DNA sequence and genome organization of genital human papillomavirus type 6b

E.Schwarz¹*, M.Dürst, C.Demankowski, O.Lattermann, R.Zech, E.Wolfspurger, S.Suhai² and H.zur Hausen¹

Institut für Virologie, Zentrum für Hygiene, Universität Freiburg, Hermann-Herder-Str. 11, 7800 Freiburg, and ²Institut für Dokumentation, Information und Statistik, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, 6900 Heidelberg, FRG

Communicated by H.zur Hausen

Received on 6 September 1983

The complete nucleotide sequence of the circular double-stranded DNA of the genital human papillomavirus type 6b (HPV6b) comprising 7902 bp was determined and compared with the DNA sequences of human papillomavirus type 1a (HPV1a) and bovine papillomavirus type 1 (BPV1). All major open reading frames are located on one DNA strand only. Their arrangement reveals that the genomic organization of HPV6b is similar to that of HPV1a and BPV1. The putative early region includes two large open reading frames E1 and E2 with marked amino acid sequence homologies to HPV1a and BPV1 which are flanked by several smaller frames. The internal part of E2 completely overlaps with another open reading frame E4. The putative late region contains two large open reading frames L1 and L2. The L1 amino acid sequences are highly conserved among analyzed papillomavirus types. By sequence comparison, potential promoter, splicing and polyadenylation signals can be localized in HPV6b DNA suggesting possible mechanisms of genital papillomavirus gene expression.

Key words: DNA sequencing/genital cancer/human papillomaviruses/genome comparison/transcription regulatory sequences

Introduction

Papillomaviruses induce benign epithelial proliferations in man and in a variety of animals (for review, see Pfister, 1983; Lancaster and Olson, 1982). They are members of the papovavirus family and contain a double-stranded circular DNA genome with a mol. wt. of 5×10^6 . Human papillomaviruses (HPV) are classified into distinct types on the basis of their DNA sequence homology (Coggin and zur Hausen, 1979); 25 different types have been characterized.

An infectious etiology of human genital cancer has been discussed for several years (for reviews, see Rotkin, 1973; zur Hausen, 1977) and human papillomaviruses have been suspected to play a role as one potential causative agent. Conversion of benign human genital warts (condylomata acuminata) into malignant tumors has been reported repeatedly (for review, see zur Hausen, 1977). In genital warts, a papillomavirus was characterized and designated as HPV6 (Gissmann and zur Hausen, 1980). The DNA of subtype HPV6b was cloned in pBR322 (de Villiers *et al.*, 1981) and in lambda L47 (Gissmann, unpublished data). Hybridi-

zation analysis of DNA from genital tumors with radioactively labelled HPV6 DNA demonstrated that the majority of condylomata acuminata contains HPV6 DNA (Gissmann *et al.*, 1982a; Gissmann *et al.*, 1983) which was also found in invasively growing non-metastasizing giant condylomata, 'Buschke-Löwenstein tumors' (Gissmann *et al.*, 1982a). In screening studies, HPV6b sequences have not been detected in malignant genital tumors (Gissmann *et al.*, 1982a, 1983). Subsequently, DNA of a different papillomavirus HPV11 was characterized in atypical condylomata of the cervix and in laryngeal papillomas and was also molecularly cloned (Gissmann *et al.*, 1982b). Hybridization revealed its presence in a small percentage of genital carcinomas (Gissmann *et al.*, 1983). Very recently, two papillomavirus DNAs from cervical carcinomas designated HPV16 and HPV18 were cloned and analyzed (Dürst *et al.*, 1983; Boshart *et al.*, in preparation). They are present in >70% of the malignant genital tumor samples analyzed. Thus, papillomavirus DNAs present in genital tumor tissues are of remarkable heterogeneity. Their strikingly different distribution patterns point to a specificity of genital HPV infections for benign and malignant tumors, respectively, which suggests that the genital papillomaviruses differ in their oncogenic potential.

As a first step in studying the molecular biology of genital papillomaviruses, we have determined the complete nucleotide sequence of HPV6b DNA. Comparison with the DNA sequences of bovine papillomavirus type 1 (BPV1, Chen *et al.*, 1982) and of human papillomavirus type 1a (HPV1a, Danos *et al.*, 1982; Clad *et al.*, 1982) revealed a similar genome organization common to all three papillomaviruses. The comparative analysis allowed the assignment of potential functions to distinct regions of HPV6b DNA.

Results and Discussion

DNA sequence and open reading frames

The entire nucleotide sequence of HPV6b DNA consists of 7902 bp. Thus it is 43 bp smaller than BPV1 DNA (7945 bp, Chen *et al.*, 1982; Danos *et al.*, 1983) and 87 bp larger than HPV1a DNA (7815 bp, Danos *et al.*, 1982, 1983).

For determination of open reading frames, the HPV6b sequence was screened for the presence of the translational termination codons TAA, TAG and TGA in all six possible reading frames. The distribution of the stop codons in the two DNA strands is shown in Figure 1. Evidently, all major open reading frames are located on only one DNA strand similar to HPV1a (Danos *et al.*, 1982) and BPV1 (Chen *et al.*, 1982). The entire nucleotide sequence of the sense strand is given in Figure 2.

By *in vitro* transformation of mouse cells, the BPV1 genome was functionally dissected into two distinct segments: a 69% *Bam*HI-*Hind*III subgenomic fragment which transforms mouse cells *in vitro* and a 31% non-transforming fragment (Lowy *et al.*, 1980). BPV1-specific transcripts present in transformed mouse cells and in BPV1 induced hamster tumors all map within the 69% segment (Heilman *et al.*, 1982; Amtmann and Sauer, 1982; Freese *et al.*, 1982).

¹Present address: Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, 6900 Heidelberg, FRG.

*To whom reprint requests should be sent.

Fig. 1. Distribution of translational termination codons in the two strands of the HPV6b DNA sequence. Each vertical bar represents one termination codon in the respective reading frame. Open reading frames potentially encoding polypeptides of >90 amino acid residues are designated E1 to E8 and L1 and L2 in correspondence with the nomenclature for HPV1a and BPV1 open reading frames (Danos *et al.*, 1983). For numbering of nucleotide positions as indicated in the middle of the figure see legend to Figure 2.

Furthermore, the 69% transforming BPV1 fragment is maintained in the transformed mouse cells as an extrachromosomal multicopy plasmid (Law *et al.*, 1981). All this indicates that the 69% fragment contains the early viral functions including the genetic elements necessary for transformation and autonomous replication. The coding segment of the 69% transforming fragment containing the open reading frames E1 to E7 and the corresponding part of HPV1a DNA have been designated as the early, E, region (Danos *et al.*, 1983). Transcripts hybridizing to the 31% fragment are only found in the peripheral parts of bovine warts where virus replication takes place (Amtmann and Sauer, 1982) and synthesis of the major structural protein VP1 with a mol. wt. of 55 000 can be directed by transcripts from this region (Engel, Heilman and Howley, cited in Danos *et al.*, 1983) indicating that this part of the BPV1 genome contains the late genes coding for the viral capsid proteins. The 31% fragment covers two large open reading frames (Chen *et al.*, 1982) which together constitute the late, L, region.

