JOURNAL OF VIROLOGY, June 1986, p. 921–936
0022-538X/86/060921-16$02.00/0

Vol. 58, No. 3

Nucleotide Sequence and Genome Organization of Human Parvovirus B19 Isolated from the Serum of a Child during Aplastic Crisis

ROSEMARY O. SHADE,[1] MATTHEW C. BLUNDELL,[1] SUSAN F. COTMORE,[2] PETER TATTERSALL,[2] AND CAROLINE R. ASTELL[1]*

Department of Biochemistry, Faculty of Medicine, University of British Columbia, Vancouver, British Columbia, V6T 1W5 Canada,[1] and Departments of Laboratory Medicine and Human Genetics, Yale University, New Haven, Connecticut 06510[2]

Received 19 November 1985/Accepted 4 February 1986

The nucleotide sequence of an almost-full-length clone of human parvovirus B19 was determined. Whereas the extreme left and right ends of this genomic clone are incomplete, the sequence clearly indicates that the two ends of viral DNA are related by inverted terminal repeats similar to those of the *Dependovirus* genus. The coding regions are complete in the cloned DNA, and the two large open reading frames which span almost the entire genome are restricted to one strand, as has been found for all other parvoviruses characterized to date. From the DNA sequence we conclude that the organization of the B19 transcription units is similar although not identical to those of other parvoviruses. In particular, we predict that the B19 genome may utilize a fourth promoter to transcribe mRNA encoding the major structural polypeptide, VP2. Analysis of the putative polypeptides confirms that B19 is only distantly related to the other parvoviruses but reveals that there is a small region in the gene probably encoding the major nonstructural protein of B19, which is closely conserved between all of the parvovirus genomes for which sequence information is currently available.

The *Parvoviridae* are a family of small DNA viruses with single-stranded, linear genomes approximately 5,000 nucleotides (nt) in length (46). Viruses of this family which infect vertebrates fall into two distinct groups. The adeno-associated viruses (AAV), also called dependoviruses, require a helper adenovirus or herpesvirus to complete their replication cycle. Members of the autonomously replicating group of parvoviruses, although capable of independent replication without the aid of a helper virus, require the host cell to exhibit a highly specific physiological and differentiated phenotype. To date, three serotypes of AAV have been described which infect human populations, but none of these are associated with human disease (7, 21). Until recently, autonomously replicating parvoviruses had only been isolated from rodents and domestic animals. These viruses are often highly pathogenic, causing lytic destruction of mitotically active cells of particular differentated types and thus inducing diseases involving rapidly proliferating tissues (such as the gut epithelium and lymphoid system) in adults (45), whereas transplacental infections of fetuses with viruses of this subgroup typically cause resorption, abortion, or teratogenesis.

The human virus B19 (previously known as serum parvoviruslike virus) has recently been definitively characterized as a parvovirus (17, 49), but DNA hybridization studies showed that, although it is not detectably related to the dependoviruses, it is only very distantly related to some other autonomously replicating members of the family (17). Although the pathogenicity of this virus was unknown at the time of the initial report (15), B19 is now known to be the causative agent of erythema infectiosum (fifth disease; 1; M. J. Anderson, S. E. Jones, S. P. Fisher-Hoch, E. Lewis, S. M. Hall, C. L. Bartlett, B. J. Cohen, P. P. Mortimer, and M. S. Perevia, Letter, Lancet i:1378, 1983) and of a type of reticulocytopenic aplastic crisis associated with several types of hemolytic anemia (e.g., homozygous sickle cell anemia, hereditary spherocytosis, pyruvate kinase deficiency, and β-thalassemia; 19, 28, 29, 32, 34, 44). Recently, B19 virus has also been shown to cause transient postinfection arthropathy in adults (35, 51) and to cross the placentas of women infected during pregnancy (11), eliciting an immune response in the developing fetuses (23). Clearly, B19 exhibits the pathogenic potential characteristically associated with an autonomous parvovirus, a proposition which is supported by the recent observation that some cells in erythropoietin-stimulated human bone marrow cultures appear to support the replication of B19 in vitro (N. S. Young, personal communication).

To learn more about the genetic strategy and evolution of this virus and its relationship to the other parvoviruses, we determined the DNA sequences of two distinct B19 viruses isolated 10 years apart, which we molecularly cloned into plasmid vectors. The first of these sequences, presented in this paper, is of a virus designated B19-Au and obtained from the serum of a child with homozygous sickle cell disease who was in the early phase of reticulocytopenic aplastic crisis. The clone which was sequenced contained an almost-full-length viral genome but lacked small regions from the viral termini which, by analogy with the other parvoviruses, are likely to be essential for self-priming during DNA replication. Thus, the clone (pYT103) analyzed in this paper is probably not an infectious form of the virus but does contain all of its coding sequences.

(A preliminary report of these data was presented at the EMBO Workshop on Parvoviruses, Grangeneuve, Switzerland, September 1985.)

* Corresponding author.

MATERIALS AND METHODS

Enzymes and chemicals. Restriction endonucleases, T4 DNA ligases, and T4 polynucleotide kinase were purchased from Bethesda Research Laboratories, Inc. DNA polymerase I (Klenow fragment) was obtained from Promega Biotec, and exonuclease III and linkers were from New England BioLabs, Inc. S1 nuclease was purchased from PL-Pharmacia. α-^{32}P-deoxyribonucleoside triphosphates (dNTPs) were obtained from Amersham Corp. α-Thio-dNTPs (αS-dNTPs) were supplied by New England Nuclear Corp. NACS-52 columns were obtained from Bethesda Research Laboratories.

Cloned viral DNA. Virus was obtained from the serum of a child with homozygous sickle cell disease during the early stages of reticulocytopenic aplastic crisis. Clinical details and a more detailed description of virus isolation and construction of the almost-full-length clone will be presented elsewhere (S. F. Cotmore et al., manuscript in preparation). Briefly, the procedures used parallel those described previously for the cloning of the B19-Wi isolate (17). The duplex viral termini were used to prime the synthesis of complementary strands by the Klenow fragment of *Escherichia coli* DNA polymerase I. The genome was cloned in two parts using an internal cut at the single *Bam*HI site in the virus and oligonucleotide linkers attached to the viral termini. In this instance, *Eco*RI linkers were used in preference to the *Bam*HI linkers used in cloning the B19-Wi genome (17).

Restriction endonuclease mapping of the viral clone. To facilitate mapping of the genome, the *Eco*RI insert from pYT103 was digested with *Bam*HI (map unit 76), and the two fragments of 3.9 and 1.4 kilobases (kb) were mapped separately by the method of Smith and Birnstiel (47). Restriction endonucleases used were *Ava*I, *Bam*HI, *Bgl*II, *Cla*I, *Eco*RI, *Eco*RV, *Hin*dIII, *Pst*I, *Pvu*I, *Pvu*II, *Sal*I, *Sst*I, *Sst*II, *Xba*I, and *Xho*I.

Sequence determination. To facilitate sequencing the genome, the major portion of the sequence of both the left-end (3.9-kb) *Eco*RI-*Bam*HI and right-end (1.4-kb) *Bam*HI-*Eco*RI fragments were recloned, and a series of nested deletions was constructed as described below.

Sequence of the right-end (1.4-kb) fragment. The 1.4-kb *Bam*HI-*Eco*RI fragment (map units 76 to 100 [see Fig. 1A]) was recloned into pUC13 and transformed into JM101 cells. Plasmid DNA from several white colonies was isolated by the mini-alkaline lysis method (26), characterized by digestion with *Bam*HI and *Eco*RI, and subsequently transformed into DH1 cells for large-scale (1 liter) growth. Plasmid DNA was purified by the alkaline lysis procedure followed by CsCl density gradient centrifugation (26). Plasmid DNA (30 μg) designated pUC13-1.4 (see Fig. 1B) was digested with *Pst*I and *Bam*HI. After phenol-$CHCl_3$ extraction and ethanol precipitation, DNA was digested with exonuclease III for 0 to 3 min (aliquots were removed at 0.5-min intervals) followed by S1 nuclease to remove the single-stranded tail. After phenol-$CHCl_3$ extraction and ethanol precipitation, DNA was incubated with DNA polymerase I (Klenow fragment) and all four dNTPs to ensure that the molecules were blunt ended. DNA was ligated overnight at 14°C and transformed into DH1 cells. Colonies from time points 0 to 3 min were grown (1.5-ml cultures), and plasmid DNA was isolated. DNA was digested with restriction enzymes (*Eco*RI and *Hin*dIII) and analyzed on 0.7% agarose gels to size the deletions. DNA from appropriate clones was further purified on NACS-52 columns as described by the manufacturer. Each clone was digested with *Eco*RI to linearize the DNA and sequenced by the enzymatic method (42) with the universal reverse primer. To obtain the end sequence (nearest the *Eco*RI linker site at nt 5107), one clone was digested with *Hin*dIII and sequenced with universal forward primer. The concentration of primer used in enzymatic sequencing of double-stranded DNA clones was 10-fold that used with single-stranded (M13) templates.

Sequence of the left-end (3.9-kb) fragment. Attempts to clone the 3.9-kb *Eco*RI-*Bam*HI fragment into M13 vectors (host cells, JM101) repeatedly met with failure. We also were unsuccessful in many attempts to clone this fragment into double-stranded pUC vectors by using JM101 or DH1 cells. Eventually the fragment was cloned into pUC13 by using JM101 cells. However, large-scale preparations of this clone could not be achieved in these cells. Hence, the DNA from a mini-alkaline lysis preparation was transformed into DH1 cells, where the plasmid appeared to be stable. Because of the initial difficulty in cloning and amplifying the plasmid, the insert was carefully mapped with several restriction endonucleases to ensure that it had not undergone rearrangement during the cloning procedure. Indeed, the 3.9-kb insert corresponds to that fragment in the original clone (pYT103) except for the extreme left-end *Eco*RI-*Hin*dIII fragment (map units 1 to 12). This fragment contains the left hairpin region, and in our pUC13 clone this region had undergone a deletion of approximately 130 base pairs. The pUC plasmid containing the 3.9-kb fragment was further altered to destroy the *Xba*I site in the multiple cloning region of the vector. This was accomplished by partial *Xba*I digestion and purification of the linear (6.6-kb) DNA on a 0.7% agarose gel. DNA, blunt-ended with DNA polymerase I (Klenow fragment) and all four dNTPs, was ligated with T4 DNA ligase and transformed into DH1 cells. Twelve clones were grown, and the plasmid DNAs were characterized by restriction digestion to identify a clone which retained the *Xba*I site at nt 477 of the insert but had lost the *Xba*I site in the multiple cloning region. This clone was designated pUC13-3.9 (see Fig. 1B) and used for sequencing.

The sequence of pUC13-3.9 was obtained by generating a series of nested deletions essentially as described above with the exception of the initial steps. DNA (15 μg) was digested with *Eco*RI, and the recessed ends were filled in with αS-dATP by using DNA polymerase I (Klenow fragment). DNA, after ammonium acetate-ethanol precipitation, was digested with *Xba*I. Following another ethanol precipitation step, the DNA was redissolved (0.25 μg/μl), and a series of nested deletion clones was constructed as described above except that exonuclease III aliquots were taken every minute for 0 to 14 min. DNAs from appropriate plasmid clones were purified on NACS-52 columns, linearized by digestion with *Bam*HI, and sequenced using the universal forward primer (10-fold concentration of that used with single-stranded DNA).

