Maps of Papillomavirus mRNA Transcripts

Most of the transcript maps have been compiled, reproduced and modified from the original LANL Human Papillomavirus Database transcript maps. We continue to add new transcript maps as they become available. We are grateful to the authors for providing us with the data.

Alison McBride

Laboratory of Viral Diseases, National Institutes of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland

----

Maps of Papillomavirus mRNA Transcripts
1995, 1996 and 1997 LANL Human Papillomavirus Database
Carl Baker[a] and Charles Calef[b]

[a] *Laboratory for Tumor Virus Biology, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892-5055*
[b] *MS K710, Los Alamos National Laboratory, Los Alamos, New Mexico 87545*

Explanation of Maps

Maps showing the open reading frames and mRNA transcripts of nine papillomaviruses are presented on the following pages. Facing each map is a brief description of the transcripts, and following the maps is a list of references from which the maps were compiled.

In each map the significant ORFs are shown in their proper reading frames as colored rectangles. At the upper left end of the rectangle are two numbers. The first corresponds to the nucleotide (nt) position of the ORF start, the first nucleotide following a stop codon. The second number records the nt position of the first ATG, which is also indicated by a dotted line within the rectangle. The position of the last nt in the stop codon of each ORF is printed at the lower right corner of the rectangles. The rectangle's colored fill begins at the ATG and extends to the end; thus it corresponds to the coding sequence, not the entire ORF. In the cases where no ATG exists in the ORF, only one number is present in the upper left corner and the rectangle is completely filled with color. Below the ORFs is a scale of the genome divided into thousands. On the scale are placed the positions of promoters (represented by arrows) and the poly(A) signals. The exact position of the poly(A) signal is indicated.

Located below the genome scale are diagrams of mRNA species, most of which are spliced. The exons are illustrated by narrow black rectangles, while the introns are indicated by black hairlines between. The numbers printed below the lines indicate the 5' and 3' termini of the RNAs, and the 5' and 3' splice junction positions. The splice junction numbers give the position of the last nucleotide in the exon before the splice and the position of the first nucleotide of the exon following the splice. Splice junctions in parentheses were deduced from the genomic sequence and have not been confirmed by cDNA sequencing. Where 5' or 3' ends of the RNAs are uncertain, no nt position is given. Superimposed on the exons are colored rectangles representing the gene, or part thereof, coded by that portion of the exon. The coding potential of each transcript is also listed at the right and includes all proteins which could be translated from a given mRNA. In that list a ^ symbol between two gene names (e.g., E1^E4) indicates a fusion product. The * symbol indicates a truncated form of a protein. Alternative splicing can join the same upstream exon to different downstream exons. The resulting truncated proteins are identified by Roman numerals, such as E6*I, E6*II, and E6*III as illustrated by HPV16R species B, C, and D (published in 1996). Occasionally an intron occurs wholly within an ORF. In such cases it is possible that the normal beginning of the ORF may be spliced to a downstream, in-frame portion of the orf. This results in a truncated version of the complete protein, bearing a normal N-terminal portion ("head") and normal C-terminal portion ("tail"), but missing the middle. The E6*I protein from species B in HPV16R is an example. Truncation can also occur when an upstream

exon is spliced to a downstream exon such that the downstream exon is translated in a reading frame that results in a stop codon soon being encountered. The result is a protein composed of a head, translated in one of the "normal" ORFs such as E6 or E1, followed by a short tail that is out of frame. We have chosen to draw this out-of-frame (OOF) tail as a gray box to emphasize it is not in the frame of one of the normal (colored) ORFs. (Note: in the 1996 maps we showed the OOF tail as a narrower bar in the color of the upstream intron to which it is fused--a convention we now feel was misleading.) An example of this situation is the E5a* protein from species q-t in HPV31. Here a short piece of normal E5a is spliced to position 5555, the region of overlap of the L2 and L1 genes. But, because of the splice position, the 5555 region is actually translated in neither the L2 or L1 frame, but in frame three which results in a stop codon being quickly encountered. The OOF bits of RNA need not always terminate the translation of a transcript. Species l of HPV31 could be translated such that 15 amino acids (aa) of E1^E4 are fused to 18 aa from the L2 region, but not in the L2 frame, which are in turn fused to a 21 aa tail in the L1 region but not the L1 frame. The next level of complexity is exemplified by the E1M fusion protein coded by species B in HPV-1a. Here the first 139 aa are normal E1 followed by 20 aa translated from RNA in the E2/E4 overlapping region of the genome. However, these 20 amino acids are neither E2 or E4 because they happen to be translated in frame 2, and consequently are shown in gray. In the same mRNA species, (species B in HPV1a), the "converse" situation is also illustrated; the E2C protein bears an OOF head (beginning at the AUG codon at position 1200 in the E1 region) joined to a normal E2 tail. The narrow gray bar in exon 1 goes with the thick (normal) blue bar in exon 2. Occasionally a very short normal head is joined to a very short OOF tail, such as in the E1* protein of species C in HPV1a. We illustrate these merely because they have the potential to be translated.