Comparison of the HPV6b DNA sequence with the HPV1a and BPV1 sequences revealed characteristic homologies in the nucleotide and deduced amino acid sequences particularly within the four large open reading frames. Based on these conserved sequences, the HPV6b sequence could be aligned with the two other papillomavirus genomes and the open reading frames could be classified as putative early (E) or putative late (L). They were designated E1 to E8 and L1 and L2 in correspondence with the nomenclature of the HPV1a and BPV1 open reading frames (Danos *et al.*, 1983) to emphasize their mutual relationship. An alignment of the HPV6b genome with HPV1a and BPV1 as depicted in Figure 3 shows clearly that the arrangement of the open reading frames in HPV6b is very similar to that of HPV1a and BPV1: both the E and L regions are covered by two large open reading frames E1, E2 and L1, L2, respectively. In addition, the E region contains several smaller open reading frames E4 to E7. The nucleotide positions and amino acid coding capacities of the open reading frames in HPV6b DNA are given in Table I.

E region

The 69% transforming fragment of BPV1 extends from the *Hind*III site at position 6959 up to the *Bam*HI site at position 4451 and includes all the open reading frames of the E region together with a presumably noncoding region which is

located between the end of L1 and the beginning of E6 (see Figure 3).

E1 is the largest open reading frame in the HPV6b genome. It extends for 1947 nucleotides and thus can encode a protein of 649 amino acids. The corresponding E1 genes of HPV1a and BPV1 are smaller and have coding capacities of only 612 and 605 amino acids, respectively (Danos *et al.*, 1983). Sequence comparison demonstrates clearly that it is the N-terminal part of the HPV6b-E1 which codes for the additional amino acids. The C-terminal 453 amino acids of HPV6b-E1 can be aligned perfectly with the corresponding parts of HPV1a and BPV1 (Figure 4A). This is also the region of maximal amino acid conservation between the E1 genes of HPV1a and BPV1 (Danos *et al.*, 1983). The E1 polypeptides may contain at least two functionally different domains, one residing in the N-terminal parts diverging in primary structure, the other in the well conserved C-terminal segments.

The evolutionary relationship of the three papillomaviruses becomes very obvious when looking at the nucleotide and encoded amino acid sequences: stretches of homologous amino acids are mostly conserved within all three E1 sequences containing only a few amino acid exchanges (e.g., compare in Figure 4A the corresponding sequences ranging from position 361 to 419 in HPV6b: in a stretch of 57 amino acids 34 residues are identical and only six positions are found where the three sequences differ completely, whereas the remaining 17 positions are conserved between either two of the three sequences). Many exchanges occur between related amino acids and amino acid differences are often due to single nucleotide changes in non-silent codon positions. On the other hand, many differences in the DNA sequences affect the third or another silent codon position and therefore do not alter the respective amino acid. In addition, some differences in the DNA sequences are not reflected at all in the amino acid sequences. This is repeatedly the case for serine which is encoded by two different groups of triplets (AGC, AGT and TCN).