The sequence of the complementary strand was obtained by cloning the 3.9-kb *Eco*RI-*Bam*HI fragment from pUC13-3.9 into M13mp10. The host cells used, DH21 cells, are DH1 cells which contain an F′, although they lack the color selection of the JM series of cells. The M13 clone obtained was designated M13mp10-3.9 (see Fig. 1B). The sequence of M13mp10-3.9 was obtained by constructing a series of nested deletions similar to that described above for pUC13-3.9. M13mp10-3.9 replicative-form DNA was digested with *Sal*I, and the recessed ends were filled in with αS-dTTP by using DNA polymerase I (Klenow fragment). To get complete protection of the *Sal*I site from exonuclease III, we found it necessary to add dGTP to the filling-in

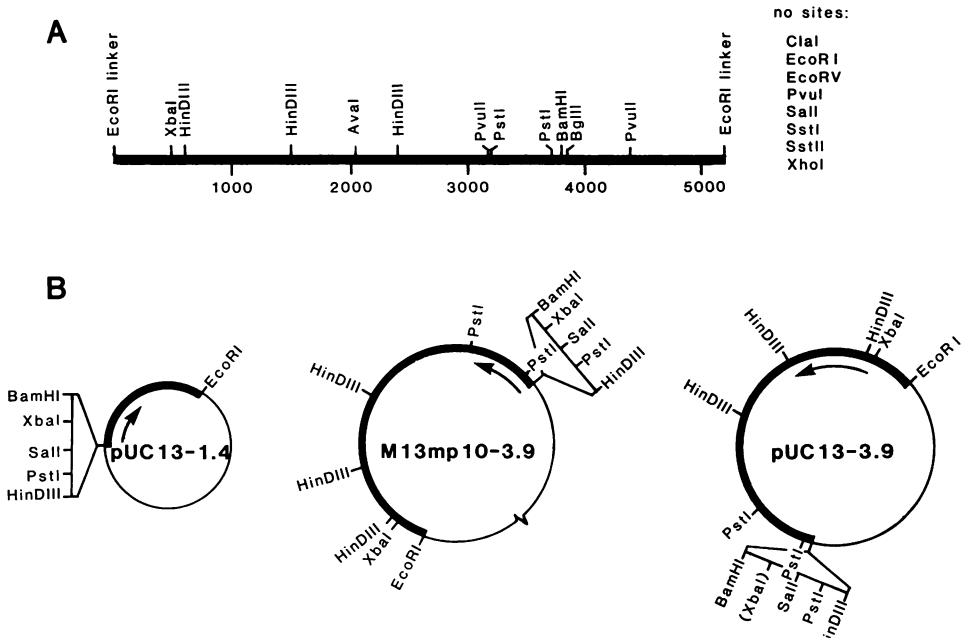


FIG. 1. (A) Restriction endonuclease map of the B19-Au genome cloned in plasmid pYT103. To sequence the genome by creating a series of nested deletions, the *Bam*HI and *Eco*RI fragments (1.4 and 3.9 kb) were cloned into pUC13 and M13mp10. (B) The arrangements of the inserts and relevant restriction endonuclease sites in these clones are illustrated below the linear map (see Materials and Methods for details). The arrows indicate the direction of exonuclease III digestion used to generate deletion clones. Although we were unable to clone the 3.0-kb fragment directly from pYT103 into M13, we were able to transfer it from our pUC13-3.9 clone into M13mp10. We believe that the left hairpin end was deleted sufficiently in the pUC clone to permit its recovery in an M13 clone.

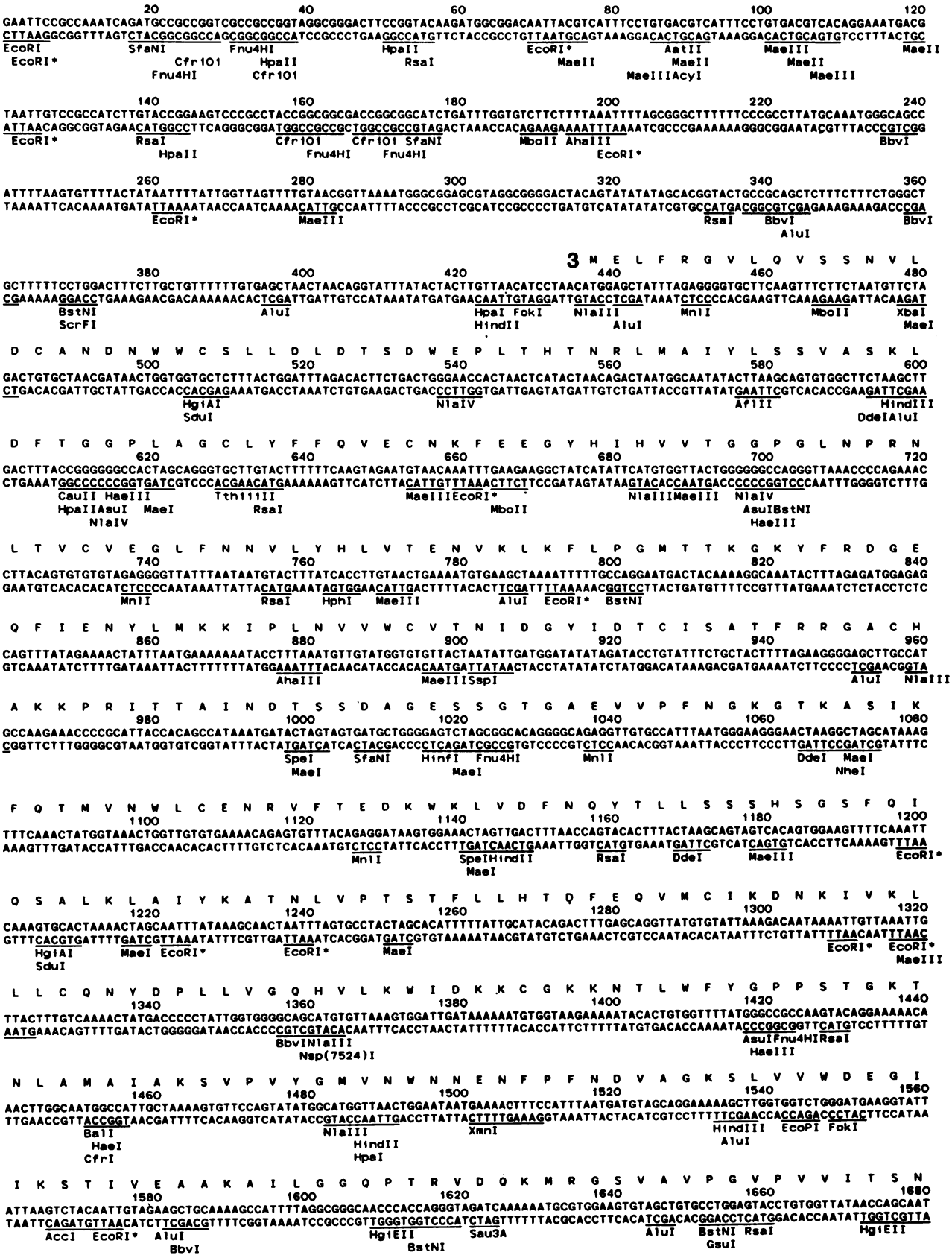
reaction. It seems probable that, without dGTP, the 3′ exonuclease activity of the polymerase digests back into the DNA before it has added the thiophosphate derivative. Hence, by adding the penultimate nucleotide, we were able to prevent this from occurring. After phenol-$CHCl_3$ extraction and ethanol precipitation, the DNA was digested with *Bam*HI, and deletion clones were constructed as described above. Single-stranded M13 DNA deletion clones were sized on a 0.7% agarose gel (100 V for 16 h), and appropriate clones were sequenced by the enzymatic procedure using universal forward primer.

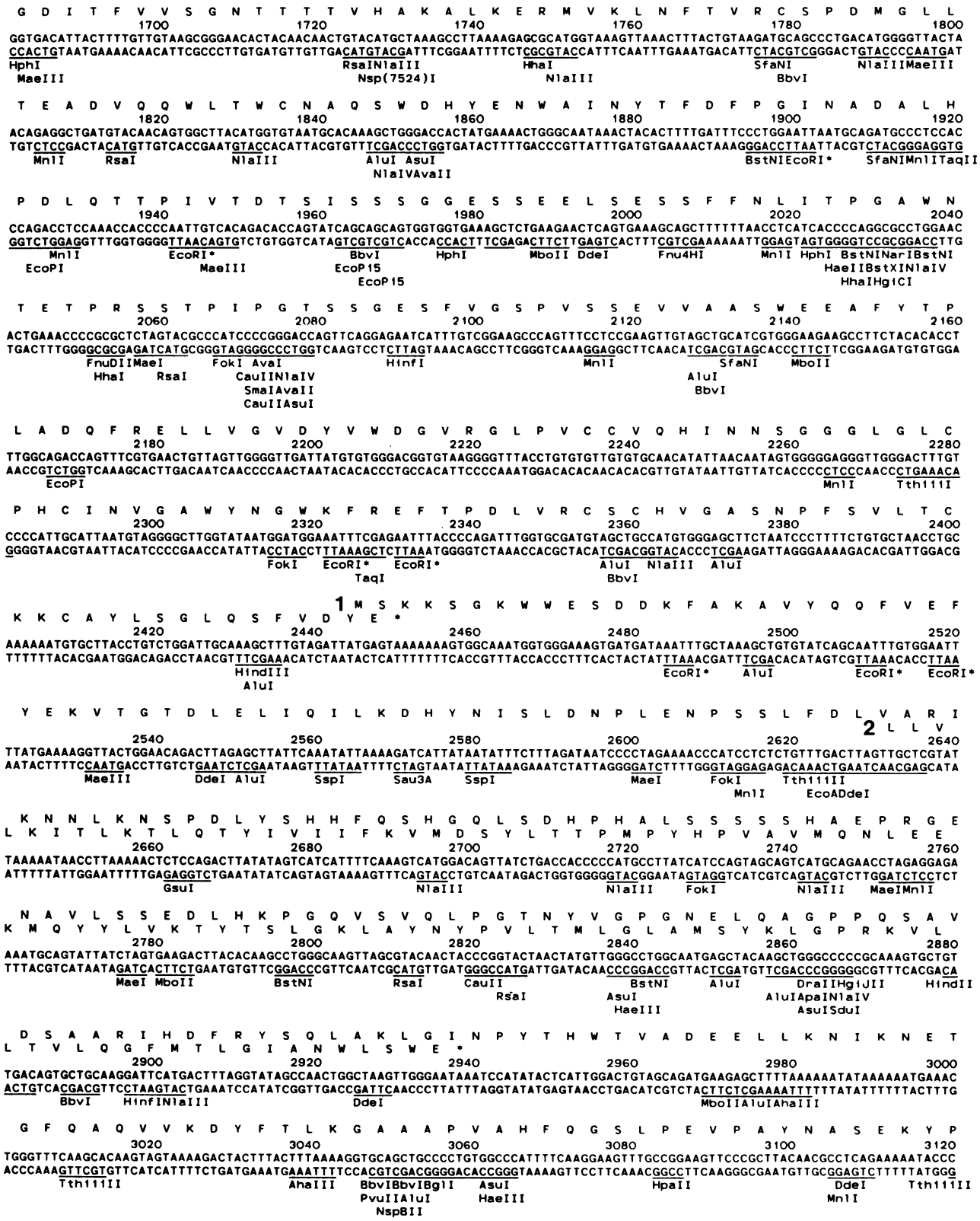
After the sequence of the above clones was obtained, several regions still remained in doubt. To repeat the sequence in these regions, synthetic deoxyribooligonucleotides (13- to 16-mers) were synthesized on an Applied Biosystems oligonucleotide synthesizer and used as primers on appropriate clone templates. To sequence through the *Bam*HI site at the junction of the 3.9- and 1.4-kb fragments, plasmid pYT103 DNA (approximately 20 μg) was digested with *Bgl*II and *Pst*I (see Fig. 1A), and the *Bgl*II ends were labeled with dGTP and [α-^{32}P]dATP using DNA polymerase I (Klenow fragment). DNA was run on a 5% acrylamide gel, and the appropriate fragment was electroeluted and sequenced by the chemical method (27).

