JOURNAL OF VIROLOGY, Sept. 1986, p. 564–573
0022-538X/86/090564-10$02.00/0

Vol. 59, No. 3

A Precise Map of Splice Junctions in the mRNAs of Minute Virus of Mice, an Autonomous Parvovirus

C. VICTOR JONGENEEL,[1]* ROLAND SAHLI,[1]† GARY K. McMASTER,[2] AND BERNHARD HIRT[1]

Swiss Institute for Experimental Cancer Research, CH-1066 Epalinges, Switzerland,[1] and Max-Planck Intitut für Immunbiologie, D-7800 Freiburg i.B., Federal Republic of Germany[2]

Received 17 March 1986/Accepted 23 May 1986

We have determined the exact splicing patterns of the mRNAs of the minute virus of mice by a combination of cDNA sequencing and S1 nuclease protection analysis. There are four virus-specific mRNA species, each coding for one of the four polypeptides identified by in vitro translation. The R1 mRNA comprises sequences from nucleotide ~200 to 2281 and from 2378 to ~4800 and codes for the NS1 protein. The R2 mRNA is derived from nucleotides ~200 to 515, 1991 to 2281, and 2378 to ~4800 and codes for the NS2 protein. Between nucleotides 1991 and 2281, the coding sequence for NS2 overlaps that of NS1, but in a different reading frame. R3 covers nucleotides ~2007 to 2281 and 2378 to ~4800 and codes for VP2. The fourth species, R3', differs from R3 by using an alternative splice donor and acceptor in the region around 47 map units (nucleotide 2400); it extends from nucleotide ~2007 to 2317 and from 2400 to ~4800 and almost certainly codes for VP1. The R2 transcript is unusual in that the intron that was removed from it (nucleotides 516 to 1990) starts with GC rather than the canonical GU. With the exception of the splice acceptor at position 2378, which is found only in rodent parvoviruses, the splice junctions are highly conserved among autonomous parvoviruses. These results show that minute virus of mice, like other small DNA viruses, uses multiple strategies to compress the coding information for several viral proteins into a short (5,104 nucleotide) genome.

Minute virus of mice (MVM) has long served as a paradigm for the genus *Parvovirus*, or autonomous parvoviruses (29). Members of this family of small, nonenveloped, single-stranded DNA viruses have been isolated from a large number of vertebrates. They depend on active division of their host cells for their own multiplication but do not require coinfection with other viruses to replicate.

MVM has a single-stranded DNA genome approximately 5,100 base pairs long, which is converted to a double-stranded replicative form after penetration of the host cell and uncoating. There are two laboratory strains of MVM, which show a strong specificity for fibroblasts (MVMp) or lymphocytes (MVMi) as their host cells (16, 31). The genomes of both strains have been sequenced, and their genetic organizations are identical (1, 2, 27). A closely related parvovirus, H-1, which has also been sequenced, shares the same genetic organization (25). MVMi, MVMp, and H-1 are essentially identical in the distribution of their open reading frames, the structure of the viral mRNAs, and the polypeptides they code for. The canine parvovirus (CPV [23]) and feline parvovirus (FPV [7]) have been partially sequenced and are closely related to each other. They show strong similarities to the murine parvoviruses, particularly in their transcriptional signals and in the region thought to code for nonstructural proteins (7).

Replicative-form DNA is the template for viral transcription, which is probably catalyzed by the host RNA polymerase II (8). The mRNAs, which have a unique polarity opposite to that of the encapsidated strand, are transcribed from two promoters, located at 4 (P4) and 39 (P39) map units (m.u.) on the viral genome (4, 20) (since the MVMi genome is 5,104 nucleotides [nt] long, 1 m.u. corresponds to 51 nt). Nuclear runoff assays have located these promoters more precisely, at nt 201 ± 5 and 2005 ± 5 (4). To function at its fullest efficiency, the P39 promoter requires a viral protein (NS1) whose mRNA is transcribed from the P4 promoter (24). Three species of mRNA, designated R1, R2, and R3, have been mapped on the MVM and H-1 genomes (Fig. 1, bottom) (8, 20). The R1 and R2 transcripts share the same promoter (P4) but differ in that R2 lacks a large block of intervening sequence between 10 and 39 m.u. The R3 mRNA, which is the most abundant, originates from P39. All three mRNAs are missing a small intron in the region between 46 and 48 m.u.

In vitro translation and other studies have shown that the MVM and H-1 genomes code for four primary translation products with partially overlapping structures (9, 25). They are coded in major part in two large open reading frames (ORF1 and ORF2; see Fig. 4) spanning the left and right halves of the genome (2, 9, 27). The capsid proteins, VP1 (83 kilodaltons [kDa]) and VP2 (64 kDa), originate from the right half (toward the 5' end of the encapsidated strand), and VP2 contains a subset of the peptides generated from VP1 (32). A large (83 kDa) nonstructural protein, NS1, is coded in a single contiguous block of sequence within the left half of the genome. A smaller (24 kDa) nonstructural protein of unknown genetic origin can be translated from viral mRNA (9). While it is reasonably clear that the VP2 and NS1 proteins are coded in single open reading frames uninterrupted by introns, there was no clear experimental evidence for the genetic origin of VP1 or NS2 at the outset of this study.

To obtain more precise information regarding the structure of the MVM mRNAs and, by extension, the coding of the viral proteins, we have isolated a collection of MVM-specific cDNA clones and analyzed them by restriction endonuclease digestion and by sequencing. In addition, we have refined the mapping of the splice junctions around 47 m.u. This information has allowed us to define precisely the structure of the viral mRNAs and to determine the coding schemes used for the NS2 and VP1 proteins.

* Corresponding author.
† Present address: Department of Pathology, Harvard Medical School, Boston, MA 02115.