The large open reading frame E1 is preceded by two smaller ones E6 and E7 which are consistently found in all three papillomavirus sequences (Figures 1 and 3). The amino acid sequence homology between the E6 sequences is poor. However, the 4-fold repetition and the homologous spacing of the tetrapeptide sequence Cys-X-X-Cys observed in HPV1a and BPV1 (Danos *et al.*, 1983) is also conserved in HPV6b. The four Cys-X-X-Cys repeats are separated by 29,

```

1  GTTAATAACA ATCTTGGTTT AAAAAATAGG AGGGACCGAA AACGGTTCAA CCGAAAAACGG TTGTATATAA ACCAGCCCTA AAATTTAGCA AACAGGCAT
101  TATGGGAAGT GCAAATGCCCT CCACGCTGTC AGGACACCAT GACCAAGCTGT CCAAGACGCTT TAATCTATCT ATGCATACGT TGCAAAATTA TGTGTGTTT
201  TGCAAGAAAT CACTGACCCAC CACTGACAGT TATTCATATG CATATAAAACA CCTATTTCGAG CTGTTTCGAG CTGCTTCGAG TGCAGCGCTC TCGCGCTGCT
301  GCCTAGAATT TCATGGAAAA ATAAACCAAT ATAGACACTT TGATTATGCT GATATATGCA CAACAGTGA AGAAGAAACT AAACAAGACA TCTTAGACGT
401  GCTAATTCGT TGCTACCTGT GTCCACAAACC GCTGTGTGAA GTAGAAAAGG TAAACACATAT ACTAACCAAG GCGCGGTTC AAGAGCTAAA TTGTACGTGG
501  AAGGGTCGGG GCCTACACTG CTGGCAACA TGCATGGAAG ACATGTTACC CTAAGGATAA CTAAGGATAA CCGTGAACCT CCAGACCCCT GAGGCTTACA
601  TTGCTATGAG CAATTAGTAG ACAGCTCAGA AGATGAGGTG GACGAAGTGG ACAGACAAGA TTCACAACCT TTAACAACAC ATTTCCAAAT AGTGACCTGT
701  TGCTGTGGAT GTGACAGCAA CGTTCGACTG GTTGTGCACT GTACAGAAAC AGACATCAGA GAAGTGC AACCTTCTGT GGGAAACCTA AACATAGTGT
801  TGCCCATCTG GCACCCGAAG ACCTAACAC GATGGCGGAC GATTCCAGTA CAGAAAATGA GGGGTCTGGG TGTACAGGAT GGGTTATGGT AGAAGCTATA
901  GTGCAACACC CAACAGGTAC ACAAATATCA GACGATGAGG ATGAGGAGGT GGAGGACAGT GGGTATGACA TGGTGGACTT TATTGATGAC AGCAATATTA
1001  CACACAATTC ACTGGAAGCA CAGGCAATTG TTAACAGGCA GGAGGCGGAC ACCCATTTAT CGACTGTGCA GGACCTAAAA CGAAAGTATA TAGGTACTCC
1101  ATATGTTAGT CCTATAAACA CTATAGCCGA GGCAGTGGAA AGTGAAATAA GTCCACGATT GGACGCCATT AAACCTTACA GAGAGCTAAA AAAGGTAAAG
1201  CGACGGCTGT TTCAAACCCG GGAACCTAAC GACAGTGGAT ATGGCTATTC TGAAGTGGAA GCTGGAACCG GAACCGAGGT AGAGAAACAT GCGCTACCGG
1301  AAAATGGGGG AGATGGCTAG GAAAAGGACA CAGGAAGGGA CATAGAGGGG GAGGAACATA CAGAGGCGGA AGCCGCCACA AACAGTGTAC GGGAGCATGC
1401  AGGCACAGCA GGAATATTGG AATTGTTAAA ATGTAAGAT TTAGGGCAG CATTACTTGG TAAGTTTAAA GAATGCTTTG GGTGCTCTTT TATAGATTTA
1501  ATTAGGCCAT TTAAAGTGA TAAACAACA GGTTTAGATT GGGTGGTAGC AGGGTGGTAG ATACATCATA GCATATCAGA GGCATTTCAA AAATTAATGT
1601  AGCCATTAA TTTATATGCA CATATACAAT TGCTGACAAA TGCATTGGGA ATGGTATTGT TAGTATTATT AGAATTAAAA GTAGAGATAC GTAGAGATAC
1701  CGTTGCACGT AACTTTGCAA CGCTATTAAA TATACCTGAA AACCAAATGT TAATAGAGCC ACCAAAAATA CAAAGTGGTG TTGCAGCCCT GTATTGGTTT
1801  CGTACAGAGA TATCAAAATG CAGTACAGTT ATAGGGGAAG CACCGAAGTG AGTACACCGC CAACACAGTA TTGAACACCG GTTGGCAGAC AGTCAGTTTA
1901  AATTAACAGA AATGGTGCAG TGGCGGTATG ATAATGACAT CCGCAGAGG GATAAAATTA GATGAATATT CATTGAATA TGCACAAAGG GGAGATTTTG ATTTCTAATG
2001  ACGAGCATT TTAATAGCA ATATGCAGGC AAAATATGTG AAAGATTGTG CAACATATGT TAGACATTAT AAACATGCAG AAATGAGGAA GATGTCTATA
2101  AAACAATGGA TAAACATATG GGGTCTTAAA ATAGAGGACA CAGGAAATTTG GAAACCAATT GTACAAATCC TACGACATCA AAATATAGAA TTCAATCTCT
2201  TTTTAAGCTA ATTTAAATTA TGGCTGCACG GTACGCCAA AAAAAACCTG ATAGCCATAG TAGGCCCTGG AGATACCTGG AAATCGTACT TTGTATGAG
2301  TTTAATAAGC TTTCTAGGAG GTACAGTTAT TAGTCATGTA AATTCCAGCA GCCATTTTTT GTTGCAACCG TTAGTAGATG CTAAGGTAGC ATTGTTAGAT
2401  GATGCAACAC AGGCATGTTG GATATATATG GATACATATA TGAGAAATTT GTTAGATGTT ATATCTATGA GTATTGACAG AAAGCATAAA TACTTGACAT
2501  TAATTAATAG TCCACCTCTG CTAGTAACGT CCAACATAGA TATTACTAAA TAAGATAAAT ATAAGTATT ACATACTAGA GTAAACAATC TTACATTTCC
2601  AAATCCATTC CTTTGTGACA GAAATGGGAA TGCAGTGTAT GACTGTCAA AATCAATTTA AATCAATTTA AAGGACTGAG TATAGTATGG AACCGTGGAC CTGAGACATT
2701  CAGGATTCTG AGGACGAGGA AGATGGAAGC AATAGGCAA CGTTTATAGT GTGTAGTATT ATATAAGCA AAACAAATGG CCTAAGGCCA CATAGGAATG CAAGTAGTGC
2801  CACAACATTC TATTGCAATG GAAATGCATG AGACATGAA GTGTATTATT ATATAAGCA AAACAAATGG CCTAAGGCCA CATAGGAATG CAAGTAGTGC
2901  CACCATTAAA GGTGTCGCAA GCAAAAGGAC ATAATGCCAT TGAATGCCAA ATGCATTTAG AATCAATTTA AAGGACTGAG TATAGTATGG AACCGTGGAC CTGAGACATT
3001  AATACAAGAA ACAAGTTATG AAATGTGGCA AACACCCTT AAACCGTGT TTAACAAACG GGGCAAAATC GTAGAAGTTA AATTTTAGTG CTGTGCAAC
3101  ATTACATAGG ATATTGGTGT ATGGACAGAT GTGTATTGTC AGGCAATGCA CACCTGGGTA AAGGTGCATA GTATGGTAGA TGTCTAAGGT ATATATTACA
3201  CATGTGGACA ATTTAAACCA TATTATGTAA ACTTTGTAAA AGAGGCAAG AAGTATGGGA GCACCAACAA TTGGGAAGTA TGGTATGCA GCACAGTATG
3301  TATGTTCTCT GCATCTGTAT CTAGCACTAC AAGAAGATA TCCATTCTGT AATCTACTAC ATACACCCCT GCACAGACCT CCACCTTGTG GTCCTCAAGT
3401  ACCAAGGAAG ACGCAGTGCA AACGCCGCTT ACAAACAGAG CACGACAGTG CCAACAGTCC CTTTGCAGCT CTTTGCAGCT CTTTGCAGCT CTTTGCAGCT
3501  ACAGTGGAAA CCACAACCTC ATCACTAACA ATCAGACACA GCACCAAGA CGGAACAACA GTAACAGTTC AGTACGCTC AGTACGCTC AGTACGCTC
3601  ATCCAATTGT TTAAGTGTG TTAGATATAG GCTAAATGAC AGACAAGAG AGAACAAGAG GCAACAGTTT TTAGATGTTG TTAGATGTTG TTAGATGTTG
3701  CCACATAAAC ATGCCATTGT AACTGTAACA TATGATAGTG AGAACAAGAG GCAACAGTTT TTAGATGTTG TTAGATGTTG TTAGATGTTG TTAGATGTTG