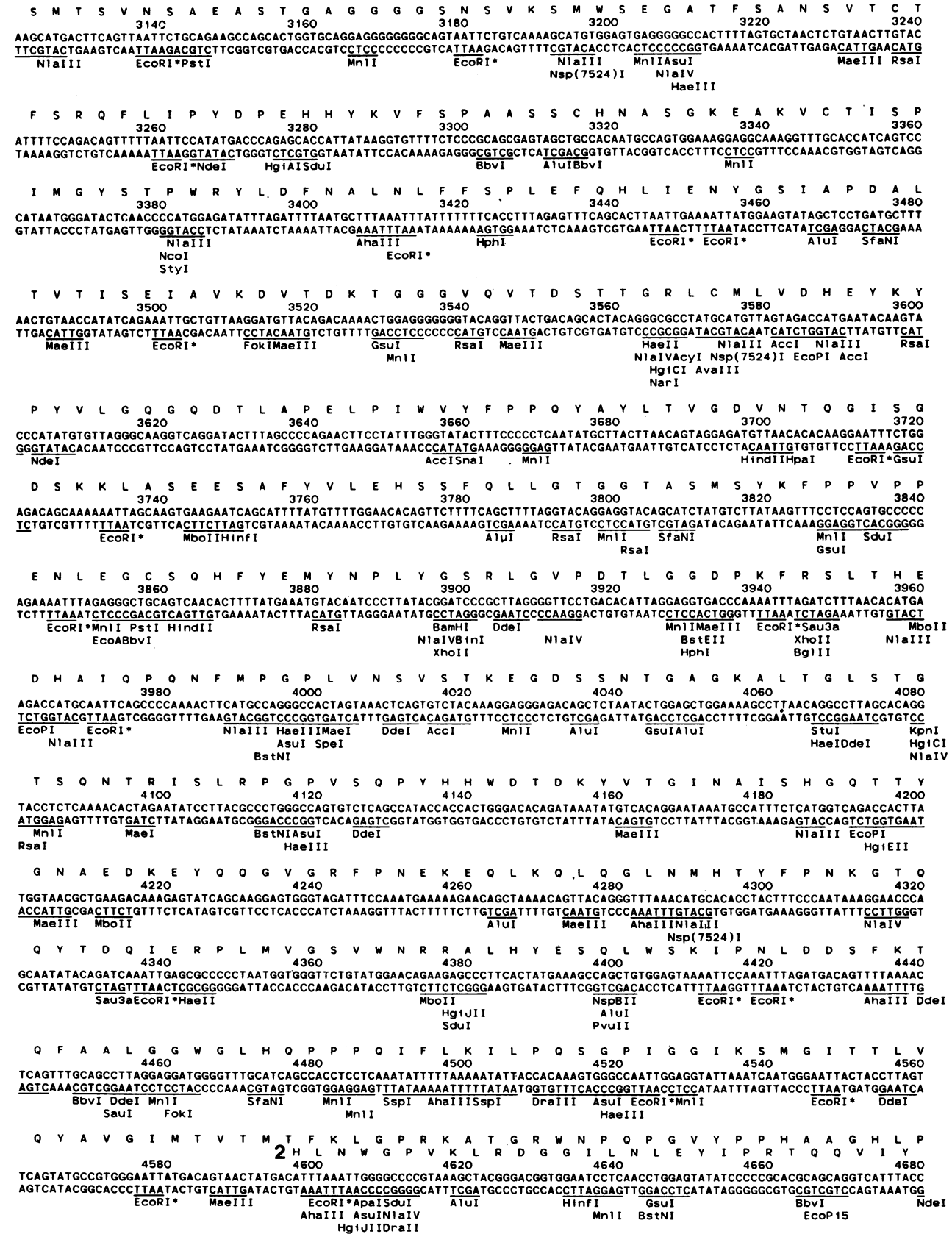
RESULTS

Restriction endonuclease map of the cloned B19-Au genome and the sequencing strategy. Figure 1A is a restriction endonuclease map of some of the six-base-pair recognition enzymes. In addition, a list of such enzymes which do not cut the DNA is given. The list is not exhaustive, but all known restriction endonuclease sites are included in Fig. 2, which lists the entire nt sequence of the pYT103 clone of B19-Au together with the amino acid sequences encoded by the major regions of the open reading frame (ORF). The sequencing strategy for the B19-Au genome is depicted in Fig. 3 and described in Materials and Methods. The sequence through the *Bam*HI site of the pYT103 clone (nt 3996) was confirmed by chemical sequencing (27), but it should be noted that in the construction of this clone viral DNA was digested with *Bam*HI and the left and right fragments were cloned separately. Thus, it is possible that a small *Bam*HI fragment may have been lost during the cloning procedure. We will not be able to exclude this possibility until viral DNA or a junction clone is sequenced.

Sequence of the left- and right-hand hairpin regions. Our difficulty with cloning the left *Eco*RI and *Bam*HI fragment was almost certainly due to the presence of the terminal inverted repeat. Hence, our pUC13-3.9 and M13mp10-3.9 clones had deleted at least 130 nt relative to the longest inverted repeat in clone pYT103. We believe that the terminal hairpin at the left end in pYT103 is deleted, even from the virion DNA. The sequence of the longest cloned left-end hairpin was obtained from pYT103 by 3′-end labeling at either the *Eco*RI (nt 1), *Xba*I (nt 477), or *Hin*dIII (nt 595) site and sequencing by the chemical method (27). This approach gave us nt 1 to 595, which overlapped with the enzymatic sequencing results by at least 450 nt (Fig. 3). The right end of the genome was also sequenced with clone pYT103. The hairpin at the right end was much shorter in clone pYT103 than that at the left end (Fig. 4), and this indicates that, in the construction of this clone, DNA polymerase I (Klenow fragment) extension ended shortly after the polymerase entered the stem portion of the hairpin. We know that the right hairpin contains a large palindrome, because analysis of






S M T S V N S A E A S T G A G G G G S N S V K S M W S E G A T F S A N S V T C T
3140 3160 3180 3200 3220 3240
AAGCATGACTTCAGTTAATTCTGCAGAAGCCAGCACTGGTGCAGGAGGGGGGGGCAGTAATTCTGTCAAAAGCATGTGGAGTGAGGGGGCCACTTTTAGTGCTAACTCTGTAACTTGTAC
TTCGTACTGAAGTCAATTAAGACGTCTTCGGTCGTGACCACGTCCTCCCCCCCCGTCATTAAGACAGTTTTCGTACACCTCACTCCCCCGGTGAAAATCACGATTGAGACATTGAACATG
NlaIII EcoRI*PstI MnlI EcoRI* NlaIII MnlIAsuI MaeIII RsaI
Nsp(7524)I NlaIV
HaeIII
F S R Q F L I P Y D P E H H Y K V F S P A A S S C H N A S G K E A K V C T I S P
3260 3280 3300 3320 3340 3360
ATTTTCCAGACAGTTTTTAATTCCATATGACCCAGAGCACCATTATAAGGTGTTTTCTCCCGCAGCGAGTAGCTGCCACAATGCCAGTGGAAAGGAGGCAAAGGTTTGCACCATCAGTCC
TAAAAGGTCTGTCAAAAATTAAGGTATACTGGGTCTCGTGGTAATATTCCACAAAAGAGGGCGTCGCTCATCGACGGTGTTACGGTCACCTTTCCTCCGTTTCCAAACGTGGTAGTCAGG
EcoRI*NdeI HgiAISduI BbvI AluIBbvI MnlI
I M G Y S T P W R Y L D F N A L N L F F S P L E F Q H L I E N Y G S I A P D A L
3380 3400 3420 3440 3460 3480
CATAATGGGATACTCAACCCCATGGAGATATTTAGATTTTAATGCTTTAAATTTATTTTTTTCACCTTTAGAGTTTCAGCACTTAATTGAAAATTATGGAAGTATAGCTCCTGATGCTTT
GTATTACCCTATGAGTTGGGGTACCTCTATAAATCTAAAATTACGAAATTTAAATAAAAAAAGTGGAAATCTCAAAGTCGTGAATTAACTTTTAATACCTTCATATCGAGGACTACGAAA
NlaIII AhaIII HphI EcoRI* EcoRI* AluI SfaNI
NcoI EcoRI*
StyI
T V T I S E I A V K D V T D K T G G G V Q V T D S T T G R L C M L V D H E Y K Y
3500 3520 3540 3560 3580 3600
AACTGTAACCATATCAGAAATTGCTGTTAAGGATGTTACAGACAAAACTGGAGGGGGGGGTACAGGTTACTGACAGCACTACAGGGCGCCTATGCATGTTAGTAGACCATGAATACAAGTA
TTGACATTGGTATAGTCTTTAACGACAATTCCTACAATGTCTGTTTTGACCTCCCCCCCATGTCCAATGACTGTCGTGATGTCCCGCGGATACGTACAATCATCTGGTACTTATGTTCAT
MaeIII EcoRI* FokIMaeIII GsuI RsaI MaeIII HaeII NlaIII AccI NlaIII RsaI
MnlI NlaIVAcyI Nsp(7524)I EcoPI AccI
HgiCI AvaIII
NarI
P Y V L G Q G Q D T L A P E L P I W V Y F P P Q Y A Y L T V G D V N T Q G I S G
3620 3640 3660 3680 3700 3720
CCCATATGTGTTAGGGCAAGGTCAGGATACTTTAGCCCCAGAACTTCCTATTTGGGTATACTTTCCCCCTCAATATGCTTACTTAACAGTAGGAGATGTTAACACACAAGGAAATTTCTGG
GGGTATACACAATCCCGTTCCAGTCCTATGAAATCGGGGTCTTGAAGGATAAACCCATATGAAAGGGGGAGTTATACGAATGAATTGTCATCCTCTACAATTGTGTTCCTTAAAGACC
NdeI AccISnaI MnlI HindIIHpaI EcoRI*GsuI
D S K K L A S E E S A F Y V L E H S S F Q L L G T G G T A S M S Y K F P P V P P
3740 3760 3780 3800 3820 3840
AGACAGCAAAAAATTAGCAAGTGAAGAATCAGCATTTTATGTTTTGGAACACAGTTCTTTTCAGCTTTTAGGTACAGGAGGTACAGCATCTATGTCTTATAAGTTTCCTCCAGTGCCCCC
TCTGTCGTTTTTTAATCGTTCACTTCTTAGTCGTAAAATACAAAACCTTGTGTCAAGAAAAGTCGAAAATCCATGTCCTCCATGTCGTAGATACAGAATATTCAAAGGAGGTCACGGGGG
EcoRI* MboIIHinfI AluI RsaI MnlI SfaNI MnlI SduI
RsaI GsuI
E N L E G C S Q H F Y E M Y N P L Y G S R L G V P D T L G G D P K F R S L T H E
3860 3880 3900 3920 3940 3960
AGAAAATTTAGAGGGCTGCAGTCAACACTTTTATGAAATGTACAATCCCTTATACGGATCCCGCTTAGGGGTTCCTGACACATTAGGAGGTGACCCAAAATTTAGATCTTTAACACATGA
TCTTTTAAATCTCCCGACGTCAGTTGTGAAAATACTTTACATGTTAGGGAATATGCCTAGGGCGAATCCCCAAGGACTGTGTAATCCTCCACTGGGTTTTAAATCTAGAAATTGTGTACT
EcoRI*MnlI PstI HindII RsaI BamHI DdeI MnlIMaeIII EcoRI*Sau3a MboII
EcoABbvI NlaIVBinI NlaIV BstEII XhoII NlaIII
XhoII HphI BglII
D H A I Q P Q N F M P G P L V N S V S T K E G D S S N T G A G K A L T G L S T G
3980 4000 4020 4040 4060 4080
AGACCATGCAATTCAGCCCCAAAACTTCATGCCAGGGCCACTAGTAAACTCAGTGTCTACAAAGGAGGGAGACAGCTCTAATACTGGAGCTGGAAAAGCCTTAACAGGCCTTAGCACAGG
TCTGGTACGTTAAGTCGGGGTTTTGAAGTACGGTCCCGGTGATCATTTGAGTCACAGATGTTTCCTCCCTCTGTCGAGATTATGACCTCGACCTTTTCGGAATTGTCCGGAATCGTGTCC
EcoPI EcoRI* NlaIII HaeIIIMaeI DdeI AccI MnlI AluI GsuIAluI StuI KpnI
NlaIII AsuI SpeI HaeIDdeI HgiCI
BstNI NlaIV
T S Q N T R I S L R P G P V S Q P Y H H W D T D K Y V T G I N A I S H G Q T T Y
4100 4120 4140 4160 4180 4200
TACCTCTCAAAACACTAGAATATCCTTACGCCCTGGGCCAGTGTCTCAGCCATACCACCACTGGGACACAGATAAATATGTCACAGGAATAAATGCCATTTCTCATGGTCAGACCACTTA
ATGGAGAGTTTTGTGATCTTATAGGAATGCGGGACCCGGTCACAGAGTCGGTATGGTGGTGACCCTGTGTCTATTTATACAGTGTCCTTATTTACGGTAAAGAGTACCAGTCTGGTGAAT
MnlI MaeI BstNIAsuI DdeI MaeIII NlaIII EcoPI
RsaI HaeIII HgiEII
G N A E D K E Y Q Q G V G R F P N E K E Q L K Q L Q G L N M H T Y F P N K G T Q
4220 4240 4260 4280 4300 4320
TGGTAACGCTGAAGACAAAGAGTATCAGCAAGGAGTGGGTAGATTTCCAAATGAAAAAGAACAGCTAAAACAGTTACAGGGTTTAAACATGCACACCTACTTTCCCAATAAAGGAACCCA
ACCATTGCGACTTCTGTTTCTCATAGTCGTTCCTCACCCATCTAAAGGTTTACTTTTTCTTGTCGATTTTGTCAATGTCCCAAATTTGTACGTGTGGATGAAAGGGTTATTTCCTTGGGT
MaeIII MboII AluI MaeIII AhaIIINlaIII NlaIV
Nsp(7524)I
Q Y T D Q I E R P L M V G S V W N R R A L H Y E S Q L W S K I P N L D D S F K T
4340 4360 4380 4400 4420 4440
GCAATATACAGATCAAATTGAGCGCCCCCTAATGGTGGGTTCTGTATGGAACAGAAGAGCCCTTCACTATGAAAGCCAGCTGTGGAGTAAAATTCCAAATTTAGATGACAGTTTTAAAAC
CGTTATATGTCTAGTTTAACTCGCGGGGGGATTACCACCCAAGACATACCTTGTCTTCTCGGGAAGTGATACTTTCGGTCGACACCTCATTTTAAGGTTTAAATCTACTGTCAAAATTTTG
Sau3aEcoRI*HaeII MboII NspBII EcoRI* EcoRI* AhaIII DdeI
HgiJII AluI
SduI PvuII
Q F A A L G G W G L H Q P P P Q I F L K I L P Q S G P I G G I K S M G I T T L V
4460 4480 4500 4520 4540 4560
TCAGTTTGCAGCCTTAGGAGGATGGGGTTTGCATCAGCCACCTCCTCAAATATTTTTAAAAATATTACCACAAAGTGGGCCAATTGGAGGTATTAAATCAATGGGAATTACTACCTTAGT
AGTCAAACGTCGGAATCCTCCTACCCCAAACGTAGTCGGTGGAGGAGTTTATAAAAATTTTTATAATGGTGTTTCACCCGGTTAACCTCCATAATTTAGTTACCCTTAATGATGGAATCA
BbvI DdeI MnlI SfaNI MnlI SspI AhaIIISspI DraIII AsuI EcoRI*MnlI EcoRI* DdeI
SauI FokI MnlI HaeIII
Q Y A V G I M T V T M T F K L G P R K A T G R W N P Q P G V Y P P H A A G H L P
2 H L N W G P V K L R D G G I L N L E Y I P R T Q Q V I Y
4580 4600 4620 4640 4660 4680
TCAGTATGCCGTGGGAATTATGACAGTAACTATGACATTTAAATTGGGGCCCCGTAAAGCTACGGGACGGTGGAATCCTCAACCTGGAGTATATCCCCCGCACGCAGCAGGTCATTTACC
AGTCATACGGCACCCTTAATACTGTCATTGATACTGTAAATTTAACCCCGGGGCATTTCGATGCCCTGCCACCTTAGGAGTTGGACCTCATATAGGGGGCGTGCGTCGTCCAGTAAATGG
EcoRI* MaeIII EcoRI*ApaISduI AluI HinfI GsuI BbvI NdeI
AhaIII AsuINlaIV MnlI BstNI EcoP15
HgiJIIDraII