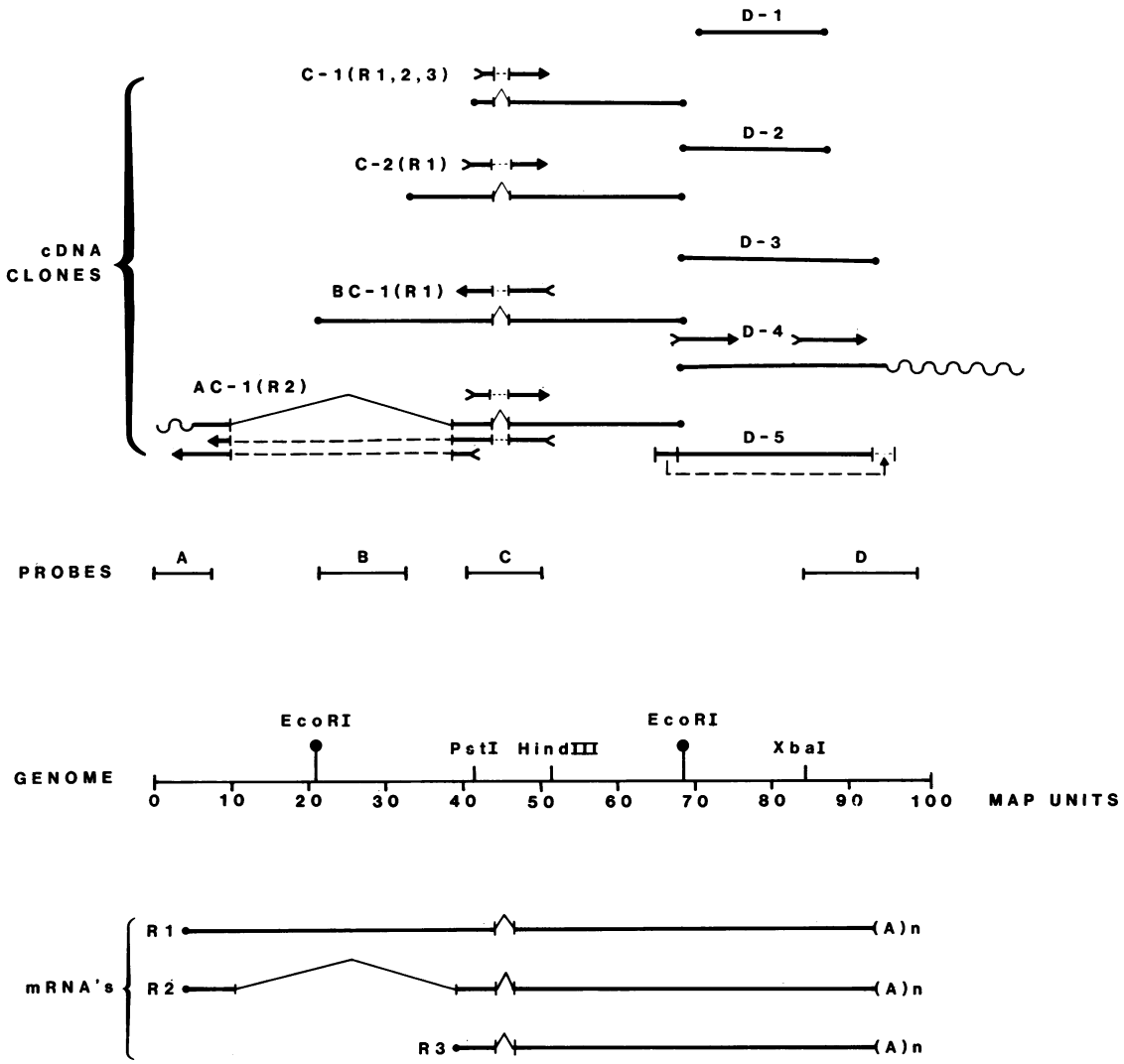

FIG. 1. Map of the cDNA clones. (Bottom) Structures of the known mRNA species (20). The representation of the MVMi genome emphasizes the two *Eco*RI sites, which mark the boundaries between potential cDNA clones, and the restriction sites used as starting points for sequencing. A, B, C, and D are the four genomic probes used in screening the cDNA library. (Top) Schematic representation of the nine cDNA clones characterized in this study. The names of the clones reflect the probes to which they hybridize; the mRNAs from which the clones were derived are indicated in parentheses (except for the clones hybridizing to probe D, which could have been derived from any of the mRNAs). Arrows mark the extent and direction of sequencing. Wavy lines represent DNA sequences of unknown origin, probably ligated to the cDNAs during the cloning procedure. In clone D-5, a short MVM sequence was ligated out of place (stippled arrow).

MATERIALS AND METHODS

Materials. Restriction enzymes were purchased from Boehringer Mannheim Biochemicals or Anglian Biotechnologies. The Klenow fragment of *Escherichia coli* DNA polymerase I, avian myeloblastosis virus reverse transcriptase, DNA polymerase I, and T4 DNA ligase were from Anglian Biotechnologies. S1 nuclease was from Sigma Chemical Co. or Boehringer Mannheim Biochemicals, cold nucleotides and dideoxynucleotides were from Pharmacia, and ^{32}P-labeled nucleotides were from Amersham Corp.

Strains. The immunosuppressive strain of MVM, MVMi, was routinely grown in EL-4 lymphoblastoid cells. The lambda gt11 cloning vector and the *E. coli* strains used for its plating and amplification are described by Huynh et al. (14). The mWB238 and mWB239 filamentous phage cloning vectors have been described previously (27) and were derived from the mWB2344 vector constructed by Barnes et al. (3).

Nomenclature and analysis of nucleotide sequences. Since the MVMi strain was used in all of the experiments described in this paper, we adhere to the nucleotide numbering of Sahli et al. (27). Because of small differences between the two strains, this numbering is not identical to the MVMp numbering used by Astell et al. (2) but agrees with the MVMi numbering of these authors (1). The differences between the sequence of MVMi used here and that of MVMp are described in detail in the article by Sahli et al. (27). For the sake of consistency with previously published reports, the locations of some features will be given in m.u. Absolute reading frames are initiated at the first three nucleotides of the sequence of MVMi. Nucleotide numbering in sequences determined by others (MVMp, H-1, CPV, FPV) follows the original reports (2, 7, 23, 25).

Computer analysis of the sequences was done on a Sirius/Victor microcomputer with the COMPSEQ package developed by A. Bairoch, University of Geneva. The pro-

gram that predicts the location of splice junctions uses the algorithm described by Staden (30). Matrix comparisons between sequences were performed by the method of Pustell and Kafatos (22).

Construction and screening of the cDNA bank. The detailed protocol used for the construction of the cDNA bank has been described elsewhere (17). As our starting material, we used oligo(dT)-cellulose-selected cytoplasmic RNA extracted from EL-4 cells 20 h after infection with MVMi. Standard techniques were used throughout, in the following sequence: (i) oligo(dT)-primed first-strand cDNA synthesis by avian myeloblastosis virus reverse transcriptase; (ii) self-primed second-strand synthesis by *E. coli* DNA polymerase I; (iii) removal of terminal hairpins by S1 nuclease, followed by repair by the Klenow fragment of DNA polymerase I; (iv) addition of *Eco*RI linkers (Pharmacia) by using phage T4 DNA ligase; (v) digestion with *Eco*RI and removal of digested linkers; (vi) ligation to *Eco*RI-digested and phosphatase-treated lambda gt11 DNA; and (vii) packaging in vitro and plating on the Y1090 *E. coli* host strain. It should be noted that we did not methylate our cDNAs with *Eco*RI methylase before digesting them with the restriction endonuclease and that the library therefore contained *Eco*RI fragments of the original cDNAs.

Screening of the library for phage vectors containing portions of the MVMi genome was done by hybridization to plate replicas on colony/plaque screen filters (New England Nuclear Corp.), used as recommended by the manufacturer. As probes, we used nick-translated DNA fragments containing known portions of the MVMi genome, purified from recombinant mWB vector replicative-form DNA. Hybridizations were at 68°C in 6× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate) for 16 h; 0.25% nonfat dry milk was used as a blocking agent (15). The filters were washed in a solution containing 1× SSC, 0.25% dry milk, 0.1% sodium dodecyl sulfate, and saturated sodium pyrophosphate (1:30) at 68°C. Two or three rounds of screening and plaque selection were used to purify recombinant phages containing the cDNA clones described below.

Subcloning and sequencing. Recombinant lambda phage DNA was extracted from 4-ml lysates grown from single positive plaques and was digested with *Eco*RI. The cDNA inserts were separated on agarose gels and purified after migration into low-temperature-gelling agarose (16). They were then subcloned into the *Eco*RI site of the mWB238 or mWB239 vectors and characterized by digestion with restriction endonucleases and by single-nucleotide sequence analyses (T tracks). Complete nucleotide sequence analyses were done on subclones of these cloned *Eco*RI fragments, by using the dideoxynucleotide chain termination method of Sanger (3, 28).