3801  TGGGATTTAT GTCACTGCAC CTATTGTAA TTTGATATAT GTAAATGTGT AAAATATATG TTAGTGGTGA TTAGTGGTGA TTAGTGGTGA TTAGTGGTGA
3901  TGTACAAATA GCTGCAGGAA CAACAGCAGC ATTCATACCT GTAAATGTGT TTAGTGGTGA TTAGTGGTGA TTAGTGGTGA TTAGTGGTGA TTAGTGGTGA
4001  GAGTTTATTG TGTACACATC TGTGCTAGTA CTAACACTGC TTTTATATTT ACTATTGTGG CTGCTATTAA GTGCTATTAA GTGCTATTAA GTGCTATTAA
4101  TACTTGTGTT TACTGTGCCC GCATTTGTATA TACTACTACTA TATTGTTACC ACACAGCAAT GATGCTAACA GTGCTAACA GTGCTAACA GTGCTAACA
4201  GGTGTTGTTG TGTATGTTG CTTTATGTTA GGGATGTTGG GGTATTATTT AGTGCACTAT AGAGCTGTAC AAGGGGATAA ACACACCAAA ATGAGGAGAG
4301  GTAACAAACA CAACGTGTAAT GATGATTATG TAACATATGA TTATACTACT GATGCTGATT ATATATATAT AGCTACACAG CTAATATCAA CATGTAAACT
4401  AACAGTGTAT GCTTTGTATA CCATGGCACA TAGTAGGGCC TACAGACGCA GTGCAGATCA AATATTAAAA TGGGGAAGTT TGGGGAAGTT TGGGGAAGTT TGGGGAAGTT
4501  TGCCCCCCAG A TGAATTTCC TAAGGTGGAG CACAACACCA CTTACAAACT TCTGCAAAAC CTTCTATTAC TAGTGGCCCT ATGGCTCGTC CTCCTGCTG
4601  GCACGGGTTG CCGCAGCTGG GGTGCTACTG GCTATGTTCC TTAATTTGAAG AATCGGCAAT CATTAAACGCA CATTAAACGCA CATTAAACGCA CATTAAACGCA
4701  GGTGGAGCCT GTGGCCCTCT CGATCCATCT CTTGAGTCTC TTAATTTGAAG AATCGGCAAT CATTAAACGCA CATTAAACGCA CATTAAACGCA CATTAAACGCA
4801  GGTGGGTTTA CAATTACATC CTCGTAACAA ACTACGCTGT CAAATTTGGA GATATACAGT ACTAGTCACA CTAAGTACAG TATATTATTA TATATTATTA
4901  TTACAGAAC TTTCTGTAACA CAACCCCAAC CACCCCGTGA GGTCAATGGA CATATATTA TTTCTGCACC CACTGTAAAG TCACACCTTA TAGAGGAAAT
5001  TCTTTAGAT ACTTTTGTGG TATCATCTAG TGATAGCGGT CCTACATCCA CTACCCCTGT TCTGCTACT TCAACGCTTA ATACATATG ATACACCTGG ATATGAAGGG
5101  CGTGCATTGC ACCAGGTGCA GGTTCACAGC CCTGCATTTC TTTCCACTCC TCAACGCTTA ATACATATG ATACACCTGG ATATGAAGGG ATATGAAGGG ATATGAAGGG
5201  GTGTACAATT TAGTCATGAT TCTATACAGA ATGACCTGTA TGAGCGTTTT ATGGAACATA TTTGTTGCA CAGACCTGCG ATTTGCGTCC ATTTGCGTCC
5301  TGTGCGGTAC AGTGCATTTG GCAACAGGCG GTCTATGCAC ACTCGCAGCG CTGCACAGGA TGATACATTT GATATTATG GATATTATG GATATTATG
5401  GCACAGGCTG CAGAAGAAAT AGAAATGCAC CCTCTTGTGG CTGCACAGGA TGATACATTT GATATTATG GATATTATG GATATTATG GATATTATG
5501  CCAACACACC TGTTCACAAAT ATATCAGATA CATATTATTC TCCACACCTT AATACAGTTA CACAACCGTG GGTGAACACG GGTGAACACG GGTGAACACG
5601  TAATGACCTG TTTTACAAAT CTGGCCCTGA TATAACTTTT CCTACTGCAC CTATGGGAAC ACCCTTAGT CCGTGAACCT CCGTGAACCT CCGTGAACCT
5701  GTTTTCATTA CAGGTTCTGG ATTTTATTTT CATCCTGCAT GGTATTTTGC ACGTAAACGC CGTAAACGTA TTCCCTTATT TTTTTCAGAT GTGGCGCCT
5801  AGGCACAGCA CAGTATATGT GCCTCTCTCT AACCCTGTAT CCAAAGTTGT TGCCACGAGT GCTTATGTTA CTTCCGACAA CATATTTTAT CATGCCAGCA
5901  GTTCTAGACT TCTTGCAGTG GGACATCTCT ATTTTTCAT AAAACCGGGT AACAACACCT GTTGTGCAAA GGTGTGCAAA GGTGTGCAAA GGTGTGCAAA
6001  GGTGTTGTTA CCAGATCTTA ACAAATTTGC ATTGCCTGAC TCGTCTCTTT TCGATCCAC AACACAACGT TTAGTATGGG CATGCACAGG CCTAGAGGTG
6101  GGCAGGGGAG AGCCATTAGG TGTGGGTGTA AGTGGACATC CTTTCTCTAA TAAATATGAT GATGTTGAAA AGGCGGAGG TGGTGGTAA CTTGGCAAGG CTTGGCAAGG
6201  AAAACAGGTT TAATGTAGGT ATGGAATTATA AGCAACACCA ATTATGATG GTTGGATGTG CCCCCTTTT GGGCGAGCAT TGGGGTAAAG GGTGGTAAAG GGTGGTAAAG
6301  TACTAATACA CTTGTACAGG CTGTTGACTG CCCGCCCTTA GAACCTATTA CCGAGTGTAT CCAGTATGAT TGAATATAT CAGATTATTT ACACAGGCTT ACACAGGCTT
6401  AATTTGCTG ATTTGCAGAC CAATAAATCA GATGTTCTTA TTGACATAT GTTTTCTTGA CATTTTTTTA ACAGGCTGCG GAGGCTGCGG GAGGCTGCGG GAGGCTGCGG
6501  ATGTTGATAG ATTTATTTT TTTCTACGGA AGGAACAAAT GTTTGCGAGA CATTTTTTTA TTAACACCCC GAGCGGCTCT TTGGTGTCTT CTGTGTTGTA CTGTGTTGTA
6601  TATAATTAAG GGTAGTGGAA ATCCGACGTC TGTAGGAGT AGTATATATG TTAACACCCC GAGCGGCTCT TTGGTGTCTT CTGTGTTGTA CTGTGTTGTA
6701  AAGCCATATT GGTACAAAA AGCCACAGGA CATAACAATG CATATTGTTG GGTATATCAA CTGTTTGTTA CTGTGTTGTA CTGTGTTGTA CTGTGTTGTA
6801  TGACATTATG TGCATCCGTA ACTACATCTT CCACATACAC CAATTCTGAT TATAATTCACA CAATGATCCA TACCTGTCAA AAGCCCACTC CTAAGTATTTG TATGATTATTT
6901  TCAATTATGT AGCATTAGAT TGTCTGCTGA AGTAATGGCC TATATTTCACA CAATGATCCA TACCTGTCAA AAGCCCACTC CTAAGTATTTG TATGATTATTT
7001  CCAATGGTGA CATTAGAAGA TACCTATAGG TATGTGCACT CACAGGCCAT TACCTGTCAA AAGCCCACTC CTAAGTATTTG TATGATTATTT TATGATTATTT
7101  ACCTTAGTTT TTGGGAGGTT AATTTTAAAG AAAAGTTTTC TAGTGAATTT TCCAGTATC GATCAGTATC TTTTGGGAGC CAAGTATTTG TATGATTATTT
7201  ACGGTCCTCT ATTCGTACAG GTGTTAAGCG CCTGCTGTTT TCCAAAGCCT CTGCTGCCCC TAAACGTAAG CCGGCCAAAA CTAAAAAGGA TATATATGTT
7301  ATATGTACTG TTAATATATG GTGTTATAGT ACTGTTATGT ATGTTGTGTT CAGTGGCTGT TGCACGCGTT TTGGTTTGA CCGGCTTTAC ACACATAAGT AATATACATG
7401  ATGTGCAATA AACAATTACC TCTTGTTACA CCCTGTGACT CAGTGGCTGT TGCACGCGTT TTGGTTTGA CCGGCTTTAC ACACATAAGT AATATACATG
7501  CACAATATAT ATATTTTGT TTAATAACT ATACTTTTAT ATTGCAAC CCCTGCCAAG TTTGTTGCTG GCTCTTAGCA TACCTTTTCC ACCAATTTT
7601  TCTCTCTTAA TCTATATAT TTTGTGCCAG TGTACCATTT CCCTGCCAAG TTTGTTGCTG GCTCTTAGCA TACCTTTTCC ACCAATTTT ACCAATTTT
7701  TGGGTTATTG CTTTACTCAT AAACCTGTCT TTTGTTTATA CTTTATGCA CTTTATGCA CTTTATGCA CTTTATGCA CTTTATGCA
7801  GCTCTTACTG TTTGGTATAC AATAACATAA AAATGAGTAA CCTAAGGTCA CACACCTGCG ACCGTTTTCG GTTATCCACA CCTACATAT
7901  TA