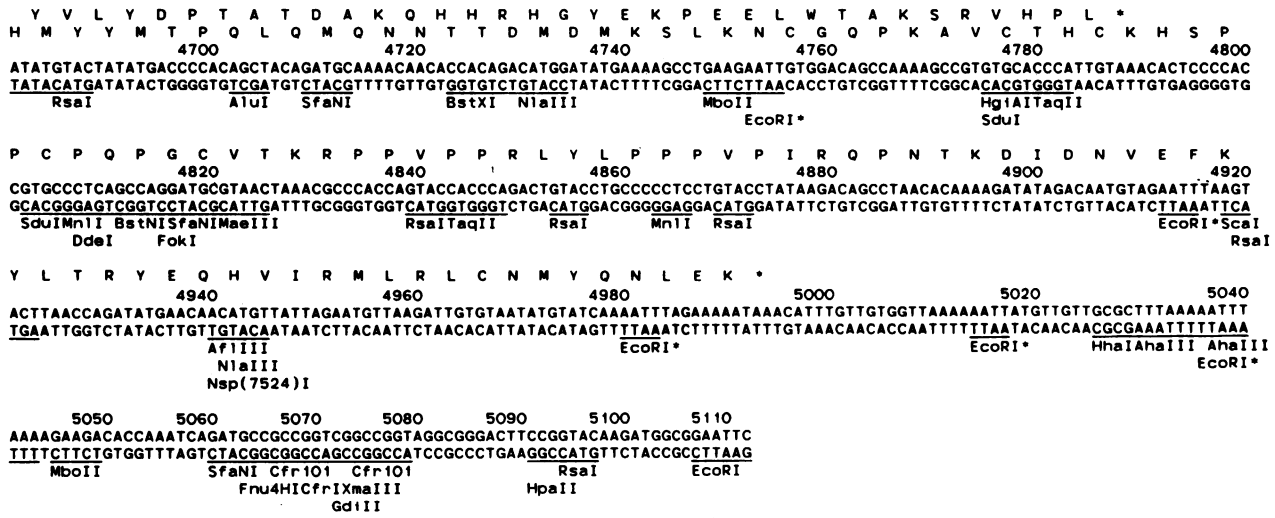


FIG. 2. The nucleotide sequence of the cloned B19-Au genome. The top line of the sequence corresponds to the complementary (plus) strand (5′ to 3′). Amino acid sequences of putative polypeptides corresponding to the major ORFs are shown above the sequence. All restriction endonuclease sites are shown below the sequence. A list of potentially relevant promotorlike sequences and downstream translation start and stop signals is given in Table 1. The sequence is complete except for ~150 nt at the left end and ~280 nt at the right end. One nt within the coding region remains in doubt. The A at position 3940 is not present in our pYT103 clone. Lack of this nt would shift the reading frame from 1 to 3, which contains numerous stop codons. Because we also sequenced a closely related isolate (B19-Wi) and found that the A is present in this clone, we believe that the A is missing in pYT103 because of a cloning artifact (M. C. Blundell and C. R. Astell, unpublished data). In any event, except for the amino acids at this junction, the predicted amino acid sequence is still valid for purposes of comparison with other viral polypeptides (Astell et al., manuscript in preparation).

a clone from another isolate, B19-Wi, indicates that the right hairpin is at least 240 nt long (M. C. Blundell and C. R. Astell, unpublished data). However, what is striking from the data presented here on the B19-Au clone is that the left and right ends are related (Fig. 4), unlike other autonomous parvoviruses characterized to date [(MVM(p) (4–6), MVM(i) (3), and H1 (38)], and similar to that seen with AAV2 (25). However, the B19 termini appear to be substantially longer than the AAV termini and much more unstable when cloned into circular plasmid vectors.

Potential protein coding regions. The final DNA sequence of the B19-Au clone was analyzed by the SEQNCE program of A. Delaney (18). Figure 5 summarizes the positions of termination codons. Like all other parvoviruses characterized to date, only one strand of the viral genome has significant blocks of ORF in frames 1, 2, and 3 (Fig. 5A). Hence, this strand of B19 almost certainly defines the sense of viral transcription (2). Comparison of the genomes of B19-Au with those of MVM(p) and MVM(i) (3, 6, 40), H1 (39), CPV (36), FPV (12) (all autonomous parvoviruses) (Fig. 5B), and AAV2 (48) (a defective parvovirus) (Fig. 5C) shows that the organization of the B19 genome is similar to that of both the autonomous parvoviruses and AAV2, but there are notable differences. Frame 1 (Fig. 5A) is open from nt 2441 to nt 4786. By analogy with other parvoviruses (3, 6, 12, 36, 39, 40), this region would be expected to code for viral structural proteins (VP1, VP2, VP3). The second codon into this ORF is a methionine codon, unlike other autonomous parvoviruses (3, 6, 12, 36, 39, 40), in which the first in-frame ATG in the major right-hand ORF is at least 500 nt downstream or in the AAV2 genome, in which the first in-frame ATG in the 2.3-kb spliced message is at nt 2870. Although it has been established that the long ORFs present in autonomous parvovirus sequences encode the majority of virus-specific polypeptide, it is now clear that mRNA splicing is also used to connect smaller, less obvious ORF segments to create contiguous protein coding sequences. For instance, in MVM and H1 one such small region exists in frame 2 (~nt 1936 to 2299 in MVM) and has recently been shown to encode the carboxy-terminal half of the smaller nonstructural protein NS2 (S. F. Cotmore and P. Tattersall, manuscript in preparation). A comparable alternate short ORF does not appear to exist in B19, FPV, or CPV (Fig. 5). However, the B19 genome has at least two other small ORFs in frame 2 which may be used. One of these is located between nt 2631 and 2936 (map units ~51 to 58). A comparable region is also open in frame 2 of MVM and H1 (map units ~47 to 53). The AAV2 genome also has several short ORFs near the center of the genome, one of which may be used instead of the major right-hand ORF to encode the amino-terminal region of VP1 (22; see the Discussion). The second minor ORF, in frame 2 of B19 at map units 88 to 97 (nt 4596 to 4991), has no obvious analog in MVM or H1.

Transcription signals. A search of the complementary strand of the B19-Au genome reveals that there are multiple potential transcription start and polyadenylation signals. The sequence TATAAA occurs at nt 1225 (map unit ~24) and nt 2986 (map unit ~58.5). A search for related TATA-like signals positioned upstream of major ORFs identified a TATTAA sequence at nt 2247, a TATAAT sequence at nt 2308, and five different TATA-like sequences at nt 257, 319, 321, 323, and 412. In addition to the consensus sequences, some of these regions have appropriately spaced upstream CAAT sequences. These sequences are summarized in Table 1. As yet we do not have information on functional promotor sequences within the B19 genome; hence, we cannot say which of these TATA boxes are used. Experiments to determine functional promoters are in progress. Comparison of TATA boxes within the MVM(p), H1, and AAV2 genomes suggests that, if all the TATA-like se-

