are the most transcriptionally active (16, 50). The major early 2.7-kb monocistronic RNA encoded by the TRL4 ORF is also found much less abundantly at IE times and relatively less abundantly at late times postinfection (20, 34). Late transcription begins with the onset of viral DNA replication from around 24 h postinfection (hpi) (48). Expression at late times is from the entire genome and results in the production of transcripts encoding virion proteins such as UL86 (9), phosphoproteins such as UL32(pp150) (25) and UL83(pp65) (37), glycoproteins such as UL55(gB) (13), and other transcripts encoding proteins of unknown function, such as UL89 (12).

As large parts of the HCMV sequence have no known function, it is important initially to make a more accurate map of transcription from predicted (and, in some cases, unexpected) ORFs within the viral genome. Eukaryotes are known to have most of their translated nuclear genes split into coding (exon) and noncoding (intron) sequences (44). It is predicted that HCMV may also encode many spliced genes (17), as do a number of other herpesviruses (1, 40). Currently, only 12 of the ORFs of HCMV have been shown to be spliced; 3 differentially at IE times (29, 51, 58) and 2 at early times (60). The donor and acceptor sequences of published splice sites, along with the consensus splice signals for viral splices (44), are shown in Table 1.

The results of our study of splicing within a number of ORFs of HCMV strain AD169 are presented here and are summarized in Fig. 1 and 2. Using the polymerase chain reaction (PCR) of cDNA with primers on either side of potential splices (reverse transcription [RT]-PCR of splices), we have been able to confirm suggestions that UL122(IE2) is spliced to a downstream exon (UL118) (22, 53). As ORFs within this region (comprising ORFs UL119 to UL115) also undergo differential splicing (31), this observation demonstrates increased complexity of the MIE region. Details of spliced RNAs from a number of other ORFs are presented. The strategy utilizing RT-PCR of splice sites has proven useful in identifying the exact locations of splice signals, although the high sensitivity of PCR means that it has not been possible to prove whether the splices found are present in mRNA or heterogeneous nuclear RNA (hnRNA). Northern (RNA) blot studies attempting to distinguish between the two possibilities are also therefore presented.

MATERIALS AND METHODS

RNA analysis. (i) RNA preparation. Cytoplasmic RNA was obtained from MRC-5 cells infected with HCMV strain AD169 at a multiplicity of infection of 10 PFU per cell, as previously described (58). The RNA from IE, early, and late times was used as template for the synthesis of first-strand cDNA in RT-PCR splicing studies. Whole-cell RNA (used in the Northern blot studies) was obtained from MRC-5 cells infected with HCMV strain AD169 at a multiplicity of infection of 10 PFU per cell by a standard guanidinium isothiocyanate and phenol chloroform method (10). For the preparation of IE RNA, cycloheximide (100 μg/ml) was

* Corresponding author.

TABLE 1. HCMV splice sites from published studies[a]

ORF	Position of:		Intron size (bp)	Sequence of:		Reference
	Don	Acc		Donor	Acceptor	
Consensus				NAGGTAAGT	NTTNTNTTTTTTTTNCAGG	44
UL36	49575	49471	103	AAGGTAAGC	TTTTTCTATTCTCTACCAGG	29
UL37	52219	50989	1,229	CCAGTAAGC	TGTTTCATTTTCTTTCTAGT	29
	50947	50842	104	CAGGTAAAA	CCGTGTCGTCTCCACGTAGG	
UL112-UL113	161345, 161781	161503, 162063, 162182	157, 281, 400	ACGGTGAGT	TGTGTCGTCCCGTCTGTAGG	60
				ACGGTGCGT	TGTTCCTCCGAATTCGCAGG	60
				ACGGTGCGT	ATCTCCCCCTGGTTTCCAGC	
UL118-UL117	166978	165761	1,216	ATGGTGATT	GATGCCGCACACGCCACAGG	31
UL116-UL115	164662	164606	55	CTGGTACGT	GCAATATATAACGTTTTAGG	31
UL123	173610	172782	827	GACGTAAGT	CCATGGGTCTTTTCTGCAGT	51
UL123	172695	172580	114	ACGGTACGT	CTATTTCTCATGTGTTTAGG	51
UL123	172396	172225	170	TCGGTAAGT	TTGTTATCCTCCTCTACAGT	51
UL122		170850	1,545	TCGGTAAGT	GTCTTCTTATCACCATCAGG	51
US6[b]	?198622	?198277		CGAGCCGCT	CGAAACTGAGCTCCCACAGG	26

[a] Donor and acceptor splice sequences are shown for previously mapped mRNAs and for consensus virus splice sequences (excluding those in reference 59; see Table 2 and Fig. 2). The positions of the last base of exon 1 (Don) and the first base of exon 2 (Acc) are underlined and are shown for the prototype genome (8).

[b] Location of precise intron-exon boundaries unknown.

added to the culture medium 1 h before infection and the cells were harvested at 12 h hpi. For early RNA, phosphonoacetic acid (100 μg/ml) was added 1 h before infection and the cells were harvested at 24 hpi. Late and mock RNAs were derived from infected and uninfected cells, respectively, cultured in parallel, and harvested at 96 hpi.

(ii) Poly(A) selection. Poly(A)$^+$ RNA was obtained by incubating the whole-cell RNA with paramagnetic polystyrene beads attached to 25-nucleotide (nt)-long, poly(T) tracts (Dynabeads, no. 610.01; Dynal, United Kingdom). All procedures were performed as recommended by the manufacturer.

(iii) Northern blots. Northern blotting was performed by a standard method modified for the use of short (21-nt) oligonucleotides as probes. Twenty micrograms of total RNA or 2 μg of poly(A)$^+$ RNA was denatured with glyoxal and dimethyl sulfoxide and then loaded into separate wells on an agarose gel and size fractionated alongside commercial RNA markers covering the size ranges 0.16 to 1.77 and 0.24 to 9.5 kb (Bethesda Research Laboratories, Bethesda, Md.). The RNA was transferred by capillary elution and then covalently linked to a nylon membrane by UV irradiation (43).

Oligonucleotides end labelled with digoxigenin (DIG; Boehringer Mannheim Biochemica, Mannheim, Germany) by using terminal deoxynucleotidyl transferase (11) were used to probe the nylon filters. The filters were prehybridized for at least 1 h at 42°C in 5× SSC (1× SSC is 150 mM NaCl plus 15 mM sodium citrate), 0.5% blocking reagent (Boehringer Mannheim Biochemica), 0.1% Sarkosyl, and 0.02% sodium dodecyl sulfate (SDS). The probe was added to 5 ml of new prehybridization solution, and hybridization was allowed to proceed for 4 h at 42°C with continuous rolling in a hybridization oven (Techne, Cambridge, United Kingdom). The probe was removed and stored at −20°C for future use. Filters were washed once at room temperature in 2× SSC–0.1% SDS and then once each in 1× SSC–0.1% SDS at 37°C for 2 min and at 55°C for 2 min. The detection reaction was performed by a standard commercial chemiluminescent protocol with anti-DIG antibody. This antibody was commercially available labelled with alkaline phosphatase that dephosphorylates the substrate AMPPD [3-(2′-spiroadamantane)-4-methoxy-4-(3′-phosphoryloxy)-phenyl-1,2-dioxetane], which then decomposes to produce a steady state of light emission (Boehringer Mannheim Biochemica). The light reaction was detected by exposure of X-ray film to the filters for from 5 min to 6 h.

Splice site determination (RT-PCR of splice sites). (i) Oligonucleotide primers and probes. Oligonucleotides were synthesized on an Applied Biosystems 380B DNA synthesizer, and the concentration was determined spectrophotometrically and then adjusted to 10 pmol/μl. Purification before their use as primers in the PCR or sequencing reaction mixtures was found to be unnecessary. Primers (21 nt) were designed to avoid long runs of adjacent G+C residues, especially at the 3′ end, where greater than 2 such residues promote mispriming. Complementary 3′ ends were avoided so that priming and amplification of the oligonucleotides alone (primer-dimers) were minimized. Oligonucleotides with sequence motifs that encouraged secondary structure (including palindromes) were avoided (5). The primers were chosen from regions 100 to 200 bases of the putative 5′ and 3′ splice sites (either side) to allow for the possible presence of alternative splice sites close to the ends of the ORF (42).