```

Fig. 2. Nucleotide sequence of HPV6b DNA. The sequence of the sense strand is given. The numbering system for the HPV6b nucleotide sequence was chosen to correspond to that for BPV1 and HPV1a which both start at the G of a *Hpa*I sequence 5'-GTTAAC-3' located in a conserved region in front of the E6 open reading frame and proceed in the 5' to 3' direction of transcription (Danos *et al.*, 1983). Because the *Hpa*I recognition site is absent in the HPV6b sequence due to a single C-T nucleotide exchange, the numbering starts at the G of the corresponding '*Hpa*I'-like sequence 5'-GTTAAT-3'.

Fig. 3. Alignment of the HPV6b genome with HPV1a and BPV1. Open reading frames are represented by open bars with the presumed translational start codons indicated by dashed lines. For numbering of nucleotide positions see legend to Figure 2. The 69% transforming fragment of BPV1 flanked by a *Bam*HI site and a *Hind*III site at positions 4451 and 6959, respectively, is set off by heavy drawing. Putative transcription elements discussed in the text are indicated by the following symbols: ○ promoter TATA-box sequences; ▲ splice acceptor sequences; ▽ splice donor sequences; ■ polyadenylation signals. The designation of open reading frames is unchanged according to Danos *et al.* (1983) although the modifications introduced into the HPV1a and BPV1 sequences very recently (see Materials and methods) fused the former open reading frames E2 and E3 to one open reading frame E2.

Table I. Open reading frames in HPV6b DNA

Open reading frame	Nucleotide positions			Size [nucleotides]		Amino acid coding capacity	
	first nucleotide	potential ATG start codon	nucleotide preceding stop codon	total	from ATG	total	from methionine
E1	715	832	2778	2064	1947	688 (77 511) ^a	649 (73 171)
E2	2696	2723	3826	1131	1104	377 (43 305)	368 (42 130)
E4	3240	3255	3581	342	327	114 (12 842)	109 (12 212)
E5a	3872	3887	4159	288	273	96 (10 945)	91 (10 360)
E5b	4003	4159	4374	372	216	124 (14 130)	72 (8406)
E6	30	102	551	522	450	174 (19 946)	150 (17 290)
E7	440	530	823	384	294	128 (14 134)	98 (10 861)
E8	7611	7746	2	294	159	98 (11 146)	53 (6198)
L1	5678	5789	7288	1611	1500	537 (59 851)	500 (55 549)
L2	4378	4423	5799	1422	1377	474 (51 115)	459 (49 338)

^aThe mol. wts. of the putative proteins are given in parentheses.

36 and again 29 intervening amino acids (data not shown). In E7, BPV1 differs from the two human papillomaviruses in two aspects: the putative gene product of BPV1 is considerably larger (127 amino acids) than that of HPV6b (98) and HPV1a (93) and the low amino acid sequence homology is restricted to the C terminus. The E7 sequences of HPV6b and HPV1a are completely conserved in 37% of their amino acid residues (Figure 4C). From the eight cysteine residues in HPV6b-E7, four are arranged in the C-terminal half in an identical manner and with the same spacing as in HPV1a and BPV1: 2-fold repetition of Cys-X-X-Cys separated by 29

amino acids. The conservation of this structural feature in both E6 and E7 sequences may indicate that the two putative gene products served related functions.

By screening the protein data bank of Dayhoff *et al.* (1983) for Cys-X-X-Cys repeats, it became apparent that this constitutes a conserved feature in certain classes of proteins including the small t-antigens of polyomaviruses which contain the extended variant Cys-X-Cys-X-X-Cys (Friedman *et al.*, 1978; Seif *et al.*, 1979), several early proteins of adenoviruses, glycoprotein hormones such as thyrotropin, metallothioneins (which are extremely rich in Cys-X-X-Cys as well as Cys-X-

Fig. 4. Amino acid sequences of the putative HPV6b gene products E1, E2, E7 and L2 in comparison with HPV1a and BPV1. **(A)** E1, **(B)** E2, **(C)** E7, **(D)** L2. Amino acid residues are given in the one-letter code. Names of open reading frames are indicated at the right side, each followed by a number which denotes the respective papillomavirus sequence: 1, HPV6b; 2, HPV1a; 3, BPV1. Amino acid positions are indicated at the left side. For E1 only the C-terminal portions are presented and for L2 only the conserved N-terminal parts. In **(B)**, the portions of the E2 genes which overlap with the E4 open reading frames are indicated by open triangles. Programs ALIGN and ALIGNSTAT were used for comparison with the following global variables: aligntype = 1, distance matrix = 4, weight 1 = 0.3, weight 2 = 0.4. Identical amino acids are indicated by asterisks. Amino acids are furthermore classified into specified groups and are denoted as follows: ▽ aromatic F, Y, W; △ aliphatic M, I, L, V; + basic R, K, H; ⊕ acid and acid amide N, D, E, Q; □ hydroxyl and small aliphatic S, T, P, A, G; ~□⊖ hydrophilic. Conserved cysteine residues are indicated by arrows.

Cys), one type of keratin (in which, however, cysteine residues are generally very abundant), protease inhibitors and iron-containing proteins such as cytochromes and ferredoxins.

E1 is followed by the second major open reading frame E2 of the E region which covers 1104 nucleotides (Table I). Interestingly, the internal third of the E2 coding DNA sequence is in addition occupied by another open reading frame E4 which uses a different triplet frame shifted by one nucleotide. Due to sequence modifications introduced only recently into the HPV1a and BPV1 DNA sequences (see Materials and methods) the overlapping gene arrangement in the E2/E4 region as observed in HPV6b is also found in the corresponding parts of HPV1a and BPV1 (see Figure 3). This is an additional strong argument that frame E4 represents a functional gene or exon. HPV6b-E2 shows significant amino acid sequence homology with HPV1a and BPV1 in the N- and C-terminal parts, whereas the homology is drastically reduced in the internal region where E4 overlaps with E2 (Figure 4B). Likewise, the three amino acid sequences of E4 are unrelated.

An overlap exists between the end of E1 and the beginning

of E2 covering 56 nucleotides. This is exactly conserved in the HPV1a and BPV1 genomes as well (Danos *et al.*, 1983). The functional importance of this conserved gene overlap is not known. It is tempting to speculate, however, that it plays a role in the temporal and coordinate regulation of papillomavirus gene expression.

