

# II

---

## ALIGNMENTS


---

Part II of this compendium contains alignments of PV coding sequences and their corresponding protein sequences. The coding sequences are presented from the first methionine codon of the reading frame, and protein sequences from the corresponding methionine. The order of the alignments is E6, E7, E1, E2, E4, E5, L1, L2, and LCR. The LCR was operationally defined as the region after the end of L1 and before the first methionine of the E6 ORF. The LCR region is presented as a nucleotide alignment only, and contains only HPVs. You can find your way in this section by looking at either the headline or the page number, both of which contain the gene name.

Within the alignments the sequences have been grouped according to the phylogenetic organization adopted by the HPV Sequence Database (see Figure III-3). This is based on phylogenetic analysis of partial L1 sequences using weighted parsimony. Note that in the alignments presented here, the Rhesus monkey papillomavirus, RhPV1, and the Pygmy Chimpanzee papillomavirus, PCPV1, are placed within the groups with which they cluster in phylogenetic analysis, A and B respectively. They were not, however, used to constitute the consensus sequences for those groups.

Each group has a consensus sequence as its first member. Within the consensus, capitalized letters indicate that the base or amino acid is present at that location in all taxa in the group, that is to say, it is completely conserved; lower case letters indicate that the base or amino acid is present in 50% or more of the group sequences; a question mark indicates that no one base or amino acid is present at that position 50% or more of the time.

The alignments have been created progressively, beginning with pairwise alignments within the groups and then proceeding to inter-group comparisons. Agreement with the consensus sequence at any location is shown by a dash (-) while gaps are indicated by dots (. . .). Blank spaces within the alignment indicate lack of sequence information over that region. Occasionally, a nucleotide sequence will contain a percentage sign, (%), which indicates that the sequence appears to contain a frameshift indel at that position.

Typically, the alignments are displayed so that all the sequences on the same page, as well as the sequences on the facing page, are homologous to one another. There are two exceptions from this general rule. At the beginnings and ends of some of the alignments, some sequences may be separated by lines of arrows (➡). This indicates that these sequences are significantly longer than the rest, and that they continue below on the same page. The other exception is found in the E4, E5 and LCR alignments. These differ from the other alignments in two respects: first, sequences on facing pages are not homologous to each other; second, these regions contain solid "separation bars" (  ) throughout the alignment, to indicate that no significant similarity exists between sequences above and below the bar, and they could not be aligned for that reason. In some cases, it seems probable that this lack of similarity may be attributable to an absence of homology between the sequences (Part III).

## Contents

### PART II Alignments

Introduction . . . . .	II-1
Contents . . . . .	II-2
E6 Protein Alignment . . . . .	II-E6-2
E6 Nucleotide Alignment . . . . .	II-E6-8
E7 Protein Alignment . . . . .	II-E7-2
E7 Nucleotide Alignment . . . . .	II-E7-8
E1 Protein Alignment . . . . .	II-E1-2
E1 Nucleotide Alignment . . . . .	II-E1-22
E2 Protein Alignment . . . . .	II-E2-2
E2 Nucleotide Alignment . . . . .	II-E2-18
E4 Protein Alignment . . . . .	II-E4-2
E4 Nucleotide Alignment . . . . .	II-E4-6
E5 Protein Alignment . . . . .	II-E5-2
E5 Nucleotide Alignment . . . . .	II-E5-4
L2 Protein Alignment . . . . .	II-L2-2
L2 Nucleotide Alignment . . . . .	II-L2-22
L1 Protein Alignment . . . . .	II-L1-2
L1 Nucleotide Alignment . . . . .	II-L1-20
LCR Nucleotide Alignment . . . . .	II-LCR-2