JOURNAL OF VIROLOGY, June 1986, p. 991–995
0022-538X/86/060991-05$02.00/0

Vol. 58, No. 3

Genome Organization and Nucleotide Sequence of Human Papillomavirus Type 33, Which Is Associated with Cervical Cancer

STEWART T. COLE* AND ROLF E. STREECK

Groupement de Génie Génétique, Institut Pasteur, 75724 Paris Cedex 15, France

Received 10 December 1985/Accepted 13 February 1986

The 7,909-nucleotide sequence of human papillomavirus type 33, which is associated with cervical cancer, has been determined and used to deduce the corresponding genome arrangement. Extensive sequence homologies and other genetic features are shared with the related oncogenic virus, human papillomavirus type 16, especially in the major reading frames. A surprising difference was found in the noncoding region of human papillomavirus type 33 as, unlike all other sequenced papillomaviruses, it contains a perfect 78-base pair tandem repeat.

Papillomaviruses are members of the papovavirus family and possess a genome of about 7,900 base pairs (bp) consisting of a covalently closed circular DNA molecule. Human papillomaviruses (HPV) are classified on the basis of their DNA sequence homology (6) and nearly 40 types have now been described. Considerable insight into HPV biology and their involvement in human disease has been attained by the application of the techniques of molecular biology. A possible role for HPVs in human cancer had been suspected for several years and was further supported by the frequent detection of HPV DNA in tumors resulting from the malignant conversion of cutaneous lesions (17) and genital warts (32). The cloning of two HPV genomes, HPV-16 and HPV-18 (3, 10), from cervical carcinomas has further stimulated research in this field. These viruses were discovered in more than 70% of the malignant genital tumors examined, and in many others HPV-16-related sequences were detected (3, 32). Among these is HPV-33, which was recently cloned from an invasive cervical carcinoma, using HPV-16 as a probe under conditions of reduced stringency (S. Beaudenon, D. Kremsdorf, O. Croissant, S. Jablonska, S. Wain-Hobson, and G. Orth, Nature (London), in press). In the present study we have determined the DNA sequence of an episomal form of HPV-33 and describe its relationship to HPV-16.

The complete 7,909-nucleotide sequence of HPV-33, determined by the M13 shotgun cloning/dideoxy sequencing approach, is presented in Fig. 1. On average, each position was sequenced 6.5 times and 92% of the sequence was obtained from both strands. In agreement with the convention for other papillomavirus sequences, the numbering begins at a site resembling the recognition sequence for *Hpa*I in the noncoding region.

An analysis of the distribution of nonsense codons shows that, as in all other sequenced papillomaviruses, the eight major open reading frames are located on the same strand (Fig. 2). Some features common to HPV-33 and HPV types 1a, 6b, and 16 together with the cottontail rabbit papillomavirus and bovine papillomavirus type 1, BPV-1 (5, 7, 8, 12, 20, 21), include the overlap between the largest open reading frames in the early region, E1 and E2, and the inclusion of E4 within the section encoding E2. The *Bgl*II site used in the molecular cloning of HPV-33 is situated within the E1/E2 overlap and, as both reading frames are unaffected, the possibility of clustered *Bgl*II sites can be excluded. Another property common to all papillomaviruses, except BPV-1, is the overlap between the L1 and L2 reading frames. Following L1 is the 892-bp noncoding region which, by analogy with BPV-1 (14, 28), undoubtedly contains the origin of replication and various transcriptional regulatory elements. The principal characteristics of the HPV-33 genome are summarized in Table 1.

HPV-16 is the only other oncogenic papillomavirus, isolated from tumors of the anogenital region, which has been completely sequenced (21). The gross features of HPV-33 resemble those of HPV-16 except that the E1 reading frame of the latter is interrupted. All of the coding sequences in HPV-33, except that of E5, are slightly shorter than their counterparts in HPV-16. This may contribute to the fact that its noncoding region, between L1 and E6 (Fig. 2), is 76 bp longer, thereby keeping the genomes nearly constant in size.

When the open reading frames were compared pairwise (Table 2), it was found that E1, E2, E6, E7, L1, and L2 displayed between 65 and 75% homology, whereas those for E4 and E5 were more divergent (about 50% homology). These findings confirm the heteroduplex analysis performed previously (Beaudenon et al., in press). A comparative study (7) of papillomavirus E1 gene products showed that the polypeptide consists of an NH_2-terminal segment, the sequence of which is highly variable, and a COOH-terminal domain of well-conserved primary structure. The longest

TABLE 1. Principal features of the HPV-33 genome

Open reading frame	Start	First ATG	Stop Codon	Predicted mol wt[a]
E6	76	109	556 TGA	17,632
E7	543	573	864 TAA	10,825
E1	867	879	2811 TGA	72,387
E2	2728	2749	3808 TAA	40,207
E4	3326		3575 TAG	9,452
E5	3842	3854	4079 TAA	9,895
L2	4198	4210	5611 TAG	50,539
L1	5516	5594	7091 TAA	55,839

[a] Calculated from the first ATG where this exists or from the start of the open reading frame.