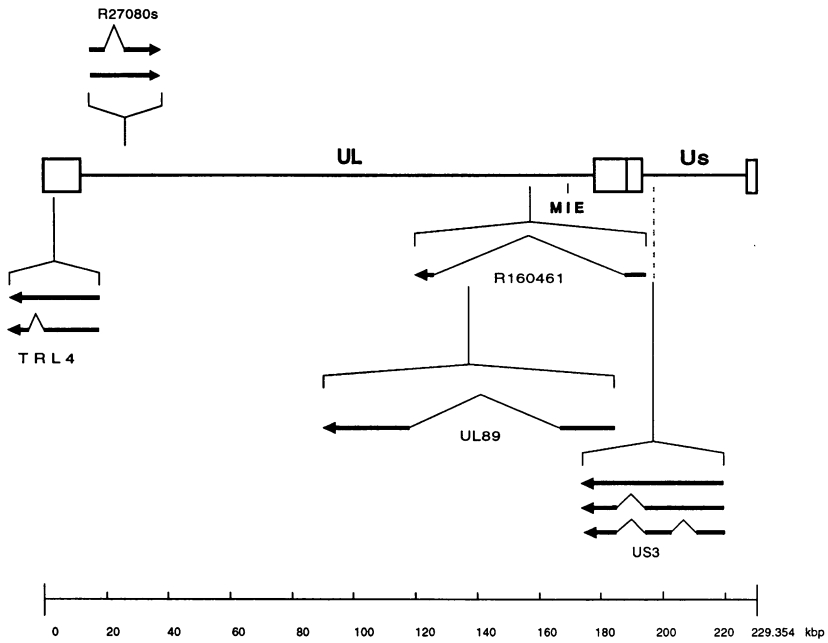

FIG. 1. Summary of splicing patterns in transcripts from the present study described in Table 2. ORFs are shown along the prototype genome (8), with the location marker at the center of the ORF. Arrow length is proportional to ORF length from the start of the ORF to the poly(A) signal and indicates the direction of coding except for the ORFs US3 and R27080s, the lengths of which are magnified four times compared with other ORF lengths. MIE, position of the major immediate early region shown in more detail in Fig. 3.

(ii) cDNA synthesis for splicing studies. First-strand cDNA synthesis was performed by RT of cytoplasmic mRNAs by using a 30-nt poly(T) primer and standard methods (19). To confirm the production of first-strand cDNA in each experiment, the incorporation of deoxynucleoside triphosphates into the newly synthesized cDNA was shown by including [α-^{32}P]dCTP in the nucleotide mix of an aliquot removed from the RT reaction mixture. The labelled first-strand cDNA was subjected to electrophoresis on a 1.4% agarose gel under denaturing conditions (50 mM NaOH) before autoradiography.

(iii) PCR. PCR conditions were altered from standard methods (42) to optimize reaction conditions for the primers and DNA used. The sensitivity of the reaction was improved by omitting gelatin completely and changing the final $MgCl_2$ concentration in the *Taq* polymerase buffer to the optimum for each set of primers, determined by titration of the magnesium concentration over the range of 1.0 to 4.0 mM (61). All reactions were carried out with DNA and cDNA in parallel to allow simple and accurate comparison of the sizes of the unspliced and spliced products. This low-concentration positive control of 1 pg of HCMV DNA was used in all PCRs. Any product in the cDNA lane smaller than that in the DNA lane was possibly spliced. These bands were cut out of a low-melting-temperature agarose gel, phenol extracted, and ethanol precipitated before being sequenced. The PCRs were carried out in polycarbonate 96-well microtiter plates in a thermal cycler (Techne Dri block; Techne) with an initial denaturation step of 94°C for 5 min followed by 30 cycles of denaturation (94°C for 1 min), annealing (55 to 60°C for 1 min), and extension (72°C for 1 to 4 min) with a final prolonged extension step (72°C for 10 min).

The procedures described here and in the preceding section are referred to in the text as RT-PCR of splice sites to distinguish them from splices identified by sequencing of late cDNAs from the cDNA library, the technique for which is described in the next section.

The cDNA library derived from late mRNAs. A library of HCMV late mRNAs that were poly(A)$^+$ selected, reverse transcribed to produce cDNA, and then inserted into plasmid pUC9 was obtained from Jon Oram, Public Health Laboratory Service, Porton Down, United Kingdom. The library was prepared from fibroblasts infected with AD169 at a multiplicity of infection of 10 PFU per cell, and the RNA was harvested at 120 hpi. cDNA was inserted into pUC9 by GC tailing. Template DNA for sequencing was prepared from this library either in microtiter plates by a standard alkaline lysis method (18) or by PCR amplification of the insert with primers flanking the polylinker region of the plasmid. The DNA was then visualized on an agarose gel to allow for accurate size estimation of each insert. Each cDNA discussed in the text was found to have a poly(A) tail and a consensus poly(A) signal. In all of the cDNAs sequenced, examination of the genomic DNA sequence revealed the presence of a consensus G/T cluster sequence downstream of the mRNA cleavage site (data not shown).

(i) Sequencing. Standard dideoxy sequencing was performed by *Taq* cycle sequencing (also called the linear PCR method) (14). The sequencing reactions were carried out with a thermal cycler able to accept samples loaded into 96-well polycarbonate microtiter plates (PHC3 Techne). Primer was end labelled with [α-^{32}P]dATP with polynucleotide kinase under standard conditions (43). Following this, 0.2 pmol of labelled primer was added to a mixture containing 1× PCR buffer (10 mM Tris [pH 8.3], 50 mM KCl, 1.5 mM $MgCl_2$, 0.01% [wt/vol] gelatin), dideoxy deoxynucleoside triphosphates (containing one of the dideoxy nucleotides ddTTP, ddCTP, ddGTP, and ddATP), and 1 U of *Taq* polymerase (Cetus Corp.). The reaction mix (total volume, 18 μl) was added to the DNA in a 96-well polycarbonate