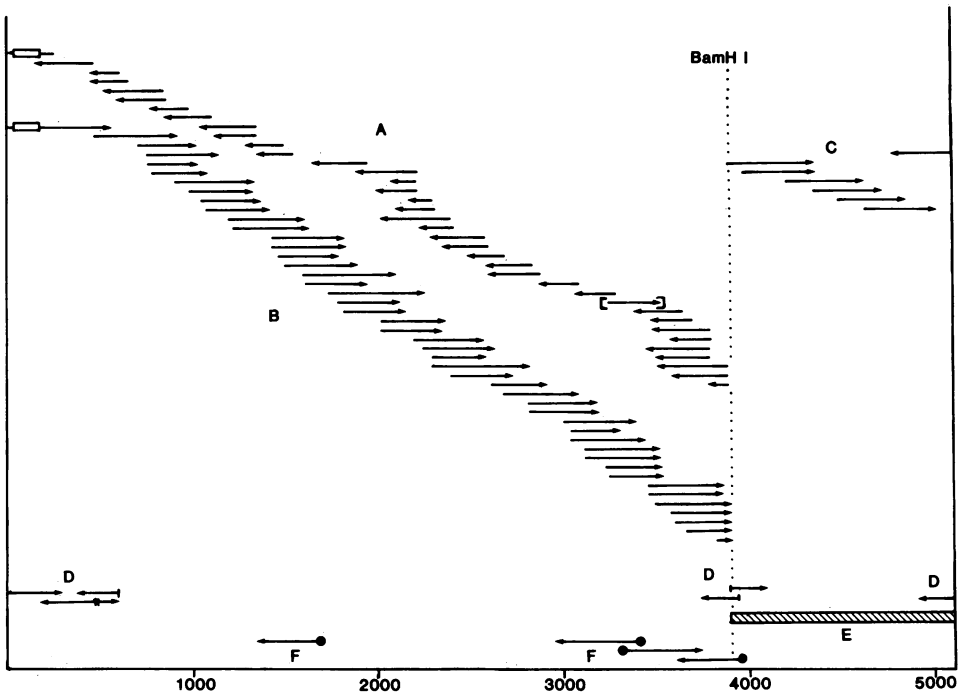


FIG. 3. Sequencing strategy for the B19-Au insert in plasmid pYT103. The sequences obtained from the deletion clones constructed in M13mp10-3.9 (A), pUC13-3.9 (B), and pUC13-1.4(C) are illustrated with arrows. Deletions of the M13mp10-3.9 clone were not obtained for the regions ~nt 1700 and ~nt 3200. This latter region was obtained in another M13 clone of opposite orientation, which is bracketed. In addition, the regions obtained by chemical sequencing are illustrated below (D). The cross-hatched area (E) corresponds to the sequence of the B19-Wi isolate obtained by chemical sequencing (Blundell and Astell, unpublished data; see the legend to Fig. 2). The sequences obtained by using specific synthetic oligonucleotides are indicated by arrows at the bottom (F). The open box at the left is the region (~130 base pairs) which is deleted in our pUC13-3.9 clone, and the sequence in this region was obtained from the original pYT103 clone.

quences in the B19 genome are functional, then the organization of this genome is somewhat different from those of either the autonomous or defective genomes.

The distribution of polyadenylation signals is also unlike that of either the autonomous or defective parvoviruses (Fig. 5). There are six polyadenylation signals in B19-Au. These occur throughout the genome at nt 1303, 1872, 2935, 4168, 4307, and 4990. The autonomous parvoviruses, MVM(p) (3) and H1 (39) have, respectively, four and three polyadenylation sites, all clustered at map units approximately 90 to 96 (Fig. 5b). The presence of an upstream AATAAA sequence is seen in the AAV2 genome (at nt 2182) and the FPV genome (at nt 483) (Fig. 5b). However, at least for AAV2, analysis of transcripts show that this upstream site is not used, as all AAV2 transcripts are coterminal at ~96 map units (13), and the AATAAA sequence is likely spliced out of mature 2.3-kb mRNA. At present nothing is known about the splicing of B19 transcripts, as viral transcripts have not been characterized because there is no in vitro cell culture system currently available for studying virus replication. However, transcription and splicing patterns for MVM(p) (33) and AAV2 (13) are included in Fig. 5.

Potential polypeptides encoded by B19-Au. While we do not as yet have data on transcripts from B19, the two largest ORFs both contain a methionine codon within the first few nt of the start of the open frame, therefore it is not unreasonable to assume that this entire region is translated. We deduced the amino acid sequence encoded by these regions as well as the two short open regions in frame 2 (Fig. 2 and 5A).

The left-hand ORF extends from nt 427 to the TAA termination codon at nt 2449. The first in-phase ATG codon occurs at nt 436, and if translation begins at this position, a protein of 671 amino acids could be synthesized. This putative polypeptide would have a slightly negative net charge, as there are 55 basic amino acids (lysine, arginine) and 65 acidic residues (aspartate, glutamate). While the hydropathy profile (Fig. 6A) shows that hydrophilic and hydrophobic regions alternate throughout the protein with no large concentration of either, its average hydropathy

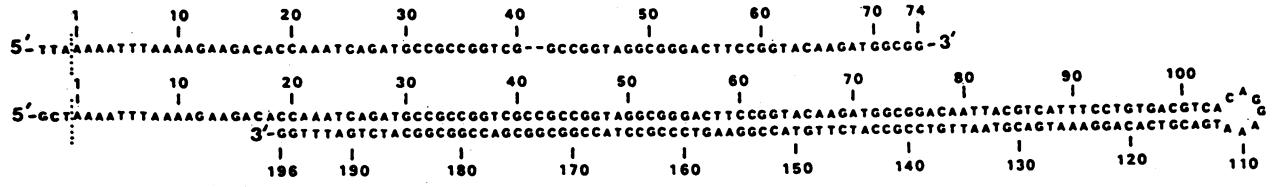


FIG. 4. The nucleotide sequences of the hairpin ends present in the pYT103 clone of the B19-Au genome. As discussed in the text, the right hairpin (top) is missing ~280 nt, and the left hairpin (bottom) is missing ~150 nt. The inverted terminal repeats begin at nt 1.

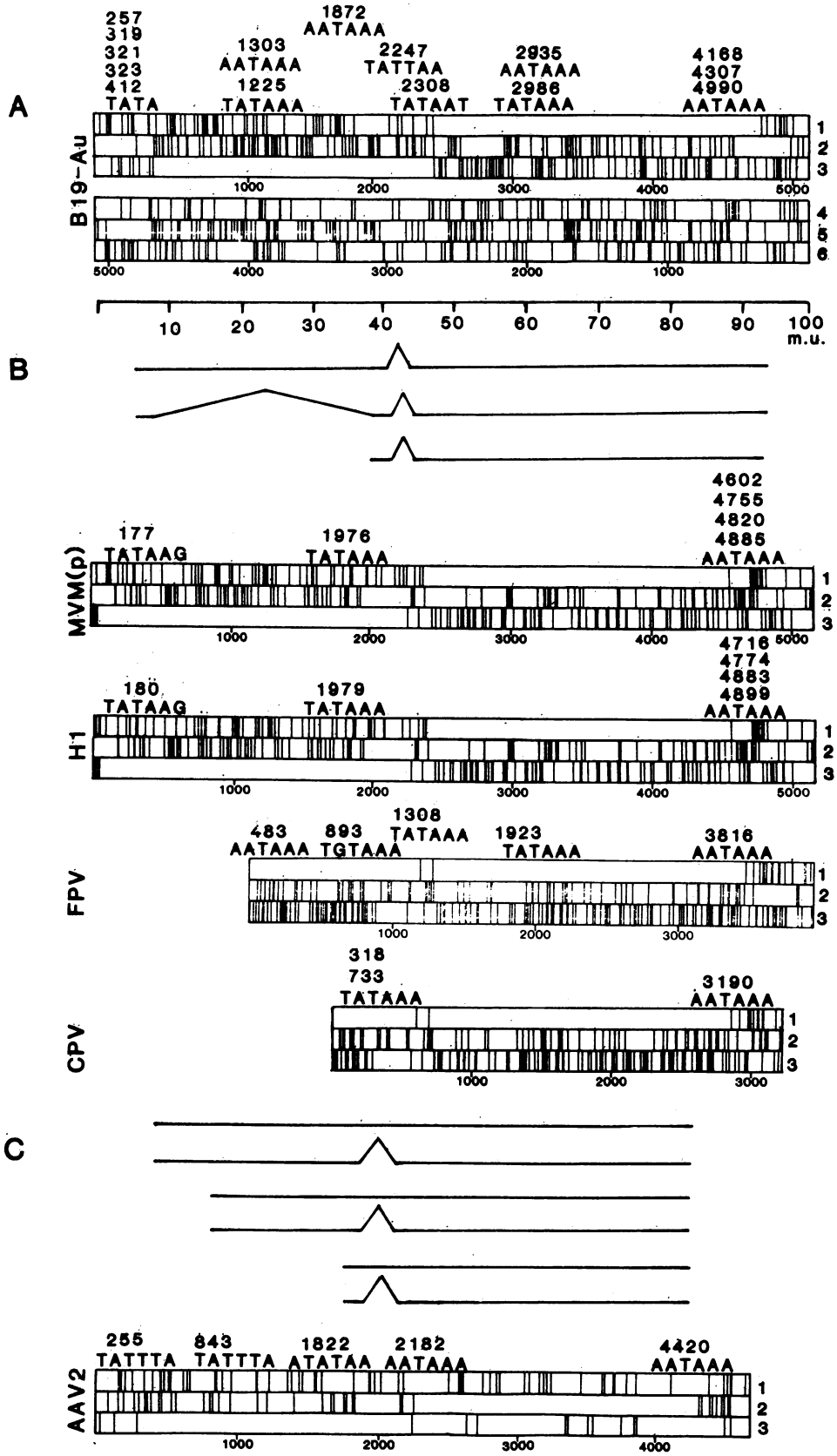


FIG. 5. Comparison of genome organization of B19-Au DNA with that of autonomous parvoviruses [MVM(p), H1, FPV, and CPV] and AAV2, a dependovirus. For reference, transcription maps for MVM(p) (33) and AAV2 (13) are illustrated above their respective ORF diagrams. m.u., Map units.

TABLE 1. Summary of promotorlike sequences in the complementary strand of the B19 genome, including relevant downstream translation initiation and termination signals

Nt no. and:			
CAAT-like sequence	TATA-like sequence	PuNNATGPu sequence	In-phase termination codon
	257 TATAATT	291 AAAATGG	303 TAG
	319 TATATATA	436 AACATGG	2445 TGA
	412 TATACT	436 AACATGG	2445 TGA
1195 CAAATT	1225 TATAAA	1288 GTTATGG	2445 TGA
	2247 TATTAA	2294 TTAATGT	2297 TAG
		2352 GCGATGT	2355 TAG
		2363 GCCATGT	2393 TAA
		2406 AAAATGT	2445 TGA
		2444 ATTATGA	4787 TAA
2283 CCATT	2308 TATAAT	2352 GCGATGT	2355 TAG
		2363 GCCATGT	2393 TAA
		2406 AAAATGT	2445 TGA
		2444 ATTATGA	4787 TAA
2950 TCATT	2986 TATAAA	3125 AGCATGA	4787 TAA

value (−0.10) is within the range of some membrane-spanning proteins. However, this protein does not contain a large hydrophobic region, suggesting a membrane association (24). The right-hand ORF begins at nt 2441 and extends to the TAA termination codon at nt 4787. The first methionine codon occurs at nt 2444, and if translation begins here, a 781-amino acid protein could be synthesized. It is noteworthy that this protein, a presumptive VP1, is larger by at least 50 amino acid residues than the VP1 polypeptides of other parvoviruses, whereas the VP1 polypeptide present in B19 virus comigrated with that of MVM in sodium dodecyl sulfate-polyacrylamide gels (Cotmore and Tattersall, unpublished data). This polypeptide would have a net negative charge with 64 basic amino acids and 76 acidic residues. The hydropathy curve (Fig. 6B) is typical of a soluble protein which has fairly evenly spaced hydrophobic and hydrophilic regions. The average hydropathy value is (−0.52) (24).

Two other short ORFs occur in reading frame 2 (Fig. 5A). One extends from nt 2631 to the TAA stop codon at nt 2937 and contains seven methionine codons. The other ORF is near the end of frame 2, beginning at nt 4596 and extending to the TAA codon at nt 4992. Whereas there are seven methionine codons within this region, it is unlikely that translation begins here. This region would be expected to be a terminal exon spliced to another ORF if it is used at all. It is interesting that the TAA termination codon at the right-hand side of this ORF is part of the most downstream polyadenylation signal.

Protein sequence comparisons with other parvoviruses. Computer analyses of putative viral polypeptides from MVM(p), AAV2, and B19-Au have identified significant regions of homology between all three of these genomes, although the overall picture is of three quite distinct and different viruses. By comparing homologous sequences between B19, MVM, and AAV2, we identified regions conserved in both the left-hand and right-hand ORFs.