* Corresponding author.

```
   1 GTAAACTATA ATGCCAAGTT TTAAAAAAGT AGGGTGTAAC CGAAAGCGGT TCAACCGAAA ACGGTGCATA TATAAAGCAA ACATTTTGCA GTAAGGTACT
 101 GCACGACTAT GTTTCAAGAC ACTGAGGAAA AACCACGAAC ATTGCATGAT TTGTGCCAAG CATTGGAGAC AACTATACAC AACATTGAAC TACAGTGCGT
 201 GGAATGCAAA AAACCTTTGC AACGATCTGA GGTATATGAT TTTGCATTTG CAGATTTAAC AGTTGTATAT AGAGAGGGAA ATCCATTTGG AATATGTAAA
 301 CTGTGTTTGC GGTTCTTATC TAAAATTAGT GAATATAGAC ATTATAATTA TTCTGTATAT GGAAATACAT TAGAACAAAC AGTTAAAAAA CCTTTAAATG
 401 AAATATTAAT TAGGTGTATT ATATGTCAAA GACCTTTGTG TCCTCAAGAA AAAAAACGAC ATGTGGATTT AAACAAACGA TTTCATAATA TTTCGGGTCG
 501 TTGGGCAGGG CGCTGTGCGG CGTGTTGGAG GTCCCGACGT AGAGAAACTG CACTGTGACG TGTAAAAACG CCATGAGAGG ACACAAGCCA ACGTTAAAGG
 601 AATATGTTTT AGATTTATAT CCTGAACCAA CTGACCTATA CTGCTATGAG CAATTAAGTG ACAGCTCAGA TGAGGATGAA GGCTTGGACC GGCCAGATGG
 701 ACAAGCACAA CCAGCCACAG CTGATTACTA CATTGTAACC TGTTGTCACA CTTGTAACAC CACAGTTCGT TTATGTGTCA ACAGTACAGC AAGTGACCTA
 801 CGAACCATAC AGCAACTACT TATGGGCACA GTGAATATTG TGTGCCCTAC CTGTGCACAA CAATAAACAT CATCTACAAT GGCCGATCCT GAAGGTACAA
 901 ATGGGGCTGG GATGGGGTGT ACTGGTTGGT TTGAGGTAGA AGCAGTCATA GAGAGAAGAA CAGGAGATAA TATTTCAGAA GATGAGGATG AAACAGCAGA
1001 TGACAGTGGC ACGGATTTAC TAGAGTTTAT AGATGATTCT ATGGAAAATA GTATACAGGC AGACACAGAG GCAGCCCGGG CATTGTTTAA TATACAGGAA
1101 GGGGAGGATG ATTTAAATGC TGTGTGTGCA CTAAAACGAA AGTTTGCCGC ATGTTCACAA AGTGCTGCGG AGGACGTTGT TGATCGTGCT GCAAACCCGT
1201 GTAGAACGTC TATTAATAAA AATAAAGAAT GCACATACAG AAAACGAAAA ATAGATGAGC TAGAAGACAG CGGATATGGC AATACTGAAG TGGAAACTCA
1301 GCAGATGGTA CAACAGGTAG AAAGTCAAAA TGGCGACACA AACTTAAATG ACTTAGAATC TAGTGGGGTG GGGGATGATT CAGAAGTAAG CTGTGAGACA
1401 AATGTAGATA GCTGTGAAAA TGTTACGTTG CAGGAAATTA GTAATGTTCT ACATAGTAGT AATACAAAAG CAAATATATT ATATAAATTT AAAGAGGCCT
1501 ATGGAATAAG TTTTATGGAA TTAGTAAGAC CATTTAAAAG TGATAAAACA AGCTGTACAG ATTGGTGTAT AACAGGATAT GGAATTAGTC CATCAGTAGC
1601 AGAAAGTTTA AAAGTATTAA TTAAACAGCA TAGTTTGTAT ACTCATTTAC AATGTTTAAC TTGCGATAGA GGAATAATAA TATTATTGTT AATTAGATTT
1701 AGGTGTAGCA AAAACAGGTT AACAGTAGCA AAACTAATGA GTAATTTATT ATCAATACCT GAAACATGTA TGGTTATAGA GCCACCAAAA TTACGGAGCC
1801 AAACATGTGC ATTGTATTGG TTTAGAACAG CAATGTCAAA CATTAGTGAT GTACAAGGTA CAACACCTGA ATGGATAGAT AGACTAACTG TTTTACAACA
1901 TAGCTTTAAT GATAATATAT TTGATTTAAG TGAAATGGTA CAGTGGGCAT ATGATAACGA GTTAACGGAC GATAGTGACA TTGCATATTA TTATGCACAA