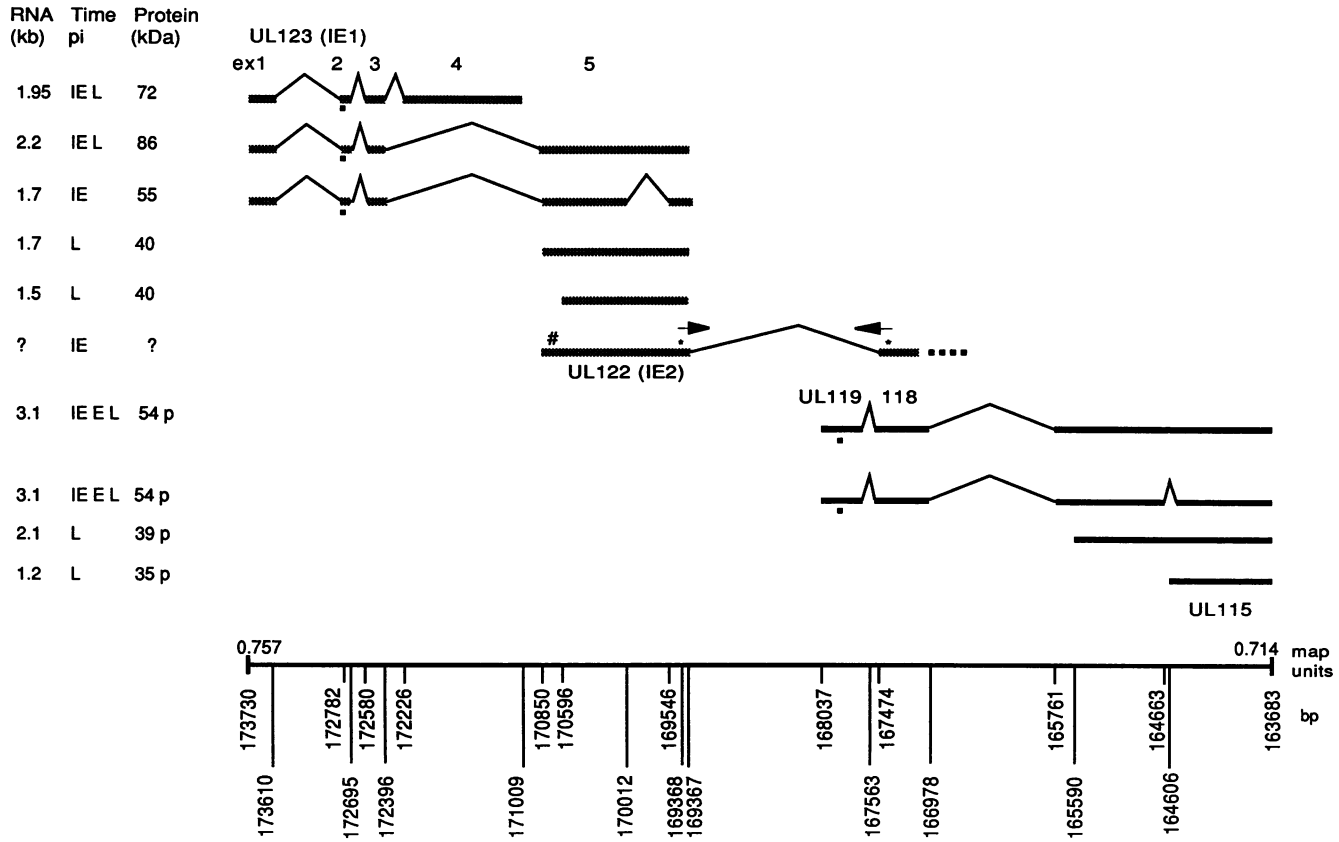

FIG. 2. Summary of the arrangement of the MIE region of HCMV. All exons and introns are drawn to scale, and the beginning and predicted end of the MIE region are shown. Donor and acceptor splice sites from the current study for the UL122(IE2)-UL118 and UL119-UL118 ORFs are indicated. The positions of all the features on the prototype genome are inverted (from the complementary strand) to be read 5′ to 3′ from left to right. The locations of the features on the AD169 genome are indicated at the foot of the figure. Numbers for splice sites apply to the last base of exon 1 and the first base of exon 2 and for ORFs apply to the start or stop of mapped transcripts. The undetermined end of the UL122(IE2)-UL118 spliced ORFs is indicated by a dashed line. Crosshatched exons are those described in references 22, 51, and 53, and solid exons are described in reference 31 and the current study (UL119/118). Exons 1 to 5 of UL123(IE1) are indicated. Exon 1 is noncoding, and exon 5 is ORF UL122(IE2). The location of the labelled oligonucleotide used in the Northern blots, the results of which are presented in Table 2 and Fig. 3c, is denoted by "#." The locations of Kozak consensus K-ATG sequences in UL123(IE1) exon 2 at 172765 to 172763 and in UL119 at position 167983 are denoted by ■. Non-Kozak consensus ATG sequences are present in UL116 at position 165474 and in UL115 at position 164530. The PCR primers used to amplify the UL122(IE2)-UL118 splice, the forward primer located at positions 169440 to 169460, and the reverse primer at positions 167320 to 167300 are denoted by solid arrows and asterisks.

microtiter plate, covered with 20 μl of paraffin oil (BDH, Poole, United Kingdom), centrifuged briefly at 1,500 rpm, and then placed in the temperature cycler. The amplification reactions consisted of 30 cycles of 95°C for 1 min, 55°C for 2 min, and 72°C for 2 min. The samples were subjected to electrophoresis on a gradient polyacrylamide gel by using standard conditions (4). The gel was dried and autoradiographed at −70°C overnight, and the sequence was read by using a sonic digitizer and the Gelin computer program (47).

(ii) Statistics. The A+T contents of introns and exons were compared by a one-tailed Student *t* test for paired data (2).

RESULTS

In order to describe more fully some of the large number of uncharacterized genes in the HCMV genome and to build on the data acquired from determining the HCMV DNA sequence, we have looked for new transcripts from predicted ORFs in several ways. Possible splice sites in HCMV were found by searching the genomic sequence with known splice donor and acceptor consensus sequences (44) and other features of the sequences immediately surrounding the intron-exon borders (36, 45). A subset of the large number of possible splices predicted was then studied by RT-PCR of splice sites with first-strand cDNA from IE, early, and late times as template. This identified splices between ORFs with close consensus donor and acceptor sites but did not yield information regarding those with poor consensus sequences or splices across very large introns (such as are found in Epstein-Barr virus [21, 38]). Information relating to transcripts with these latter characteristics (TRL4s, R27080s, and R160461 in Table 2 and Fig. 1) was obtained by sequencing cDNAs derived from the late-cDNA library.

By these techniques, 6 new splices were defined (Table 2 and Fig. 1) and 21 other putative splices (Table 3) were found not to occur in HCMV strain AD169 under the conditions used in this study. The sizes of the newly defined transcripts were determined by probing Northern blots with labelled oligonucleotides (Fig. 3). Data shown in Table 2 regarding the poly(A) signals and associated sequences were derived from sequencing cDNAs from the cDNA library, which was produced by RT of infected-cell RNA. The functional signif-

TABLE 2. Splice donor and acceptor sequences mapped in present study with positions on prototype genome of HCMV[a]

ORF	Position of:		Splice sequence				Transcript characteristic				
	Start	K ATG	Donor	Position[b]	Acceptor	Position	Intron size (bp)	Time pi[c]	Size (kb)	Poly(A) signal (sequence and position)	RNA cleavage site
TRL4s	4435		ACGGTGAAT	3323	CCCCACCTCGGCATGCCGG	3253	69	E++, L+++	2.7, 2.7	AATAAA, 2113	2092
R27080s	27080	27108	CAGGTAAAC	27193	GTTATCGTGTTTTTGCAGC	27277	83	L+++	0.4	ATTAAA, 27549	27574
R160461	160461	159668	CAGGTAGGT	159632	GGGTTTCTTCTCTTGCAGG	155103	4,528	L++	1.1	AATAAA, 154829	154795
UL89	138803	138389	AACGTGAGT	137502	TGTCTCCTCTCTACACAGA	133599	3,902	E+, L++	3.2, 3.2	AATAAA, 131390[d]; AATAAA, 129513[d]	
UL119/118	168037	167983	AAGGTAAGT	167563	TATGAATTTTATCCACAGG	167474	88	IE+[e], E+, L+++	3.1; 4.2, 3.1; 4.2, 3.1, 2.1, 1.2	AATAAA, 163683[e]	
UL122/118	170878[d]	170599[d]	TCAGTAAGT	169368	TATGAATTTTATCCACAGG	167474	1893	IE++, L++	7.5, 7.0, 5.0, 2.2, 1.7; 2.2, 1.7	AATAAA, 163683[e]	
US3iii	194767[f]	194690[f]	CAGGTGAGG	194295 (−126)	CTGCGATTTATTTAACAGG	194126	168	IE, E	0.74[f], 0.74	AATAAA, 193918[f]	193890[f]
US3iv	194767	194690	GTGGTATCG, CAGGTGAGG	194607 (−454), 194295 (−126)	TCAATCTTACATGGACAGA, CTGCGATTTATTTAACAGG	194454, 194126	152,168	IE, E	0.58[f], 0.58	AATAAA, 193918[f]	193890[f]

[a] Size of the transcripts and the pattern of temporal expression established by probing Northern blots with strand-specific oligonucleotide probes are shown. The cleavage site for poly(A) addition and the consensus poly(A) signal with the respective positions on the genome are shown when these were sequenced as cDNAs. When no cleavage site has been sequenced, the predicted cleavage site is shown.

[b] Position of the underlined base on the prototype genome (8). For splice sequences, this represents the position of the last nucleotide of exon 1 and that of the first nucleotide of exon 2.

[c] +, an estimate of the band intensity on the Northern blots shown in Fig. 3. pi, postinfection; E, early; L, late.

[d] Position predicted from genomic sequence.

[e] Data from reference 31.

[f] Data from references 59 and 55.

TABLE 3. mRNA splices predicted from genomic sequence but not found on sequencing products after PCR of cDNA

ORF	Donor sequence	Position[a]	Acceptor sequence	Position	Characteristics of PCR products			
					Actual size (bp)[b]	Time[c]	Size (bp) of predicted product	
							Splice[d]	Full length[e]
TRL13/14	GAGGTAATC	11162	ATAAAATGTGCGAATTAGG	11284	300	E, L	119	300
UL1/4	ACGGTAATT	12391	CAATATTTGATCGTGAGG	13456	1,160	E, L	115	1,160
UL1/5	ACGGTAATT	12391	GCGCTAACATGTTTCTAGG	14022	1,700 and 700	E, L	99	1,730
UL1/8	ACGGTAATT	12391	TGATGTGCTTTTTATCAGG	16212	0/0	E/L	238	4,040
UL4/5	CTTGTTACA or GCTGTAGCT	13899 or 13957	GCGCTAACATGTTTCTAGG	14022	260	L	137	571 or 260
UL4/8	GAGGTTCTT	13659	TGATGTGCTTTTTATCAGG	16212	2,860	L	415	2,860
UL6/8	CTGGTTTGT	15452	TGATGTGCTTTTTATCAGG	16212	0/1,100	E/L	240	1,100
UL7/8	TTGGTAGGT	16110	TGATGTGCTTTTTATCAGG	16212	380	E/L	277	380
UL33	CGGTCAGT	44151	AAAGCGGTACGTGGAAAGG, CACGTCACAGAAACAAAGG	44338, 43901	200, 620, 620	IE, E, L	437 or 537 or 99	623
US27	CAGGTAAGC	217755	GTGTAATGCTTTTTACAGG	217871	800, 200, and 16	IE, E, L	143	199
UL120/120	GTGGTAAGT	168605	CGCCCTTTTTGGCCTCAGG	168444	320	L	113	320
UL120/118	GTGGTAAGT	168605	TATGAATTTTATCCACAGG	167474	0	IE, L	219	1,350
UL123D1/118	GACGTAAGT	173610 (EX1)	TATGAATTTTATCCACAGG	167474	0	IE, L	265	6,401
UL123D2/118	ACGGTACGT	172695 (EX2)	TATGAATTTTATCCACAGG	167474	0	IE, L	230	5,451
UL123D3/118	TCGGTAAGT	172396 (EX3)	TATGAATTTTATCCACAGG	167474	0	IE, L	323	5,150
UL123D4/118	CAGTAAACT	171009 (EX4)	TATGAATTTTATCCACAGG	167474	0	IE, L	285	3,820
UL130/129	CAGTAAACT	175734	AGCCCGTCGCGCGCGCAGG	175644	310, 0	IE, L, E	287	310

