VIROLOGY 212, 562–573 (1995)

Analysis of the Complete Nucleotide Sequences of Goose and Muscovy Duck Parvoviruses Indicates Common Ancestral Origin with Adeno-Associated Virus 2

ZOLTÁN ZÁDORI,[1] RAYMUND STEFANCSIK,[2] TIBOR RAUCH,[3] and JÁNOS KISARY[4]

Research Department, Phylaxia-Sanofi Veterinary Biologicals Co. Ltd., H-1107 Budapest, Hungary

Received May 2, 1995

The complete nucleotide sequences of two parvoviruses isolated from goose and muscovy duck were determined. The two virus genomes share 81.9% nucleotide sequence identity, indicating that they are closely related. The coding regions are bracketed by inverted terminal repeats containing palindromes. This is similar to the genome organization of human parvoviruses, adeno-associated virus 2, and B19. Amino acid sequence comparison shows that the closest relative of the goose and muscovy duck parvoviruses is adeno-associated virus 2. This is surprising, because the goose and muscovy duck parvoviruses do not require any helper virus for productive replication, suggesting that adeno-associated virus 2 has been derived from a helper-independent ancestor. © 1995 Academic Press, Inc.

INTRODUCTION

Parvoviruses are among the smallest animal viruses and have linear, single-stranded genomes. Parvoviruses have been isolated from vertebrates and invertebrates (Tijssen, 1990). Parvovirus virions are nonenveloped, have icosahedral symmetry, and are 18–26 nm in diameter (Berns, 1990). The virions each contain a single DNA molecule of about 5–6 kilobases in length, in which a long single-stranded coding region is bracketed by shorter terminal palindromic regions capable of folding into hairpin duplexes. The terminal hairpins serve as origins of replication and provide *cis* signals for packaging (Tattersall and Cotmore, 1990; Carter *et al.*, 1990b). According to Astell (1990), parvoviruses can be classified based on the properties of the terminal hairpins. Type A viruses (e.g., AAV-2 and B19) have inverted terminal repeats containing palindromes, and viruses with type B genomes (e.g., MVM, H1, LuIII, etc.) have palindromes at the left and right termini, the sequences of which are not related. The genome of vertebrate parvoviruses contains two major open reading frames, one encoding for the capsid proteins and the other coding for a family of pleiotropic regulatory proteins (Rhode and Iversen, 1990).

Parvoviruses replicate in the nucleus. Due to the limited number of viral genes, parvovirus replication depends on several functions provided by the host cell (Tattersall and Gardiner, 1990). In the case of the members of *Dependovirus*, effective parvovirus replication may require the presence of a helper virus, usually adenovirus or herpesvirus (Carter, 1990).

In the mid-sixties a highly fatal disease of goslings and muscovy ducklings was described (Derzsy, 1967). The causative agent of the disease was classified as a parvovirus (Kisary and Derzsy, 1974). The goose parvovirus (GPV) was described as a small particle of 20–22 nm in diameter with a hexagonal outer appearance. GPV was found to replicate in primary goose embryo fibroblast cells with the formation of Cowdry type A intranuclear inclusion bodies (Kisary and Derzsy, 1974). Another parvovirus isolated from muscovy ducks with clinical signs of Derzsy's disease was shown to be closely related to GPV by Southern hybridization (Zádori *et al.*, 1994). Both GPV and muscovy duck parvovirus (MDPV) replicate autonomously and encapsidate strands of opposite polarities in equal amounts, if they are propagated *in embryo* (Zádori *et al.*, 1994).

A fair number of parvoviruses have been molecularly cloned and sequenced, including members of the family which parasitize insects and mammals (for references see Rhode and Iversen, 1990; Diffoot *et al.*, 1993; Dumas *et al.*, 1992; Bando *et al.*, 1987). However, there is no information available on the genome organization of avian parvoviruses at the level of nucleotide sequence. Therefore the phylogenetic relationship between avian and the well-characterized mammalian parvoviruses has not been established. We were interested in obtaining the complete nucleotide sequence of GPV and MDPV, because this information would allow us to define the evolutionary relationship between the avian and mammalian parvoviruses. On the other hand, GPV and MDPV

[1] Present address: Veterinary Medical Research Institute, Hungarian Academy of Sciences, H-1581 Budapest, P.O.B. 18, Hungary.

[2] Present address: Program in Cell, Molecular & Developmental Biology, Tufts University, Sackler School of Biomedical Sciences, 136 Harrison Avenue, Boston, MA 02111.

[3] Present address: Institute of Biochemistry, BRC of the Hungarian Academy of Sciences, 6701 Szeged, Temesvári krt. 62., Hungary.

[4] To whom reprint requests should be addressed at Research Department, Phylaxia-Sanofi Veterinary Biologicals Co. Ltd., H-1107 Budapest, *Szállás u.* 5., Hungary. Fax: (36-1) 260-3889

0042-6822/95 $12.00

differ in several ways. They have different host preferences, there are antigenic differences between them (Kisary, unpublished observations), and the genomes show distinct restriction site patterns (Zádori *et al.*, 1994). Therefore, we also wanted to determine how much GPV and MDPV have diverged.

MATERIALS AND METHODS

Enzymes and reagents

Restriction endonucleases and T4 DNA polymerase were purchased from MBI Fermentas (Vilnius, Lithuania). For DNA sequencing, we purchased T7 DNA polymerase and reagents from Pharmacia Biotech (Sollentuna, Sweden).

Bacterial cells and plasmids

Escherichia coli strains XL1-Blue {*Δ(mcrA)183, Δ(mcrCB-hsdSMR-mrr)173, endA1, supE44, thi-1, recA, gyrA96, relA1, lac,* λ−, [*F', proAB, lacI*^{q}*ZΔM15, Tn10, (tet*^{r}*)*]} *and SURE {el4*$^-$*(mcrA), Δ(mcrCB-hsdSMR-mrr)171, endA1, supE44, thi-l, gyrA96, relA1, lac, recB, recJ, sbcC, umuC:Tn5 (kan*^{r}*), uvrC,* [*F' proAB, lacI*^{q}*ZΔM15, Tn10, (tet*^{r}*)*]} from Stratagene (La Jolla, CA) were used for plasmid propagation. For DNA cloning the vector pBluescript II (SK$^-$) (Stratagene, La Jolla, CA) was used.

Virus propagation and preparation of viral DNA

For molecular cloning, virion DNA was isolated from specific pathogen-free muscovy duck (*Cairina moschata*) and goose (*Anser anser*) embryonated eggs infected with strain FM of MDPV and virulent B strain of GPV as previously described (Zádori *et al.*, 1994). As MDPV and GPV propagated *in embryo* encapsidate both negative and positive DNA strands, we obtained double-stranded nucleic acid.

Molecular cloning and sequencing

MDPV virion DNA was digested with either restriction enzyme *Sac*I or *Hin*dIII. Both of these enzymes cut the MDPV genome at a single site. The resulting two DNA fragments were ligated into the *Eco*RV–*Sac*I or *Eco*RV–*Hin*dIII site, respectively, of pBluescript II (SK$^-$). Upon transformation of the ligation mixture into XL1-Blue competent bacteria, we obtained all of the expected recombinant clones; however, they contained deletions in the inverted terminal repeats (ITRs). In contrast, we were able to obtain recombinant *Hin*dIII clones when we prepared competent bacteria from *E. coli* strain SURE. We also cloned internal *Eco*RI and *Pst*I fragments into pBluescript II SK$^-$. We generated exonuclease III–mung bean nuclease nested deletions in the clones harboring the *Sac*I "A" and "B" fragments for sequencing. We also created subclones with restriction endonucleases *Hin*dIII, *Stu*I, *Kpn*I, and *Dra*I into pBluescript II (SK$^-$). We determined the nucleotide sequences by the dideoxynucleotide chain termination method by applying either T7 DNA polymerase and S^{32}-labeled dATP for conventional sequencing gels or fluorescein-labeled primers for automated DNA sequencing with the ALF DNA Sequencer of Pharmacia (Sweden). We used the SK, KS, M13–20, and reverse oligonucleotide primers for DNA sequencing.

GPV virion DNA was cut with *Hin*dIII, and the resulting end fragments, "A" and "C," were ligated to the *Eco*RV–*Hin*dIII site of the vector pBluescript II (SK$^-$). The *Hin*dIII "B" fragment was ligated to the *Hin*dIII site of the same vector. We obtained recombinants of correct sites by using the competent bacteria SURE. Internal GPV genomic *Pst*I fragments were also cloned into pBluescript II (SK$^-$). We generated exonuclease III–mung bean nuclease nested deletions in the *Hin*dIII clones for sequencing. We determined the nucleotide sequences of the GPV clones as described above for MDPV.