2001 CTTGCAGATT CAAATAGTAA TGCTGCTGCA TTTTTAAAAA GTAACTCACA AGCAAAAATA GTAAAGGACT GTGGAATAAT GTGTAGACAT TATAAAAAAG
2101 CAGAAAAACG TAAAATGTCA ATAGGACAAT GGATACAAAG TAGATGTGAA AAAACAAATG ATGGAGGAAA TTGGAGACCA ATAGTACAGT TGTTAAGATA
2201 TCAAAACATT GAATTTACAG CATTTTTAGG TGCATTTAAA AAGTTTTTAA AAGGTATACC AAAAAAAAGC TGTATGCTAA TTTGTGGACC AGCAAATACA
2301 GGAAAGTCAT ATTTTGGAAT GAGTTTAATA CAGTTTTAA AAGGGTGTGT TATATCATGT GTAAATTCTA AAAGTCACTT TTGGTTGCAG CCATTATCAG
2401 ATGCAAAAAT AGGAATGATA GATGATGTAA CGCCAATAAG TTGGACATAT ATAGATGATT ACATGAGAAA TGCGTTAGAT GGAAATGAAA TTTCAATAGA
2501 TGTGAAACAT AGGGCATTAG TGCAATTAAA ATGTCCACCA CTGCTTCTTA CCTCAAATAC AAATGCAGGC ACAGACTCTA GATGGCCATA TTTACATAGT
2601 AGATTAACAG TATTTGAATT TAAAAATCCA TTCCCATTTG ATGAAAATGG TAACCCAGTG TATGCAATAA ATGATGAAAA TTGGAAATCC TTTTTCTCAA
2701 GGACGTGGTG CAAATTAGAT TTAATAGAGC AAGAGGACAA GGAAAACCAT GGAGGAAATA TCAGCACGTT TAAATGCAGT GCAGGAGAAA ATACTAGATC
2801 TTTACGAAGC TGATAAAACT GATTTACCAT CACAAATTGA ACATTGGAAA CTGATACGCA TGGAGTGTGC TTTATTGTAT ACAGCCAAAC AAATGGGATT
2901 TTCACATTTA TGCCACCAGG TGGTGCCTTC TTTGTTAGCA TCAAAGACCA AAGCATTTCA AGTAATTGAA CTACAAATGG CATTAGAGAC ATTAAGTAAA
3001 TCACAGTATA GTACAAGCCA ATGGACATTG CAACAAACAA GCTTAGAGGT GTGGCTTTGT GAACCACCAA AATGTTTTAA AAAACAAGGA GAAACAGTAA
3101 CTGTGCAATA TGACAATGAC AAAAAAAATA CAATGGATTA TACAAACTGG GGTGAAATAT ATATTATAGA GGAAGATACA TGTACAGTAG TTACAGGGAA
3201 AGTAGATTAT ATAGGTATGT ATTATATACA TAACTGTGAA AAGGTATATT TTAAATATTT TAAAGAGGAT GCTGCAAAGT ATTCTAAAAC ACAAATGTGG
3301 GAAGTACATG TGGGTGGTCA GGTAATTGTT TGTCCTACGT CTATATCTAG CAACCAAATA TCCACTACTG AAACTGCTGA CATACAGACA GACAACGATA
3401 ACCGACCACC ACAAGCAGCG GCCAAACGAC GACGACCTGC AGACACCACA GACACCGCCC AGCCCCTTAC AAAGCTGTTC TGTGCAGACC CCGCCTTGGA
3501 CAATAGAACA GCACGTACTG CAACTAACTG CACAAACAAG CAGCGGACTG TGTGTAGTTC TAACGTTGCA CCTATAGTGC ATTTAAAAGG TGAATCAAAT
3601 AGTTTAAAAT GTTTAAGATA CAGATTAAAA CCTTATAAAG AGTTGTATAG TTCTATGTCA TCCACCTGGC ATTGGACCAG TGACAACAAA AATAGTAAAA
3701 ATGGAATTGT AACTGTAACA TTTGTAACTG AACAGCAACA ACAAATGTTT TTAGGTACCG TAAAAATACC ACCTACTGTG CAAATAAGTA CTGGATTTAT
3801 GACATTATAA GTGTACATCA CAAGCCAATA TGTGCTGCTA ATTGTATATA ACCATGATAT TTGTTTTTGT ATTATGTTTT ATATTGTTTT TATGCTTATC