[a] Position of the underlined bases given (the last base of possible exon 1 and the first base of exon 2).
[b] From PCR of cDNA.
[c] E, early; L, late.
[d] Predicted from the genomic sequence.
[e] Predicted from PCR of genomic DNA.

icance of the new sequences found remains uncertain at this time, as no homologies were found to known genes on screening the ORFs defined against the SwissProt, the Protein Identification Resource, or a herpesvirus data base with the FastA program for global alignments of protein sequences (32). It is uncertain whether the sequenced cDNA splices shown in Table 2 and Fig. 2 and 3 are from mRNAs (and hence encode protein products) or hnRNAs (that are not transported into the cytoplasm and are not expressed as proteins), as RT-PCR was used to generate several of them [UL122(IE2)-UL118, UL119-UL118, UL89, and US3]. Also, given that the splicing of a particular transcript and perhaps the intron-exon boundaries may vary between strains and under different conditions of growth (36, 45), the putative splice sites shown in Table 3 may yet be found to occur in other strains and under growth conditions different from those used here. The single round of PCR used here detected approximately 100 to 1,000 molecules of target cDNA. The variation in sensitivity was most likely a result of the different sequence characteristics of the different primer sets used. In all cases, the PCR was repeated at least once, and all PCR products of the appropriate size seen on agarose gel electrophoresis were sequenced.

(i) Splicing between UL122(IE2) and the downstream ORF UL118. The MIE region of HCMV is known to undergo complex differential splicing, as summarized in Fig. 2. Spliced transcripts within the ORFs UL123 (IE1 or MIE region 1) and UL122 (IE2, IE exon 5, or MIE region 2) and downstream ORFs (UL119-UL115) have been studied in detail (22, 31, 51, 53, 56). We specifically aimed to determine whether the UL122(IE2) ORF was spliced to ORFs within the downstream IE3 region, particularly to those ORFs known to undergo differential splicing (UL119-UL115).

We mapped the donor and acceptor splice sites of a transcript from IE times that splices together the UL122(IE2) and UL118 ORFs (Fig. 2) using PCR of first-strand cDNA (referred to as RT-PCR of splices in Materials and Methods). The donor [UL122(IE2)] and acceptor

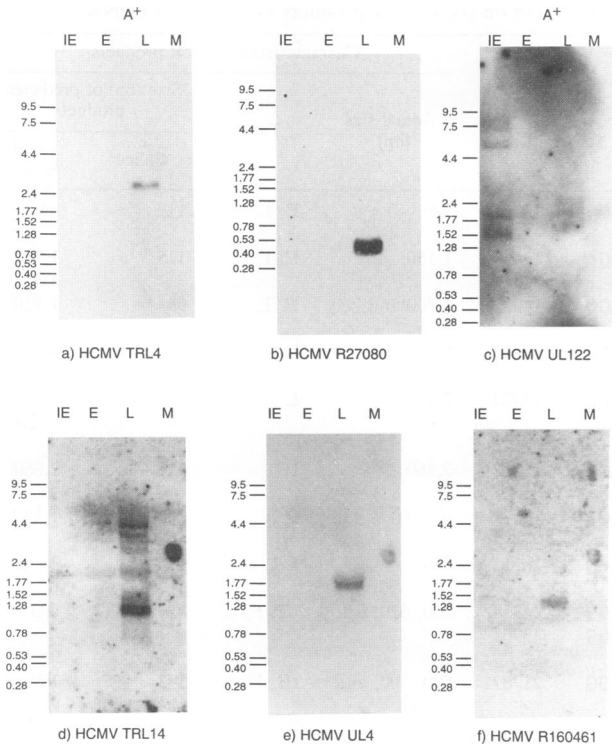

FIG. 3. Northern blots of RNA prepared at IE, early (E), and late (L) times postinfection or from mock-infected (M) fibroblasts were probed with DIG-labelled oligonucleotides as described in Materials and Methods. Molecular size standards (in kilobase pairs) are at the left of each panel. A+ indicates that poly(A)-selected RNA (2 μg per lane) was added to each lane of blots a and c. Total cell RNA (20 μg per lane) was added to each lane of blots in panels b, d, e, and f. The probes used were located at positions 2860 to 2880 (a), 27430 to 27450 (b), 170770 to 170790 (c), 11390 to 11410 (d), 14060 to 14080 (e), and 160420 to 160440 (f) on the prototype genome (8).

(UL118) sequences within these ORFs (shown in Table 2) closely match the currently accepted consensus viral splice sequence (Table 1). The spliced transcript was not found in RNAs from early or late times, although this result should be interpreted with caution, particularly as the UL122(IE2)-UL118 spliced transcript is probably in low abundance (51). The qualitative PCR used here may not delineate all transcripts; the expression of mRNAs from the viral genome may vary with different conditions of viral growth, different RNA preparations studied, and differing infections of cells. The PCR product containing the UL122(IE2)-UL118 splice was constructed from reverse-transcribed, poly(A)-selected cytoplasmic RNA. However, PCR is so sensitive (we routinely detect 50 to 400 molecules with a single round of amplification) that the spliced transcript could be derived from the nucleus or the cytoplasm. It is therefore uncertain whether the transcript represents mRNA or hnRNA and whether it is a single transcript or one of a family of RNAs.

Northern blots probed with an oligonucleotide complementary to the 5′ end of UL122(IE2) (# in Fig. 2) confirm the presence of the previously identified RNAs of 2.2 and 1.7 kb (51) as well as larger IE RNAs 7.5, 7.0, and 5.0 kb in length (Fig. 3c). RNAs of 2.2 and 1.7 kb were present at late times but at a much lower abundance (Fig. 3c).

The possibility that other transcripts containing splices

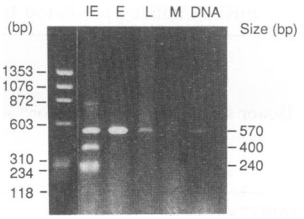

FIG. 4. PCR products generated from the US3 region with poly(T)-primed cDNAs. Primers 5′ complementary and 3′ contiguous to the genomic sequence were 21-mers, located outside the splice sites but within the longest cDNA. Lanes: 1, ϕX174 *Hae*III-digested molecular size standards; 2 to 4, cDNA from IE, early (E) and late (L) times postinfection; 5, material from mock-infected cells; 6, 1 pg of HCMV genomic DNA, all amplified by PCR.

between UL123(IE1) or UL122(IE2) and other ORFs downstream exist was further studied. Earlier experiments using primers located within exon 3 and exon 4 of the MIE region confirmed the precise location of the intron-exon boundaries previously determined by S1 nuclease mapping (53). Separate PCRs were performed with 5′ primers located within the four exons of UL123(IE1) and 3′ primers located within UL118. No evidence of splicing between UL123(IE1) and UL118 was found (Table 3). Furthermore, no splices between UL120 [located downstream of UL122(IE2)] and UL118 were found (Table 3). The UL120 ORF contains donor sequences of moderate consensus at positions 168650, 168536, and 168605 and acceptor sequences of moderate consensus at positions 168471, 168444, and 168471 (Table 3). There was no evidence of transcripts with splices between UL120 and UL118 found with these donor sites (Table 3).

The two adjacent ORFs located downstream of the MIE region (UL119 and UL118) have sequence features of incomplete glycoprotein genes. They were found to be spliced at the predicted consensus splice sequences (Table 2), as has been shown by nuclease mapping techniques (31). The ORF UL119 has features of a glycoprotein signal and UL118 has those of a glycoprotein anchor sequence, and splicing produces an ORF predicted to encode a complete glycoprotein (8).

(ii) Spliced transcripts from the US3 ORF. The US3 ORF extends from positions 194133 to 194924 on the complementary strand of HCMV strand AD169. The RNA start is at position 194799, and the consensus poly(A) signal is at positions 193918 to 193923. We aimed to establish the accurate position of the intron-exon boundaries of US3 by RT-PCR of splices. Two different sets of primers were used (which amplified the sequences between positions 194100 and 194669 or 194000 and 194639), and identical results were obtained with both sets of primers (Fig. 4). The acceptor splice sites predicted from the genomic sequence were shown to lie at positions 194454 and 194126 (Table 2). No other acceptor sites were found to be used in mRNA splicing in all cDNAs sequenced. The donor splice sites used were at position 194294 (as previously predicted) and at position 194607, the latter shown to lie 3 bases downstream of the predicted site (58). No other donor sites were used in the splicing reactions. Products from a typical RT-PCR of the US3 region are shown in Fig. 4. Using this technique, we determined that the donor splice site at position 194295 and the acceptor site at position 194126 (Table 2) were used to encode a transcript (Fig. 1), a determination which had not been possible by S1 nuclease techniques (58).