In order to sequence the ITRs of MDPV and GPV, we generated exonuclease III–mung bean nuclease nested deletions and subcloned the terminal *Sph*I fragments. The restriction enzyme cuts the ITRs into two fragments.

Sequence analysis softwares

For sequence analysis the Wisconsin sequence analysis package version 8 of the Genetics Computer Group (Madison, WI) was employed. Transcription factor recognition sequences were searched with the program Findpatterns in the Transcription Factor Database (Ghosh, 1990).

Nucleotide sequence accession number

The complete DNA sequences of MDPV and GPV have been submitted to GenBank and assigned Accession Nos. U22967 and U25749, respectively.

RESULTS

Nucleotide sequence

We succeeded in cloning and sequencing of the entire genomes of GPV and MDPV (Fig. 1). The genome of GPV (5106 bp) is somewhat shorter than that of MDPV (5132 bp). The difference in length between the two genomes can be attributed to short deletions in the ITRs and to a 5-bp deletion in the GPV chromosome shortly after the first stop codon of the capsid gene (from nucleotide 4639 of GPV). By comparing the coding regions of the two genomes, we can say that there is not any deletion; however, there are numerous substitutions. Otherwise, the two genomes show high similarity. They share 81.9% nucleotide sequence identity.

We must note that we experienced difficulties in sequencing the intact "bubble" region (Fig. 1) of the ITRs. We started to sequence the inserts of the plasmids carrying the intact ITRs. We could not obtain unambiguous nucleotide sequences of the plus or the minus DNA

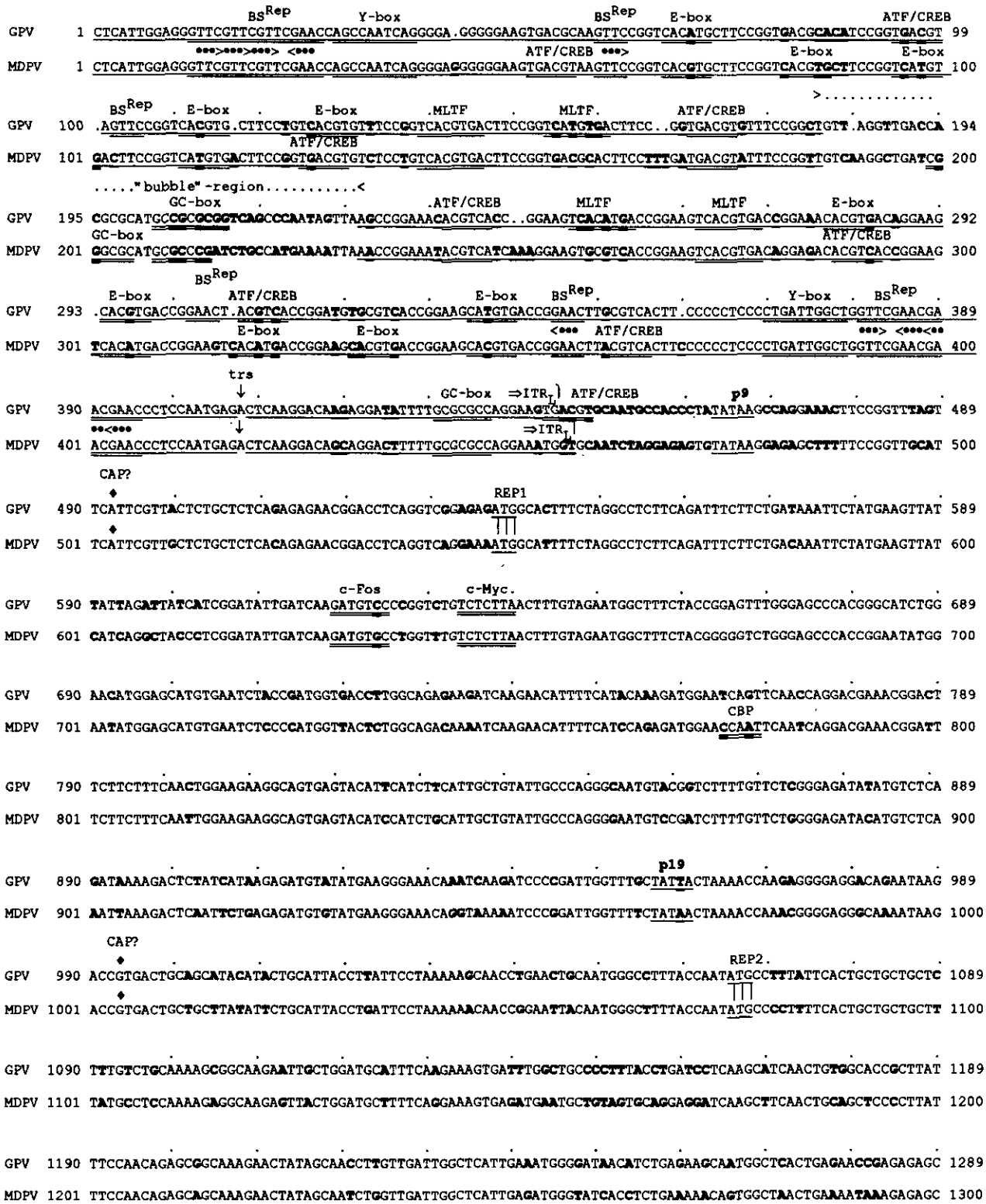

FIG. 1. Complete nucleotide sequences of GPV and MDPV. The nucleotide sequences of the plus strands are shown. The differences between the two sequences are in boldface. The regions corresponding to the inverted terminal repeats (ITRs), except those of the "bubble regions," are underlined. The putative Rep recognition sequences (BS^{Rep}) and the transcription factor binding sites (Y-box, E-box, ATF/CREB, MLTF, GC-box, c-Fos, c-Myc, CBP, and AP-1) are double underlined. The vertical arrows indicate the putative terminal resolution sites (trs). Vertical lines (| | |) indicate start and stop codons of the putative proteins REP1, REP2, VP1, VP2, and VP3. The promoters p9, p19, and p41 are underlined. The marks ♦ and "CAP" designate putative transcription start sites. The putative polyadenylation and splice donor and acceptor sites (*) are also indicated.