Considerable differences between HPV6b and the two other papillomavirus genomes are detected in the region between E2 and the first open reading frame of the L region, L2 (see Figure 3). This region extends for 552 bp in HPV6b, 335 bp in BPV1 and only 100 bp in HPV1a. In BPV1, it contains the C-terminal part of the small open reading frame E5 which starts overlapping with the end of E2. HPV1a-E5 just starts within the 100-bp segment, but its main body overlaps with L2. The HPV1a and BPV1 E5 frames which are assumed to be non-homologous contain no methionine codon in their 5'-terminal regions (Danos *et al.*, 1983). The 552-bp region of HPV6b is completely covered by two small open reading frames designated E5a and E5b (Figure 3). E5a contains a 5'-terminal ATG codon and thus can encode a protein of 91 amino acid residues. The amino acid sequence includes

Table II. Putative transcription elements in HPV6b DNA in comparison with HPV1a and BPV1

A. PROMOTER							
'CAT'-box region			'TATA'-box		cap site		
	position	sequence	position	sequence	position	sequence	
HPV6b	7898	TTATAGTTAATAACAATC	54	TATATAA	98	CATTATG ^d	
HPV1a	7808	TTGTTGTTAACTACCATC	68	TATAAAA	101	CAGATG	
BPV1	7941	TTGTTGTTAACAATAATC	58	TATAAAA	88	CACATG	
B. SPLICE ACCEPTOR SEQUENCES				C. SPLICE DONOR SEQUENCES			
	sequence	position splice site	following gene	distance ^b	sequence	position splice site	location
HPV6b	ATCTGCGCACCGAAG / A	820/21	E1	11	CAG / GTAGAG	1279/80	in E1
HPV1a	CTGTGCACCCTACAG / C	801/02		10	CAG / GTAGAT	1231/32	
BPV1	TTGTGAATCTCGGAG / C	844/45		4	CAG / GTACAG	1235/36	
	ACCTTTTAAACTCAG / A	806/07		42			
HPV6b	ATGTTTTTTTGAAAG / A	2678/79	E2	44	AAG / GTGAAT	3596/97	behind E4
HPV1a	ATCTTTCTTTGAAAG / G	2544/45		47	CAG / GTACAG	3592/93	
BPV1	ATCTTTTTTTGTAAG / G	2557/58		50	CAG / GTAAAG	3618/19	
HPV6b	TAAACTTTGTAAAAG / A	3242/43	E4	12	AAT / GTAAGT	7368/69	noncoding
HPV1a	TGCTGTAAATTACAG / G	3163/64		17	AAA / GTGAGC	7385/86	region
BPV1	GCATTACTCTGTAAG / A	3173/74		17	AAG / GTAAGT	7137/38	
HPV6b	TTATATTTGTAACAG / T	4405/06	L2	17	^a the E6 start codon is underlined ^b distance between splice site and start codon of following gene		
HPV1a	CCACATACCTTGAAG / C	3877/78		18			
BPV1	TTATATGGTTTAAAG / C	4171/72		14			
	ATTCCCCCTGCTCAG / A	4154/55		31			
HPV6b	CCTTATTTTTTTTCAG / A	5788/89	L1	0			
HPV1a	TATAATGTTTTTCAG / A	5430/31		0			
BPV1	TAATTTTTTTTGCAG / A	5608/09		0			
HPV6b	CTGCAACGACCATAG / A	141/42	E7	388			
HPV1a	AATCCGGACCGTCAG / A	129/30		399			
BPV1	CCAATCCATTCTCAG / G	130/31		348			
D. SMALL INTRON							
	splice donor sequence	position splice site	splice acceptor sequence		position splice site	intron size	
HPV6b	AAG / GTAAAA	449/50	AATTGTACGTGGAAG / G		503/04	54	
HPV1a	AGG / GTGTGC	285/86	GTTTTATATTATCAG / G		331/32	46	
BPV1	CAG / GTGAGG	344/45	CAAAACACTTGATAG / G		344/45	40	

a stretch of 13 amino acids extremely rich in leucine residues Leu-Thr-(Leu)₃-Tyr-(Leu)₃-Try-(Leu)₃ which is also detected, only slightly modified, in the C-terminal region of the BPV1-E5 amino acid sequence (Leu)₅-Phe-(Leu)₃-Phe-Phe-Leu-Val. The most abundant RNA species in BPV1-transformed cells map to the 3'-terminal part of the E region and cover the open reading frames E4, E5 and the C-terminal half of E2 (Heilman *et al.*, 1982; Chen *et al.*, 1982). Whether the transforming gene function of BPV1 resides in one of these open reading frames, however, is not known.

In HPV1a and BPV1, a polyadenylation signal AATAAA (for review, see Proudfoot, 1982) is located downstream from the E region genes (Danos *et al.*, 1983, see Figure 3). The 3' ends of viral transcripts from BPV1-transformed mouse cells (Heilman *et al.*, 1982) and from HPV1-infected COS monkey cells (Chow and Broker, 1983) have been mapped at corresponding sites. The HPV6b sequence contains no classical AATAAA polyadenylation sequence in the expected area. A related sequence ATTAAAA, however, is located at position 4554 (Figures 2 and 3) which is known to substitute the AATAAA polyadenylation site in certain eukaryotic genes (Hagenbüchle *et al.*, 1980; Jung *et al.*, 1980).

L region

The L region contains the two large open reading frames L1 and L2 which are expected to code for the viral structural polypeptides. HPV6b-L1 shares considerable amino acid sequence homology with the L1 genes of both HPV1a and BPV1 throughout the entire amino acid sequences (50% and 48% identical amino acids, respectively). Simultaneous comparison of all three L1 sequences leaves an identity in 176 amino acid residues (= 35%), and conservative amino acid changes, as specified in the legend to Figure 4, occur in the majority of the other positions (data not shown). It is noteworthy that six cysteine residues out of 11 (HPV6b) and eight (HPV1a, BPV1), respectively, are located at identical positions in the three L1 sequences.

In contrast to L1, the L2 sequence of HPV6b shares only limited homology with HPV1a-L2 and even less with BPV1-L2. Sequence conservation is mainly restricted to the N- and C-terminal portions of the amino acid sequences. The N-terminal 61 amino acids of HPV6b-L2 match perfectly with both HPV1a (61% homology) and BPV1 (67%) and the homology between HPV6b and HPV1a can be extended for a further 13 amino acids (Figure 4D). Remarkable features are

a cluster of basic residues at the N terminus and the glycine-rich heptapeptide sequence Gly-Gly-Leu-Gly-Ile-Gly-Thr at position 55–61 (HPV6b) which is completely conserved. The conserved sequence feature at the C terminus is a cluster of six basic amino acids downstream from a tripeptide sequence Leu-His-Pro. As observed already by Danos *et al.* (1982, 1983), a basic stretch at the C terminus is also a characteristic feature of the minor capsid proteins VP2 and VP3 of polyoma viruses. Based on the strikingly different degrees of sequence conservation within L2, it may be speculated that the N-terminal parts are involved in some interaction with the structurally also conserved L1 polypeptides, e.g., in the assembly of viral particles, whereas the structurally diverging parts may have some host or tissue specific functions.