3901 CTTATTATTA CGTCCTTTAA TACTTTCCAT TTCTACCTAT GCTTGGTTGC TGGTGTTGGT ATTGCTGCTT TGGGTGTTTG TGGGATCTCC TTTAAAAATT
4001 TTTTTTTGCT ATTTGTTGTT TTTATATTTA CCAATGATGT GTATTAATTT TCATGCACAG CATATGACAC AACAAGAGTA ATGTATATAC ATGTATATAT
4101 TGTTTGTATA TATGTGCACA TGGTGGTGTT TTAACATTGT TGTTGTTATT TTAGTTTTTT TTTTTTGTA TTACTAATAA ATACCTTTAT ATTTTAGCAG
4201 TGTATTATTA TGAGACACAA ACGATCTACA AGGCGCAAGC GTGCATCTGC AACACAACTA TACCAAACAT GCAAGGCCAC AGGCACCTGC CCACCCGATG
4301 TTATTCCTAA AGTGGAAGGA AGTACCATAG CAGATCAAAT TCTTAAATAT GGCAGTTTAG GGGTTTTTTT TGGTGGTTTA GGTATTGGCA CAGGCTCTGG
4401 TTCAGGTGGA AGGACTGGCT ATGTACCTAT TGGTACTGAC CCACCTACAG CTGCAATCCC CTTGCAGCCT ATACGTCCTC CGGTTACTGT AGACACTGTT
4501 GGACCTTTAG ACTCGTCTAT AGTGTCATTA ATAGAAGAAA CAAGTTTTAT AGAGGCAGGT GCACCAGCCC CATCTATTCC TACACCATCA GGTTTTGATG
4601 TTACTACATC TGCAGATACT ACACCTGCAA TTATTAATGT TTCATCTGTT GGGGAGTCAT CTATTCAAAC TATTTCTACA CATTTAAATC CCACATTTAC
4701 TGAACCATCT GTACTACACC CTCCAGCGCC TGCAGAAGCC TCTGGACATT TTATATTTTC TTCCCCTACT GTTAGCACAC AAAGTTATGA AAACATACCA
4801 ATGGATACCT TTGTTGTTTC CACAGACAGT AGTAATGTAA CATCAAGCAC GCCCATTCCA GGGTCTCGCC CTGTGGCACG CCTTGGTTTA TATAGTCGCA
4901 ATACCCAACA GGTTAAGGTT GTTGACCCTG CTTTTTTAAC ATCGCCTCAT AAACTTATAA CATATGATAA TCCTGCATTT GAAAGCTTTG ACCCTGAAGA
5001 CACATTACAA TTTCAACATA GTGATATATC ACCTGCTCCT GATCCTGACT TTCTAGATAT TATTGCATTA CATAGGCCTG CTATTACATC TCGTAGACAT
5101 ACTGTGCGTT TTAGTAGAGT AGGTCAAAAA GCCACACTTA AAACTCGCAG TGGTAAACAA ATTGGAGCTA GAATACATTA TTATCAGGAT TTAAGTCCTA
5201 TTGTGCCTTT AGACCACACC GTGCCAAATG AACAATATGA ATTACAGCCT TTACATGATA CTTCTACATC GTCTTATAGT ATTAATGATG GTTTGTATGA
5301 TGTTTATGCT GACGATGTGG ATAATGTACA CACCCCAATG CAACACTCAT ACAGTACGTT TGCAACAACA CGTACCAGCA ATGTGTCTAT ACCTTTAAAT
5401 ACAGGATTTG ATACTCCTGT TATGTCTGGC CCTGATATAC CTTCCCCTTT ATTTCCCACA TCTAGCCCAT TTGTTCCTAT TTCGCCTTTT TTTCCTTTTG
5501 ACACCATTGT TGTAGACGGT GCTGACTTTG TTTTACATCC TAGTTATTTT ATTTTACGTC GCAGGCGTAA ACGTTTTCCA TATTTTTTTA CAGATGTCCG
5601 TGTGGCGGCC TAGTGAGGCC ACAGTGTACC TGCCTCCTGT ACCTGTATCT AAAGTTGTCA GCACTGATGA ATATGTGTCT CGCACAAGCA TTTATTATTA
5701 TGCTGGTAGT TCCAGACTTC TTGCTGTTGG CCATCCATAT TTTTCTATTA AAAATCCTAC TAACGCTAAA AAATTATTGG TACCCAAAGT ATCAGGCTTG
5801 CAATATAGGG TTTTTAGGGT CCGTTTACCA GATCCTAATA AATTTGGATT TCCTGACACC TCCTTTTATA ACCCTGATAC ACAACGATTA GTATGGGCAT