S1 nuclease mapping of intron-exon boundaries. Strand-specific probes for the mapping of S1 endonuclease-resistant hybrids (5) were prepared from recombinant mWB239 filamentous phage containing fragments of MVMi genomic DNA by the method of Burke (6). Gel-purified ^{32}P-labeled probe (80,000 cpm) was mixed with 20 μg of total cytoplasmic RNA extracted from MVMi-infected EL-4 cells, and the mixture was freeze-dried, suspended in 50 μl of hybridization buffer (0.4 M NaCl, 40 mM PIPES [piperazine-*N*,*N*′-bis(2-ethanesulfonic acid); pH 6.8], 1 mM EDTA, 80% formamide), and allowed to anneal for 3 h at 49°C. The annealing mixture was quenched with 200 μl of ice-cold S1 buffer (0.25 M NaCl, 30 mM sodium acetate [pH 4.5], 1 mM zinc acetate). S1 nuclease (2,000 U; Boehringer Mannheim) was added, and digestion was allowed to proceed for 30 min at 45°C. After ethanol precipitation, the S1-protected fragments were separated in standard sequencing gels and detected by autoradiography of the fixed and dried gels.

RESULTS

Isolation and characterization of cDNA clones. Since MVM encodes several mRNAs with extensively overlapping structures, we needed to characterize a reasonable number of cDNA clones to be able to reconstruct the exact structures of the mRNAs from which they were derived. We also wanted to be able to confirm tentative coding assignments by determining the antigenic specificities of the polypeptides coded by the individual cDNA clones. Therefore, we used the lambda gt11 vector developed by Huynh et al. (14), which allows the construction of relatively complex libraries while also directing the expression of the cloned cDNA under control of the *E. coli lac* promoter. In the present paper, we will not discuss the characteristics of the polypeptides expressed from our cloned cDNAs.

Our strategy for the isolation and characterization of cDNA clones was based on the mapping data (20; our unpublished observations) defining the three major mRNA species (R1, R2, and R3) shown in Fig. 1. The structures of all of our cDNA clones can be interpreted on this basis. The existence of a fourth mRNA, R3′, was inferred from experiments described later in this report and can be ignored in the context of the description of the cDNA clones.

After amplification as plate lysates on *E. coli* Y1090 cells (these cells were R^+, which may have caused an approximately fivefold observed loss in the complexity of the library), the library had a total complexity of about 2×10^4 independently produced clones, of which about 20 were MVM specific. This complexity is rather low but still sufficient for the characterizations described here.

We used four different cloned genomic fragments of MVMi to probe our cDNA library (Fig. 1, above the line representing the viral genome). The probes were as follows: A, 0 to 8 m.u.; B, 21 to 33 m.u.; C, 40 to 53 m.u.; and D, 87 to 99 m.u. Since our library contained *Eco*RI fragments of the original cDNAs and since the MVMi genome contains two *Eco*RI sites at 21 and 69 m.u., we could predict the following hybridization patterns for different classes of cDNA clones. (i) Clones derived from the R1 mRNA should fall into three classes, hybridizing to probes A, B plus C, or D, since both *Eco*RI sites are present in full-length cDNAs derived from this mRNA. (ii) Clones derived from the R2 mRNA should fall into two classes, of which the first has the distinct property of hybridizing to both the A and the C probes but not to the B probe, since the *Eco*RI site at 21 m.u. and the region covered by probe B are in the intron spliced out of this mRNA; the second class hybridizes to probe D. (iii) Clones derived from the R3 mRNA (the most abundant) should fall into two classes, hybridizing to probes C or D.

In the course of the present work, we characterized nine different cDNA clones (Fig. 1) with the following hybridization characteristics: (i) one clone (BC-1) hybridizing to probes B and C, and therefore derived from the R1 mRNA; (ii) one clone (AC-1) hybridizing to probes A and C but not B, and therefore derived from the R2 mRNA; (iii) two clones (C-1 and C-2) hybridizing to probe C only, which could be derived from any mRNA (although one of them has to come from R1; see below); (iv) five clones (D-1 through D-5) hybridizing to probe D only, derived from the region downstream of the *Eco*RI site at 69 m.u., where all mapped mRNA species overlap.

The exact structures of these nine cDNA clones were established by a combination of restriction endonuclease digestion and T-track analysis and are shown in the top part of Fig. 1. It should be noted that two of the clones (AC-1 and D-4) contained sequences that were not found in the MVM genome (Fig. 1; wavy lines). The most likely explanation is that ligation between host and MVM cDNAs occurred during the linker addition step and that these structures are therefore artifacts of the cloning procedure. One clone (D-5) also contained a rearrangement of sequences around the *Eco*RI site at 69 m.u., which may have been generated in a similar fashion.

Structure of the R1 mRNA. The R1 mRNA is a 4.8-kilobase species with a single intron removed around 46 to 48 m.u. We isolated two cDNA clones that could be traced unambiguously to this mRNA, BC-1 and C-2. The left end of BC-1 was at the *Eco*RI site at 21 m.u., while that of C-2 (which has not been mapped precisely) was around 33 m.u., well upstream of the promoter and splice acceptor site at 39 m.u. In agreement with the mRNA mapping data, we do not have any evidence for additional introns in these clones. However, since we sequenced only portions of the clones, we cannot rule out the existence of very small introns.

To find the exact boundaries of the small splice (46 to 48 m.u.) in the R1 mRNA, we subcloned fragments of BC-1 and C-2 into the mWB238 and mWB239 vectors and sequenced them. We sequenced BC-1 from the *Hin*dIII site at position 2651 (52 m.u.) in the direction of the mRNA 5′ end and C-2 from the *Pst*I site at position 2130 (42 m.u.) in the opposite direction, thereby determining the sequence of the splice site on both strands in two independently derived clones. In both clones, nt 2281 and 2378 were joined together (Fig. 2), so that the intron spanned nt 2282 to 2377. The splice donor site is just 1 nt downstream of the stop codon terminating the first large open reading frame (ORF1; see Fig. 4), which is consistent with the predicted structure of the NS1 protein. There are stop codons in all three reading frames within 45 nt of the splice acceptor site. In particular, the first codon of ORF2 (the large open reading frame covering the right half of the MVM genome) lies 6 nt downstream of the splice acceptor.

Structure of the R2 mRNA. The R2 mRNA (3.3 kilobases) has the same 5′ end as the R1 mRNA but differs by having a large intron (from 10 to 40 m.u.) removed. One of the cDNA clones that we isolated, AC-1, had a structure consistent with its being derived from the R2 mRNA, since it contained sequences hybridizing to probes A and C but not B and terminated at the *Eco*RI site at 69 m.u. We also found that AC-1 contained approximately 200 base pairs that were not derived from MVM (see above), but we do not think that this invalidates any of our conclusions.

We prepared subclones of AC-1 which would allow us to read the sequences of both splice sites (i) from the *Hin*dIII site at position 2651 towards the mRNA 5′ end, (ii) from the *Pst*I site at position 2130 in the same direction, and (iii) from the same *Pst*I site towards the mRNA 3′ end. Sequencing of these subclones showed that nt 515 was joined to 1991, and nt 2281 was joined to 2378 (Fig. 2). Since the small intron was identical to the one described for the R1 mRNA, it will not be discussed further. The large intron spans nt 516 to 1990, consistent with the mRNA mapping. The intron spliced out of this mRNA starts with GC rather than the canonical GU. A portion of a sequencing gel that spanned the splice junction is shown in Fig. 2B to document the transition between the exon sequences. It should be noted that the sequence at the position homologous to the 5′ end of the intron is GC in all three rodent parvoviruses, which were sequenced independently (Fig. 2A shows the nucleotides that differed from MVMi in H-1 and MVMp). Therefore, the occurrence of GC in this position is not unique to MVMi and is unlikely to have arisen from a sequencing error.