119 nucleotides downstream from the end of L1, the HPV6b sequence contains a polyadenylation signal AATAAA at position 7407 (Figures 2 and 3).

Non-coding region and putative transcription elements

The region located between L1 and E6 contains no significant open reading frame in HPV1a and BPV1 and is therefore assumed to be non-coding in both genomes (Danos *et al.*, 1983). The corresponding segment in the HPV6b sequence contains a small open reading frame designated E8 (Figure 1 and Table I). However, since the first methionine codon is located almost in the middle of the coding region and since this open reading frame has no equivalent in the two other papillomavirus DNA sequences, it may be non-functional. Hence, this intervening region is probably a non-coding region in HPV6b as well. It extends for 715 bp (positions 7292–101). The non-coding regions in HPV1a and BPV1 are larger, 982 bp and 943 bp, respectively (Danos *et al.*, 1983). In analogy to the genome organization of other viruses such as polyoma viruses, this non-coding region is suspected to harbor several *cis*-acting DNA elements including the origin of viral DNA replication and the promoter for early gene transcription.

Directly following the L1 termination codon, the non-coding region of HPV6b contains an A+T-rich segment in which the monotonous repetition of the dinucleotide sequence thymine-purine is only slightly disturbed. This segment includes a tandem repetition of the 24 nucleotide sequence 5'-TATATGT^ATATATGTACTGTTAT^A-3' at positions 7292 and 7316 and extends up to the polyadenylation signal AATAAA at position 7407. The non-coding region of HPV1a also begins with a long A+T-rich segment, but the repetition of the T^A-G dinucleotide is much less pronounced and is not at all evident in the corresponding part of the BPV1 non-coding region.

The nonanucleotide sequence 5'-TTGCACGCG-3' is repeated twice at positions 7450 and 7466 and the octanucleotide sequence 5'-TGCCAAAGT-3' is repeated four times from position 7625 to 7680 with an equal spacing of 11 nucleotides between the outer and inner repeat units. This segment is followed by a G+C-rich inverted repeat (positions 7681–7705). Direct and inverted repeats are also characteristic features of the HPV1a and BPV1 non-coding regions (Danos *et al.*, 1983), but their arrangements and primary structures exhibit no obvious homologies to the sequence elements in HPV6b DNA.

The putative early promoter with three of its elements can be localized in front of the E6 open reading frame. It shares significant sequence homologies with HPV1a and BPV1.

There is a TATA-box sequence at position 64, an appropriate cap site directly in front of the E6 methionine codon and a conserved sequence around the GTTAAT 'HpaI'-like site (Table IIA) which may represent the so-called CAT-box an element of RNA polymerase II promoters which is located 70–90 nucleotides upstream from the mRNA start point (Benoist *et al.*, 1980; Efstratiadis *et al.*, 1980).

Analysis of the BPV1-specific mRNAs in transformed mouse cells has shown that they all share a common 3' terminus, but have different 5' ends. The latter have been mapped at different positions within the BPV1 E region in good agreement with the positions of the various open reading frames (Heilman *et al.*, 1982; Chen *et al.*, 1982). Assuming that all early transcripts are initiated from a common promoter and that their size heterogeneity is produced by splicing different RNA bodies to a common leader, the open reading frames should be preceded by splice acceptor sequences. We have used the splice acceptor consensus sequence (C)₁₁N^CAG/G described by Mount (1982; in which the dinucleotide AG has to be strictly conserved directly adjacent to the splice site and is not allowed to occur in the pyrimidine-rich tract) to screen the HPV6b DNA sequence together with the HPV1a and BPV1 sequences for potential splice acceptor signals located at homologous positions within the genomes. By this comparison, it was possible to define putative splice acceptor sequences in front of the open reading frames E1, E2, E4, E7, L1 and L2. The sequences are compiled in Table IIB. Their positions are consistent with the mapping data for the 5' ends of BPV1-specific mRNAs (Heilman *et al.*, 1982) and for the splice sites of HPV1 mRNAs produced in COS monkey cells (Chow and Broker, 1983). One of these HPV1-specific RNAs was shown to be composed of three discontinuous segments. The two exon-intron boundaries (splice donor sites) were mapped 210 bp downstream from one *EcoRI* site and near the *BamHI* site, respectively (Chow and Broker, 1983). Using the splice donor consensus sequence CAG/GT^AAGT (Mount, 1982) for screening, the two putative splice donor sequences can be detected in the HPV1a DNA sequence. Moreover, similar donor signals are also present in the HPV6b and BPV1 sequences at corresponding positions. One is located ~400 nucleotides downstream from the E1 start codon, the other directly following the end of E4 (Table IIC), indicating that these splice sites are actually used in the biogenesis of papillomavirus transcripts. In addition, the DNA region covered by the E6 open reading frame harbors a putative small intron. The nucleotide sequences around the intron boundaries are given in Table IID.

A further splice donor site correspondingly found in all three papillomavirus sequences is located downstream from the L1 gene in the non-coding region (Table IIC). A TATAAA promoter sequence and an ACA cap site are present just in front of this splice donor signal in BPV1 DNA and transcription initiated at that promoter would give rise to a leader of only seven bases (Chen *et al.*, 1982). In HPV6b and HPV1a, the corresponding putative promoters may be located in the A+T-rich region behind the L1 termination codon which would result in larger leader sequences for the human papillomaviruses. It is not known whether this and the other putative promoter discussed earlier are both actually used. They may be active at different stages in papillomavirus development. It should be mentioned, that in all three papillomavirus DNA sequences the splice donor site is located upstream from the polyadenylation signal presumably used in

L gene transcription; the two sites are separated from each other by 11 bases (BPV1), 33 bases (HPV1a) and 32 bases (HPV6b), respectively. This close association of two different transcription signals may be fortuitous, however, it may point to some mechanism involved in the coordinate regulation of papillomavirus gene expression.

Conclusions

HPV6 is the first representative of the group of human papillomaviruses present in tumors of the genital tract which has been identified and characterized (Gissmann and zur Hausen, 1980; de Villiers *et al.*, 1981). Recent studies on the distribution of this virus in genital lesions reveal that it is almost exclusively found in benign warts and in non-metastasizing tumors (Gissmann *et al.*, 1982a, 1983), whereas DNA of other papillomavirus types resides preferentially in malignant tumors (Dürst *et al.*, 1983; Boshart *et al.*, in preparation) suggesting that the various genital papillomaviruses may differ in their oncogenic potential. An etiological role of these papillomaviruses in the development of genital cancer is not proven by the presence of their DNAs in the tumor biopsy samples. This specific association, however, is a strong argument for their involvement. The DNA sequence and genome organization of the 'low risk' genital papillomavirus HPV6b will provide a basis for characterizing and comparing the genome organization of those 'high risk' papillomavirus DNAs and for identifying their putative oncogenic functions.