Left-hand ORF. A region located in the middle of the ORF is highly conserved (Fig. 7). Using protein homology as a guide, the nt sequences were lined up below the protein sequences. No gaps were necessary in the lineups to maximize the amino acid homologies, although in several places leaving gaps in the nt lineups improved homology slightly. The levels of nt homology in this 425-nt stretch are 52% between B19 and AAV2, 52% between B19 and MVM, and 53% between MVM and AAV2. At the protein level, these homologies are 51, 41, and 51%, respectively. However, if one looks at shorter regions within this sequence, the homologies are much higher. For example, at the protein level, there is a region within the B19 genome between nt 1486 and 1521 (Fig. 7, arrows A and B), where 11 of the 12 amino acids are identical between B19 and AAV2 (92% homology). The same region in MVM has 8 of 12 amino acids identical to the B19 sequence (67% homology), but interestingly, at the nt level MVM and B19 share an 18-base stretch of perfect homology within this region, while the longest perfect match between AAV and B19 is 8 nt. In general, many changes at the nt level occur at the third position of codons and do not alter the protein sequence. Other nt changes often result in conservative amino acid changes, and if conservative changes are permitted, the level of amino acid homology is much higher. For comparison, the sequences of H1, CPV, and FPV are included in Fig. 7.

Right-hand ORF. The right-hand ORF encoding the capsid genes can most easily be considered as two distinct segments: the VP1-specific region and the region common between VP1 and VP2. In AAV this is particularly useful because a large segment of the VP1-specific region may be located in a different ORF from the VP2 sequence (22). Amino acid sequence homology between B19, AAV2, and MVM supports this suggestion. The presumed VP1-specific regions of all three viruses contain a highly conserved 30-amino-acid sequence (Fig. 4, nt 2815 to 2905 in B19 frame 1, nt 2341 to 2431 in AAV2 frame 1, and nt 2407 to 2496 in MVM(p) frame 1). At the protein level, homology in this region is 47% between B19 and AAV2, 50% between B19 and MVM, and 70% between MVM and AAV2. Interestingly, the homology between AAV2 and the other two viruses stops at this point, while the homology between MVM and B19 continues for a significant distance (50% homology over 44 amino acids to nt 2947 in B19 and 38% homology over 81 amino acids to nt 3058 in B19).

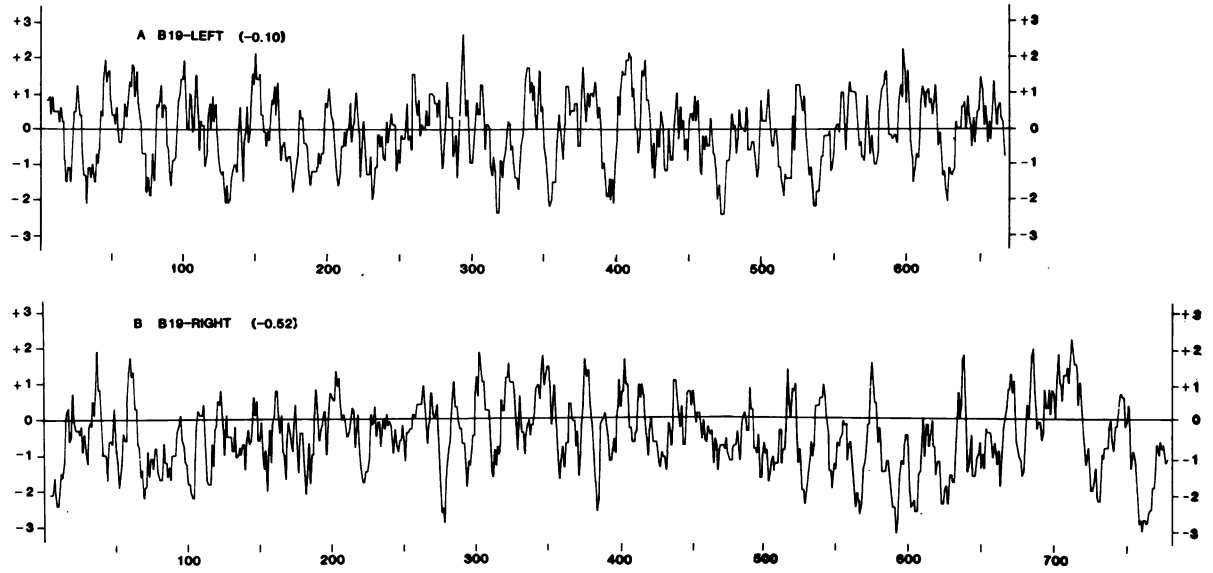


FIG. 6. Hydropathy curves for putative polypeptides corresponding to the major left-hand (A) and right-hand (B) ORFs. Average hydropathy values are given in brackets. Numbers on the abcissas indicate amino acid positions.

In the VP2-specific region, the most striking protein homologies are seen between AAV2 and B19. For example, over a region of 29 amino acids [nt 3368 to 3454 in frame 1 of B19, nt 3023 to 3109 in frame 2 of AAV2, and nt 3109 to 3195 in frame 1 of MVM(p)], the exact homologies are 50% (AAV-B19), 28% (MVM-B19), and 34% (AAV-MVM). In the rest of the VP2 gene, B19 and AAV have at least three regions of between 30 and 40 amino acids where the exact homology is between 40 and 50%, while the homologies between MVM and B19 are substantially lower in each case. Unlike other autonomous parvoviruses for which sequence information is available, B19 does not contain a characteristic glycine-rich region just downstream of the ATG presumed to initiate VP2.

DISCUSSION

The sequence of an almost-full-length clone of the B19-Au genome is presented. Portions of the genome that are not included are an estimated 280 nt of the hairpin at the right end and an estimated 150 nt of the hairpin at the left end. We believe that the transcription signals and coding regions are complete; hence, these can be compared with those of autonomous [MVM(p)] and helper-dependent (AAV2) parvoviruses.

Whereas the structures of the right and left hairpins from the cloned B19-Au are incomplete, they do indicate that, unlike all autonomous parvoviruses characterized to date, the B19-Au ends are extensively homologous and may in fact contain inverted terminal repeats similar to those seen in AAV2. The termini, which are presumed to be identical except for a difference of two C residues which are missing in the stem of the right-hand hairpin relative to the left-hand hairpin (Fig. 4), are extremely GC rich, although the sequence itself is not related to any known autonomous or helper-dependent terminal hairpins. Moreover, if our size estimates are correct, then at around 330 nt these palindromes are much longer than any other parvoviral termini characterized to date. Like the AAVs, which also have related terminal sequences, B19 packages both plus and minus DNA strands into virions with approximately equal efficiencies (17, 51), whereas those autonomous parvoviruses which have been fully sequenced (MVM [6] and H1 [39]) have different terminal sequences and package the minus strand predominantly. The termini of another autonomous parvovirus, LuIII, have not yet been characterized; however, this virus packages both plus and minus strands in approximately equal amounts (8). Thus, the B19 sequence presented in this paper remains consistent with the suggestion that there is a specific nt sequence or three-dimensional DNA structure which directs the packaging of only one or either strand into virions. The difference between the left and right hairpins of B19 is an interesting one in that it has been shown with AAV2 that alteration of one hairpin end of the AAV2 genome can result in a correction mechanism in which presumably the intact end acts as a template for repair of the altered (nonfunctional) end (41). If such a mechanism operates in B19 replication, the fact that a two-base-pair difference can exist in the stem region implies that the system which recognizes topological features of the hairpin ends is insensitive to a minor sequence or length difference or both in the stem of the hairpin. A further implication of the structure of the B19 terminal inverted repeats is that, as with AAV2, it is possible that this genome may integrate within the host chromosome under conditions in which lytic viral replication is blocked (10, 14).

Like all other parvoviruses, the coding region of the B19 genome is confined to one strand, by definition the plus or complementary strand. The entire genome from map units ~8 to 98 is open, although the reading frame shifts at approximately map units 45 to 50 from frame 3 to frame 1 (Fig. 5A). By comparison with other autonomous parvoviruses and AAV2, the left-hand ORF would be expected to encode nonstructural polypeptides, and the right-hand ORF would be expected to encode viral polypeptides. Procaryotic expression studies confirm this organization; antibodies raised against a sequence encoded in frame 3 of B19 between nt 1072 and 2044 identify a 71,000-dalton nonstructural protein in B19-infected fetal bone marrow, whereas antibodies against the amino acid sequence specified in frame 1 between nt 2897 and 3749 recognize both the

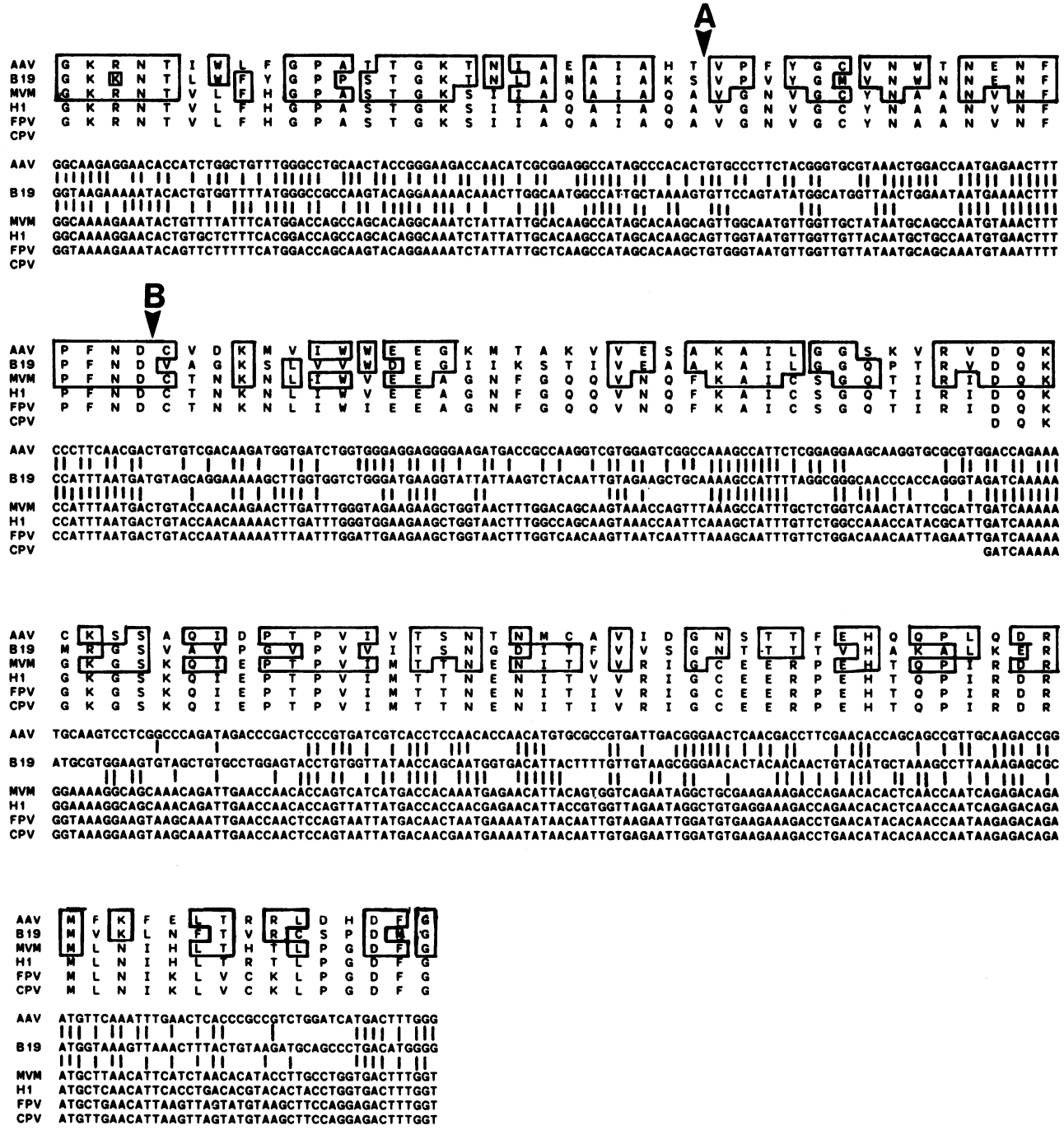


FIG. 7. Homology between putative NS1 protein of B19 and NS1 polypeptides from helper-dependent (AAV2) and -independent (MVM) parvoviruses. A region of conserved amino acid sequence is shown along with the corresponding nucleotide sequence. Nucleotide homology between B19 and AAV2 or MVM is indicated by short vertical bars. At the protein level, homology between B19, AAV2, and MVM is indicated by boxes enclosing the homologous amino acids. The corresponding sequences from other related autonomous parvoviruses are shown for comparison. The nucleotide sequences in this figure begin at nt 1293 (AAV2 [48]), nt 1390 (B19-Au [this paper]), nt 1428 [MVM(p) (6)], nt 1431 (H1 [39]), nt 1 (CPV [36]), and nt 345 (FPV [12]). Arrow A marks nt 1486, and arrow B marks nt 1521 in the B19 sequence.

