Assessing Recombination in HPV, Part I

Aaron L. Halpern

MS K710, Los Alamos National Laboratory, Los Alamos, NM 87545

Introduction

Given the relatively common situation in which a single sample contains DNA from two or more HPV types (or variants of a single type), the possibility arises of recombination between strains of HPV. The issue of the existence, or lack thereof, of recombination is relevant to diagnostic, therapeutic and epidemiologic concerns, as well as to taxonomic endeavors and the study of papillomavirus natural history. From a clinical standpoint, the existence of recombinant strains could raise difficulties in vaccination through the generation of novel immunotypes. Use of a single short region of the virus for olignoucleotide probes or sequences could fail to correctly characterize infections with respect to oncogenicity. Uncritical phylogenetic analyses performed on recombinant sequences could lead to the impression of novel, relatively isolated branches. Similar issues have recently become a concern in the study of the human immunodeficiency virus (HIV) [14, 15, 16, 17].

The use of stringent hybridization assays to type HPV isolates offers some evidence to the effect that PV isolates are not frequently recombinants between two distinct types. Indeed, in an early treatment of the issue, Kremsdorf and colleagues [8] argued that the diversity of EV-related HPV types did not result from recombination, since the hybridization patterns of nonoverlapping segments of an HPV genome generally agreed rather well with one another. This result has frequently been cited to support the claim that recombination between HPV types is rare or nonexistent.

Despite this, the possibility of the existence of recombinant HPVs is suggested by some available data, including: a) reported sequences of HPV16 variants which could be intratype recombinants [12, 18]; b) an isolate of a novel HPV type (HPV77) which has an unusual pattern of sequence similarity over its E6, E7 and L1 coding sequences [2]; c) the biological viability of hypothetical HPV recombinants, as suggested by studies in which chimeric proteins are constructed and shown to maintain various biological properties [5].

In light of these observations, we present here the results of a search for PV recombinants among the sequences of the papillomavirus types available in the database. We have recently developed a tool for rapidly scanning a sequence alignment for recombinant or mosaic sequences. However, before going into the details of our methods and results, we will present more clearly the data, mentioned briefly above, for which a case for recombination might be made.

Some possible HPV recombinants

Based both on comparisons of PV types and on comparison of variants of a single type, we can say that the rate of divergence between different PV ORFs is such that $E7 < L1 < E6 < L2$. That is, between a pair of sequences, E7 is generally least diverged (measured for example as % dissimilarity), L1 next least, E6 somewhat more, with L2 being the most divergent.

In data on the L1 and L2 coding sequences (cds) of HPV16 variants from Trinidad, for several samples (T3, T17, and T49), the sequence of the L1 cds differed from that of the reference clone at 7 positions (roughly 0.5%) [12]. In contrast, in the L2 cds, which, given the above observation, we would expect to be more diverged than the L1 cds, the sequence was identical to that of the reference clone. While the dissimilarity for L1 is quite low, the pattern is surprising. A similar pattern is seen for isolate T45 from the same paper, which differs from the reference sequence at 4 positions in L1 but is identical to the reference clone in L2. Possible explanations are multiple, including: recombination between the reference strain and an as yet unknown variant (the L1 sequence is unlike any other published L1 sequence that we are aware of); the chance concentration of several recent mutations in the L1 cds; an unknown and unusual set of pressures leading to divergence of L1 or conservation of L2; an unrecognized multiple infection, with differential amplification of L1 and L2 fragments such that the sequence of L1 reflects one strain and the sequence of L2 another; or sample contamination. One additional point of interest regarding these isolates is that the L2 sequences for T3, T17, T49 and T45 all agree with the reference sequence at nt 4938 [11]; with the exception of isolate S83 from Pushko *et al*, all other published HPV16 L2 sequences differ have an “A” at this position in contrast to the reference “G” [7, 20].