5901 GTGTAGGCCT TGAAATAGGT AGAGGGCAGC CATTAGGCGT TGGCATAAGT GGTCATCCTT TATTAAACAA ATTTGATGAC ACTGAAACCG GTAACAAGTA
6001 TCCTGGACAA CCGGGTGCTG ATAATAGGGA ATGTTTATCC ATGGATTATA AACAAACACA GTTATGTTTA CTTGGATGTA AGCCTCCAAC AGGGGAACAT
6101 TGGGGTAAAG GTGTTGCTTG TACTAATGCA GCACCTGCCA ATGATTGTCC ACCTTTAGAA CTTATAAATA CTATTATTGA GGATGGTGAT ATGGTGGACA
6201 CAGGATTTGG TTGCATGGAT TTTAAAACAT TGCAGGCTAA TAAAAGTGAT GTTCCTATTG ATATTTGTGG CAGTACATGC AAATATCCAG ATTATTTAAA
6301 AATGACTAGT GAGCCTTATG GTGATAGTTT ATTTTTCTTT CTTCGACGTG AACAAATGTT TGTAAGACAC TTTTTTAATA GGGCTGGTAC ATTAGGAGAG
6401 GCTGTTCCCG ATGACCTGTA CATTAAAGGT TCAGGAACTA CTGCCTCTAT TCAAAGCAGT GCTTTTTTTC CCACTCCTAG TGGATCAATG GTTACTTCCG
6501 AATCTCAGTT ATTTAATAAG CCATATTGGC TACAACGTGC ACAAGGTCAT AATAATGGTA TTTGTTGGGG CAATCAGGTA TTTGTTACTG TGGTAGATAC
6601 CACTCGCAGT ACTAATATGA CTTTATGCAC ACAAGTAACT AGTGACAGTA CATATAAAAA TGAAAATTTT AAAGAATATA TAAGACATGT TGAAGAATAT
6701 GATCTACAGT TTGTTTTTCA ACTATGCAAA GTTACCTTAA CTGCAGAAGT TATGACATAT ATTCATGCTA TGAATCCAGA TATTTTAGAA GATTGGCAAT
6801 TTGGTTTAAC ACCTCCTCCA TCTGCTAGTT TACAGGATAC CTATAGGTTT GTTACCTCTC AGGCTATTAC GTGTCAAAAA ACAGTACCTC CAAAGGAAAA
6901 GGAAGACCCC TTAGGTAAAT ATACATTTTG GGAAGTGGAT TTAAAGGAAA AATTTTCAGC AGATTTAGAT CAGTTTCCTT TGGGACGCAA GTTTTTATTA
7001 CAGGCAGGTC TTAAAGCAAA ACCTAAACTT AAACGTGCAG CCCCCACATC CACCCGCACA TCGTCTGCAA AACGCAAAAA GGTTAAAAAA TAACACTTTG
7101 TGTAATTGTG TTATGTTGTT GTTTTGTTCT GTCTATGTAC TTTGTGTTGT TGTGTTGTGT TGTTGTTTGT TTTTTGTGTA TGTGTTACAA TGTATGTTAT
7201 GTTGTATGTT ACTGTGTTTG TTTTATGTGT ACTTGTTTGT GTGCATGTTC TATGTACTTG TCAGTTTCCT GTTTGTGTAT ATGTTAATAA AACATTGTGT
7301 GTATTTGTTA AACTATTTGT ATGTATGTTA TGTATATGGG TGTACCTATA TGAGTAAGGA GTTGTATTGC TTGCCCTACC CTGCATTGCA ATGTACCTAC
7401 CTTTATTTCC CTATATTTGT AGTACCTACA TGTTTAGTAT TGCTTTACCT TTTGACATAC TAGTGTCCAT ATTGTACAAT TTCCTCCATT TTGTATGCCT
7501 AACCGTTTTC GGTTACTTGG CATACATACC CTATGACATT GGCAGAACAG TTAATCCTTT TCTTTCCTGC ACTGTGTTTG TCTGTACTTG CTGCATTGGC
7601 ATACATACCC TATGACATTG GCAGAACAGT TAATCCTTTT CTTTCCTGCA CTGTGTTTGT CTGTACTTGC TGCATTGACT CATATATACA TGCAGTGCAA
7701 TTGCAAAATA CTTAATTGTA CTAATAGTTT ACACATGCTT TTAGGCACAT ATTTTTACTT TACTTTCAAA CCTTAAGTGC AGTTTTGGCT TACACAATTG
7801 CTTTGTATGC CAAACTATGC CTTGTAAAAG TGAGTCACTA CCTGTTTATT ACCAGGTGTG GACTAACCGT TTTAGGTCAT ATTGGTCATT TATAATCTTT
7901 TATATAATA
```