83,000- and 58,000-dalton capsid proteins (Cotmore et al., manuscript in preparation). The TATA-like sequences (Fig. 5A) positioned at nt 257, 319, 321, 323, and 412 likely function in initiation of transcription from approximately map unit 7. Further examination (Table 1) indicates that there is no upstream CAAT box associated with any of these sequences. If the TATA box at nt 257 is functional and transcription originates ~30 nt downstream, then the first ATG codon occurs at nt 291 but is followed by an in-frame TAG codon at nt 303. The TATA box at nt 412 is also not likely to be used, as the ATG codon at nt 436 near the start of the left-hand ORF would be very close to the 5′ end of the

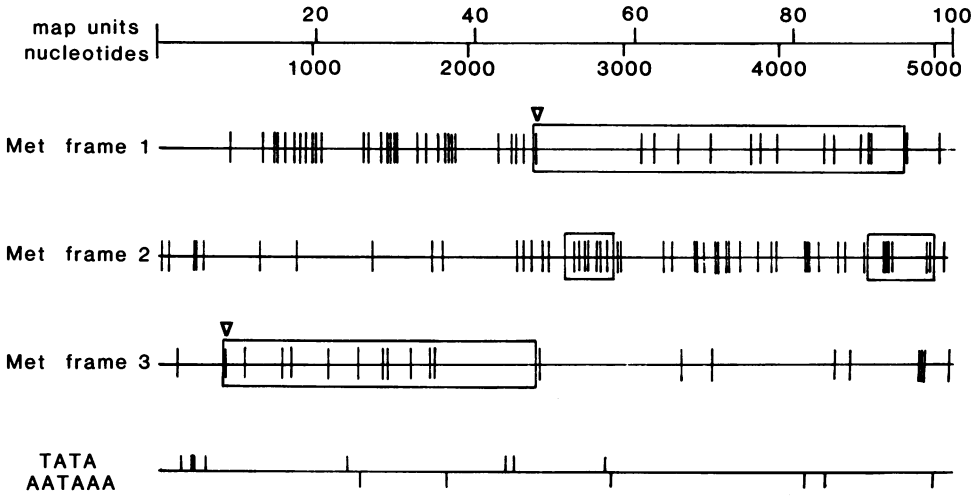


FIG. 8. Summary of methionine codons in reading frames 1, 2, and 3 of the B19-Au genome. Methionine codons are indicated by vertical bars, and for comparison TATA sequences and polyadenylation signals (AATAAA) are included below (see also Table 1 and Fig. 5A). Major ORFs are enclosed in boxes, and open triangles indicate the positions of the first methionine codons at the beginnings of the major left- and right-hand ORFs.

transcript. Hence, we favor the idea that the TATA sequence at nt 319 is the functional left-hand promotor, in which case transcription would begin at ~nt 350 (map unit ~6.5). The first downstream ATG codon at nt 436 is followed by an ORF ending with TGA at nt 2445.

The right half of the genome may be transcribed from a promoter at map unit ~45. Two TATA-like sequences are present at nt 2247 and 2308 (Fig. 5A; Table 1). The second, at nt 2308, is preceded by the sequence CCATT at nt 2283. The methionine codons downstream of these promoters are at nt 2325, 2363, 2406, and 2444 (Table 1). The first three are followed by in-frame termination codons within 50 nt; hence, they are either nonfunctional or provide leader peptides (by mRNA splicing) for the primary translation product (VP1), as has been suggested for other autonomous parvoviruses (3, 6, 12, 36, 39). This may also be the case with AAV2; however, the precise coding region for the AAV2 VP1 protein is uncertain (22). If none of these ATG signals in B19 (nt 2352, 2363, or 2406) is used, the one at nt 2444 is the second codon into the right-hand ORF which terminates with the TAA codon at nt 4787. If this ATG signal at nt 2444 is used, it is interesting that the left- and right-hand coding regions overlap by four nt.

A TATAAA sequence occurs at nt 1225 (~map unit 24) and, if functional, does not have a comparable promoter in other autonomous parvoviruses, although the AAV2 genome has a functional promoter within the middle of the left-hand ORF (Fig. 5C) at map unit 18 (nt 843). This B19 TATA-like sequence is preceded by an upstream CAAATT sequence nt 1195. There is also a TATAAA sequence at nt 2986 in the B19 genome (Fig. 5A) which also may be functional (see below), and it is preceded by an upstream TCATT sequence at nt 2950. A similar signal is seen in the FPV genome at nt 1923, although it is not present in the closely related CPV genome (12, 36).

Although we lack information on splicing of B19 transcripts, both large ORFs have ATG codons near the beginning; hence, we deduced a protein sequence of 671 amino acids for the left-hand region and one of 781 amino acids for the right. In all parvoviruses analyzed to date, the right-hand ORF has been shown to encode a nested set of structural proteins in which the amino acid sequence of the smaller proteins is entirely contained within the carboxy-terminal region of the largest. Similarly, procaryotic expression studies cited previously have shown that in B19 at least two structural polypeptides (with apparent molecular weights of 83,000 and 58,000) utilize a region of the genome between nt 2897 and 3749. Comparison of the genomes of MVM(p) (6), MVM(i) (3), H1 (39), CPV (36), and FPV (12) shows that the sequence near the small splice at map unit ~45 is highly conserved, and a search for splice site donor and acceptor positions has identified two donor sites (at nt 2280 and 2316) and one acceptor site (at nt 2399; see reference 3). In addition, recent cDNA sequencing studies have identified a second acceptor site at nt 2377 (C. V. Jongeneel, G. McMaster, R. Sahli, and B. Hirt, EMBO Workshop on Parvoviruses, abstr. no. P11, p. 64, 1985). Splicing from the second donor site (nt 2316) to the second acceptor site (nt 2399) is believed to produce an mRNA encoding VP1 (3). An alternate splice (the first donor site, nt 2280, to the acceptor site at nt 2377) removes upstream ATG sequences (in all three reading frames) such that the in-frame ATG codon encountered is approximately 675 nt from the start of the mRNA. Studies with H1 proteins have shown that this methionine codon is the translation start position for VP2 (31), and such an mRNA would therefore program the synthesis of the VP2 molecule, which is the major protein species produced by the virus. A number of cDNA clones with this splice junction have recently been identified (Jongeneel et al., EMBO Workshop on Parvoviruses, 1985). In comparison, the B19-Au genome contains many ATG codons (in all reading frames) between the presumed start of the transcript (~nt 2340), the first methionine codon in the ORF (nt 2444, frame 1), and the second in-phase methionine codon, which occurs 681 nt later at nt 3125 (Fig. 8). In B19, an alternate method for synthesis of VP2 in the same reading frame but several hundred nt downstream of the start of VP1 might be to utilize an additional promoter (e.g., the TATAAA sequence at nt 2986). This strategy would avoid the problem (seen with VP2 synthesis in MVM) of having the

ribosome ignore three ATG codons in the leader of the mRNA and recognize the fourth ~675 nt from the start of the messenger. In B19 the problem is accentuated with possibly 16 ATG codons before the in-frame one at nt 3125 (Fig. 8). Further analysis of transcription and coding regions within the B19 genome must await in vitro studies identifying functional promoters and splicing patterns.

The hydropathy curve of the right-hand ORF shows that hydrophilic and hydrophobic regions alternate through the protein with no large concentration of either (Fig. 6). This is in marked contrast to the situation in many of the autonomous parvoviruses (36), where the VP1-specific region of the capsid gene is markedly hydrophilic and contains a high concentration of basic residues which are thought to interact with and stabilize DNA in the intact virus particle. In all of the autonomous viruses for which sequence information is available, this highly charged region of VP1 is separated from the bulk of the capsid gene by a glycine-rich sequence of some 23 residues beginning approximately 22 residues downstream of the AUG initiating VP2. However, the right-hand major ORF of B19 does not contain such a sequence. In AAV, the precise coding region for the amino terminus of VP1 (protein A) is still uncertain. Janik et al. (22) have shown that this polypeptide is likely initiated within the small intervening sequence contained in the unspliced 2.6-kb RNA species (Fig. 5). There are three ATG codons (nt 1919, 1967, and 2213) in frame 2 followed by in-phase termination codons (nt 2165, 2177, and 2258). Janik et al. (22) proposed that one of these ATG codons may be the initiation codon for VP1 and downstream terminators are either read through or VP1 is translated from an mRNA species which is spliced to remove these terminator codons. This mRNA has so far not been detected, but if the splice were extremely small, this RNA may comigrate in gels with the 2.6-kb unspliced transcript. Another possibility is that the amino terminus of VP1 is encoded in either of the other two reading frames and that VP1 is generated by a frameshift to frame 2 (22). Based on our observations on amino acid homology among the equivalent regions of the VP1 proteins of AAV2, MVM, and B19 (see Results), we suggest that the VP1 protein of AAV2 is initiated in an alternate reading frame (frame 1), possibly at the ATG at nt 2203, and a small, so far undetected splice joins reading frame 1 to frame 2 beyond nt 2431. VP3 (protein C) of AAV2 is initiated at the ATG codon at nt 2810, and it is believed that VP2 (protein B) begins at a novel ACG codon at nt 2615 (9). Peptide homology searches suggest that the amino-terminal region of the B19 VP1 protein is not encoded in an alternate reading frame (although there is an alternative frame open in this region), since in B19 the 30-amino-acid sequence (nt 2815 to 2905) which is homologous in the VP1-specific region of B19, MVM, and AAV2 is located in the continuous ORF in B19.

Perhaps the most striking feature revealed by such searches is that B19 is as different from MVM(p) and AAV2 as those two viruses are from each other. Given that the overall homology between the three viruses is sufficiently widespread and extensive as to suggest divergent rather than convergent evolution, this would suggest that the ancestors of the three modern viruses separated at approximately the same point in evolutionary time. DNA hybridization studies make it clear that there are other members of the autonomous parvovirus genus, such as bovine parvovirus, which are only very distantly related to the KRV-type group (which includes KRV, MVM, H1, LuIII, FPV, and PPV). However, such studies also fail to reveal detectable nt homology between B19 and bovine parvovirus (17), and antibodies raised against bacterial peptides expressing B19 NS1 and capsid sequences fail to recognize bovine parvovirus proteins synthesized in vitro in a rabbit reticulocyte lysate system programmed with mRNA from bovine parvovirus-infected cells (Cotmore et al., manuscript in preparation). Since B19 appears to replicate in erythropoietin-stimulated human bone marrow cultures without the aid of a helper virus (Young, personal communication), it is probably an autonomous virus, suggesting that this genus contains viruses which are physically as different from each other as they are from the dependoviruses (AAVs). Clearly physical characteristics such as unique versus identical termini cannot be used to assign new viruses to a particular genus, and recent experiments suggest that even the ultimate distinction of their being able to replicate without the aid of a helper virus may require redefinition, since it seems that AAV5 may be able to replicate in some mutagen-treated cells in vitro without the help of another virus (43).

At the present time, the B19 genome appears to have features in common with both autonomous and helper-dependent viruses. Computer analyses of putative viral polypeptides from MVM(p), AAV2, and B19-Au have also identified extensive regions of homology between the polypeptides encoded by all three of these genomes.