The second example involves a recently identified isolate which has been designated HPV77 [2]. The sequence of isolate VS93-1-G, the source of the reference clone of HPV77, is more similar to HPV29 over each of the E7, E6 and L1 cds than to any other known type. However, in L1 and E6, HPV77 and HPV29 are substantially diverged (86% and 83% similar on the nucleotide level respectively), while in E7 they are 98% similar. Though this pattern of divergence, E7 $<$ L1 $<$ E6, is as observed elsewhere, the magnitude of the difference between E7 and the other cds is surprising. For comparison, HPV7 and HPV40 are 89%, 87% and 83% similar in E7, L1, and E6; HPV33 and HPV58 are 88%, 86% and 85% similar in the same regions. Again, recombination is not the only explanation possible for these data, although the fact that the sequence is from a clone rather than direct PCR sequencing makes misidentification of a multiple infection implausible. Additional isolates with sequences of fragments of L1 closely matching that of VS93-1-G were obtained in the same study; sequence analysis of other ORFs from these additional samples may shed some light on whether isolate VS93-1-G represents a widespread, stable HPV strain, or whether it is an isolated result of detecting an ephemeral recombinant strain.

One of the difficulties in interpreting these two examples lies in the fact that the potential recombinant sequences are very similar or identical to other isolates in one region, but no equally closely related sequences can be identified for the other regions. If indeed they are recombinants, only one of the donor strains (or a close relative) has been identified in each case. Another example from the literature of HPV16 variants may serve to illustrate the situation in which two potential donors are known. Smits and colleagues presented sequence data from various fragments of several isolates of HPV16 variants [18]. Taken in combination with data from other papers [1, 3, 20], several of the Smits *et al* sequences from the Barbados appear to be potential recombinants; as before, various explanations are possible. However, in this case, multiple donor lineages may be identified. Sequence data for the relevant Barbados samples and from some samples from other studies which may serve as background are given in Figure 1.

For the background samples, changes relative the the reference sequence in one region are correlated with changes in other regions. Based on variation in a fragment of the LCR, HPV16 variants have been classified into four major clusters, E (European), AA (Asian/American), Af1 (African 1) and Af2 (African 2); the AA, Af1 and Af2 classes together form a coherent phylogenetic group [6], and will jointly be referred to as the Ax classes below. Variations in E6, E7, E5, L2 and L1 have been found to be largely compatible with the LCR classification [1, 4, 20]: specific changes in one region are correlated with specific changes in another. Likewise, sequence from any of these regions suffices to classify a sample into one of the variant classes in a fashion consistent with the LCR sequence. Additional data supports similar results in E2 [19].

Examination of the Smits *et al* sequences shows that the classifications obtained from the E7 and E5 regions of the samples in question do not correspond to those based on the LCR, L1 and L2 sequences (see Figure 1). Notably, in E7, changes at nt 789 and 795 relative to the reference genome seen in these samples have otherwise been seen only in Ax isolates; in E5, changes at nt 3857 are likewise otherwise markers of Ax isolates. At a finer level of classification, several of the Barbados samples have an “A” at nt 3867, generally a marker of an Af1 isolate; these same samples have other Af1 markers at nt 3990 and 4041. Samples BT7, BT8 and BT15 have a “G” at nt 3990, a variant base otherwise seen in AA or Af2 isolates. BT7 has a “C” at nt 4058, otherwise a marker of Af2 isolates, while BT8 and BT15 have an “A” at nt 4016, otherwise seen only in AA isolates. In contrast, none of the samples in question contain certain changes characteristic of Ax isolates at various positions in L2, L1 and the LCR, namely nt 4280 in L2, nt 6558 in L1 and several positions in the LCR; similarly, the samples which matched the Af1 markers in E5 lack other changes characteristic of Af1 isolates at nt 4307 (L2) and nt 6567 and 6576 (L1). Those samples which match the Af2 or AA patterns in E5 likewise do not have changes in the LCR which are characteristic of other variants of these classes (e.g. the presence of a “C” at nt 7484).

All of this serves simply to illustrate that there is something unusual about these Barbados samples. Recombination certainly could result in such a pattern, although other possibilities raised above must be considered as well.

With these examples of possible HPV recombination in mind, we turn to an automated search for anomalous patterns in the sequences of PV types. (The data from the examples just discussed can not be analyzed by the method described below. The low level of divergence among HPV16 variants make the method an inappropriate means of detecting recombinants among variants. As for HPV77, only a fragment of L1 sequence is currently available.)