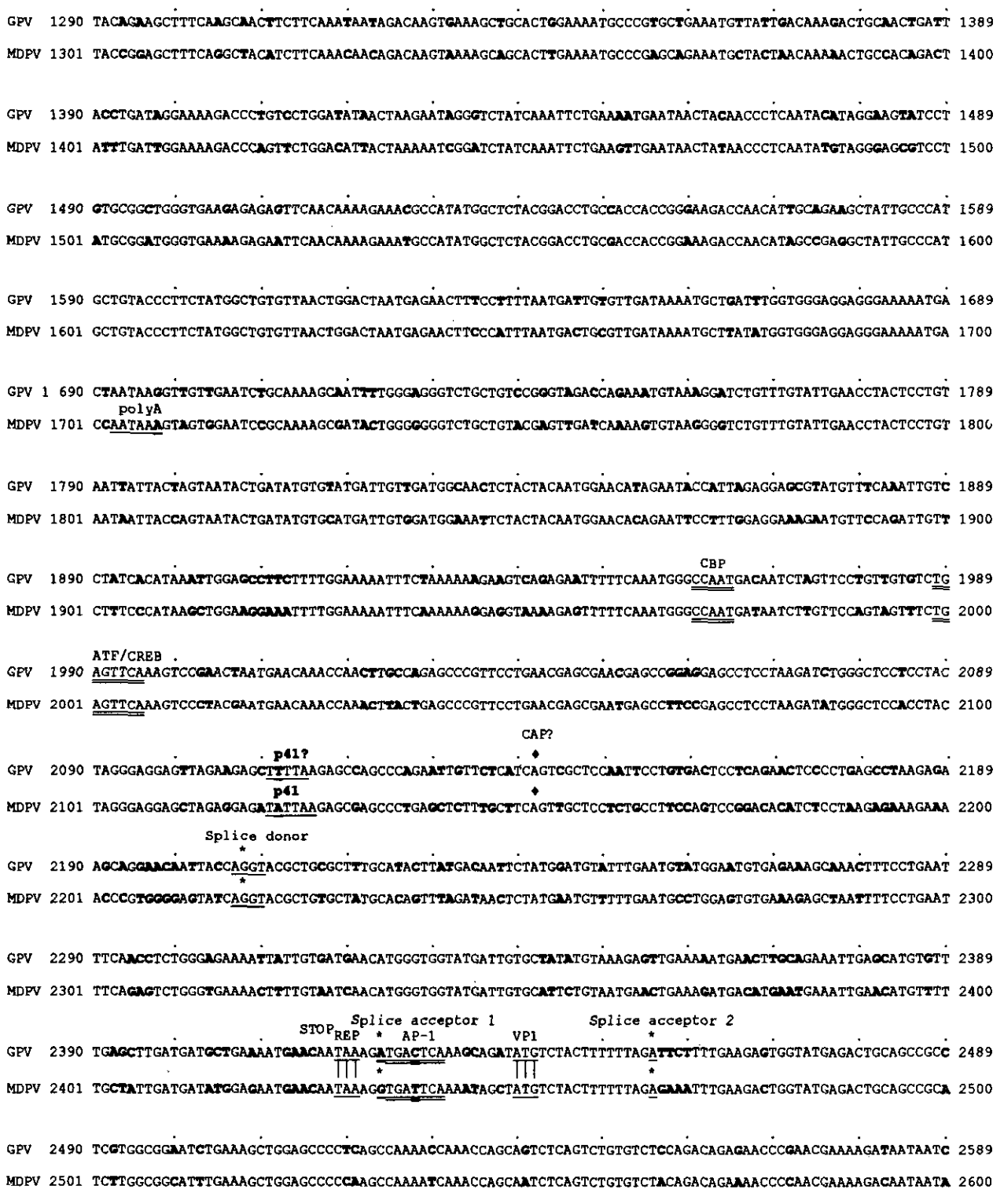

FIG. 1—*Continued*

strands corresponding to the "bubble" region. Therefore, we generated subclones of the intact ITRs by cutting them in half with restriction endonuclease *Sph*I. The restriction endonuclease *Sph*I cuts the GPV and MDPV ITRs at nucleotide 202 and 4907, and 208 and 4924, respectively. The subclones obtained this way gave unambiguous sequence data for both the plus and minus DNA strands by using the suitable pBluescript oligonucleotide primers. The analysis of the sequences of the subclones revealed that the sequencing difficulties in the complete ITR-containing clones were caused not by the high GC content of the "bubble" region, but by the inter-

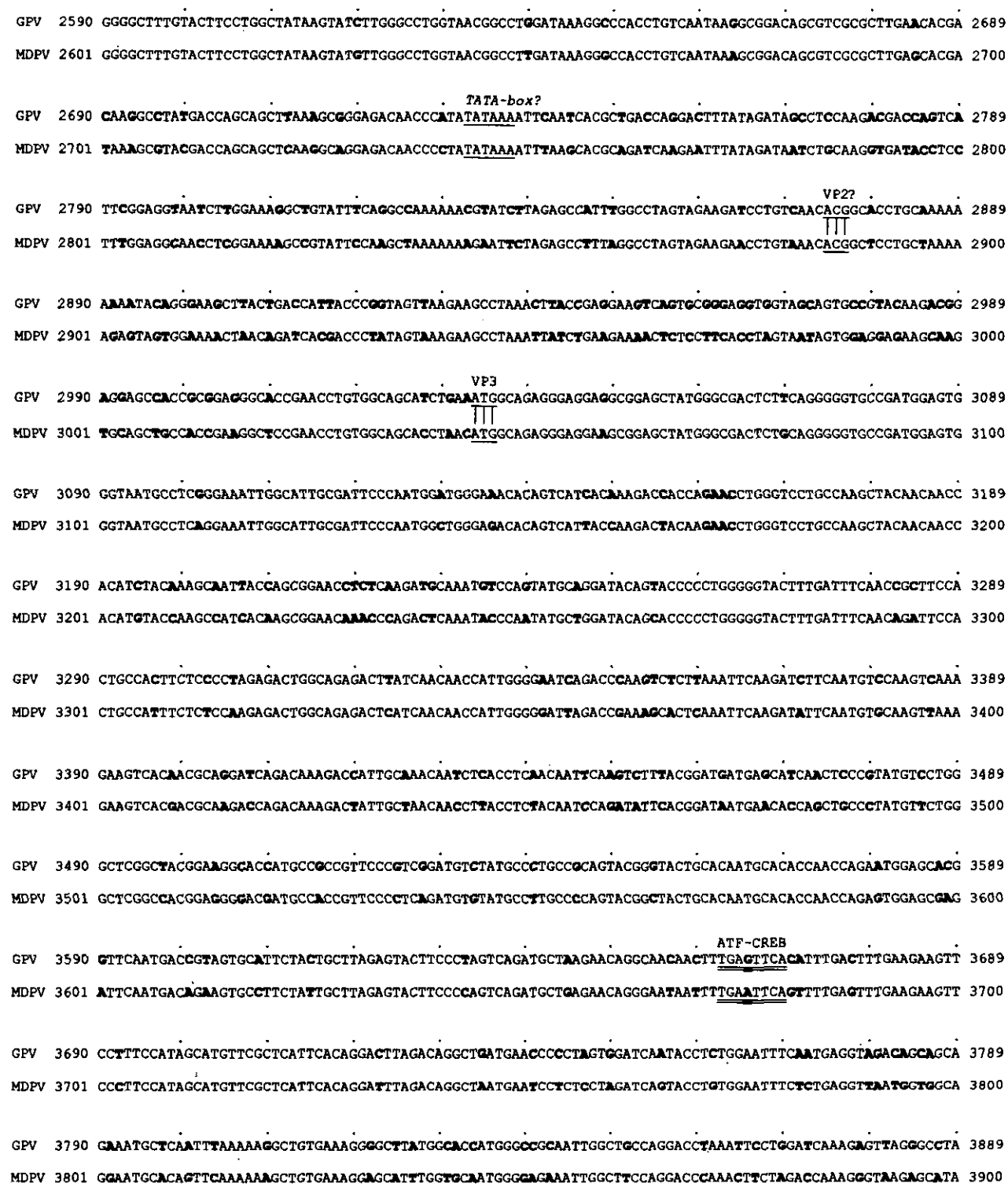

FIG. 1—*Continued*

conversion between the "flip" and "flop" orientation. The interconversion took place while the full-length ITR clones were propagated in *E. coli*, in spite of the fact that we used the strain SURE, which is deficient in the major recombination genes. Thus, even if we started DNA purification from a single, newly transformed colony, we obtained a mixture of plasmids carrying the "bubble" region in both "flip" and "flop" orientation, as judged by the *Sph*I restriction enzyme cleavage patterns (data not shown). Also, about 30% of the *Sph*I-generated subclones contained the opposite orientation of the region compared to that of the rest of the analogous clones. In Fig.