Computer searches have identified a highly conserved 145-amino-acid sequence located in the middle of the NS1 gene of all three viruses. The function of the NS1 protein is not fully understood; however, it seems to be a multifunctional polypeptide. The gene product of the major left-hand ORF has been shown to be necessary for viral DNA replication (20, 50) as well as transactivation of the P38 promotor used for transcription of mRNA encoding viral structural proteins (37). In addition, it has been proposed that the protein may be the nicking enzyme responsible for resolving dimer replicative-form molecules (5). Sebring and Rose have observed an accumulation of AAV DNA concatemeric intermediates when AAV polypeptide synthesis is restricted (E. Sebring and J. Rose, personal communication). Presumably, the highly conserved amino acid sequence identified in Fig. 7 has a very specific role to play in at least one of these functions. A computer search of the data base for relatedness of the B19 sequence with other proteins failed to recognize any significant homologies.

The observation that AAV2 and MVM both share a region of 50% nt sequence homology with B19 seems inconsistent with the finding that B19 DNA hybridizes weakly to MVM DNA but not AAV2 DNA (17). However, analysis of the nucleotide sequence of this region presented in Fig. 7 shows that, whereas there are at most 9 nt in a row in the B19 sequence that match perfectly with the AAV2 sequence, this same region shares an 18-nt region of perfect homology with the MVM sequence. Since the formation of stable DNA-DNA hybrids in short regions of homology relies on stretches of perfect homology which are intact over a minimum stable length rather than a numerically equivalent but more dispersed type of homology, the sequence comparisons of B19 with MVM and AAV2 provide a cautionary note with regard to correlating low-level cross-hybridization, or lack thereof, with evolutionary relatedness.

In support of our sequence data we have now cloned the region from the *Xba*I site (nt 477) to the *Kpn*I site (nt 4079) from B19-Au and sequenced the region spanning the *Bam*HI site (nt 3896) in two different clones. This experiment confirmed that the sequence across the *Bam*HI site is correct (and that no small *Bam*HI fragment was lost during the original cloning). Also, in both clones the sequence at nt

3937 is AAAATTT (Fig. 2). Hence, our pYT103 clone is missing one A residue at nt 3940, presumably because of a cloning artifact (see the legend to Fig. 2).

ACKNOWLEDGMENTS

We thank C. Newton and C. Beard for advice on generating nested deletions for sequencing and G. M. Tener for generously supplying the αS-dNTPs. Also, we thank Masako Williams, Susan Heming, and Mae Soo for typing the manuscript.

R.O.S. is supported by a Killam postdoctoral fellowship and M.C.B. is supported by a MacMillan Family university graduate fellowship. This work was funded by grants from the BCHCRF and the Medical Research Council of Canada to C.R.A. and Public Health Service grant CA29303 to P.T. from the National Institutes of Health.

LITERATURE CITED

1. **Anderson, M. J., E. Lewis, I. M. Kidd, S. M. Hall, and B. J. Cohen.** 1984. An outbreak of erythema infectiosum associated with human parvovirus infection. J. Hyg. **93**:85–93.
2. **Armentrout, R., R. Bates, K. Berns, B. Carter, M. Chow, D. Dressler, W. Fife, W. Hauswirth, G. Hayward, G. Lavelle, S. Rhode, S. Straus, P. Tattersall, and D. Ward.** 1978. A standard nomenclature for restriction endonuclease fragments, p. 523–526. *In* D. Ward and P. Tattersall (ed), Replication of mammalian parvovirus. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
3. **Astell, C. R., E. M. Gardiner, and P. Tattersall.** 1986. DNA sequence of the lymphotropic variant of minute virus of mice, MVM(i), and its comparison with the DNA sequence of the fibrotropic prototype strain. J. Virol. **57**:656–659.
4. **Astell, C. R., M. Smith, M. B. Chow, and D. C. Ward.** 1979. Structure of the 3′ hairpin termini of four rodent parvovirus genomes: nucleotide sequence homology at origins of DNA replication. Cell **17**:691–792.
5. **Astell, C. R., M. Thomson, M. B. Chow, and D. C. Ward.** 1983. Structure and replication of minute virus of mice DNA. Cold Spring Harbor Symp. Quant. Biol. **47**:751–762.
6. **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.
7. **Bantel-Schaal, U., and H. zur Hausen.** 1984. Characterization of the DNA of a defective human parvovirus isolated from a genital site. Virology **134**:52–63.
8. **Bates, R. C., C. E. Snyder, P. T. Banerjee, and S. Mitra.** 1984. Autonomous parvovirus LuIII encapsidates equal amounts of plus and minus DNA strands. J. Virol. **49**:319–324.
9. **Becerra, S. P., J. A. Rose, M. Hardy, B. M. Baroudy, and C. W. Anderson.** 1985. Direct mapping of adeno-associated virus capsid proteins B and C: a possible ACG initiation codon. Proc. Natl. Acad. Sci. USA **82**:7919–7927.
10. **Berns, K. I., A. K. M. Cheung, J. M. Ostrove, and M. Lewis.** 1982. Adeno-associated virus latent infection, p. 249. *In* B. W. J. Mahy, A. C. Mirson, and G. K. Darby (ed.), Virus persistence. Cambridge University Press, New York.
11. **Brown, T., A. Anand, L. D. Ritchie, J. P. Clewley, and T. M. S. Reid.** 1984. Intrauterine parvovirus infection associated with hydrops fetalis. Lancet **ii**:1033–1034.
12. **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.
13. **Carter, B. J., C. A. Laughlin, and C. J. Marcus-Sekura.** 1984. Parvovirus transcription, p. 153–207. *In* K. I. Berns (ed.), The parvoviruses. Plenum Publishing Corp., New York.
14. **Cheung, A. K. M., M. D. Hoggan, W. W. Hauswirth, and K. I. Berns.** 1980. Integration of the adeno-associated virus genome into cellular DNA in latently infected human Detroit 6 cells. J. Virol. **33**:739–748.
15. **Cossart, Y. E., B. Cant, A. M. Field, and D. Widdows.** 1975. Parvovirus-like particles in human serum. Lancet **i**:72–73.
16. **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.
17. **Cotmore, S. F., and P. Tattersall.** 1984. Characterization and molecular cloning of a human parvovirus genome. Science **226**:1161–1165.
18. **Delaney, A.** 1982. A DNA sequence handling program. Nucleic Acids Res. **10**:61–67.
19. **Duncan, J. R., C. B. Potter, M. D. Cappelini, J. B. Kurtz, M. J. Anderson, and D. J. Weatherall.** 1983. Aplastic crisis due to parvovirus infection in pyruvate kinase deficiency. Lancet **ii**:14–16.
20. **Hermonat, P. L., M. A. Labow, R. Wright, K. I. Berns, and N. Muzyczka.** 1984. Genetics of adeno-associated virus: isolation and preliminary characterization of adeno-associated virus type 2 mutants. J. Virol. **51**:329–339.
21. **Hoggan, M. D.** 1970. Adenovirus associated viruses. Prog. Med. Virol. **12**:211–239.
22. **Janik, J. J., M. M. Huston, and J. A Rose.** 1984. Adeno-associated virus proteins: origin of the capsid components. J. Virol. **52**:591–597.
23. **Knott, P. D., G. A. C. Welply, and M. J. Anderson.** 1984. Serologically proved intrauterine infection with parvovirus. Br. Med. J. **289**:1660.
24. **Kyte, J., and R. F. Doolitte.** 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157**: 105–132.
25. **Lusby, E., K. H. Fife, and K. I. Berns.** 1980. Nucleotide sequence of the inverted terminal repetition in adeno-associated virus DNA. J. Virol. **34**:402–409.
26. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
27. **Maxam, A., and W. Gilbert.** 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA **74**:560–564.
28. **Mortimer, P. P.** 1983. Hypothesis: the aplastic crisis of hereditary spherocytosis is due to a single transmissible agent. J. Clin. Pathol. **36**:445–448.
29. **Mortimer, P. P., R. K. Humphries, J. G. Moore, R. H. Purcell, and N. S. Young.** 1983. A human parvovirus-like virus inhibits haematopoietic colony formation *in vitro*. Nature (London) **302**:426–429.
30. **Muller, D. E., and G. Siegl.** 1983. Maturation of parvovirus LuIII in a subcellular system. I. Optimal conditions for *in vitro* synthesis and encapsidation of viral DNA. J. Gen. Virol. **64**:1043–1054.
31. **Paradiso, P. R., K. R. Williams, and R. L. Constantino.** 1984. Mapping of the amino terminus of the H-1 parvovirus major capsid protein. J. Virol. **52**:77–81.
32. **Pattison, J. R., S. E. Jones, J. Hodgson, L. R. Davis, J. M. White, C. E. Stroud, and L. Murtaza.** 1981. Parvovirus infections and hypoplastic crisis in sickle-cell anemia. Lancet **i**:664–665.
33. **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.
34. **Rao, K. R., A. R. Patel, M. J. Anderson, J. Hodgson, S. E. Jones, and J. R. Pattison.** 1983. Infection with parvovirus-like virus and aplastic crisis in chronic hemolytic anemia. Ann. Intern. Med. **98**:930–932.
35. **Reid, D. M., T. Brown, T. M. S. Reid, J. A. N. Rennie, and C. J. Eastmond.** 1985. Human parvovirus-associated arthritis: a clinical and laboratory description. Lancet **ii**:422–425.
36. **Rhode, S. L., III.** 1985. Nucleotide sequence of the coat protein gene of canine parvovirus. J. Virol. **54**:630–633.
37. **Rhode, S. L., III.** 1985. *trans*-Activation of parvovirus P_{38} promoter by the 76K noncapsid protein. J. Virol. **55**:886–889.
38. **Rhode, S. L., III, and B. Klaassen.** 1982. DNA sequence of the 5′ terminus containing the replication origin of parvovirus replicative form DNA. J. Virol. **41**:990–999.
39. **Rhode, S. L., III, and P. R. Paradiso.** 1983. Parvovirus genome:

nucleotide sequence of H-1 and mapping of its genes by hybrid-arrest translation. J. Virol. **45**:173–184.

40. **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**:3617–3632.
41. **Samulski, R. J., A. Srivastava, K. I. Berns, and N. Muzyczka.** 1983. Rescue of adeno-associated virus from recombinant plasmids: gene correction within the terminal repeats of AAV. Cell **33**:135–145.
42. **Sanger, S., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
43. **Schlehofer, J. R., R. Heilbronn, B. Georg-Fries, and H. zur Hausen.** 1983. Inhibition of initiator-induced SV40 gene amplification in SV40-transformed chinese hamster cells by infection with a defective parvovirus. Int. J. Cancer **32**:591–595.
44. **Sergeant, G. R., J. M. Topley, K. Mason, B. E. Sergeant, J. R. Pattison, S. E. Jones, and R. Mohamed.** 1981. Outbreak of aplastic crisis in sickle-cell anemia associated with parvovirus-like agent. Lancet **ii**:595–598.
45. **Siegl, G.** 1984. Biology and pathogenicity of autonomous parvoviruses, p. 297–362. *In* K. I. Berns (ed.), The parvoviruses. Plenum Publishing Corp., New York.
46. **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.
47. **Smith, H. O., and M. L. Birnstiel.** 1976. A simple method for DNA restriction site mapping. Nucleic Acids Res. **3**:2387–2398.
48. **Srivastava, A., E. W. Lusby, and K. I. Berns.** 1983. Nucleotide sequence and organization of the adeno-associated virus 2 genome. J. Virol. **45**:555–564.
49. **Summers, J., S. E. Jones, and M. J. Anderson.** 1983. Characterization of the genome of the agent of erythrocyte aplasia permits its classfication as a human parvovirus. J. Gen. Virol. **64**:2527–2532.
50. **Tratschin, J.-D., I. L. Miller, and B. J. Carter.** 1984. Genetic analysis of adeno-associated virus: properties of deletion mutants constructed in vitro and evidence for an adeno-associated virus replication function. J. Virol. **51**:611–619.
51. **White, D. G., P. P. Mortimer, D. R. Blake, A. D. Woolf, B. J. Cohen, and P. A. Bacon.** 1985. Human parvovirus arthropathy. Lancet **ii**:419–421.