(iii) RNAs encoded by the TRL4 ORF (also known as the

major early transcript). No protein product of the most abundantly transcribed early gene of HCMV (the major early RNA) has been found to date in cells infected with HCMV. Large numbers of full-length cDNAs encoded by the TRL4 ORF were sequenced from the late-cDNA library, consistent with the high transcriptional activity of this region at late as well as early times. Indeed, the full-length TRL4 cDNA represented the largest number of late-cDNA clones sequenced (data not shown). A spliced cDNA was sequenced from the late-cDNA library that was 3′ coterminal with the TRL4 ORF (TRL4s in Table 2 and Fig. 1). The spliced transcript was present in 2 of the 100 cDNA clones sequenced. The spliced region was in the 3′-noncoding part of the RNA (positions 3323 to 3353 of strain AD169), and splicing removed a short (68-bp) intron (Table 2). The donor site matched the consensus sequence (Table 1) in 6 of 9 positions. The acceptor sequence was unusual in that the nucleotide at the −2 position was a cytosine rather than the consensus guanine (Table 2). The polypyrimidine tract upstream of the acceptor splice site consisted mainly of cytosine residues (in 5 of 9 positions) where the consensus sequence would normally be thymidine residues (Table 2). The data for the TRL4 cDNAs sequenced from the late-cDNA library showed that the ORF encodes a minimum of two 3′-coterminal RNAs (the full-length TRL4 and the spliced TRL4s). Northern blots with seven different oligonucleotide probes located along the TRL4 ORF showed one predominant band of 2.7 kb (the full-length transcript) present most abundantly at late times (Fig. 3a).

(iv) Transcripts from the RL11 family. Several ORFs studied for mRNA splicing were predicted from the genomic sequence to belong to a gene family (RL11) encoded by ORFs near the UL-TRL junction. A single cDNA sequenced from the late cDNA library encoded both the TRL13 and TRL14 ORFs in a 3′-coterminal manner. No spliced cDNA was found either in the cDNA library or by RT-PCR of IE, early, and late first-strand cDNAs (Table 3). The late cDNA encoding TRL13 and TRL14 started at position 9574 and had a consensus AATAAA sequence at positions 11714 to 11719 and a cleavage site [for poly(A) addition] at position 11731. Northern blots probed with a labelled oligonucleotide within TRL14 showed at least five bands present at late times (Fig. 3d). The late 1-kb transcript was most abundant, and a transcript of 2.3 kb (the same size as the sequenced cDNA) was one of five transcripts found at late times. The RNAs previously detected with probes complementary to the TRL-UL junction are visible in Fig. 3d, and the 4.5-kb transcript shown in Fig. 3d was also visible in the same published Northern blots (6). Whether the start sites of the 3.4- and 4.5-kb RNAs are within the TRL13 and TRL14 (TRL13/14) ORFs is uncertain. They may all terminate at the same 3′ end, as the oligonucleotide probe used in this Northern blot was located toward the 3′ end of TRL14 (at position 11390) and there is only one poly(A) consensus site (which is known from sequencing the complete cDNA) that could be used by TRL13 and TRL14 within the next 3 kb downstream.

None of the ORFs UL1, UL4, UL6, and UL7 was found to be spliced to the first consensus splice acceptor downstream within UL8 under the conditions used for this study (Table 3). The UL4 ORF was not found to be spliced to UL5. A cDNA contiguous with UL4 was sequenced and found to have a consensus poly(A) signal (AATAAA, positions 14747 to 14752) with the poly(A) tail added at position 14766. The genomic sequence contains a downstream consensus G/T cluster (between positions 14784 and 14798). As this is the first consensus poly(A) signal downstream of both the UL4 and UL5 ORF stop codons, this transcript represents the previously described UL4 ORF (6) and further shows that UL4 and UL5 are 3′ coterminal. A Northern blot probed with a labelled oligonucleotide from the 5′ end of the UL4 and UL5 ORFs showed a single band of 1.7 kb, the same size (Fig. 3e) as UL4 (6). There was no evidence on the Northern blot of a separate transcript encoded by UL5.

(v) The new R160461 spliced transcript. A spliced transcript (R160461 in Table 2 and Fig. 1) containing a large intron (of 4,528 nt) was sequenced from three cDNAs found in the late-cDNA library. R160461 was encoded within the middle of the UL segment from the complementary strand (Fig. 1). The size of this transcript found on Northern blotting (1.1 kb) corresponded to that found from sequencing the cDNA (1.14 kb in Table 2). Northern blots probed with an oligonucleotide complementary to exon 1 showed a single RNA species present only at late times (Fig. 3f). Upstream of the start of the late cDNA is the sequence TATTTATA, which has homology to the TATA sequence of eukaryotic promoters that begins at position 160493. There are no CCAAT promoter sequences present in this region on the strand encoding R160461 (although the sequence CATCAT is present downstream of the TATA site starting at position 160485). The protein predicted to be encoded by sequence across the splice junction is 31 amino acids long, has a nonconsensus ATG, and has a glycine residue that is conserved after splicing of the mRNA. The spliced 1.1-kb mRNA of R160461 has the following amino acid sequence: MPSQSAAQLPVR*G*TTRRLSASRGDTADRGNG. The locations of the transcript on the genome is shown in Fig. 1, and the DNA sequence of the splice site is shown in Table 2. The amino acid residue encoded across the splice junction is shown in boldface type with adjacent asterisks. As the entire cDNA was sequenced, no other splicing within this mRNA occurs. Examination of the genomic sequence showed that longer ORFs that cross the splice do not contain a methionine residue.

(vi) The new R27080s spliced transcript. Two cDNA species, one spliced (R27080s in Table 2) and the other the contiguous full-length product (R27080), were sequenced from the positive strand of the prototype genome (Fig. 1). They differed in size by an 83-nt intron (495 nt unspliced and 412 nt spliced). These transcripts were encoded from within a 3.7-kb region between ORFs UL20 and UL25 that has not previously been predicted to be coding (8). A wide, apparently single band was seen at late times (Fig. 3b) on probing Northern blots with a labelled oligonucleotide complementary to sequences common to both spliced and unspliced transcripts. The small size of the intron makes it likely that differentiation between the spliced and unspliced products on Northern blotting was not feasible, and they may both be represented in the wide band of 400 to 500 bp shown in Fig. 3b. The mRNAs identified from Northern blots (of 400 to 500 bp) and by sequencing the late cDNAs (412 and 495 bp) were similar in size. The spliced R27080s ORF has a potential promoter (TATTTAA) at positions 27050 to 27056 and a Kozak consensus methionine (AgcATGG) at positions 27105 to 27111. The predicted amino acid sequence of R27080s has a hydrophobic amino terminus, with leucine residues preceded by basic amino acids (RR). The spliced 0.4-kb mRNA of R27080s has the following amino acid sequence: MARRL WILSLLAVTLTVALAAPSQKSKR*S*VTVEQPSTSA DGSNTTPSKNVTLSQGGSTTDGDEDYSGEYDVLIT DGDGSEHQQPQKTDEHKENQAKENEKKIQ. The locations of the transcript on the genome is shown in Fig. 1, and

the DNA sequences of the splice site is shown in Table 2. Potential N-linked glycosylation sites (NXT/S) are underlined. The amino acid residue encoded across the splice junction is shown in boldface type with adjacent asterisks. This arrangement does not have interleucine spacing consistent with that of a leucine zipper, nor does it have the typical cluster-spacer-cluster arrangement. In the spliced R27080s transcript, the stop signal at positions 27195 to 27198 is replaced by a serine residue (Table 2), and the subsequent sequence contains two potential glycosylation sites of the form NXT/S (where X represents any amino acid). In comparison, the stop signal is still present in the unspliced R27080 transcript, and the amino acid sequence of the predicted protein, MARRLWILSLLAVTLTVALAAPSQK SKRR, is considerably shortened (29 compared with 105 amino acids). No homology was found between the putative gene products of R27080s or R27080 and any proteins in the EMBL, PIR, and SwissPROT data bases. However, given that the virus encodes this mRNA and that splicing of the R27080s mRNA has removed a stop codon, further study of protein expression by these two transcripts is likely to be of interest.

(vii) Other potentially spliced transcripts of HCMV. Several other homologs of herpesvirus genes were studied by RT-PCR of splices. The intron-exon boundaries for the HCMV homolog (UL89) of the major spliced transcript of herpesviruses (12) were determined by the techniques described above for splicing studies and found to match those predicted from the genomic sequence. The donor and acceptor sequences were found to be separated by a 3,902-bp intron (Table 2).

Two of the three G protein-coupled receptor homologs (US27 and UL33) predicted to be encoded by HCMV (7) have moderate consensus splice donor and acceptor sequences. The US27 sequences are located at the 5′ end of the ORF, and the UL33 sequences are located toward the 3′ end of the ORF. Under the conditions of study used here, no evidence of internal splicing of the US27 or UL33 ORF was found (Table 3), in agreement with previous data from Northern blots (57). Given the sensitivity of the RT-PCR procedures used here, it is unlikely that these potential splice sites are used under ordinary conditions of viral growth. The DNA sequence of the third G protein-coupled receptor homolog (US28) does not have any consensus splice sequences.