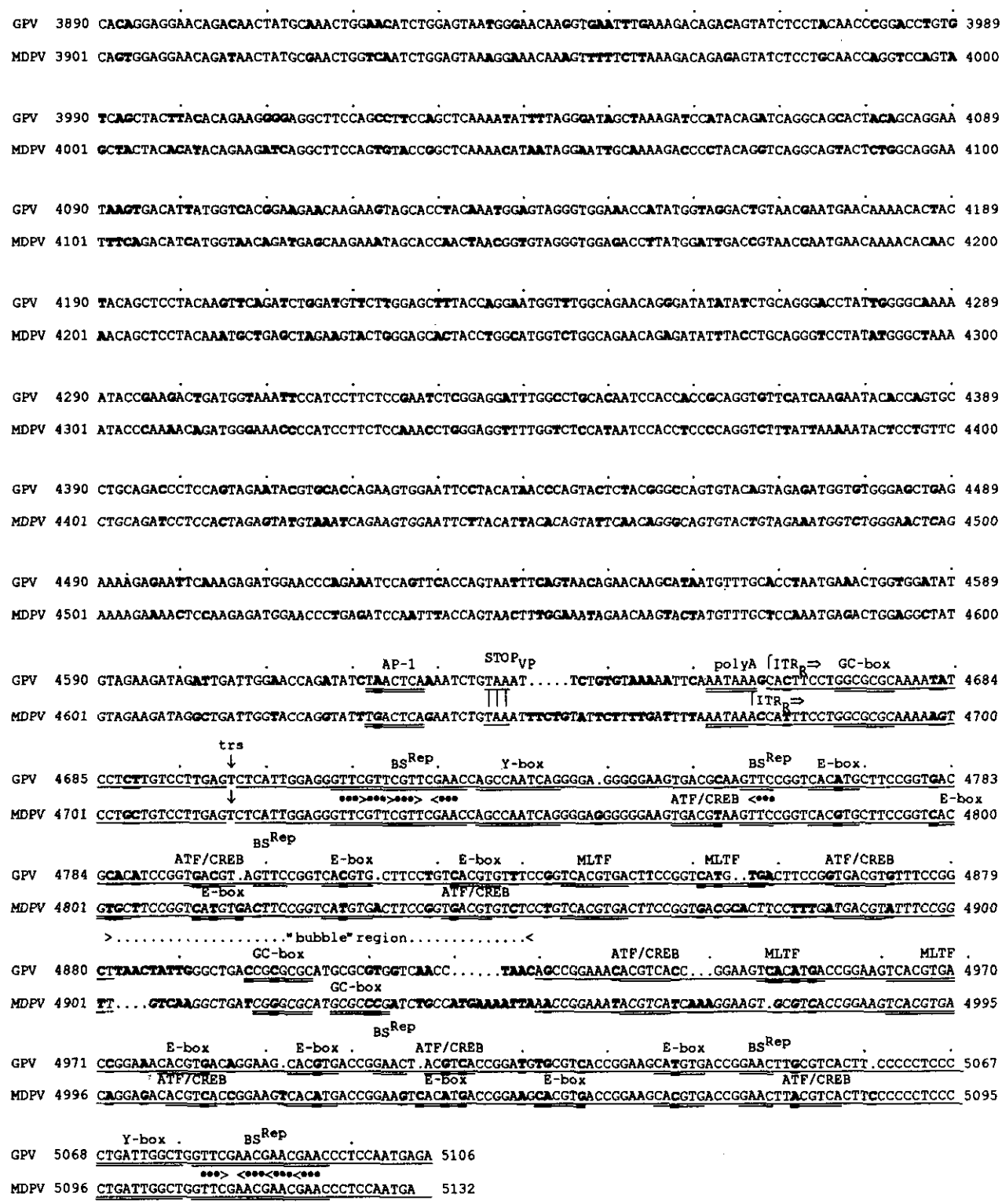

FIG. 1—*Continued*

1 we show the more frequent forms of the ITRs regarding the "flip" and "flop" configurations.

Structure of the terminal palindromes

The sequence analysis revealed that the left and the right ends of the GPV and MDPV genomes consist of identical inverted repeats that are 444 and 457 nucleotides in length, respectively. The distal 407 nucleotides of GPV and 418 nucleotides of MDPV can fold up to form a U-shaped double-stranded hairpin structure. The hairpins of the GPV and MDPV genomes consist of a 181- and a 186-bp stem without mismatches and a "bub-

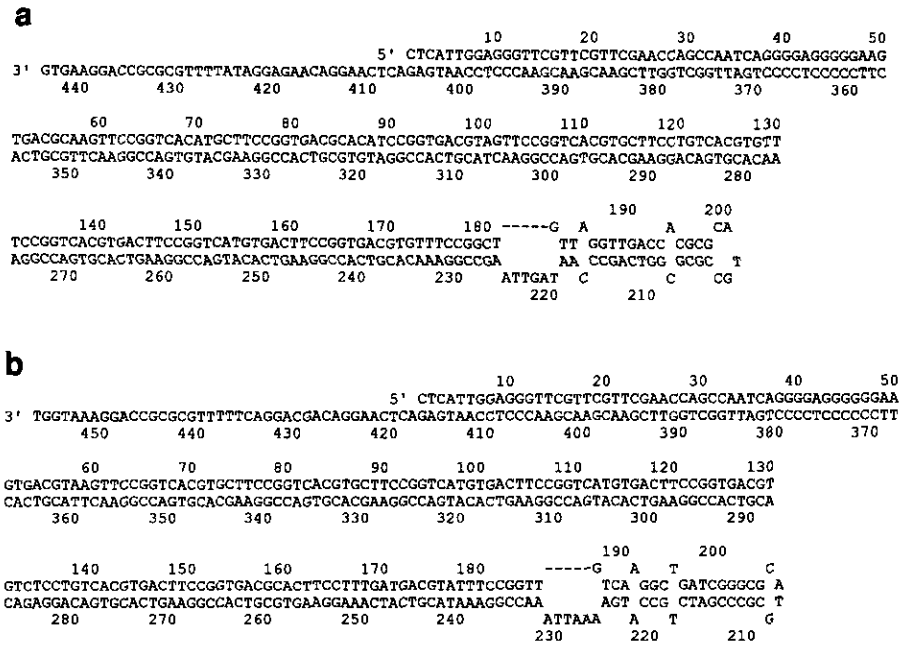

FIG. 2. Structure of the left ITRs of (a) GPV and (b) MDPV genomes. The positive strand is shown.

ble" region (Fig. 1) of 44 and 45 bp, forming an arrow-like structure as drawn in Fig. 2. The restriction endonuclease *Sph*I cuts in the "arrowhead," thereby forming a quasi-symmetrical center. The shape and size of the ITRs of GPV and MDPV are reminiscent of those of human parvovirus B19 (Deiss *et al.*, 1990).

The ITR also contains numerous short repeats of 4–10 nucleotides in length. For example, the heptanucleotide "TTCCGGT" or its degenerated form is repeated 18 times in the ITR of GPV. Many other repeated motifs match the recognition sequences of known transcription factors (Fig. 1).

We identified the putative terminal resolution sites (trs) by analogy with other parvoviruses (Astell, 1990) at nucleotides 407 and 4700 of GPV and nucleotides 418 and 4716 of MDPV (Fig. 1).

Assignment of coding domains

There are two major open reading frames (ORFs) in both the GPV and the MDPV genomes. Interestingly, the two ORFs are located in the same reading frame. A diagram of the stop codons and ATG codon in each of the three reference frames of GPV and MDPV plus-strand DNA is shown in Fig. 3. When the minus-strand DNA was analyzed, like that in all of the vertebrate parvoviruses studied, no ORFs of significant sizes were found.

By analogy with the genomic organizations reported for other parvoviruses, and also by the amino acid homologies observed for the putative proteins, we assign the left ORF of GPV and MDPV to code for the nonstructural proteins, while the right ORF codes for the three capsid proteins VP1–3 (see below).

Analysis and comparison of the putative nonstructural proteins

We compared the amino acid sequences of the largest putative proteins coded by the left ORFs of GPV and MDPV to one another and to the nonstructural proteins of other parvoviruses. We found that the polypeptides encoded by the left ORFs of GPV and MDPV show high similarity and that the amino acid sequences share 90.6% identity. To our surprise, the closest relative of GPV and MDPV among other parvoviruses is the human parvovirus AAV-2. By using the GAP program of the GCG computer program package, we found 61.7% similarity between the left ORF encoded polypeptide of GPV and AAV-2, and the similarity is 61.8% between the same proteins of MDPV and AAV-2. Therefore, we designated the GPV and MDPV nonstructural proteins REP in analogy with the homologous proteins of AAV-2. On the basis of the homology between GPV, MDPV, and AAV-2 (Fig. 4), we assume that there are two potential translational start sites for the REP proteins. The first ATG initiates the largest putative nonstructural protein REP1 at nucleotide 537 of GPV and at nucleotide 548 of MDPV. The second ATG that initiates the putative REP2 polypeptides is located at nucleotide 1065 of GPV and at nucleotide 1076 of MDPV (Fig. 1). The left ORF is terminated with an ocher codon at nucleotide 2418 of GPV and at nucleotide 2429 of MDPV (Fig. 1). We looked for the presence of conserved protein sequence motifs of known function in the putative REP polypeptides of GPV and MDPV. We found the "P loop" ATP/GTP binding motif (GXXXXGK[TS], Saraste *et al.*, 1990), which is also highly conserved among parvovirus and papovavirus polypeptides (Astell *et al.*, 1987). It is worth noting that the amino acid sequence of

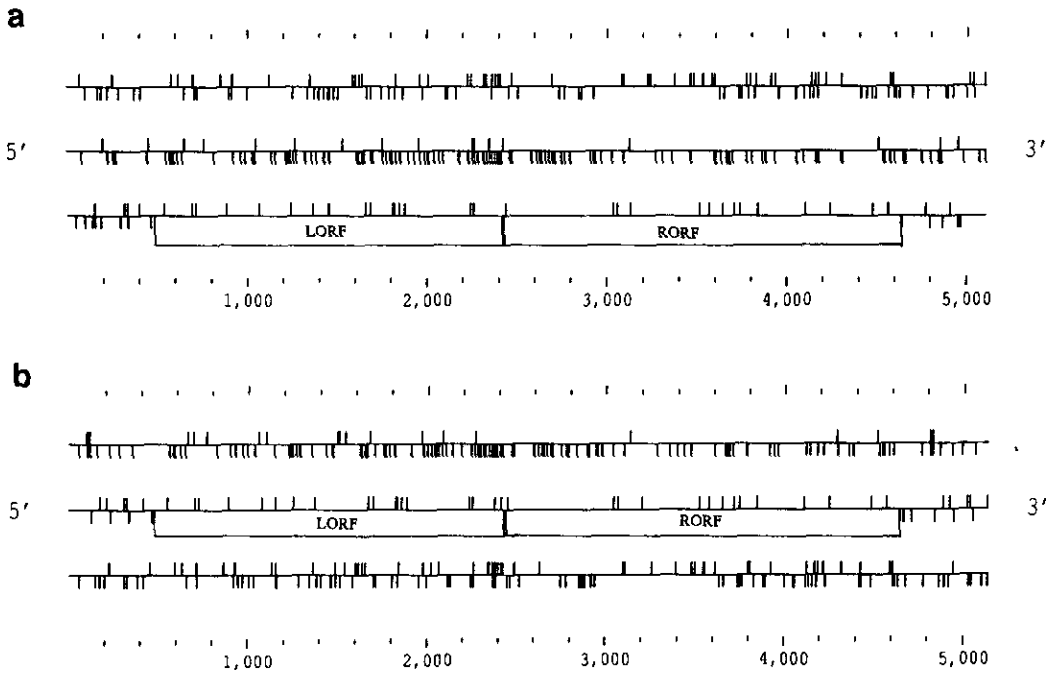

FIG. 3. Diagram of the open reading frames (ORFs) of the genomes of (a) GPV and (b) MDPV. The plus strands are shown. The top vertical bars indicate ATG codons, and the bottom bars indicate stop codons. The left ORFs (LORF) and right ORFs (RORF) are shown by open blocks.