We conclude that in the R2 mRNA, the N-terminal portion of ORF1 (84 amino acids from the first methionine at nt 262) is spliced in phase to most of an open reading frame (ORF3) extending from nt 1938 to 2300. As the small intron starts at nt 2282, the stop codon of ORF3 (at nt 2300) is spliced out, and the predicted protein actually terminates at position 2396. The predicted molecular weight for a protein coded in this fashion (nt 262 to 515, 1991 to 2281, and 2378 to 2396) is 21,671, consistent with the observed molecular weight of the small nonstructural MVM-coded protein, NS2.

Structure of the R3 mRNA. The R3 mRNA (3.0 kilobases) is the most abundant one found in MVM-infected cells. It originates at a promoter at 39 m.u. and lacks sequences from 46 to 48 m.u. Unfortunately, there are no criteria that would allow us to distinguish between a cDNA clone derived from the R3 mRNA and an incomplete cDNA originating from R1 or R2. However, on the basis of relative mRNA abundance (20), one is more likely to isolate clones originating from R3 than from R1 or R2. Therefore, we decided to sequence clone C-1, which terminates in the region where one would expect the 5′ end of the R3 mRNA.

The small splice in C-1 joined nt 2281 and 2378, exactly as found for AC-1, C-2, and BC-1. As mentioned above, there were termination codons in all three frames within 45 nt of the splice acceptor. In addition, there was no methionine not followed in the same frame by a termination codon until nt 2795. Therefore, if this clone is indeed derived from the R3 mRNA, it can code only for the VP2 viral capsid protein, which originates at the methionine at position 2795 (in the H-1 virus; reference 19).

Structure of the 3′ terminal region of the mRNAs. We isolated five cDNA clones (D-1 through D-5) originating from the region between the *Eco*RI site at 69 m.u. and the polyadenylation site. Careful restriction endonuclease mapping, T-track determination, and partial sequencing of these clones revealed no additional introns downstream of the *Eco*RI site at position 3523. To get an idea of the location of the poly(A) addition site, we sequenced one clone (D-4) from the *Xba*I site (position 4343) towards the mRNA 3′ end. Our sequencing gel did not allow us to read all the way to the poly(A) site, but we could ascertain that it was located downstream of the first AATAAA signal, which is located at position 4603. Since the stop codon terminating the open reading frame for VP1 or VP2 is at position 4556, it is clear that the mRNA covers the entire open reading frame.

Fine-structure mapping of the region around 46 to 48 m.u. As stated before, the complexity of our library was very low, making it unlikely that we would be able to find a cDNA clone generated from an mRNA coding for VP1, which is about five times less abundant than VP2. Since it is clear from the data presented above that the R2 mRNA does not code for VP1 and that the splicing pattern of the presumed R3 mRNA does not allow it to code for a protein of the size of VP1, the most likely coding scheme for this protein involves the use of alternative splice junctions in the region around 47 m.u. A computer search (30) for potential splice junctions in this area revealed the existence of two additional potential donor sites, at nt 2318 and 2336, and a potential acceptor site at nt 2399. If either of the donors was used in conjunction with the acceptor at nt 2399, an mRNA could be generated with a Met initiator codon (nt 2287 through 2289)

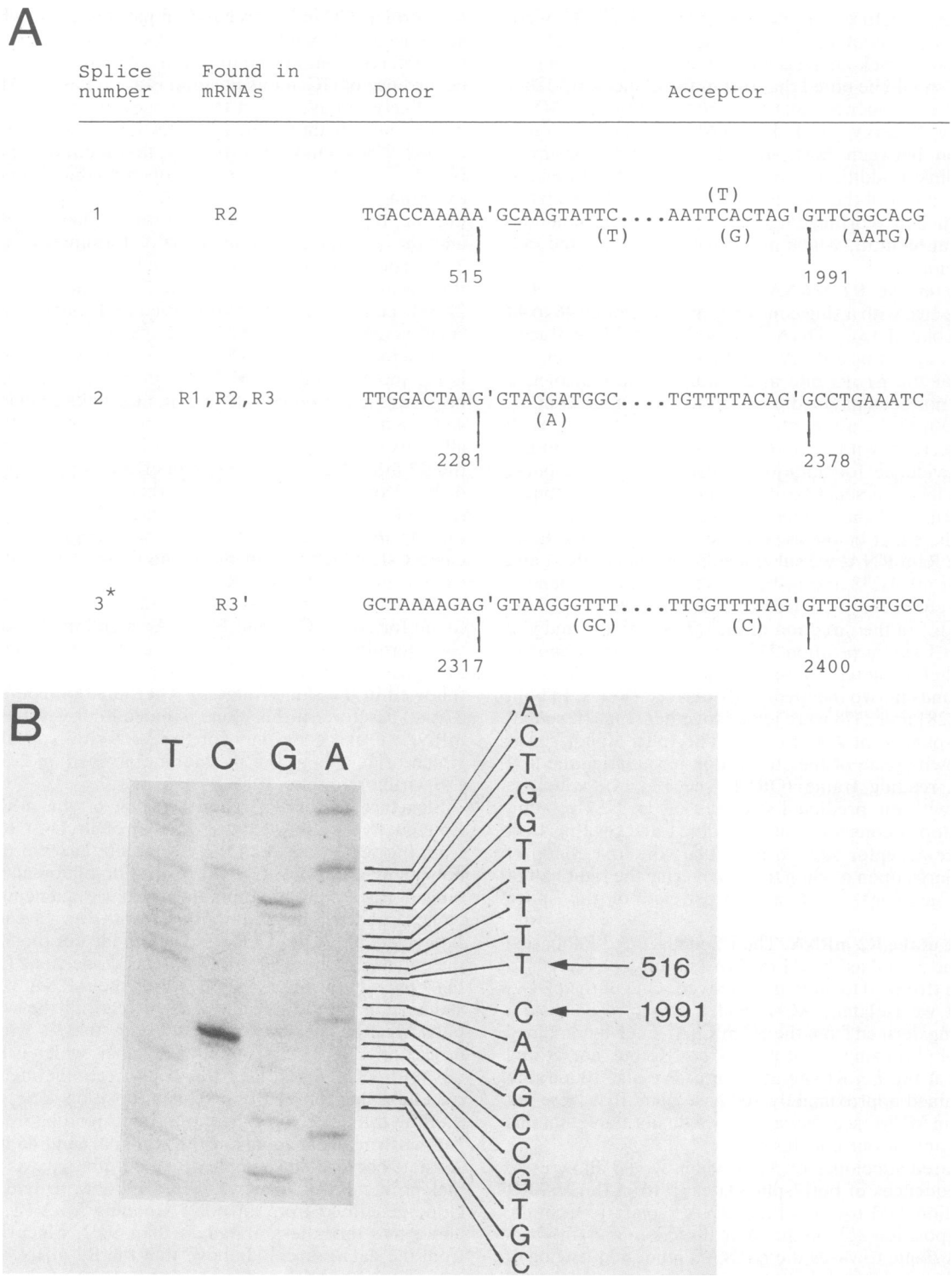

FIG. 2. Sequences of the splice junctions of MVMi. (A) Nucleotides shown in parentheses above the line are those that differ from MVMi in the sequence of MVMp, while those below the line are different in H-1. The asterisk indicates that splice number 3 was not sequenced but deduced from S1 mapping experiments and computer predictions. (B) Part of a sequencing gel documenting the transition between exons in the R2 mRNA (clone AC-1). The strand read on the gel is the encapsidated strand, which is complementary to the coding strand shown in panel A.