DISCUSSION

In attempting to assess whether mRNA splicing or the use of different transcriptional start and stop sites produces multiple transcripts from a given genomic sequence, nuclease protection assays have not always been able to distinguish between the two alternatives in HCMV (26, 58). It has previously been possible to distinguish discontinuities in the mRNAs by nuclease protection, but the source of this discontinuity has not always been evident, particularly when there are multiple spliced mRNAs produced from differential splicing within an ORF (58). Because of the complexity of splicing within the MIE region, it has not been possible so far to identify the features of all of the differentially spliced transcripts, particularly the larger 9.5-, 7.5-, and 4.4-kb mRNAs (49, 52). The use of PCR primers designed to amplify possible splice sites in a region of uncertainty (RT-PCR of splices) has allowed us to define precisely the intron-exon boundaries of some known spliced mRNAs [UL89, UL123(IE1)-UL122(IE2), and US3] as shown in Fig. 1 and 2, to characterize new spliced transcripts from the UL122(IE2)-UL118 ORFs within the MIE region (Fig. 2), and to dismiss other mRNA splices (Table 3) that have been uncertain (57). The negative RT-PCR results (shown in Table 3 and Fig. 5) should be interpreted with caution, as detection of HCMV transcripts may differ with conditions of infection, viral strains studied, mRNA transcript abundance, and the experimental conditions used to produce the RNA (23, 55).

All of the cDNAs sequenced from the late-cDNA library (R27080s, R160461, and TRL4s), some of those produced by PCR of first-strand cDNAs by using RT-PCR of splices (UL118, UL119, and UL89), and those listed in Table 3 were produced by RT of unblocked, late cytoplasmic RNA. In all of these experiments, no spliced mRNA was found by RT-PCR of IE, early, or mock RNA. Therefore, since these splices were identified with RNA from untreated cells, they were not an artifact of the cell culture conditions (55). The US3 and UL122(IE2)-UL118 splices were the only splices sequenced solely from RT-PCR of cDNA made from cytoplasmic RNA produced at IE times.

Whether the transcripts identified in Fig. 1 and 2 represent mRNA or hnRNA is uncertain. The small size of the intron present in TRL4s and R27080s (Table 2) has not allowed electrophoretic separation (and hence identification) of the full-length RNAs from the spliced RNAs on Northern blots (Fig. 3). RT-PCR of splice sites is so sensitive (in our hands detecting 100 to 1,000 molecules of DNA, depending upon the primer set used) that RT-PCR study of the splices present in TRL4s and R27080s does not clarify this problem, as it would potentially detect mRNA and hnRNA.

The MIE region has previously been divided into three coding regions: IE1, IE2, and IE3 (54). These three subdivisions correspond to the ORFs UL123(IE1) and UL122(IE2) and to downstream undefined ORFs (IE3). ORFs present within the broadly defined region 3 (0.709 to 0.728 map units of strain Towne, corresponding to positions 162612 to 166970 of strain AD169) have previously been shown to undergo differential splicing (31). Furthermore, it has previously been noted that UL122(IE2) has a consensus splice donor sequence toward the 3′ end of the ORF, and the results of nuclease protection assays suggest an mRNA splice between UL122(IE2) and sequences downstream (22, 53) which then terminate at a poly(A) site approximately 1.6 to 2 kb downstream (49, 52). Our finding that UL122 (IE2 or MIE exon 5) encodes an mRNA spliced via a donor site at position 169368 to the splice acceptor at the start of UL118 at position 167474 (Table 2 and Fig. 2) suggests that at least one (and possibly more) of the predicted low-abundance UL122(IE2) transcripts (51, 53) arise from the UL122(IE2)-UL118 ORFs. Consistent with this hypothesis, TATA sequences are present at positions 170948 (TATATTATA) and 170998 (TATATATATAT) upstream of the transcription initiation site of UL122(IE2) (at position 170916) (48). The predicted size of a UL122(IE2)-UL118-spliced mRNA would be a minimum of 2.4 kb and a maximum (from presently available data) of 4.5 kb (Fig. 2). It is uncertain which mRNAs represent this spliced product, because there are a number of large, undefined mRNAs found on Northern blots studied with probes complementary to the IE2 and IE3 regions (Fig. 3c) (31, 51). Northern blots analyzed with probes complementary to the UL122(IE2) region identify large (9.5- and 4.4-kb) undefined mRNAs that originate from within region IE2 or IE3 (51). Another 4.2-kb mRNA found arising within the UL119-UL115 ORFs (Fig. 2 and Table 2) is found at IE times, disappears at early times, and is found again at late times (31). Using a probe located near the 5′ end

of UL122(IE2) at positions 170770 to 170790, we have demonstrated the known IE mRNAs of 2.2 and 1.7 kb (Fig. 3c) encoded by UL123(IE1) and UL122(IE2) exons (51, 53). We have also shown large mRNAs of around 7.5, 7.0, and 5.0 kb, one of which may represent the UL122(IE2)-UL118 spliced mRNA(s). These may correspond to those mRNAs found previously (31, 51), although in our study they were most abundant at IE rather than late times and their sizes were slightly different from those previously reported (Fig. 3). What has been shown is that the IE2 region (UL122) is linked via mRNA splicing to UL118 (which may represent part of the IE3 region) and may thus encode large transcripts that from previous evidence are known to originate from within the UL122(IE2) ORF (51). Whether this UL122(IE2)-UL118 spliced transcript is involved in some way in the control of later processes of infection by HCMV is worthy of further investigation. In this context, it is notable that UL115 is spliced via UL116 to UL118 (Fig. 2) (31) and that the UL115 ORF encodes a functional homolog of gL which forms a stable complex with gH (UL75). This gH-gL aggregate makes up the glycoprotein complex gcIII which is expressed on the cell surface (27). Some products of the UL122(IE2) region are already known to be involved in transactivation and autoregulation of the MIE genes (39, 49, 51). Additional predicted transcripts from within the MIE region which have so far not been proven to exist are not shown in Fig. 2. These RNAs were predicted to be a 1.4-kb RNA [present at IE times, encoded by UL123(IE1) exon 1 plus exon 2 plus exon 3 plus UL122(IE2) and containing a splice within the UL122(IE2) region] (53), a 2.25-kb RNA [present at IE times, coded by UL122(IE2) plus downstream exons containing a splice within the UL122(IE2) region] (22), and a 1.7-kb RNA [present at late times, coded by the unspliced UL122(IE2) ORF] (53).

We have sequenced two transcripts from the TRL4 ORF, one of which was the known unspliced ORF encoding a 2.7-kb RNA, previously named the major early transcript (20, 34), and the other of which was an uncharacterized spliced mRNA. The unspliced cDNAs from this region were the most abundant within the library (data not shown) and represent a significant amount of the sequencing effort. In attempting to sequence new transcripts from cDNA libraries, it will be important in the future to first screen the library by using a probe for TRL4 in order to avoid sequencing large numbers of identical noninformative recombinant clones. To date, no protein product has been found to be expressed by the TRL4 ORF, and we found no homology of the predicted product of either the full-length or spliced mRNA to protein sequences in the data base. This was most likely because the splice is within the region of the ORF predicted to be noncoding. In addition to the 2.7-kb RNA (encoded by TRL4), smaller RNAs of 1.2 and 1.3 kb (23) and 1.3 kb (20) have been found on Northern blots analyzed with probes complementary to the TRL region of the genome. There was no indication of whether the smaller RNAs resulted from 3′-coterminal overlapping mRNAs, and we found no evidence that they resulted from pre-mRNA splicing (Table 3). Another 1.2-kb transcript with a different 3′ terminus found in the same study (34) is now known to be encoded by the TRL7 ORF (26). Probes overlapping the majority of the TRL region also detect several other RNAs of 4.4, 3.6, 3.3, and 1.8 kb (34). The 3.6-kb RNA reported previously was detected only at late times by a probe that detected the major 2.7-kb (TRL4) and 1.8-kb RNAs. This may represent the larger (approximately 4-kb) transcript present as a faint band on some Northern blots probed with oligonucleotides from this region (data not shown). If this were a minor transcript and had an S1 nuclease-sensitive site, then it may not have been detected in other studies. The TRL4 2.7-kb transcript may have an S1 nuclease site (34). If the 4-kb mRNA were the spliced transcript (TRL4s in Table 2), this would explain why splicing was not detected by conventional techniques if the S1 nuclease site of TRL4 was cleaved during mapping studies. To determine whether further minor transcripts were encoded by the TRL4 ORF, Northern blots probed with seven different labelled oligonucleotides located along the TRL4 ORF were performed (data not shown). These all showed one band of 2.6 kb (the full-length transcript) and a much fainter one of 4.0 kb (Fig. 3a). Given that the majority of the TRL region is nonessential for growth in cell culture (41), the number of RNAs detected from this region suggest that further study of the function of transcripts encoded by the TRL segment is warranted.