Methods

In the absence of recombination, the relative degrees of divergence among a set of lineages will—convergence, conservation, mutational saturation and random fluctuations aside—be the same in different portions of the genome. Different degrees of selectional pressure on conserved and variable regions may mean that the absolute amount of divergence between a given pair of sequences will differ substantially from region to region, but if A and B are more similar in one region than are A and C in that same region, the same relation will hold in another region (above caveats excepted, again).

With recombination, this no longer holds true. A recombinant sequence will reflect the sequences of its donors, and may be most similar to one sequence in one region and to another sequence in another region. Likewise, convergent evolution or lack of divergence in some lineages, may result in similar **mosaic** patterns of sequence relatedness. This suggests a simple test for detecting mosaic sequences of various kinds: given a set of background or reference sequences, we may ask whether a given sequence of interest is consistently most like a certain sequence or set of sequences in the background set over different portions of the sequence, or whether it is most like one set in some regions and most like another set (or other sets) in other regions.

This test is implemented in a program (RIP) which was written to scan alignments of HIV1 sequences for recombinants and other mosaic sequences. The program works by comparing a region or **window** of an alignment, the **background** alignment, against the corresponding portion of a sequence of interest, the **query** sequence. In each such window, it is determined which sequence or set of sequences in the background alignment are most similar to the query sequence, and whether the difference in similarity is great enough to confidently declare the query sequence to be more closely related to the matching sequence set than to the remaining sequences. The window is moved along the alignment one position at a time and a record is kept of the best matches for each window. The length of sequence contained by the window, the **window size**, may be varied; shorter windows will potentially be able to detect shorter stretches of mosaicism, while longer windows will potentially detect mosaicism which arose longer ago, and has been partially masked by subsequent mutations. This is analogous to the use of different “word sizes” in dot-matrix analysis: short word sizes can identify short repeats in a sequence which would be missed by a larger word size, but a larger word size (with lower stringency) can pick up larger, imperfect matches. The degree of a match is determined as the number of matching positions between the query and a background sequence, or between the query and the consensus sequence for a set of related sequences.

A normal result, given a background set which contains relatively close relatives of a query sequence as well as additional sets of less closely related sequences, will be for the query sequence to match the same set of sequences in every window, at least for those windows in which a closest match can confidently be determined. (In many windows, it may not be possible to determine a closest match with confidence, because of insufficient or excessive divergence in the region covered by the window, or because of the “noise” added by the random nature of mutations.) Confidence in a best match is estimated by calculating a z-score for the comparison of the best match (number of mismatches/number of positions in the window) to the next best match. In order to reduce the number of false positive results due to the large number of tests involved (one for each window, for each sequence being tested), a 99% confidence cutoff was used for the analyses reported below. Several other details of the program, such as the treatment of gaps in the alignment and the treatment of invariant positions, can be chosen by the user; specifics are described further in the program documentation and other references [13, 17].

We examined PV sequences in two different ways, using a variety of specific program settings. Alignments of the E6, E7, E1, E2, L2 and L1 sequences of all available PV types for which sequence of at least one complete ORF was available were combined into one master alignment for nucleotide sequences and one for protein sequences.

Two different tests were performed, one for detecting mosaics showing greatest similarity, in different regions, to two (or more) of the phylogenetically defined groups used elsewhere in this compendium (**intergroup** analysis), and one for detecting mosaics similar to two (or more) types drawn from a single group (**intragroup** analysis).

In the intergroup case, each type is successively removed from the master alignment, consensus sequences are determined for each group, and the program is run to determine which consensus sequence is most similar to the query sequence in each window. Sequences are added back into the background alignment after they have been used as query sequences.

In the intragroup analysis, each phylogenetically group is treated separately. Within a group, each sequence is successively removed and compared to the remaining sequences of the group, and then returned to the background set.

To provide a set of positive controls, two artificially created chimeric sequences were evaluated. One chimeric sequence was created by concatenating the first half of the HPV16 sequence to the second half of the HPV6b sequence; this provides a chimeric sequence composed of sequences taken from two different groups. Another was created by concatenating half of the HPV6b sequence to the other half of the HPV44 sequence, providing a chimera composed of two different types from the same group. Analyses were done with the original sequences (HPV6b, HPV16 and HPV44) present in the background alignment, and with the original sequences removed from the alignment.