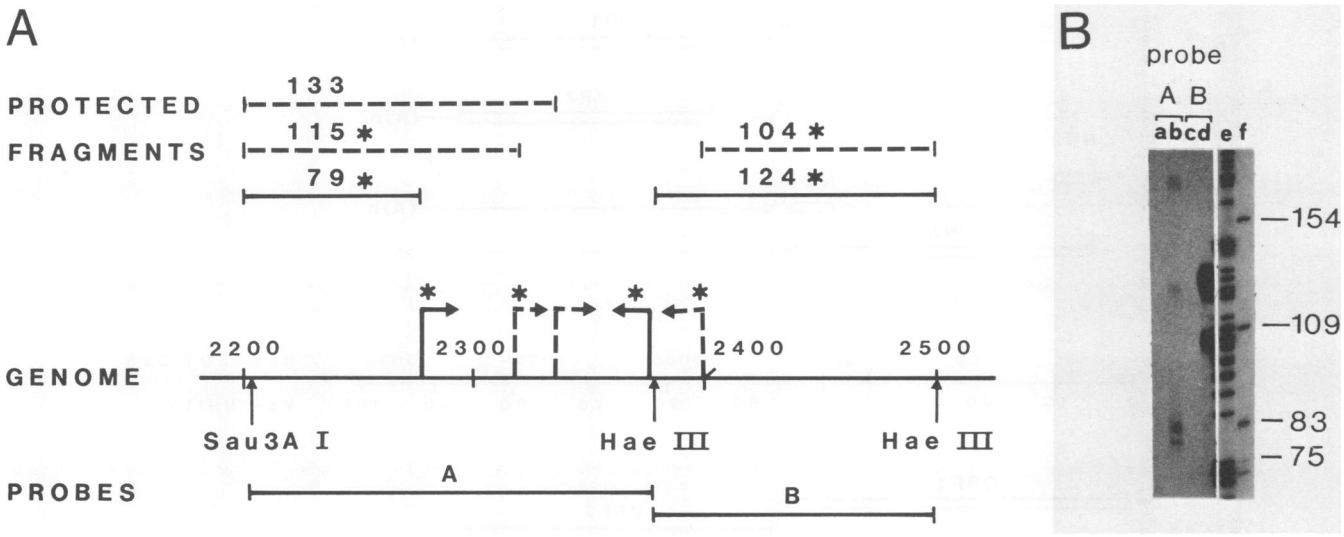

FIG. 3. S1 nuclease mapping of the region between 46 and 48 m.u. (A) Strategy for the analysis. The splice junctions known from the cDNA sequencing are shown in solid lines, while those predicted by computer analysis are shown in dashed lines (arrowheads point toward the intron in all cases). The representation of predicted protected fragments (above) uses the same convention. The probes used for hybridization are shown below. Asterisks mark those splice junctions and protected fragments whose existence was verified experimentally. (B) Results of one of the S1 mapping experiments. The samples shown in lanes a and b were obtained by hybridizing RNA with probe A, while those in lanes c and d were obtained with probe B. In lanes a and c, the probe was hybridized to RNA from uninfected EL-4 cells, while in lanes b and d, it was hybridized to RNA extracted from cells 30 h after infection with MVMi. Lane e shows a T track of the single-stranded DNA used to prepare probe B, while lane f contains restriction endonuclease-generated fragments used as size markers (indicated in nucleotides to the right of the gel).

spliced in the proper frame to ORF2, the large open reading frame extending to nt 4556 and shared with VP2.

To test whether these alternative splice sites are actually used, we analyzed nuclease S1-resistant hybrids generated by annealing viral RNA to the two genomic probes depicted in Fig. 3A. Probe A extends from nt 2203 to 2378 and allows the detection of alternative donor sites. mRNAs containing the known splice donor (nt 2282) will protect a 79-nt fragment, while the predicted mRNAs should protect fragments 115 or 133 nt long. Probe B (nt 2378 through 2501) was designed to detect splice acceptors downstream from the known site (nt 2377). The known mRNAs (or unspliced RNAs) will protect a 124-nt fragment, while the predicted VP1 mRNA should protect a 104-nt fragment.

Figure 3B shows that probe A protected fragments approximately 79, 115, and 175 nt long (exact sizes cannot be determined from the gels, since digestion with S1 nuclease does not generate uniquely sized protected fragments). These fragments represent hybrids of the probe with RNAs with the known splice donor, RNAs with the first predicted donor, and unspliced RNA (or contaminating replicative-form DNA), respectively. There was no protected fragment 133 nt long. From these data, we conclude that only one of the additional donors, the one at nt 2318, is used. Probe B protected fragments about 124 and 104 nt long, corresponding to the known and the predicted splice acceptors. With both probes, the hybrid originating from the known spliced mRNA gave a stronger signal, as would be expected if the less-abundant mRNA coding for VP1 were the only species in which the alternative splice junctions were used. The heterogeneity of the protected fragments precludes mapping of the junctions to the single-nucleotide level, but known constraints on possible splice junctions make it extremely unlikely that the junctions we detected are not the ones predicted from the sequence.

Therefore, we feel that we have identified a fourth viral mRNA. We call this R3′, because it has the same overall structure as the R3 mRNA and because we wish to distinguish it from the R4 mRNA described by Pintel et al. (20). The R3′ mRNA is generated by a splicing event linking the second donor (nt 2317) to the second acceptor (nt 2400), contains an initiation codon linked in phase to ORF2, and almost certainly codes for VP1. We cannot exclude the existence of alternative mRNAs in which the first donor is linked to the second acceptor or the second donor is linked to the first acceptor. However, neither of these potential mRNA species could code for VP1.

DISCUSSION

Genetic organization of MVMi. The data presented in this paper complete the transcription map of MVMi and, by extension, those of the closely related parvoviruses MVMp and H-1, providing a precise genetic origin to the four known proteins coded by these viruses. As is usually found in eucaryotic organisms and the viruses infecting them, each protein has its own mRNA. The squeezing of four proteins into less than 5,000 bases of coding sequence is accomplished in several ways (Fig. 4): (i) sharing of a large block of coding sequence by VP1 and VP2, (ii) alternative splicing around 47 m.u. to provide a start codon connected to an open reading frame for VP1, (iii) overlapping out-of-phase coding regions in the C-terminal region of NS1 and NS2, and (iv) sharing of the N-terminal region of NS1 and NS2. Thus, the strategies used by the autonomous parvoviruses are strongly reminiscent of those described for the better-studied papovaviruses, simian virus 40 and polyomavirus (for a review, see reference 13). The assignment of blocks of coding sequence to the four virally coded proteins (Fig. 4) is supported by the following data.