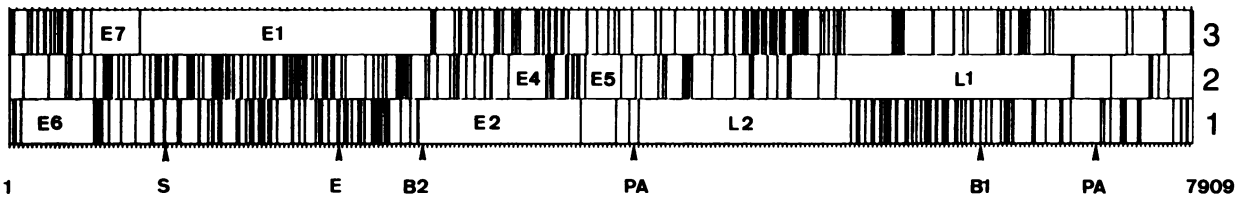

FIG. 2. Distribution of the major reading frames in the HPV-33 genome. The reading frames were identified by comparison with other HPV sequences, and the stop codons are represented as vertical bars. Also indicated are the locations of unique restriction sites (S, *Sma*I; E, *Eco*RV; B2, *Bgl*II; B1, *Bgl*I) and the likely polyadenylation signals (PA) for the early and late transcripts. In addition to these, six other potential PA sites (AATAAA) were detected at positions 862, 1215, 1221, 2666, 5837, and 6239.

stretch of perfect sequence homology, 33 nucleotides (positions 1275 to 1307, Fig. 1) is found near the 5′ end of the E1 reading frame in a region encoding the variable domain of the polypeptide. Several other regions of complete identity (19 to 28 nucleotides) were detected elsewhere in E1, and also in E2, L2, and L1. As many of these sequences are not found in the genomes of other HPVs, such as HPV-1a and HPV-6b, this raises the possibility that the corresponding oligonucleotides could be produced and used as diagnostic hybridization probes for screening biopsy material from potentially tumorigenic lesions.

The papillomavirus gene products may be divided into those believed to play a purely structural role, L1 and L2, and those required for viral propagation and persistence. The results of a comparison of the probable products of the major reading frames from HPV-33, -16, and -6b are summarized in Table 2. As expected, there is strong identity between the oncogenic HPV-33 and -16, particularly for the proposed E1, E6, E7, L2, and L1 proteins. When conservative substitutions are included the homology between the two L1 polypeptides increases to 90%, suggesting that the corresponding capsids must be antigenically related. In contrast, significantly weaker homologies were detected when the analysis was extended to include the benign genital wart-forming HPV-6b (Table 2). Comparison of the HPV-16 proteins with those of HPV-6b revealed slightly more homology than was found with HPV-33, suggesting a closer evolutionary relationship.

The noncoding region of HPV-33 displays several unique properties and bears only weak resemblance to its homolog in HPV-16. Shortly after the L1 stop codon is a 223-bp stretch of DNA (positions 7097 to 7320, Fig. 1) which is unusually rich in thymine plus guanine (79%) and includes the putative polyadenylation signal for the late transcripts. Contained within this segment of the noncoding region are two copies of a 19-bp direct repeat (with one mismatch) and seven copies of the motif TTGTRTR (where R is A or G). The latter is also found seven times in the corresponding region of HPV-16, suggesting that it may represent a recognition site for proteins involved in replication. It should be noted that nascent replication forks have been localized in this region of the BPV-1 genome (28) and that the origin of replication of the Epstein-Barr virus consists of a family of repeated sequences (31).

A 12-bp palindrome (ACCG. . . .CGGT) that occurs exclusively in the noncoding region of all papillomavirus genomes examined was recently reported by Dartmann et al. (K. Dartmann, E. Schwarz, L. Gissmann, and H. Zur Hausen, Virology, in press). Three copies were found in the HPV-33 genome (Fig. 3) and these occupy the same positions in the noncoding region of HPV-16. A role for the palindrome as a possible control site for the early promoter was proposed (4, 14; Dartmann et al., in press), and indirect support is provided by our finding that the noncoding regions of HPVs, such as HPV-33, do not display the clustered arrangement of recognition sites for the promoter-specific activation factor Sp1 (11). This is in direct contrast to the situation in another papovavirus, simian virus 40 (SV40) (11, 13).

The most striking feature of HPV-33 is a perfect 78-bp tandem repeat located 200 bp after the putative origin of replication (Fig. 3). No other repeats of this size or sequence have been described in the genomes of other papillomaviruses. The presumed early promoter for HPV-33 is located about 300 bp downstream from the tandem repeat, and the characteristic promoter elements (4) could be iden-

TABLE 2. Comparison of HPV proteins

Protein	% Homology[a] of HPV:		
	33 vs 16	33 vs 6b	16 vs 6b
E6	65 (70)	36 (51)	37
E7	61 (69)	55 (60)	56
E1	61 (69)	50 (60)	53
E2	53 (65)	46 (58)	45
E4	52 (55)	39 (46)	48
E5	40 (52)	39 (43)	33
L2	64 (66)	52 (58)	53
L1	81 (75)	68 (69)	71

[a] Expressed as percent homology after alignment with the program of reference 30. Values in parentheses represent percent nucleotide sequence homology.