this motif (GPATTGKT) is identical in GPV, MDPV, and AAV-2 (Fig. 4). As shown in Fig. 5, a region spanning an active-site tyrosine, which is implicated in the covalent attachment of the murine parvovirus MVM nonstructural protein to the 5′ end of the genome during replication (Cotmore *et al.*, 1995), is conserved in GPV, in MDPV, and also in each of the vertebrate parvoviruses. A putative zinc binding motif, which was recognized in the car-

```
MDPV   1 MAFSRPLQISSDKFYEVIIRLPSDIDQDVPGLSLNFVEWLSTGVWE..PTGIWNMEHVNLPMVTLADKIKNIFIQRWNQFNQDETDFFF.QLEEGSEYIH  97
 GPV   1 MALSRPLQISSDKFYEVIIRLSSDIDQDVPGLSLNFVEWLSTGVWE..PTGIWNMEHVNLPMVTLAEKIKNIFIQRWNQFNQDETDFFF.QLEEGSEYIH  97
AAV-2  1 MP.........GFYEIVIKVPSDLDGHLPGISDSFVNWVAEKEWELPPDSDMDLNLIEQAPLTVAEKLQRDFLTEWRRVSKAPEALFFVQFEKGESYFH  90

MDPV  98 LHCCIAQGNVRSFVLGRYMSQIKDSILRDVYEGKQVKIPDWFSITKTKR..GGQNKTVTAAYILHYLIPKKQPELQWAFTNMPLFTAAALCLQKRQELLD 195
 GPV  98 LHCCIAQGNVRSFVLGRYMSQIKDSIIRDVYEGKQIKIPDWFAITKTKR..GGQNKTVTAAYILHYLIPKKQPELQWAFTNMPLFTAAALCLQKRQELLD 195
AAV-2 91 MHVLVETTGVESMVLGRFLSQIREKLIQRIYRGIEPTLPNWFAVTKTRNGAGGGNKVVDECYIPNYLLPKTQPELQWAWTNMEQYLSACLNLTERKRLVA 190

MDPV 196 ...AFQESEMGAVVQRDQASTAAPLISNRAAKNYSNLVDWLIEMGITSEKQWLTENKESYRSFQATSSNNRQVKAALENARAEMLLTKTATDYLIGKDPV 292
 GPV 196 ...AFQESDLAAPLPDPQASTVAPLISNRAAKNYSNLVDWLIEMGITSEKQWLTENRESYRSFQATSSNNRQVKAALENARAEMLLTKTATDYLIGKDPV 292
AAV-2 191 QHLTHVSQTQEQNKENQNPNSDAPVIRSKTSARYMELVGWLVDKGITSEKQWIQEDQASYISFNAASNSRSQIKAALDNAGKIMSLTKTAPDYLVGQQPV 290

                                                  ^^^^^^^^
MDPV 293 LDITKNRIYQILKLNNYNPQYVGSVLCGWVKREFNKRNAIWLYGPATTGKTNIAEAIAHAVPFYGCVNWTNENFPFNDCVDKMLIWWEEGKMTNKVVESA 392
 GPV 293 LDITKNRVYQILKMNNYNPQYIGSILCGWVKREFNKRNAIWLYGPATTGKTNIAEAIAHAVPFYGCVNWTNENFPFNDCVDKMLIWWEEGKMTNKVVESA 392
AAV-2 291 EDISSNRIYKILELNGYDPQYAASVFLGWATKKFGKRNTIWLFGPATTGKTNIAEAIAHTVPFYGCVNWTNENFPFNDCVDKMVIWWEEGKMTAKVVESA 390

MDPV 393 KAILGGSAVRVDQKCKGSVCIEPTPVIITSNTDMCMIVDGNSTTMEHRIPLEERMFQIVLSHKLEGNFGKISKKEVKEFFKWANDNLVPVVSEFKVPTNE 492
 GPV 393 KAILGGSAVRVDQKCKGSVCIEPTPVIITSNTDMCMIVDGNSTTMEHRIPLEERMFQIVLSHKLEPSFGKISKKEVKEFFKWANDNLVPVVSEFKVRTNE 492
AAV-2 391 KAILGGSKVRVDQKCKSSAQIDPTPVIVTSNTNMCAVIDGNSTTFEHQQPLQDRMFKFELTRRLDHDFGKVTKQEVKDFFRWAKDHVVEVEHEFYVKKGG 490

MDPV 493 QTKLTEPVPERANEPSEPPKIWAPPTREELEEILRASPELFASVAPLPSSPDTSPKRKKTRGEYQVR..CAMHSLDNSMNVFECLECERANFPEFQSLGE 592
 GPV 493 QTNLPEPVPERANEPEEPPKIWAPPTREELEELLRASPELFSSVAPIPVTPQNSPEPKRSRNNYQVR..CALHTYDNSMDVFECMECERANFPEFQPLGE 592
AAV-2 491 AKKRPAPSDADISEPKRVRESVAQPSTSDAEASINYADRYQNKCSRHVGMNLMLFPCRQCERMNQNSNICFTHGQK......DCLECFPVS..ESQPVSV 590
  B19 594                                          * *         * *  VRGLPVCCVQHINNSGGGLGLCPHCINVGAMYNGVKFR 624
                                                                    * *         * *

MDPV 591 ......NFCNQHGWYD.......CAFCNELKDDMNEIEHVFAIDDMENEQ 627
 GPV 591 ......NYCDEHGWYD.......CAICKELKNELAEIEHVFELDDAENEQ 627
AAV-2 583 VKKAYQKLCYIHHIMG..KVPDACTACD.LVNVD.........LDDCIFEQ 621
  B19 625 EFTPDLVRCSCHVGASNPFSVLTCKKCAYLSGLQSF.......VDYE  671
           * *         * *
```

FIG. 4. Amino acid sequence alignment of REP proteins of GPV, MDPV, and AAV-2 (Srivastava *et al.*, 1983) and the carboxyl terminus of the nonstructural protein of B19 (Shade *et al.*, 1986). The region conserved in every vertebrate parvovirus nonstructural protein is underlined. The amino acids, which are not conserved in the four parvoviruses, are boldface and italic. The NTP binding motif is indicated by . The cysteins and histidins of the putative zinc fingers are marked with *.

```
MVM     193 tkkdy.....vkcvlfgnmiayyfltkkki 217
CPV     195 tkkdy.....tkmvhfgnmiayyfltkkki 219
FPV     195 tkkdy.....tkmvhfgnmiayyfltkkki 219
H-1     193 tkkdy.....vkcvlfgnmiayyflskkki 217
LuIII   193 tkkdy.....vkcvcfgnmvayyfltkkki 217
PPV     194 tkkdy.....vkmthfgnmiayyflnkkrk 218
GPV     142 tktkr..ggqnktvtaa.yilhylipkkqp 168
MDPV    142 tktkr..ggqnktvtaa.yilhylipkkqp 168
AAV2    135 tktrngagggnkvvdec.yipnyllpktqp 163
B19     126 tkgky......frdgeq.fienylm.kkip 147
BPV      63 kcrkangtlvaqaingtefitrymlpknrk  92
                          *
```

FIG. 5. Amino acid sequence alignment of a region of parvovirus nonstructural proteins [minute virus of mice (MVMi), Astell *et al.*, 1986; canine parvovirus (CPV), Parrish *et al.*, 1988; feline panleukopenia virus (FPV), Carlson *et al.*, 1985; H-1, Rhode and Paradiso, 1983; LuIII, Diffoot *et al.*, 1993; porcine parvovirus (PPV), Vasudevacharya *et al.*, 1990; AAV-2, Srivastava *et al.*, 1983; B19, Shade *et al.*, 1986; bovine parvovirus (BPV), Chen *et al.*, 1986] implicated in the covalent attachment to the 5′ end of parvovirus DNA. A conserved tyrosine is indicated by boldface letters and with *. A highly conserved isoleucine is shown by boldface type. The lysine-rich regions flanking the conserved Tyr are boxed.

boxyl terminus of the AAV-2 REP proteins (Carter *et al.*, 1990a), can also be found in the REP polypeptides of GPV, MDPV, and B19 (Fig. 4). It is interesting that all of these viruses posses type A genomes, regarding their terminal hairpins. This amino acid sequence motif is missing from the nonstructural proteins of bovine parvovirus and the parvoviruses belonging to the rat virus-like group.