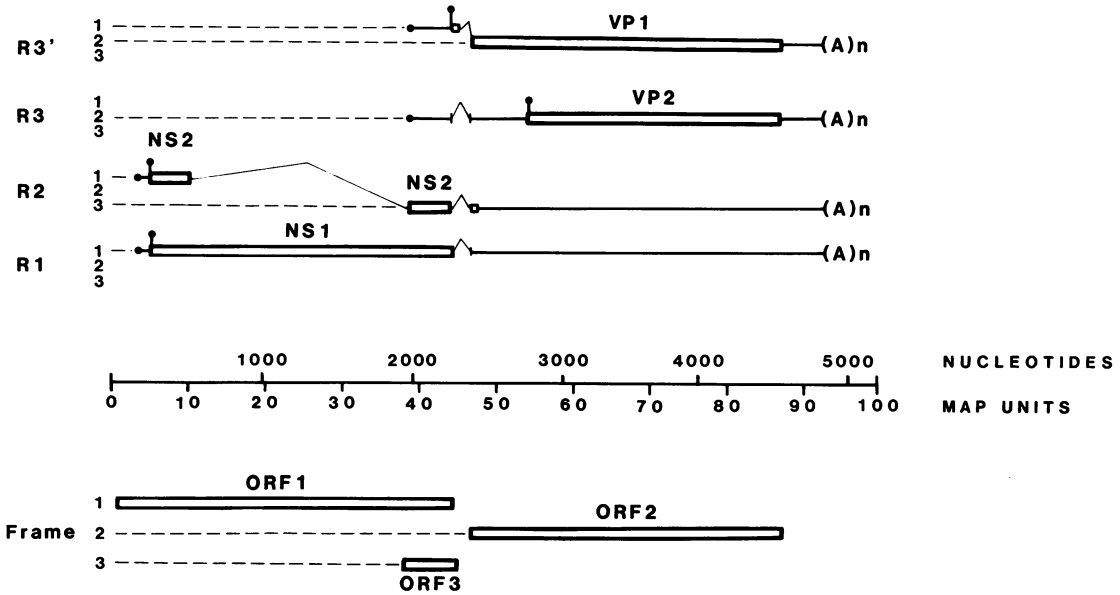

FIG. 4. Coding schemes for the four MVMi viral proteins. (Bottom) The three largest open reading frames found in MVMi and the frames in which they are read. (Top) The manner in which blocks of coding sequence are assembled in the four viral mRNAs to make up the sequences of the four viral polypeptides. The position of the blocks of coding sequence indicates from which frame they are read, and broken lines help in the alignment. Vertical markers indicate the positions of the initiation codons used for each protein, and knobs mark the cap sites of the mRNAs.

NS1. The most compelling evidence for the assignment of NS1 to the large open reading frame (ORF1) in the left half of the genome comes from hybrid-arrested or hybrid-selected in vitro translation experiments (9, 25; our unpublished observations), which show that genomic DNA originating from the large intron removed from the R2 mRNA specifically selects or arrests translation of the NS1 polypeptide. This establishes the R1 mRNA as the only possible message for NS1. Since the stop codon ending ORF1 maps just upstream of the splice donor in R1, the C terminus of NS1 is also determined. The predicted molecular weight of NS1 (76,140) is consistent with published values (83,000) if one keeps in mind that (in the H-1 virus) NS1 is phosphorylated in vivo (18).

NS2. The sequencing of clone AC-1 provides very strong evidence that NS2 is coded by splicing the N-terminal region of ORF1 to most of ORF3, which overlaps the C-terminal portion of ORF1. The predicted polypeptide (molecular weight, 21,671) coded by the mRNA from which clone AC-1 originated would have a molecular weight close to the one observed for NS2 (24,000; reference 9). Strong corroborating evidence comes from the work of Cotmore and Tattersall (9a), who have shown that antibodies raised against fusion proteins containing a portion of ORF3 in the proper frame also precipitate NS2. The sequence of the AC-1 clone also allows us to predict that the six C-terminal amino acids of NS2 are coded in reading frame 3, downstream of the acceptor for the small splice of the R2 mRNA.

VP1. The fact that VP1 shares many tryptic peptides with VP2 (31) makes it almost certain that the two proteins share a large portion of ORF2. Since there is no initiation codon in ORF2 upstream of nt 2795, it was originally thought that the R2 mRNA might code for VP1 (2, 20). As shown above, the R2 mRNA actually codes for NS2. Also, an analysis of cells harboring bovine papillomavirus vectors containing portions of the MVM genome shows that VP1 is produced in cell lines that do not express the R1 or R2 mRNAs (21). Our data indicate that the mRNA coding for VP1 closely resembles the R3 mRNA and could not have been distinguished from R3 with the rather long probes used in the initial mapping experiments (12, 20). For this reason, we chose to call it R3'. Since we have not sequenced a cDNA derived from R3', we cannot be absolutely sure of the exact location of its splice junctions, but the combination of computer analysis (which also correctly predicted the junctions used for the other mRNAs) and S1 mapping gives us good confidence in our assignments. In the proposed coding scheme for VP1, the initiation codon is brought from reading frame 1 (nt 2287) by using splice junctions different from those used for VP2. VP1 uses almost all of ORF2, since the splice acceptor falls only 16 nt from the stop codon that marks the left border of ORF2. The molecular weight predicted for a polypeptide coded in this fashion (80,222) agrees very well with observed values (83,000).

VP2. The coding scheme for VP2 is already well established, since it has been shown to originate from ORF2 and since the N terminus of the H-1 virus VP2 has been mapped by protein sequencing (19). Our sequencing data indicate that in the R3 mRNA there are no Met codons not immediately followed by stop codons until the initiator codon used for VP2; therefore, VP2 is almost certainly coded by this mRNA. The predicted molecular weight of VP2 (64,611) is in excellent agreement with experimental values (e.g., 64,000).

We have no evidence for the existence of a 1.8-kilobase mRNA (R4) of unknown origin that was detected in blot hybridizations by Pintel et al. (20). Since we can now account for the origin of all four known virally coded polypeptides from the three larger mRNA species, it seems likely that this small RNA represents a degradation product or a gel artifact. We have occasionally seen a band at the same position in blots of RNA preparations contaminated with large amounts of rRNA.

An unusual splice junction in the R2 mRNA. A very interesting point arises from our sequencing of the splice