Analyses were performed with the following program settings for both the inter- and intragroup analyses. Columns in the alignment containing a gap in either the query sequence or the background sequences were excluded (“stripped”) in all runs. Consensus sequences were determined using the most commonly observed base or residue at each position, with no minimum threshold of occurrence; in case of a tie, one of the equally common bases or residues is arbitrarily selected. As noted above, a confidence level of 99% was required for reporting a positive result, although the use of multiple tests makes it likely that a few false positives would emerge even at this level of stringency. Runs were performed both including and excluding invariant sites from the analysis. Invariant sites are are noninformative regarding relatedness or similarity, whether they are invariant because the positions are under selectional pressure or because, by chance, no mutations have occurred there [9]. A window composed of 50 informative sites will contain more information regarding similarity of the query to the background sequences than a 50-site window composed of a mixture of invariant and informative sites; however, it will also require a longer stretch of sequence, and thus be less sensitive to short mosaic patterns. For nucleotide sequences, window sizes of 50, 100, and 200, both including and excluding invariant positions, were tested; window sizes of 300 and 500 were also tested including invariant positions. Protein sequences were tested with a window size of 100, both including and excluding invariant positions.

Results

RIP analyses of nucleotide sequences and protein sequences under all the conditions above generally yielded similar results. Except as noted, results presented below are based on evaluating the nucleotide sequences with a window size of 100 positions, ignoring columns containing any gaps, but including invariant positions. All positive results under any condition are presented.

The intergroup analyses yielded no indication of recombination or other mosaic patterns. The results of this comparison are summarized in Table 1. In most cases, short stretches of nucleotide sequence (100 bases) from various regions of the genome were sufficient to identify a given type as a member of the same group to which it had previously been assigned by phylogenetic analyses; it is thus consistent with the observation that there is a high correlation between evolutionary distances from different reading frames [10]. In some cases, the similarity of a given sequence to the members of the group or supergroup to which it has been assigned is not sufficiently greater than its similarity to other groups to result in a confident assignment. Certain sequences (HPV54, HPV61, CgPV1) were not identified as members of any particular group, in agreement with their phylogenetic isolation among the sequences in the background set; others were either not identified as members of a particular group or identified as such in only a small number of windows, indicating either that they are relatively peripheral members of a group (e.g. RhPV or HPV60) or that the group is not terribly coherent (e.g. HPV41, ROPV and CRPV from supergroup E). In contrast, the artificial chimeric sequence created from HPV16 and HPV6b was identified as a mosaic, and more clearly so for larger windows; as expected, the sequence created from HPV6b and HPV44, both members of the A10 group, was not identified as an intergroup hybrid.

For the intragroup analyses, most types again showed no sign of a recombinant or other mosaic pattern. Results are summarized in Table 2. A somewhat larger number of types could not be confidently assigned a closest relative, which may result when a sequence is roughly equally related to two or more sequences, as occurs when the closest relatives are more closely related to each other than to the query sequence (e.g. HPV57, which is roughly equally closely related to HPV2a and HPV27, which are more closely related to each other than to HPV57). With window sizes of 50 and 100, a few positive results were given by the analyses of the nucleotide alignment. These positive results are summarized in Table 3. No positive results were obtained for larger window sizes (200, 300, or 500 bases), nor for analysis of the protein alignment. In contrast, the artificial chimeric sequence created from HPV6b and HPV44 was clearly identified as a mosaic for all window sizes tested (50, 100, 200, 300, 500 bases), including or excluding invariant positions, either when the original HPV6b and HPV44 sequences were retained or when they were removed from the background alignment.

Discussion

The positive results in the intragroup analyses may reflect some evidence of recombination. However, positive results were observed only with small window sizes (50 and 100 bases), and could not be extended to larger windows. The positive results presented in Table 3 appear to be in the “gray zone” in which it is impossible to say with confidence whether they are true or false positives.