By comparing the amino acid sequences of parvovirus proteins, we found that B19 and bovine parvovirus are evolutionary closer to GPV and MDPV than they are to the rat virus-like group (data not shown).

Analysis and comparison of the putative capsid proteins

Both GPV and MDPV virions contain three capsid proteins. Le Gall-Reculé and Jestin (1994) reported that the MDPV capsid proteins have apparent molecular masses of 91, 78, and 58 kDa as determined by SDS–PAGE. Our data are somewhat different; we observed by SDS–PAGE three MDPV capsid proteins with molecular masses of 85, 70, and 58 kDa (data not shown). We also determined the apparent molecular masses of the GPV capsid proteins, and we found that they have M_r 87,000, 72,000, and 60,000 (data not shown). The analysis of the right ORF of GPV revealed two ATG codons, from which the translation of the capsid proteins VP1 and VP3 could initiate. The calculated molecular masses for VP1 and VP3 are 81.3 and 60.0 kDa. Thus, by analogy with the translation initiation of the AAV-2 capsid proteins and by the observed 72-kDa molecular mass, we presume that the translation of the capsid protein VP2 initiates from an ACG codon at nucleotide 2874. The calculated molecular mass for GPV VP2 is 65.0 kDa, which is between the M_r of VP1 and VP3. The same is true for MDPV. We found two ATGs, one for VP1 and the other for VP3, but VP2 could initiate only from an ACG triplet at nucleotide 2885. Thus, the calculated molecular masses for MDPV VP1, VP2, and VP3 are 81.3, 65.0, and 60.0 kDa, respectively. The position of the unusual VP2 initiator is conserved in the sequence of the ORF of the capsid proteins of GPV, MDPV, and AAV-2, providing further support for the hypothesis that the VP2 polypeptides of GPV and MDPV start from an ACG triplet (Fig. 6).

We compared the amino acid sequences of the VP1 polypeptides of GPV and MDPV to each other and to the capsid proteins of other parvoviruses. The VP1 polypeptides of GPV and MDPV share 87.7% amino acid sequence identity. Thus, the capsid proteins are more divergent than the nonstructural proteins, possibly due to the evolutionary pressure by the host immune systems. We found that, like in the case of the REP proteins, the VP1 capsid protein of AAV-2 is the closest relative of waterfowl parvovirus capsid proteins (AAV-2 VP1 shows 70.2% amino acid sequence similarity to GPV VP1 and 70.3% amino acid sequence similarity to MDPV VP1).

Tsao *et al.* (1991) have determined the three-dimensional structure of canine parvovirus (CPV) by X-ray crystallography. By using computer-aided sequence comparison and the three-dimensional structure of the CPV capsid, Chapman and Rossman (1993) have aligned the capsid protein amino acid sequences of several parvoviruses, among them that of AAV-2. By using this alignment, the probable regions of the AAV-2 capsid protein that are exposed to the surface can be predicted. The homologous regions of the VP3 polypeptides of GPV and MDPV, which are supposedly exposed to the surface of the capsid, are more divergent from each other than from the rest of the molecule (Fig. 6). The largest divergence between the GPV and MDPV capsid polypeptides is located between amino acids 162 and 178, that is, between the start codons of VP2 and VP3 (Fig. 6). We also recognized the conserved "PGY" motif (Chen *et al.*, 1986; Dumas *et al.*, 1992) in both GPV and MDPV VP1 at amino acid 56.

Potential regulatory elements

A search for TATA boxes in the GPV and MDPV genomes revealed three potential promoters designated p9, p19, and p41 (see Fig. 1). They precede the start codons of polypeptides REP1, REP2, and VP1, respectively. The p41 promoter of GPV is atypical to some extent, having the nucleotide sequence TTTTAA. We still presume that it could be functional, because its location is analogous to that of the p40 of AAV-2 (Rhode and Iversen, 1990; Carter *et al.*, 1990a). There is a TATA box-like element starting at nucleotide 2732 in the GPV genome and at nucleotide 2743 in the MDPV genome. Whether this TATA box is functional we do not know, especially because it is located downstream of the first ATG of the right ORF.

We identified polyadenylation signals just before the right ITRs of GPV and MDPV at nucleotides 4655 and 4671, respectively. There is also a sequence matching

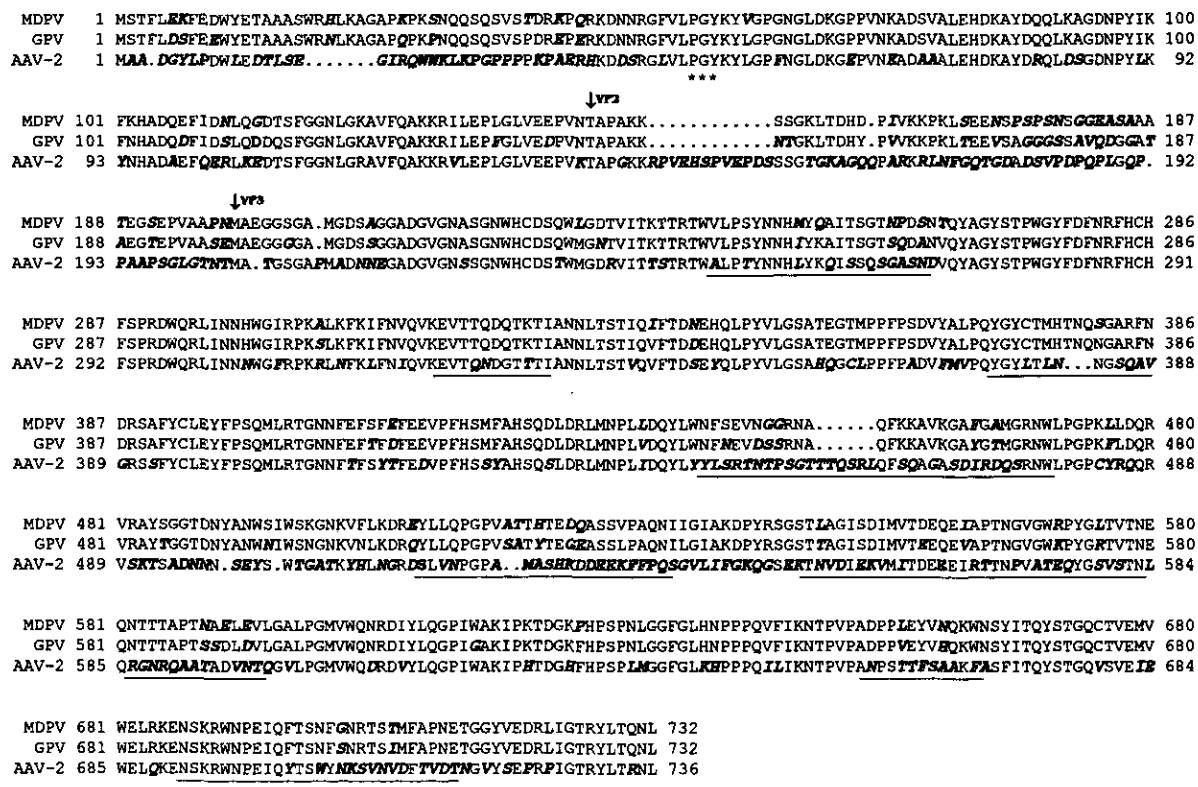

FIG. 6. Amino acid sequence alignment of the VP proteins of GPV, MDPV, and AAV-2. The reported sequence of the capsid gene of AAV-2 was corrected according to Trempe and Carter (1988) and Ruffing *et al.* (1994). The first amino acids of VP2 and VP3 are indicated by arrows. The amino acids, which are different in the three-capsid proteins, are boldface and italic. The conserved motif "PGY" is marked by *. The regions supposedly exposed to the surface of the capsid are underlined.

the consensus polyadenylation signal starting at nucleotide 1703 in the MDPV genome that is missing from the GPV chromosome. We found a potential splice donor site, which is in analogous position with that of AAV-2 (Trempe and Carter, 1988), at nucleotide 2207 of GPV and nucleotide 2218 of MDPV. Two alternative splice acceptor sites can be found at nucleotides 2423 and 2455 of GPV and nucleotides 2434 and 2466 of MDPV (Fig. 1).