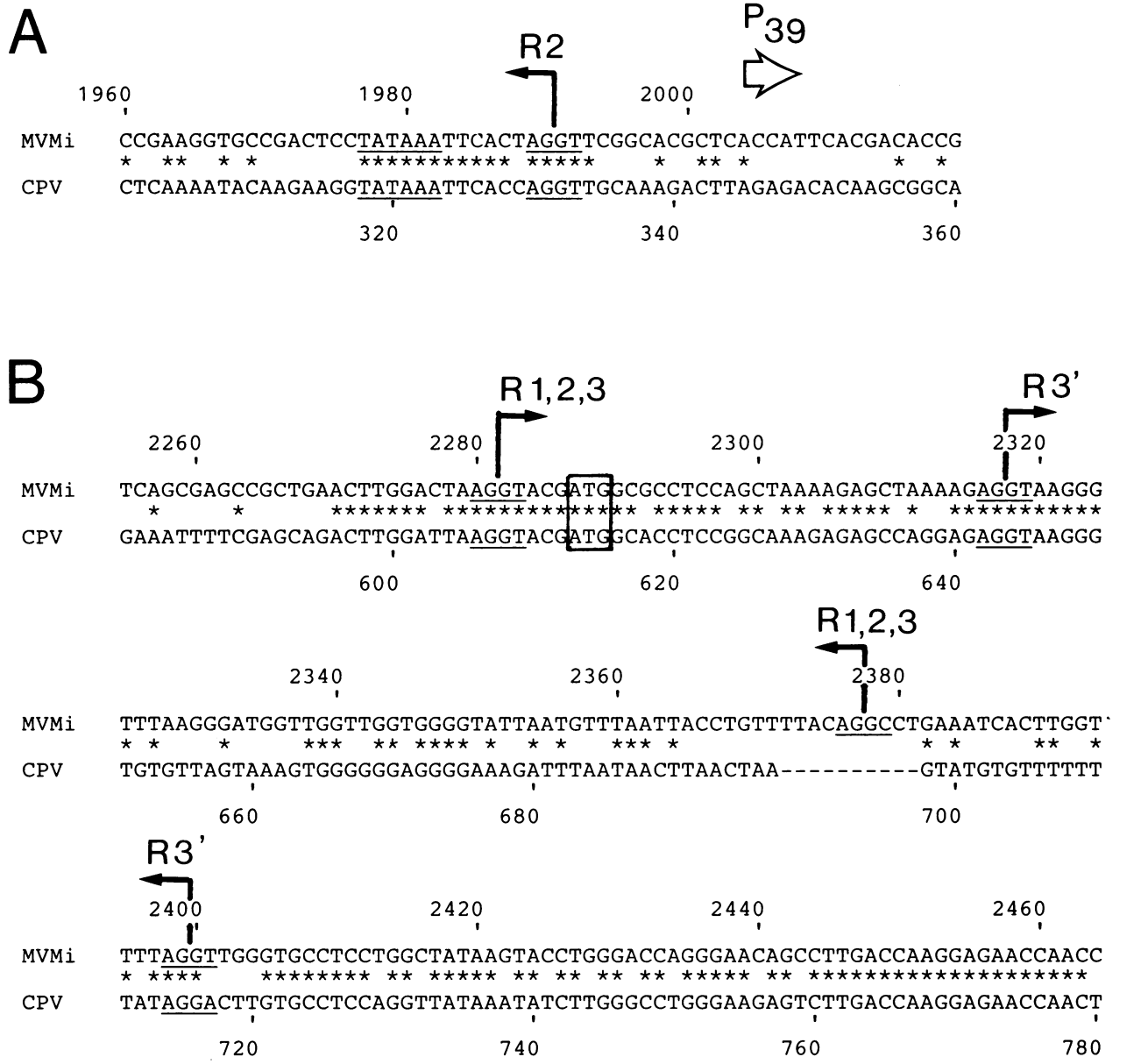

FIG. 5. Comparison of the 40-m.u. (A) and 46- to 48-m.u. (B) regions of MVMi and CPV. The sequence of CPV was taken from reference 23, and that of MVMi was taken from reference 27. Asterisks mark the positions where nucleotides are identical to the two sequences (a background noise of one position in four being identical is expected in such a representation). The TATA boxes and splice junctions (hypothetical in the case of CPV) are underlined, and the intron-exon junctions of MVMi are marked with arrows. The hollow arrow marks the position of the cap site for the mRNAs initiated at the P39 promoter (1). The ATG codon that is probably used to initiate synthesis of VP1 is boxed. A 10-nt gap (dashes) has been introduced in the CPV sequence to maximize homology (7).

junction in the R2 mRNA. The intron removed from this mRNA starts with the nucleotides GC, rather than the GU found in nearly all of the introns sequenced so far. To our knowledge, the only other documented instance of a naturally occurring intron starting with GC is found in the chicken and duck α^D-globin genes (10, 11). The region corresponding to the MVMi exon-intron junction has been sequenced independently by three groups in three rodent parvovirus genomes (1, 2, 25, 27), and assuming that the splice junctions are found in homologous positions in all three viruses, all three introns start with GC. Unfortunately, this region is still missing from the published sequences of CPV and FPV.

Interestingly, replacement of the C at the second position of the intron by a U generates a sequence to which the computer assigns a very high probability of serving as a splice donor. Therefore, it seems that all of the elements are present to create a good splice donor, except the canonical U. This may be relevant to the fact that a spliced (R2) and unspliced (R1) form of the RNA must coexist within the infected cell and that the splicing reaction cannot be too efficient lest no R1 mRNA be produced at all. In fact, the probability score assigned to the splice acceptor site (nt 1991) was rather low, even though this site is conserved in all sequenced autonomous parvoviruses (Fig. 2 and 5). On the other hand, the splice donor and acceptor sequences used in

the 46- to 48-m.u. region all showed high scores, and very little unspliced RNA coming from this region can be detected. The consensus lariat branch point sequence (pyrimidine-X-pyrimidine-T-purine-A-pyrimidine) described by Ruskin et al. (26) is found within a 15- to 50-nt distance of all three splice acceptors. If it is true that this consensus sequence fully describes acceptable branch points, then the two splice acceptors at nt 2378 and 2400 must share the same branch point (TATTAAT, nt 2350 through 2356).

Comparison of MVMi with other parvoviruses. The question arises of whether the intron-exon structure of the MVMi mRNAs exactly matches that of its close relatives MVMp and H-1. We strongly believe that it does, for the following reasons: (i) the sequences of the three viruses are so closely related that they are very unlikely to use blocks of coding sequence in a significantly different manner; (ii) the sequences corresponding to the MVMi splice junctions are essentially identical in all three viruses; (iii) the positioning of initiation codons, stop codons, and open reading frames is identical, and the presumed splice junctions would connect them in exactly the same way in the three viruses; and (iv) splice junctions predicted by the computer fall in the same places. Therefore, we are convinced that our mapping of the MVMi splice sites can be extended to the two other well-characterized rodent parvoviruses.

We also compared the sequences of the splice junctions of MVMi with sequences found at homologous positions in the more distantly related canine and feline parvoviruses. As pointed out by others (1, 7, 23), there are some striking homologies between the rodent parvoviruses and CPV and FPV in the region between 40 and 50 m.u. When the matrix comparison method of Pustell and Kafatos (22) is used, three areas of strong homology (at the nucleotide level) between MVMi and CPV are evident in this region: (i) a short stretch around the TATA box and the splice acceptor for the R2 mRNA, which are immediately adjacent to each other; (ii) a longer region around 46 m.u., which covers both of the splice donor sites; and (iii) a very strong homology in the region of VP1 which is not found in VP2 (the N-terminal part of ORF2). A side-by-side comparison of the sequences of MVMi and CPV is shown in Fig. 5. Because of the close relatedness of MVMi, MVMp, and H-1 on the one hand and CPV and FPV on the other hand, the same conclusions apply to paired comparisons between any members of the two groups.

The homology at the TATA boxes (Fig. 5A) is not too surprising if promoters are found in homologous positions in the two genomes. More interestingly, the conservation of the R2 splice acceptor suggests that a protein homologous to NS2 could be coded for by CPV and FPV. In both of these viruses, there is an open reading frame overlapping ORF1 and extending 235 nt downstream from the putative R2 splice acceptor. However, in contrast to the murine parvoviruses, this open reading frame terminates in a stop codon (nt 1142 in FPV; nt 567 in CPV) before the first splice donor (nt 1182 in FPV; nt 607 in CPV).

In the 46- to 48-m.u. region (Fig. 5B), the strong conservation of splice donors does not apply to the splice acceptor sites: the first site in MVMi (nt 2378) has no counterpart in CPV or FPV. A search for potential splice acceptors in this region of CPV and FPV reveals a single site with a high score (nt 1291 in FPV; nt 716 in CPV), which is homologous to the second acceptor of the rodent parvoviruses. Several groups (1, 7, 23) have argued that the arrangement of conserved splice donors and acceptors suggests a general scheme for the structure of the VP1 and VP2 mRNAs in autonomous parvoviruses, i.e., the splicing of two alternative donors (nt 2281 and 2317 in MVMi) to the same acceptor (nt 2400 in MVMi). Our data show that this is not the case for MVMi or, by extension, for the other rodent parvoviruses. On the other hand, in the absence of a second acceptor, this scheme looks very plausible for CPV and FPV. In fact, in CPV and FPV, the presence of a stop codon in ORF3 before the first splice donor (see above) may be linked to the existence of a single-splice acceptor, since the acceptor used in the R2 mRNA of MVMi (which positions the stop codon for NS2) is not present.