FIG. 1. Nucleotide sequence of HPV-33. Position 1 on the circular genome corresponds to an "*Hpa*I-like" sequence found by alignment with HPV-6b. The source of HPV-33 was plasmid p15-5 (Beaudenon et al., submitted for publication) which consists of an episomal HPV-33 genome, linearized with *Bgl*II, cloned in a pBR322 derivative. A library of random DNA fragments (400 to 800 bp) was prepared in M13mp8 (15) after sonication of p15-5, essentially as described previously (27). DNA sequencing was performed by the modified dideoxy chain termination method (2, 18, 19). A small part of the noncoding region was found to be absent or underrepresented in the M13 library (>300 clones), and its sequence was obtained directly from p15-5 by the method of Smith (23). Briefly, restriction fragments isolated from 2 "complementary" M13 clones were used to prime DNA synthesis on templates prepared from p15-5 which had been linearized with a restriction enzyme and then treated with exonuclease III (200 U/pmol of DNA for 1 h at 22°C). DNA sequences were compiled and analysed with the programs of Staden (25, 26).

FIG. 3. Principal features of the noncoding region. A section of the noncoding region from positions 7500 to 114 is shown. The 78-bp tandem repeats are overlined, and those regions resembling the Z-DNA-forming element of the SV40 enhancer are indicated. Potential promoter elements are denoted by stars, and the three copies of the 12-bp palindrome are enclosed between two rows of dots.

tified (Fig. 3). The size, position, and arrangement of the 78-bp repeats in the HPV-33 genome suggest that they may function as enhancers of viral transcription. Tandem repeats of 72, 73, and 68 bp have been located near the early promoter of SV40 (1, 4, 13), in the long terminal repeat of Moloney murine sarcoma virus (9), and in the BK virus genome (22) and shown to enhance transcription from *Pol*II-dependent promoters in a *cis*-active manner. From mutagenesis of the SV40 enhancer (13, 29) and sequence comparisons of characterized transcriptional activators, a consensus enhancer sequence was derived. This structure could not be detected in the 78-bp repeat, but a potential Z-DNA-forming region was uncovered. Z-DNA is believed to attract regulatory molecules to eucaryotic promoters and a Z-DNA antibody-binding site has been demonstrated within the SV40 enhancer (16). The sequence to which this antibody binds is also found, albeit with a single mismatch, in the putative HPV-33 enhancer (positions 7520 to 7527 and 7599 to 7606, Fig. 1 and 3).

The proposed HPV-33 enhancer shows no extended sequence homology to the well-characterized enhancers or to other papillomavirus regulatory regions. However, it has recently been demonstrated that an enhancer-like element is located in the noncoding region of BPV-1 and that it requires the E2 product for activation (24). These findings support our proposal that the 78-bp tandem repeats could have enhancer function and may indicate that the relatively low homology (Table 2) between the E2 proteins of HPV-33 and -16 reflects a specificity for the corresponding enhancer/regulatory regions. Experiments are currently in progress to substantiate some of these possibilities.

We thank Brigitte Lemercier for help with some of the experiments, Sylvie Beaudenon and Gérard Orth for the cloned HPV-33 genome and for communicating unpublished results, and Olivier Danos for stimulating discussions.