Materials and methods

HPV6b DNA and DNA sequence analysis

DNAs of the plasmids Amp2 and 21 containing the 3.5×10^6 and the 1.6×10^6 BamHI-EcoRI restriction fragments of HPV6b, respectively, cloned into pBR322 (de Villiers *et al.*, 1981) were kindly provided by L.Gissmann or prepared as described in Maniatis *et al.* (1982). They were used for most part of the DNA sequence analysis. HPV6b DNA linearized either at the single BamHI or EcoRI site and cloned into lambda L47 was also prepared by L.Gissmann. This DNA was further subcloned into pBR322 and was used in the sequence analysis for the final unambiguous assembly of the two parts of the HPV6b DNA sequence determined by using Amp2 and 21 DNA.

Terminal labelling of restriction fragments either at their 5' or 3' ends and the partial chemical cleavage reactions for DNA sequencing were performed as described by Maxam and Gilbert (1980).

DNA sequence data

For sequence comparison, the published DNA sequences of HPV1a (Danos *et al.*, 1982; Clad *et al.*, 1982) and BPV1 (Chen *et al.*, 1982) were used with the modifications given by Danos *et al.* (1983). Furthermore, a G residue was added into the BPV1 sequence at position 3444, a modification which joined the separate open reading frames E2 and E3 to one continuous open reading frame E2 (A.Stenlund, personal communication). Re-inspection in the corresponding region of the two published HPV1a sequences led to the insertion of one additional nucleotide each (a G at position 576 in the nucleotide sequence of the DNA sense strand in Clad *et al.*, 1982, which is now position 3306 and a C at position 3302, O.Danos, personal communication, respectively). The corrected nucleotide sequence from position 3300 to 3310 is 5'-CTCCCAGGACA-3' and results in the fusion of open reading frames E2 and E3 to give one continuous open reading frame E2.

Computer analysis

Programs developed by Staden (1980) and Kröger and Kröger-Block (1982, and unpublished for the modified versions) were used for computer-aided analysis of the nucleotide sequence data. Comparative analysis of amino acid sequences was performed by using the programs ALIGN (Smith *et al.*, 1981, for comparison of two sequences) and ALIGNSTAT (Kröger and Osterburg, 1983, for simultaneous comparison of three sequences).

Acknowledgements

We are grateful to L.Gissmann for cloning of HPV6b DNA in phage lambda and for providing plasmid DNA. We thank U.K.Freese for his stimulating cooperation in the computer analysis and M.Kröger and A.Kröger-Block for

making available to us the modified versions of their computer programs. We also thank A.Stenlund and O.Danos for communicating the modifications of the BPV1 and HPV1a DNA sequences. This work was supported by the Deutsche Forschungsgemeinschaft (SFB 31: Tumorentstehung und -entwicklung).

References

- Amtmann, E. and Sauer, G. (1982) *J. Virol.*, **43**, 59-66.
- Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980) *Nucleic Acids Res.*, **8**, 127-142.
- Chen, E.Y., Howley, P.M., Levinson, A.D. and Seeburg, P.H. (1982) *Nature*, **299**, 529-534.
- Chow, L.T. and Broker, T.R. (1983) in Pearson, M.L. and Sternberg, N.L. (eds.), *Gene Transfer and Cancer*, Raven Press, NY, in press.
- Clad, A., Gissmann, L., Meier, B., Freese, U.K. and Schwarz, E. (1982) *Virology*, **118**, 254-259.
- Coggin, J.R. and zur Hausen, H. (1979) *Cancer Res.*, **39**, 545-546.
- Danos, O., Katinka, M. and Yaniv, M. (1982) *EMBO J.*, **1**, 231-236.
- Danos, O., Engel, L.W., Chen, E.Y., Yaniv, M. and Howley, P.M. (1983) *J. Virol.*, **46**, 557-566.
- Dayhoff, M.O., Hunt, L.T., Barker, W.C., Orcutt, B.C., Yeh, L.S., Chen, H.R., George, D.G., Blomquist, M.C. and Johnson, G.C. (1983) *Protein Sequence Data Base*, (release June 1983), Natl. Biomed. Res. Found., Washington, DC.
- de Villiers, E.-M., Gissmann, L. and zur Hausen, H. (1981) *J. Virol.*, **40**, 932-935.
- Dürst, M., Gissmann, L., Ikenberg, H. and zur Hausen, H. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 3812-3815.
- Efstathiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., DeRiel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Baralle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980) *Cell*, **21**, 653-668.
- Freese, U.K., Schulte, P. and Pfister, H. (1982) *Virology*, **117**, 257-261.
- Friedmann, T., Doolittle, R.F. and Walter, G. (1978) *Nature*, **274**, 291-293.
- Gissmann, L. and zur Hausen, H. (1980) *Int. J. Cancer*, **25**, 605-609.
- Gissmann, L., de Villiers, E.-M. and zur Hausen, H. (1982a) *Int. J. Cancer*, **29**, 143-146.
- Gissmann, L., Diehl, V., Schultz-Coulon, H.-J. and zur Hausen, H. (1982b) *J. Virol.*, **44**, 393-400.
- Gissmann, L., Wolnik, L., Ikenberg, H., Koldovsky, U., Schnürch, H.G. and zur Hausen, H. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 560-563.
- Hagenbüchle, O., Bovey, R. and Young, R.A. (1980) *Cell*, **21**, 179-187.
- Heilman, C.A., Engel, L., Lowy, D.R. and Howley, P.M. (1982) *Virology*, **119**, 22-34.
- Jung, A., Sippel, A.E., Grez, M. and Schütz, G. (1980), *Proc. Natl. Acad. Sci. USA*, **77**, 5759-5763.
- Kröger, M. and Kröger-Block, A. (1982) *Nucleic Acids Res.*, **10**, 229-236.
- Kröger, M. and Osterburg, G. (1983) *Comp. Programs Biomed.*, **16**, 68-69.
- Lancaster, W.D. and Olson, C. (1982) *Microbiol. Rev.*, **46**, 191-207.
- Law, M.F., Lowy, D.R., Dvoretzky, I. and Howley, P.M. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 2727-2731.
- Lowy, D.R., Dvoretzky, I., Shober, R., Law, M.F., Engel, L. and Howley, P.M. (1980) *Nature*, **287**, 72-74.
- Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) *Molecular Cloning*, published by Cold Spring Harbor Laboratory Press, NY, pp. 86-96.
- Maxam, A.M. and Gilbert, W. (1980) *Methods Enzymol.*, **65**, 499-560.
- Mount, S.M. (1982) *Nucleic Acids Res.*, **10**, 459-472.
- Pfister, H. (1983) *Rev. Physiol. Biochem. Pharmacol.*, in press.
- Proudfoot, N. (1982) *Nature*, **298**, 516-517.
- Rotkin, I. (1973) *Cancer Res.*, **33**, 1353-1365.
- Seif, I., Khoury, G. and Dhar, R. (1979) *Cell*, **18**, 963-977.
- Smith, T.F., Waterman, M.S. and Fitch, W.M. (1981) *J. Mol. Evol.*, **18**, 38-46.
- Staden, R. (1980) *Nucleic Acids Res.*, **8**, 3673-3694.
- zur Hausen, H. (1977) *Curr. Top. Microbiol. Immunol.*, **78**, 1-30.