Since the arrangement of open reading frames and initiation codons is essentially the same in the two groups of autonomous parvoviruses for which we have sequence information, the differences outlined above probably represent minor variations on a common theme. The adeno-associated viruses (dependoviruses [29]) share many aspects of genetic oganization with the autonomous parvoviruses (8). It will be interesting to find out whether some of the schemes of protein coding described for MVMi will also hold true for these distant cousins.

ACKNOWLEDGMENTS

This work was supported by grants from the Swiss National Science Fund.

We thank B. Bentele for maintaining the cells and virus stocks and N. Thompson for valuable technical advice.

LITERATURE CITED

1. **Astell, C. R., E. M. Gardiner, and P. Tattersall.** 1986. DNA sequence of the lymphotropic variant of minute virus of mice, MVM(i), and comparison with the DNA sequence of the fibrotropic prototype strain. J. Virol. **57**:656–669.
2. **Astell, C. R., M. Thomson, M. Merchlinsky, and D. C. Ward.** 1983. The complete DNA sequence of minute virus of mice, an autonomous parvovirus. Nucleic Acids Res. **11**:999–1018.
3. **Barnes, W. M., M. Bevan, and P. H. Son.** 1983. Kilosequencing: creation of an ordered nest of asymmetric deletions across a large target sequence carried on phage M13. Methods Enzymol. **101**:98–122.
4. **Ben-Asher, E., and Y. Aloni.** 1984. Transcription of minute virus of mice, an autonomous parvovirus, may be regulated by attenuation. J. Virol. **52**:266–276.
5. **Berk, A. J., and P. A. Sharp.** 1977. Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. Cell **12**:721–732.
6. **Burke, J. F.** 1984. High-sensitivity S1 mapping with single-stranded [^{32}P]DNA probes synthesized from bacteriophage M13 mp templates. Gene **30**:63–68.
7. **Carlson, J., K. Rushlow, I. Maxwell, F. Maxwell, S. Winston, and W. Hahn.** 1985. Cloning and sequence of DNA encoding structural proteins of the autonomous parvovirus feline panleukopenia virus. J. Virol. **55**:574–582.
8. **Carter, B. J., C. A. Laughlin, and C. J. Marcus-Sekura.** 1983. Parvovirus transcription, p. 153–207. *In* K. I. Berns (ed.), The parvoviruses. Plenum Publishing Corp., New York.
9. **Cotmore, S. F., L. J. Sturzenbecker, and P. Tattersall.** 1983. The autonomous parvovirus MVM encodes two non-structural proteins in addition to its capsid polypeptides. Virology **129**:333–343.

9a. **Cotmore, S. F., and P. Tattersall.** 1986. Organization of nonstructural genes of the autonomous parvovirus minute virus of mice. J. Virol. **58**:724–732.

10. **Erbil, C., and J. Niessing.** 1983. The primary structure of the duck α^D-globin gene: an unusual 5′ splice junction sequence. EMBO J. **2**:1339–1343.
11. **Fischer, H. D., J. B. Dodgson, S. Hughes, and J. D. Engel.** 1984. An unusual 5′ splice junction is efficiently utilized in vivo. Proc.

Natl. Acad. Sci. USA **81**:2733–2737.
12. **Green, M. R., R. M. Lebowitz, and R. G. Roeder.** 1979. Expression of the autonomous parvovirus H-1 genome: evidence for a single transcriptional unit and multiple spliced polyadenylated transcripts. Cell **17**:967–977.
13. **Griffin, B. E.** 1981. Structure and genomic organization of SV40 and polyoma virus, p. 71–123. *In* J. Tooze, (ed.), DNA tumor viruses. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
14. **Huynh, T. V., R. A. Young, and R. W. Davis.** 1985. Constructing and screening cDNA libraries in lambda gt10 and lambda gt11, p. 49–78. *In* D. M. Glover (ed.), DNA cloning, vol. I. IRL Press, Oxford.
15. **Johnson, D. A., J. W. Gautsch, J. R. Sportsman, and J. H. Elder.** 1984. Improved technique utilizing nonfat dry milk for analysis of proteins and nucleic acids transferred to nitrocellulose. Gene Anal. Techniques **1**:3–8.
16. **McMaster, G. K., P. Beard, H. D. Engers, and B. Hirt.** 1981. Characterization of an immunosuppressive parvovirus related to the minute virus of mice. J. Virol. **38**:317–326.
17. **Nielsen, P. J., G. K. McMaster, and H. Trachsel.** 1985. Cloning of eukaryotic protein synthesis initiation factor genes: isolation and characterization of cDNA clones encoding factor eIF-4A. Nucleic Acids Res. **13**:6867–6880.
18. **Paradiso, P. R.** 1984. Identification of multiple forms of the noncapsid parvovirus protein NCVP1 in H-1 parvovirus-infected cells. J. Virol. **52**:82–87.
19. **Paradiso, P. R., K. R. Williams, and R. L. Costantino.** 1984. Mapping of the amino terminus of the H-1 parvovirus major capsid protein. J. Virol. **52**:77–81.
20. **Pintel, D., D. Dadachanji, C. R. Astell, and D. C. Ward.** 1983. The genome of minute virus of mice, an autonomous parvovirus, encodes two overlapping transcription units. Nucleic Acids Res. **11**:1019–1038.
21. **Pintel, D., M. J. Merchlinsky, and D. C. Ward.** 1984. Expression of minute virus of mice structural proteins in murine cell lines transformed by bovine papillomavirus-minute virus of mice plasmid chimera. J. Virol. **52**:320–327.
22. **Pustell, J., and F. C. Kafatos.** 1984. A convenient and adaptable package of computer programs for DNA and protein sequence management, analysis and homology determination. Nucleic Acids Res. **12**:643–655.
23. **Rhode, S. L., III.** 1985. Nucleotide sequence of the coat protein gene of canine parvovirus. J. Virol. **54**:630–633.
24. **Rhode, S. L., III.** 1985. *trans*-Activation of parvovirus P_{38} promoter by the 76K noncapsid protein. J. Virol. **55**:886–889.
25. **Rhode, S. L., III, and P. R. Paradiso.** 1983. Parvovirus genome: nucleotide sequence of H-1 and mapping of its genes by hybrid-arrested translation. J. Virol. **45**:173–184.
26. **Ruskin, B., A. R. Krainer, T. Maniatis, and M. R. Green.** 1984. Excision of an intact intron as a novel lariat structure during pre-mRNA splicing in vitro. Cell **38**:317–331.
27. **Sahli, R., G. K. McMaster, and B. Hirt.** 1985. DNA sequence comparison between two tissue-specific variants of the autonomous parvovirus, minute virus of mice. Nucleic Acids Res. **13**:3617–3633.
28. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain termination inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
29. **Siegl, G., R. C. Bates, K. I. Berns, B. J. Carter, D. C. Kelly, E. Kurstak, and P. Tattersall.** 1985. Characteristics and taxonomy of parvoviridae. Intervirology **23**:61–73.
30. **Staden, R.** 1984. Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res. **12**:505–519.
31. **Tattersall, P., and J. Bratton.** 1983. Reciprocal productive and restrictive virus-cell interactions of immunosuppressive and prototype strains of minute virus of mice. J. Virol. **46**:944–955.
32. **Tattersall, P., A. J. Shatkin, and D. C. Ward.** 1977. Sequence homology between the structural polypeptides of minute virus of mice. J. Mol. Biol. **111**:375–394.