The ORFs within the UL segment of the genome immediately adjacent to the TRL region have been previously predicted to encode genes belonging to the RL11 gene family (8). This family consists of 14 ORFs, including TRL14, UL1, UL4, UL5, UL6, UL7, and UL8. The UL4 gene encodes an early structural glycoprotein (gp48) (6). Transcription from this region consists of the known UL4(gp48) transcripts of 1.5, 1.35, and 1.85 kb (6). We detected the 1.5-kb RNA using an oligonucleotide probe located within the 3′-noncoding region of the sequenced UL4(gp48)-UL5 late cDNA. No transcripts are known to be encoded from the UL1, UL5, UL6, UL7, or UL8 ORF (41), and at least two transcripts of 2.7 and 3.4 kb are encoded from the region to the left of UL4. These last uncharacterized transcripts were detected previously with a long probe made from plasmid containing part of the TRL region (6). In fact, there are also two additional, larger transcripts visible in the published Northern blot which are not detected with probes further away from the TRL region. The region around the TRL-UL junction is therefore also transcriptionally active, consistent with data from mapping studies of the entire genome (16).

The sequence characteristics of the mRNA splices in HCMV were assessed. In lower eukaryotes, sequence analysis of the donor splice site shows that the A residue at the +3 position is highly conserved in greater than 90% of sequences (15). This conservation is not so markedly a feature of the donor splice sites of HCMV (55% in the current study of 20 different donor sites have an A residue in the +3 position) nor of viral sequences generally (70% of viral sequences in reference 44). Also, the introns of most nonvertebrate eukaryotes have a significantly higher mean A+T content (85%) than do the neighboring exons (64%), especially in the 30 nt preceding the 5′ splice site and particularly for small introns (44). This has functional significance for at least one group of organisms (15). The A+T content was calculated for HCMV introns known not to contain exons on the same strand, described in Tables 1 and 2 (present in the ORFs UL123 intron 2 and intron 3, UL119-UL118, TRL4s, R27080s, US3, UL36, and UL37). The A+T content of these introns (53%) was found to be significantly higher ($P < 0.001$) than that in the adjacent exons (42%) for which sequence data were available, although this difference was much less marked than has been noted for eukaryotes. The polypyrimidine tract present within the intron upstream of the 3′ splice site is essential for normal spliceosome assembly in eukaryotes (45). The acceptor sequence of the HCMV splice within the TRL4s ORF was unusual in this regard, as the nucleotide at the −2 position was a cytosine (C) rather than the consensus

adenine (A) (Table 2). This has not been noted in published virus splice acceptor sequences, although it has been described for nonconforming acceptor sequences in eukaryotes in which the terminal AG dinucleotide of the intron is replaced by the trinucleotide CAC (24, 44). The polypyrimidine tract upstream of the splice site within the TRL4 ORF consisted mainly of C residues (in 5 of 11 positions) rather than the consensus thymidine (T) residues (Table 2). This sequence contains 2 of 14 T residues, while the consensus contains 11 T residues of 14 total residues. This may reflect the overall high G+C content of HCMV. It is the presence of either pyrimidine residue in these positions near an acceptable branch point that is the most important feature in the formation of the spliceosome complex (45), and hence substitution of C for T in these positions would be logical in a G+C-rich viral genome. The increased C content of the splice acceptor site has implications for future predictions of further splice sites in HCMV and other G+C-rich genomes.

A number of herpesvirus genes are known to encode 3′-coterminal transcripts, including those of herpes simplex virus type 1 (33, 35, 59), Epstein-Barr virus (21), as well as HCMV (6, 30, 51, 60). By comparison, Epstein-Barr virus encodes several Epstein-Barr virus nuclear antigen proteins expressed during latency that are derived from a long primary transcript by means of alternate splicing and the use of alternative poly(A) sites. These transcripts have very large introns with relatively small exons spliced together from a large part of the genome (28). Whether a similar situation exists for HCMV is unknown, although we have demonstrated two transcripts with introns of 3.9 and 4.5 kb in length (Table 2).

Spliced transcripts are likely to be translated, as by comparison the majority of spliced genes in eukaryotes encode proteins (44). Given the large number of ORFs that HCMV has been predicted to encode and the large size of the genome, there is likely to be further redefinition of the viral ORFs, particularly by the analysis of spliced transcripts.

ACKNOWLEDGMENTS

We thank Jon Oram for supplying the cDNA library used in this study and Tony Minson and John Sinclair for critical reading of the manuscript.

William D. Rawlinson was supported by scholarships from the Royal Australasian College of Physicians and the Sir Robert Menzies Memorial Trust.

REFERENCES

1. **Alford, C. A., and W. J. Britt.** 1990. Cytomegalovirus, p. 1981–2010. *In* B. N. Fields et al. (ed.), Virology. Raven Press, New York.
2. **Bahn, A. K.** 1972. Basic medical statistics, p. 154–156. Grune & Stratton, New York.
3. **Bankier, A. T., S. Beck, R. Bohni, C. M. Brown, R. Cerny, M. S. Chee, C. A. Hutchison III, T. Kouzarides, J. A. Martignetti, E. Preddie, S. C. Satchwell, P. Tomlinson, K. M. Weston, and B. G. Barrell.** 1991. The DNA sequence of the human cytomegalovirus genome. DNA Sequence: J. DNA Sequencing and Mapping **2:**1–12.
4. **Bankier, A. T., K. M. Weston, and B. G. Barrell.** 1987. Random cloning and sequencing by the M13/dideoxynucleotide chain termination method. Methods Enzymol. **155:**51–93.
5. **Bell, J.** 1989. The polymerase chain reaction. Immunol. Today **10:**351–355.
6. **Chang, C. P., D. H. Vesole, J. Nelson, M. B. A. Oldstone, and M. F. Stinski.** 1989. Identification and expression of a human cytomegalovirus early glycoprotein. J. Virol. **63:**3330–3337.
7. **Chee, M., S. Satchwell, E. Preddie, K. Weston, and B. G. Barrell.** 1990. Human cytomegalovirus encodes three G-protein coupled receptor homologues. Nature (London) **344:**774–777.
8. **Chee, M. S., A. T. Bankier, S. Beck, R. Bohni, C. M. Brown, R. Cerny, T. Horsnell, C. A. Hutchison III, T. Kouzarides, J. A. Martignetti, S. C. Satchwell, P. Tomlinson, K. M. Weston, and B. G. Barrell.** 1990. Analysis of the protein coding content of the sequence of human cytomegalovirus strain AD169. Curr. Top. Microbiol. Immunol. **154:**125–169.
9. **Chee, M. S., S.-A. Rudolph, B. Plachter, B. G. Barrell, and G. Jahn.** 1989. Identification of the major capsid protein gene of human cytomegalovirus. J. Virol. **63:**1345–1353.
10. **Chomczynski, P., and N. Sacchi.** 1987. Single step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal. Biochem. **162:**156–159.
11. **Church, G. M., and S. Kieffer-Higgins.** 1988. Multiplex DNA sequencing. Science **240:**185–188.
12. **Costa, R. H., K. G. Draper, T. J. Kelly, and E. K. Wagner.** 1985. An unusual spliced herpes simplex virus type 1 transcript with sequence homology to Epstein-Barr virus DNA. J. Virol. **54:**317–328.
13. **Cranage, M. P., T. Kouzarides, A. T. Bankier, S. C. Satchwell, K. W. Weston, P. Tomlinson, B. G. Barrell, H. Hart, A. C. Minson, and G. L. Smith.** 1986. Identification of the human cytomegalovirus glycoprotein B gene and induction of neutralizing antibodies via its expression in recombinant vaccinia virus. EMBO J. **5:**3057–3063.
14. **Craxton, M.** 1991. Linear amplification sequencing, a powerful method for sequencing DNA. Methods **3:**20–26.
15. **Csank, C., F. M. Taylor, and D. W. Martindale.** 1990. Nuclear pre-mRNA introns: analysis and comparison of intron sequences from Tetrahymena thermophila and other eukaryotes. Nucleic Acids Res. **18:**5133–5141.
16. **DeMarchi, J. M.** 1981. Human cytomegalovirus DNA: restriction enzyme cleavage maps and map locations for immediate-early, early, and late RNAs. Virology **114:**23–38.
17. **Elliott, R. M., N. E. Crook, U. Desselberger, R. Hull, and D. J. McGeoch.** 1991. Some highlights of virus research in 1990. J. Gen. Virol. **72:**1761–1779.
18. **Gibson, T. J., and J. E. Sulston.** 1987. Preparation of large numbers of plasmid DNA samples in microtiter plates by the alkaline lysis method. Gene Anal. Tech. **4:**41–44.
19. **Gorman, K. B., and R. A. Steinberg.** 1989. Simplified method for selective amplification and direct sequencing of cDNAs. BioTechniques **7:**326–328.
20. **Greenaway, P. J., and G. W. G. Wilkinson.** 1987. Nucleotide sequence of the most abundantly transcribed early gene of human cytomegalovirus strain AD169. Virus Res. **7:**17–31.
21. **Henessy, K., and E. Kieff.** 1985. A second nuclear protein is encoded by EBV in latent infection. Science **227:**1238–1240.
22. **Hermiston, T. W., C. L. Malone, P. R. Witte, and M. F. Stinski.** 1987. Identification and characterization of the human cytomegalovirus immediate-early region 2 gene that stimulates gene expression from an inducible promoter. J. Virol. **61:**3214–3221.
23. **Hutchinson, N. I., R. T. Sondermeyer, and M. J. Tocci.** 1986. Organization and expression of the major genes from the long inverted repeat of the human cytomegalovirus genome. Virology **155:**160–171.
24. **Jackson, I. J.** 1991. A reappraisal of non-consensus mRNA splice sites. Nucleic Acids Res. **19:**3795–3798.
25. **Jahn, G., T. Kouzarides, M. Mach, B.-C. Scholl, B. Plachter, B. Traupe, E. Preddie, S. C. Satchwell, B. Fleckenstein, and B. G. Barrell.** 1987. Map position and nucleotide sequence of the gene for the large structural phosphoprotein of human cytomegalovirus. J. Virol. **61:**1358–1367.
26. **Jones, T. R., and V. P. Muzithras.** 1991. Fine mapping of transcripts expressed from the US6 gene family of human cytomegalovirus strain AD169. J. Virol. **65:**2024–2035.
27. **Kaye, J. F., U. A. Gompels, and A. C. Minson.** 1992. Glycoprotein H of human cytomegalovirus (HCMV) forms a stable complex with the HCMV UL115 gene product. J. Gen. Virol. **73:**2693–2698.
28. **Kieff, E., and D. Liebowitz.** 1990. Epstein-Barr virus and its