One case was analyzed in greater detail. The results summarized in Table 3 suggest that HPV10 may be a mosaic of sequence related to HPV28 and HPV3. In Figure 2, two regions in which HPV10 appears to be most closely related to HPV28 and to HPV3, respectively, are shown. The difference in similarities is fairly dramatic: if these fragments were representative of larger regions, it would be hard to dismiss a hypothesis of recombination. However, taken in context of the complete nucleotide sequence alignment, these may simply reflect chance concentrations of (mis)matching positions. False positives are most likely to result when a query sequence is roughly equidistant from the two closest sequences; in this case, we may expect to find that there is a random pattern of similarity between the query sequence and the two background sequences in different windows. HPV3 and HPV28 appear, in phylogenetic analyses, to be more closely related to one another than to HPV10, from which they are roughly equidistant. Figure 3a shows the degree to which HPV10 is more similar to HPV3 or HPV28 in each window of the nucleotide alignment on which the analyses were performed. Two short regions exceed the 99% confidence cutoff (dashed lines) scored as significantly more related to one or the other background sequence, resulting in a positive result for HPV10, but the figure as a whole suggests a rather random pattern in which the two peaks may simply be outliers. The corresponding plot for the artificial HPV6b/HPV44 chimera against HPV11 and HPV55 (Figure 3b) much more convincingly illustrates the mosaic character of the sequence.

The analysis can reveal facts about the sequences other than the existence (or lack thereof) of recombinants in a dataset. It can also be used to validate the phylogenetic isolation of sequences which appear to have no close relatives in a phylogenetic tree; as noted before, a recombinant sequence which is not recognized as such may appear to constitute such an isolated lineage. A few isolated sequences, notably HPVs 34 and 54, and CgPV among Supergroup A, and perhaps HPVs 24 and 49 among Group

B1, stand out in PV phylogenies. The results presented in Tables 1 and 2 confirm that these sequences are indeed relatively isolated among the PV types analyzed here.

To conclude, little or no recombination can be observed in the currently available sequences of PV types. However, two caveats should appended to this conclusion. Firstly, the set of sequences for PV types is far from unbiased, and indeed the bias may be against recombinants, should they exist. This is because once an isolate is known to hybridize to a sufficient degree to a known type, it is often not investigated further. Moreover, a sample which hybridizes to sufficient degree to two (or more) types may be declared a multiple infection and again set aside. Thus, potential recombinants may be passed over as not being of interest. Second, in order to pick up a recombinant, the method employed here requires that the database being examined contain not only the recombinant sequence but also the sequences of (relatively close relatives of) two "source" strains, that is, the strains which donated portions of the original recombinant genome. Cases of possible recombinants such as that of HPV77, discussed above, would not be detected.

References

[1] S.Y. Chan, H.U. Bernard, C.K. Ong, S.P. Chan, B. Hofmann, and H. Delius. Phylogenetic analysis of 48 papillomavirus types and 28 subtypes and variants: a showcase for the molecular evolution of DNA viruses. *Journal of Virology*, 66(10):5714–25, 1992.

[2] E.M. de Villiers. Information regarding HPV-77 was kindly provided by the Human Papillomavirus Reference Center, Heidelberg.

[3] D. Eschle, M. Durst, J. ter Meulen, J. Luande, HC. Eberhardt, M. Pawlita, and L. Gissmann. Geographical dependence of sequence variation in the E7 gene of human papillomavirus type 16. *Journal of General Virology*, 73(7):1829–32, 1992.

[4] Y. Fujinaga, K. Okazawa, A. Nishikawa, Y. Yamakawa, M. Fukushima, I. Kato, and K. Fujinaga. Sequence variation of human papillomavirus type 16 E7 in preinvasive and invasive cervical neoplasias. *Virus Genes*, 9(1):85–92, 1994.

[5] D.V. Heck, C.L. Yee, P.M. Howley, and K. Munger. Efficiency of binding the retinoblastoma protein correlates with the transforming capacity of the E7 oncoproteins of the human papillomaviruses. *Proceedings of the National Academy of Sciences of the United States of America*, 89(10):4442–6, 1992.

[6] L. Ho, S.Y. Chan, R.D. Burk, B.C. Das, K. Fujinaga, J.P. Icenogle, T. Kahn, N. Kiviat, W. Lancaster, P. Mavromara-Nazos, et al. The genetic drift of human papillomavirus type 16 is a means of reconstructing prehistoric viral spread and the movement of ancient human populations. *Journal of Virology*, 67(11):6413–23, 1993.