We made a computer search for the recognition sequences of known transcription factors in the genomes of GPV and MDPV. These motifs for DNA binding proteins can be found at highest densities in the ITRs (see Fig. 1), implicating that they might have some role in regulating the transcription or replication of the GPV and MDPV genomes.

DISCUSSION

As intact virus can be rescued from a recombinant plasmid carrying a full-length clone of the MDPV nucleotide sequence shown in Fig. 1, the sequence appears to represent the biologically fully active genome (Zádori *et al.*, manuscript in preparation). The nucleotide sequence of GPV also seems complete, by comparing it to that of MDPV.

There is some heterogeneity at the termini of the genomes. We observed that in our sequence the right ITR is one nucleotide longer (GPV) or shorter (MDPV) relative to the left ITR. This heterogeneity at the end bases could be due to cloning artifacts or it may result from a flexible terminal resolution site. A 1- to 2-base heterogeneity at either end of the AAV-2 genome has been described previously (Fife *et al.*, 1977). It has also been reported that the site where the NS1 protein of MVM cuts its substrate can be flexible *in vitro* (Cotmore and Tattersall, 1994). Minor products of the *in vitro* trs endonuclease reaction carried out by the purified REP78 (Im and Muzyczka, 1992; Owens and Carter, 1992; Chiorini *et al.*, 1994b) and REP68 proteins (Snyder *et al.*, 1993) can also be observed; however, these minor products may have been the result of some contaminations.

While other authors experienced difficulties in sequencing the ITRs of parvoviruses B19 and JcDNV because of GC compression (Deiss *et al.*, 1990; Dumas *et al.*, 1992), we found that interconversion between the "flip" and "flop" orientations can also cause a problem in obtaining correct sequence data. The observation that "flip-flop" interconversion can occur even in bacterial cells which are deficient in the major recombination genes indicates that the ITRs are strong recombination hot spots. It also warns that caution should be taken while propagating similar parvovirus DNA ends in bacteria, because false results could be obtained if one disre-

gards the possible different biological activities of the ITRs with "flip" and "flop" orientation, for example, in an *in vitro* replication assay. This "flip-flop" interconversion in bacterial cells is indeed artifactual, but it can take place during the replication of a double-stranded circular plasmid, thereby suggesting that "hairpin transfer" type replication is not essential for this process. Probably recombination can also produce ITRs with "flip" and "flop" orientation during parvovirus replication in the native host.

The NS1 polypeptide binds to a $[ACCA]_2$ motif 17 nucleotides apart from its endonucleotic cleavage site (Cotmore *et al.*, 1995). The same sequence motif in the left ITR in similar positions can be found in parvovirus genomes of the members of the rat group-like parvoviruses (see the ITR sequences in Astell, 1990). The AAV-2 REP68 protein binds to a GCTC repeat motif 16 nucleotides apart from the trs (Chiorini *et al.*, 1994a; McCarty *et al.*, 1994). There is a similar tetranucleotide repeat motif 12 nucleotides apart from the putative trs of the GPV and MDPV genomes (Fig. 1). Taking into account the high similarity among the GPV, MDPV, and AAV-2 REP proteins and the conserved position of the tetranucleotide motifs, it can be hypothesized that the REP1 proteins of GPV and MDPV bind to this GTTC-containing repeat (BS^{Rep}, Fig. 1). An experimental test is required to substantiate this hypothesis.

By comparison to the consensus splice acceptor site, acceptor site 2 (see Fig. 1) is in a more favorable context than acceptor site 1 (Alberts *et al.*, 1989). The possible translation products of the processed mRNA spliced to acceptor site 2 are VP2 and VP3. As VP3 is the major capsid protein of GPV and MDPV, it is likely that splice acceptor site 2 is indeed preferred *in vivo*, in a similar manner as in the case of AAV-2 (Trempe and Carter, 1988).

Our search for transcription factor binding sites has revealed several interesting motifs. We found among others sequences matching E box and Y box elements, as well binding sites of transcription factors MLTF and ATF/CREB (Anthony-Cahill *et al.*, 1992; Mitchell and Tjian, 1989; Chodosh *et al.*, 1986). The protein factors recognizing these elements have been implicated in the regulation of transcription or genomic replication of the parvoviruses AAV-2 (Chang *et al.*, 1989) and MVM (Gu *et al.*, 1995; Cotmore and Tattersall, 1994). We have obtained preliminary data that the MDPV left ITR is able to activate transcription in a *Saccharomyces cerevisiae* transcription assay system *in vivo*, indicating that the factors binding to this region are evolutionary conserved (R. Stefancsik, unpublished observations). As the E boxes, Y boxes, MLTF, and ATF/CREB recognition sequences lie in the ITRs, there might also be cellular factors recognizing these elements in the native host of GPV and MDPV.

The high similarity between the GPV and MDPV genomes suggests that they did not diverge from their common ancestor long ago on an evolutionary time scale. The density of nucleotide differences show peaks in the p9 and the "bubble" region. This could be due to differences in the cellular factors which regulate the life cycles of the two viruses.

Our results demonstrate that the closest relative of GPV and MDPV is the human parvovirus AAV-2. This is rather surprising, because neither GPV nor MDPV requires helper virus for effective replication. Our data suggest that AAV-2 may have evolved from an ancestral virus which replicated helper independently. The observations that in special circumstances AAV-2 can replicate without helper virus provides further support to this hypothesis (Yakobson *et al.*, 1987).

As the evolutionary divergence is smaller between AAV-2 and the waterfowl parvoviruses than between, e.g., AAV-2 and human parvovirus B19, the data presented in this paper limit the value of the hypothesis of Bando *et al.* (1987), who proposed that parvoviruses have evolved in a host-dependent manner. One must take into account the possibility of vertical transfer of parvoviruses between different host species when analyzing evolutionary relationships. Thus, some parvoviruses can be closer to each other on a phylogenetic tree than their respective hosts. Our protein comparison data also suggest that Astell's (1990) classification of parvoviruses based on the properties of the terminal hairpins may represent an evolutionary relationship among the viruses falling into the same category; therefore, it would be useful to incorporate these categories into the current taxonomy of parvoviruses.

REFERENCES

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. (1989). "Molecular Biology of the Cell," 2nd ed. Garland, New York/London.

Anthony-Cahill, S. J., Benfield, P. A., Fairman, R., Wasserman, Z., Brenner, S. L., Stafford, W. F., Altenbach, C., Hubbel, W. L., and DeGrado, W. F. (1992). Molecular characterization of helix–loop–helix peptides. *Science* **255,** 979–983.

Astell, C. R., Gardiner, E. M., and Tattersall, P. (1986). DNA sequence of the lymphotropic variant of minute virus of mice, MVM(i), and comparison with the DNA sequence of the fibrotropic strain. *J. Virol.* **57,** 656–669.

Astell, C. R., Mol, C. D., and Anderson, W. F. (1987). Structural and functional homology of parvovirus and papovavirus polypeptides. *J. Gen. Virol.* **68,** 885–893.

Astell, C. R. (1990). Terminal hairpins of parvovirus genomes and their role in DNA replication. *In* "Handbook of Parvoviruses" (P. Tijssen, Ed.), Vol. 1, pp. 59–79. CRC Press, Boca Raton, FL.

Bando, H., Kusuda, J., Gojobori, T., Maruyama, T., and Kawase, S. (1987). Organization and nucleotide sequence of a densovirus genome imply a host-dependent evolution of the parvoviruses. *J. Virol.* **61,** 553–560.

Berns, K. I. (1990). Parvoviridae and their replication. *In* "Virology" (B. M. Fields and D. M. Knipe, Eds.), 2nd ed., pp. 1743–1763. Raven Press, New York.

Carlson, J., Rushlow, K., Maxwell, I., Maxwell, F., Winston, S., and Hahn, W. (1985). Cloning and sequence of DNA encoding structural proteins of the autonomous parvovirus feline panleukopenia virus. *J. Virol.* **55,** 574–582.

Carter, B. J. (1990). Adeno-associated virus helper functions. *In* "Handbook of Parvoviruses" (P. Tijssen, Ed.), Vol. 1, pp. 255–282. CRC Press, Boca Raton, FL.

Carter, B. J., Trempe, J. P., and Mendelson, E. (1990a). Adeno-associated virus gene expression and regulation. *In* "Handbook of Parvoviruses" (P. Tijssen, Ed.), Vol. 1, pp. 227–254. CRC Press, Boca Raton, FL.