replication, p. 1889–1920. *In* B. N. Fields et al. (ed.), Virology. Raven Press, New York.
29. **Kouzarides, T., A. T. Bankier, S. C. Satchwell, E. Preddie, and B. G. Barrell.** 1988. An immediate early gene of human cytomegalovirus encodes a potential membrane glycoprotein. Virology **165:**151–164.
30. **Leach, F. S., and E. S. Mocarski.** 1989. Regulation of cytomegalovirus late-gene expression: differential use of three start sites in the transcriptional activation of ICP36 gene expression. J. Virol. **63:**1783–1791.
31. **Leatham, M. P., P. R. Witte, and M. F. Stinski.** 1991. Alternate promoter selection within a human cytomegalovirus immediate-early and early transcription unit (UL119-115) defines true late transcripts containing open reading frames for putative viral glycoproteins. J. Virol. **65:**6144–6153.
32. **Lipman, D. J., and W. R. Pearson.** 1985. Rapid and sensitive protein similarity searches. Science **227:**1435–1441.
33. **Mach, M., T. Stamminger, and G. Jahn.** 1989. Human cytomegalovirus: recent aspects from molecular biology. J. Gen. Virol. **70:**3117–3146.
34. **McDonough, S. H., S. I. Staprans, and D. H. Spector.** 1985. Analysis of the major transcripts encoded by the long repeat of human cytomegalovirus strain AD169. J. Virol. **53:**711–718.
35. **McGeoch, D. J., M. A. Dalrymple, A. Dolan, D. McNab, L. Perry, P. Taylor, and M. D. Challberg.** 1988. Structures of herpes simplex virus type 1 genes required for replication of virus DNA. J. Virol. **62:**444–453.
36. **Mengeritsky, G., and T. F. Smith.** 1989. New analytical tool for analysis of splice site sequence determinants. Comput. Appl. Biosci. **5:**97–100.
37. **Pande, H., T. D. Lee, M. A. Churchill, and J. A. Zaia.** 1990. Structural analysis of a 64-kDa major structural protein of human cytomegalovirus (Towne): identification of a phosphorylation site and comparison to pp65 of HCMV (AD169). Virology **178:**6–14.
38. **Petti, L., J. Sample, F. Wang, and E. Kieff.** 1988. A sixth Epstein-Barr virus nuclear protein (EBNA3B) is expressed in latently infected growth-transformed lymphocytes. J. Virol. **62:**2173–2178.
39. **Pizzorno, M. P., M. A. Mullen, Y. N. Chang, and G. S. Hayward.** 1991. The functionally active IE2 immediate-early regulatory protein of human cytomegalovirus is an 80-kilodalton polypeptide that contains two distinct activator domains and a duplicated nuclear localization signal. J. Virol. **65:**3839–3852.
40. **Rawlins, D. R., G. Milman, S. D. Hayward, and G. S. Hayward.** 1985. Sequence specific DNA binding of the Epstein Barr virus nuclear antigen 1. Cell **42:**859–868.
41. **Ripalti, A., and E. S. Mocarski.** 1991. The products of human cytomegalovirus genes UL1-UL7, including gp48, are dispensable for growth, p. 63–80. *In* M. P. Landini (ed.), Progress in cytomegalovirus research. Excerpta Medica, Elsevier, Amsterdam.
42. **Saiki, R. K., D. H. Gelfland, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich.** 1988. Primer directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science **238:**487–491.
43. **Sambrook, J., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
44. **Senapathy, P., M. B. Shapiro, and N. L. Harris.** 1990. Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. Methods Enzymol. **183:**252–278.
45. **Smith, C. W., E. B. Parro, J. G. Potton, and B. Nadal-Ginard.** 1989. Scanning from an independently specified branch point defines the 3′ splice site of mammalian introns. Nature (London) **342:**243–247.
46. **Spector, D. H., K. M. Klucher, D. K. Rabert, and D. A. Wright.** 1990. Human cytomegalovirus early gene expression. Curr. Top. Microbiol. Immunol. **154:**21–45.
47. **Staden, R.** 1984. A computer program to enter DNA gel reading data into a computer. Nucleic Acids Res. **12:**499–503.
48. **Stamminger, T., and B. Fleckenstein.** 1990. Immediate-early transcription regulation of human cytomegalovirus. Curr. Top. Microbiol. Immunol. **154:**3–19.
49. **Stamminger, T., E. Puchtler, and B. Fleckenstein.** 1991. Discordant expression of the immediate-early 1 and 2 gene regions of human cytomegalovirus at early times after infection involves posttranscriptional processing events. J. Virol. **65:**2273–2282.
50. **Staprans, S. I., D. K. Rabert, and D. H. Spector.** 1988. Identification of sequence requirements and *trans*-acting functions necessary for regulated expression of a human cytomegalovirus early gene. J. Virol. **62:**3463–3473.
51. **Stenberg, R. M., A. S. Depto, J. Fortney, and J. Nelson.** 1989. Regulated expression of early and late RNAs and proteins from the human cytomegalovirus immediate-early gene region. J. Virol. **63:**2699–2708.
52. **Stenberg, R. M., D. R. Thomsen, and M. F. Stinski.** 1984. Structural analysis of the major immediate-early gene of human cytomegalovirus. J. Virol. **49:**190–191.
53. **Stenberg, R. M., P. R. Witte, and M. F. Stinski.** 1985. Multiple spliced and unspliced transcripts from human cytomegalovirus immediate-early region 2 and evidence for a common initiation site within immediate-early region 1. J. Virol. **56:**665–675.
54. **Stinski, M. F., D. R. Thomsen, R. M. Stenberg, and L. C. Goldstein.** 1983. Organization and expression of the immediate-early genes of human cytomegalovirus. J. Virol. **46:**1–14.
55. **Tenney, D. J., and A. M. Colberg-Poley.** 1991. Human cytomegalovirus UL 36-38 and US3 immediate-early genes: temporally regulated expression of nuclear, cytoplasmic, and polysome-associated transcripts during infection. J. Virol. **65:**6724–6734.
56. **Wathen, M. W., and M. F. Stinski.** 1982. Temporal patterns of human cytomegalovirus transcription: mapping the viral RNAs synthesized at immediate-early, early, and late times after infection. J. Virol. **41:**462–477.
57. **Welch, A. R., L. M. McGregor, and W. Gibson.** 1991. Cytomegalovirus homologs of cellular G protein-coupled receptor genes are transcribed. J. Virol. **65:**3915–3918.
58. **Weston, K.** 1988. An enhancer element in the short unique region of human cytomegalovirus regulates the production of a group of abundant immediate early transcripts. Virology **162:**406–416.
59. **Weston, K., and B. G. Barrell.** 1986. Sequence of the short unique region, short repeats and part of the long repeat of human cytomegalovirus. J. Mol. Biol. **192:**177–208.
60. **Wright, D. A., S. I. Staprans, and D. H. Spector.** 1988. Four phosphoproteins with common amino termini are encoded by human cytomegalovirus AD169. J. Virol. **62:**331–340.
61. **Yang, J. L., V. M. Maher, and J. J. McCormick.** 1989. Amplification and direct nucleotide sequencing of cDNA from the lysate of low numbers of diploid human cells. Gene **83:**347–354.