LITERATURE CITED

1. **Benoist, C., and P. Chambon.** 1981. In vivo sequence requirements of the SV40 early promoter region. Nature (London) **290**:304–310.
2. **Biggin, M. D., T. J. Gibson, and G. F. Hong.** 1983. Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. USA **80**:3963–3965.
3. **Boshart, M., L. Gissmann, H. Ikenburg, A. Kleinheinz, W. Scheurlen, and H. Zur Hausen.** 1984. A new type of papillomavirus DNA, its presence in genital cancer biopsies and in cell lines derived from cervical cancer. EMBO J. **3**:1151–1157.
4. **Breathnach, R., and P. Chambon.** 1981. Organization and expression of eukaryotic split genes coding for proteins. Annu. Rev. Biochem. **50**:349–383.
5. **Chen, Y., P. M. Howley, A. D. Levinson, and P. M. Seeburg.** 1982. The primary structure and genetic organization of the bovine papillomavirus type 1 genome. Nature (London) **299**:529–534.
6. **Coggin, J. R., and H. Zur Hausen.** 1979. Workshop on papillomaviruses and cancer. Cancer Res. **39**:545–546.
7. **Danos, O., I. Giri, F. Thierry, and M. Yaniv.** 1984. Papillomavirus genomes: sequences and consequences. J. Invest. Dermatol. **83**:7S–11S.
8. **Danos, O., M. Katinka, and M. Yaniv.** 1982. Human papillomavirus la complete DNA sequence: a novel type of genome organization among papovaviridae. EMBO J. **1**:231–236.
9. **Dhar, R., W. L. McClements, L. W. Enquist, and G. F. Vande-Woude.** 1980. Nucleotide sequence of integrated Moloney sarcoma provirus long terminal repeats and their host and viral functions. Proc. Natl. Acad. Sci. USA **77**:3937–3941.
10. **Durst, M., L. Gissmann, H. Ikenburg, and H. Zur Hausen.** 1983. A papillomavirus DNA from a cervical carcinoma and its prevalence in cancer biopsy samples from different geographic regions. Proc. Natl. Acad. Sci. USA **80**:3812–3815.
11. **Dynan, W. S., and R. Tijan.** 1985. Control of eukaryotic messenger RNA synthesis by sequence-specific DNA-binding proteins. Nature (London) **316**:774–778.
12. **Giri, I., O. Danos, and M. Yaniv.** 1985. Genomic structure of the cottontail rabbit (Shope) papillomavirus. Proc. Natl. Acad. Sci. USA **82**:1580–1584.
13. **Gruss, P., R. Dhar, and G. Khoury.** 1981. Simian virus 40 tandem repeated sequences as an element of the early promoter. Proc. Natl. Acad. Sci. USA **78**:943–947.
14. **Howley, P. M., Y. C. Yang, and M. S. Rabson.** 1985. The molecular biology of bovine papillomaviruses, p. 67–81. *In* P. W. J. Rigby and N. M. Wilkie (ed.), Viruses and cancer. Society for General Microbiology Symposium 37. Cambridge University Press, Cambridge.
15. **Messing, J., and J. Vieira.** 1982. A new pair of M13 vectors for selecting either DNA strand of double digest restriction fragments. Gene **19**:269–276.
16. **Nordheim, A., and A. Rich.** 1983. Z-DNA formation in the enhancer region of supercoiled SV40, p. 45–50. *In* Y. Gluzman and T. Shenk (ed.), Enhancers and eukaryotic gene expression. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
17. **Orth, G., M. Favre, F. Breitburd, S. Jablonka, S. Obalek, M. Jarzabek-Chorzelska, and G. Rzesa.** 1980. Epidermodysplasia verruciformis: a model for the role of papillomaviruses in human cancer. Cold Spring Harbor Conf. Cell Prolif. **7**:259–282.
18. **Sanger, F., A. R. Coulson, B. G. Barrel, A. J. H. Smith, and B. A. Roe.** 1980. Cloning in single stranded bacteriophage as an aid to rapid DNA sequencing. J. Mol. Biol. **143**:161–178.
19. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
20. **Schwartz, E., M. Dürst, C. Demenkowski, O. Lattermann, R. Zech, E. Wolfsperger, S. Suhai, and H. Zur Hausen.** 1983. DNA sequence and genome organization of genital human papillomavirus type 6b. EMBO J. **2**:2361–2368.
21. **Seedorf, K., G. Krammer, M. Dürst, S. Suhai, and W. G. Röwekamp.** 1985. Human papillomavirus type 16 DNA sequence. Virology **145**:181–185.
22. **Seif, I., G. Khoury, and R. Dhar.** 1979. The genome of human papovavirus BKV. Cell **18**:963–977.
23. **Smith, A. J. H.** 1979. The use of exonuclease III for preparing single stranded DNA for use as a template in the chain terminator sequencing method. Nucleic Acids Res. **6**:831–848.

24. **Spalholz, B. A., Y. C. Yang, and P. M. Howley.** 1985. Transactivation of a bovine papillomavirus transcriptional regulatory element by the E2 gene product. Cell **42**:183–191.
25. **Staden, R.** 1979. A strategy of DNA sequencing employing computer programs. Nucleic Acids Res. **6**:2601–2610.
26. **Staden, R.** 1980. A new computer method for the storage and manipulation of DNA gel reading data. Nucleic Acids Res. **8**:3673–3694.
27. **Wain-Hobson, S., P. Sonigo, O. Danos, S. Cole, and M. Alizon.** 1985. Nucleotide sequence of the AIDS virus, LAV. Cell **40**:9–17.
28. **Waldeck, W., F. Rosl, and H. Zentgraf.** 1984. Origin of replication in episomal bovine papilloma virus type 1 DNA isolated from transformed cells. EMBO J. **3**:2173–2178.
29. **Weiher, H., M. Konig, and P. Gruss.** 1983. Multiple point mutations affecting the simian virus 40 enhancer. Science **219**:626–631.
30. **Wilbur, W. J., and D. J. Lipman.** 1983. Rapid similarity searches of nucleic acid and protein data banks. Proc. Natl. Acad. Sci. USA **80**:726–730.
31. **Yates, J., N. Warner, D. Reisman, and B. Sugden.** 1984. A *cis*-acting element from the Epstein-Barr viral genome that permits stable replication of recombinant plasmids in latently infected cells. Proc. Natl. Acad. Sci. USA **81**:3806–3810.
32. **Zur Hausen, H.** 1985. Genital papillomavirus infections, p. 83–90. *In* P. W. J. Rigby and N. M. Wilkie (ed.), Viruses and cancer. Society for General Microbiology Symposium 37. Cambridge University Press, Cambridge.