Carter, B. J., Mendelson, E., and Trempe, J. P. (1990b). AAV DNA replication, integration, and genetics. *In* "Handbook of Parvoviruses" (P. Tijssen, Ed.), Vol. 1, pp. 169–226. CRC Press, Boca Raton, FL.

Chang, L., Shi, Y., and Shenk, T. (1989). Adeno-associated Virus P5 promoter contains an adenovirus E1A inducible element and a binding site for the major late transcription factor. *J. Virol.* **63,** 3479–3488.

Chapman, M. S., and Rossmann, M. G. (1993). Structure, sequence and function correlations among parvoviruses. *Virology* **194,** 491–508.

Chen, K. C., Shull, B. C., Moses, E. A., Lederman, M., Stout, E. R., and Bates, R. C. (1986). Complete nucleotide sequence and genome organization of bovine parvovirus. *J. Virol.* **60,** 1085–1097.

Chiorini, J. A., Wiener, S. M., Owens, R. A., Kyöstiö, S. R. M., Kotin, R. M., and Safer, B. (1994a). Sequence requirements for stable binding and function of Rep68 on the adeno-associated virus type 2 inverted terminal repeats. *J. Virol.* **68,** 7448–7457.

Chiorini, J. A., Weitzman, M. D., Owens, R. A., Urcelay, E., Safer, B., and Kotin, R. M. (1994b). Biologically active Rep proteins of adeno-associated virus type 2 produced as fusion proteins in *Escherichia coli*. *J. Virol.* **68,** 797–804.

Chodosh, L. A., Carthew, R. W., and Sharp, P. A. (1986). A single polypeptide possesses the binding and transcription activities of the adenovirus major late transcription factor. *Mol. Cell. Biol.* **6,** 4723–4733.

Cotmore, S., and Tattersall, P. (1994). An asymmetric nucleotide in the parvoviral 3′ hairpin directs segregation of a single active origin of DNA replication. *EMBO J.* **13,** 4145–4152.

Cotmore, S. F., Christensen, J., Nüesch, J. P. F., and Tattersall, P. (1995). The NS1 polypeptide of the murine parvovirus minute virus of mice binds to DNA sequences containing the motif $[ACCA]_{2-3}$. *J. Virol.* **69,** 1652–1660.

Deiss, V., Tratschin, J. D., Weitz, M., and Siegl, G. (1990). Cloning of the human parvovirus B19 genome and structural analysis of its palindromic termini. *Virology* **175,** 247–254.

Derzsy, D. (1967). A viral disease of goslings. *Acta Vet. Acad. Sci. Hung.* **17,** 443–448.

Diffoot, N., Chen, K. C., Bates, R. C., and Lederman, M. (1993). The complete nucleotide sequence of parvovirus LuIII and localization of a unique sequence possibly responsible for its encapsidation pattern. *Virology* **192,** 339–345.

Dumas, B., Jourdan, M., Pascaud, A., and Bergoin, M. (1992). Complete nucleotide sequence of the cloned infectious genome of *Junonia coenia* densovirus reveals an organization unique among parvoviruses. *Virology* **191,** 202–222.

Fife, K. H., Murray, K., and Berns, K. I. (1977). Structure and nucleotide sequence of the terminal regions of adeno-associated virus DNA. *Virology* **78,** 475–487.

Ghosh, D. (1990). A relational database of transcription factors. *Nucleic Acids Res.* **18,** 1749–1756.

Gu, Z., Plaza, S., Perros, M., Cziepluch, C., Rommelaere, J., and Cornelis, J. J. (1995). NF-Y controls transcription of the minute virus of mice P4 promoter through interaction with an unusual binding site. *J. Virol.* **69,** 239–246.

Im, D., and Muzyczka, N. (1992). Partial purification of adeno-associated virus Rep78, Rep52 and Rep40 and their biochemical characterization. *J. Virol.* **66,** 1119–1128.

Kisary, J., and Derzsy, D. (1974). A viral disease of goslings IV. Characterization of the causal agent in tissue culture system. *Acta Vet. Acad. Sci. Hung.* **24,** 287–292.

Le Gall-Reculé, G., and Jestin, V. (1994). Biochemical and genomic characterization of muscovy duck parvovirus. *Arch. Virol.* **139,** 121–131.

McCarty, D. M., Pereira, D. J., Zolotukhin, I., Zhou, X., Ryan, J. H., and Muzyczka, N. (1994). Identification of linear DNA sequences that specifically bind the adeno-associated virus Rep protein. *J. Virol.* **68,** 4988–4997.

Mitchell, P. J., and Tjian, R. (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245,** 371–378.

Owens, R. A., and Carter, B. J. (1992). *In vitro* resolution of adeno-associated virus DNA hairpin termini by wild-type Rep protein is inhibited by a dominant-negative mutant of Rep. *J. Virol.* **66,** 1236–1240.

Ranz, A. I., Manclus, J. J., Diaz-Aroca, E., and Casal, J. I. (1989). Porcine parvovirus: DNA sequence and genome organization. *J. Gen. Virol.* **70,** 2541–2553.

Reed, A. P., Jones, E. V., and Miller, T. J. (1988). Nucleotide sequence and genome organization of canine parvovirus. *J. Virol.* **62,** 266–276.

Rhode, S. L., III, and Paradiso, P. R. (1983). Parvovirus genome: Nucleotide sequence of H1 and mapping of its genes by hybrid arrested translation. *J. Virol.* **45,** 173–184.

Rhode, S. L., and Iversen, P. (1990). Parvovirus genomes: DNA sequences. *In* "Handbook of Parvoviruses" (P. Tijssen, Ed.), Vol. 1, pp. 31–57. CRC Press, Boca Raton, FL.

Ruffing, M., Heid, H., and Kleinschmidt, J. A. (1994). Mutations in the carboxy terminus of adeno-associated viru 2 capsid proteins affect viral infectivity: Lack of an RGD integrin-binding motif. *J. Gen. Virol.* **75,** 3385–3392.

Saraste, M., Sibbald, P. R., and Wittinghofer, A. (1990). The P-loop: A common motif in ATP- and GTP-binding proteins. *TIBS* **15,** 430–434.

Shade, R. O., Blundell, M. C., Cotmore, S. F., Tattersall, P., and Astell, C. R. (1986). Nucleotide sequence and genome organization of human parvovirus B19 isolated from the serum of child during aplastic crisis. *J. Virol.* **58,** 921–936.

Snyder, R. O., Im, D., Ni, T., Xiao, X., Samulski, R. J., and Muzyczka, N. (1993). Features of the adeno-associated virus origin involved in substrate recognition by the viral rep protein. *J. Virol.* **67,** 6096–6104.

Srivastava, A., Lusby, E. W., and Berns, K. I. (1983). Nucleotide sequence and organization of the adeno-associated virus 2 Genome. *J. Virol.* **45,** 555–564.

Tattersall, P., and Cotmore, S. F. (1990). Reproduction of autonomous parvovirus DNA. *In* "Handbook of Parvoviruses" (P. Tijssen, Ed.), Vol. 1, pp. 123–140. CRC Press, Boca Raton, FL.

Tattersall, P., and Gardiner, E. M. (1990). Autonomous parvovirus–host–cell interactions. *In* "Handbook of Parvoviruses" (P. Tijssen, Ed.), Vol. 1, pp. 111–121. CRC Press, Boca Raton, FL.

Tijssen, P. (1990). Introduction. *In* "Handbook of Parvoviruses" (P. Tijssen, Ed.), Vol. 1, pp. 3–7. CRC Press, Boca Raton, FL.

Trempe, J. P., and Carter, B. J. (1988). Alternate mRNA splicing is required for synthesis of adeno-associated virus VP1 capsid protein. *J. Virol.* **62,** 3356–3363.

Tsao, J., Chapman, M. S., Agbandje, M., Keller, W., Smith, K., Wu, H., Luo, M., Smith, T. J., Rossmann, M. G., Compans, R. W., and Parrish, C. R. (1991). The three-dimensional structure of canine parvovirus and its functional implications. *Science* **251,** 1456–1464.

Vasudevacharya, J., Basak, S., Srinivas, R. V., and Compans, R. W. (1990). The complete nucleotide sequence of an infectious clone of porcine parvovirus, strain NADL-2. *Virology* **178,** 611–616.

Yakobson, B., Koch, T., and Winocour, E. (1987). Replication of adeno-associated virus in synchronized cells without the addition of a helper virus. *J. Virol.* **61,** 972–981.

Zádori, Z., Erdei, J., Nagy, J., and Kisary, J. (1994). Characteristics of the genome of goose parvovirus. *Avian Pathol.* **23,** 359–364.