[7] R. Kirnbauer, J. Taub, H. Greenstone, R. Roden, M. Durst, L. Gissmann, D.R. Lowy, and J.T. Schiller. Efficient self-assembly of human papillomavirus type 16 L1 and L1-L2 into virus-like particles. *Journal of Virology*, 67(12):6929–36, 1993.

[8] D. Kremsdorf, M. Favre, S. Jablonska, S. Obalek, L.A. Rueda, M.A. Lutzner, C. Blanchet-Bardon, P.C. Van Voorst Vader, and G. Orth. Molecular cloning and characterization of the genomes of nine newly recognized human papillomavirus types associated with epidermodysplasia verruciformis. *Journal of Virology*, 52(3):1013–8, 1984.

[9] W.-H. Li and D. Graur. *Fundamentals of Molecular Evolution*. Sinauer Associates, Inc., Sunderland, Massachusetts, 1991.

[10] G. Myers, H.U. Bernard, H. Delius, M. Favre, J. Icenogel, van Ranst M., and C. Wheeler. *Human Papillomaviruses: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences*. Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, New Mexico, 1994.

[11] Nucleotide Sequence Numbers. Nucleotide positions are reported in terms of the revised reference sequence presented elsewhere in this compendium as HPV16R.

[12] P. Pushko, T. Sasagawa, J. Cuzick, and L. Crawford. Sequence variation in the capsid protein genes of human papillomavirus type 16. *Journal of General Virology*, 75(4):911–6, 1994.

[13] RIP. Program documentation available on request from the HPV Sequence Database. Parties interested in applying the analysis presented here to their own data should contact the database., 1995.

[14] D. Robertson, B. Hahn, and P.M. Sharp. Recombination in AIDS viruses. *Journal of Molecular Evolution*, 40(3):249–59, 1995.

[15] E.C. Sabino, E.G. Shpaer, M.G. Morgado, B.T. Korber, R.S. Diaz, V. Bongertz, S. Cavalcante, B. Galvao-Castro, J.I. Mullins, and Mayer A. Identification of human immunodeficiency virus type 1 envelope genes recombinant between subtypes B and F in two epidemiologically linked individuals from Brazil. *Journal of Virology*, 68(10):6340–6, 1994.

[16] M.O. Salminen, J.K. Carr, D.S. Burke, and F.E. McCutchan. Identification of breakpoints and intergenotypic recombinants of HIV-1 by bootscanning. *AIDS Research and Human Retroviruses*, 11:1423–1425, 1995.

[17] A.C. Siepel, A.L. Halpern, C. Macken, and B.T.M. Korber. A computer program designed to rapidly screen for HIV-1 intersubtype recombinant sequences. *AIDS Research and Human Retroviruses*, 11:1413–1416, 1995.

[18] H.L. Smits, K.F. Traanberg, M.R. Krul, P.R. Prussia, C.L. Kuiken, M.F. Jebbink, J.A. Kleyne, R.H. van den Berg, B. Capone, A. de Bruyn, et al. Identification of a unique group of human papillomavirus type 16 sequence variants among clinical isolates from Barbados. *Journal of General Virology*, 75(9):2457–62, 1994.

[19] C.M. Wheeler. Personal communication.

[20] T. Yamada, C.M. Wheeler, A.L. Halpern, A.-C.M. Stewart, A. Hildesheim, B.B. Rush, and S.A. Jenison. Human papillomavirus type 16 variant lineages in united states populations characterized by nucleotide sequence analysis of the E6, L2 and L1 coding segments. *Journal of Virology*, In press.

Table I-A

Type	Group	Windows
HPV32	A1	1417
HPV42	A1	1484
HPV3	A2	2002
HPV28	A2	2168
HPV10	A2	1678
HPV29	A2	824
HPV2a	A4	3024
HPV27	A4	3301
HPV57	A4	2168
HPV26	A5	37
HPV51	A5	44
HPV30	A6	908
HPV53	A6	1061
HPV56	A6	1314
HPV66	A6	1224
HPV18	A7	158
HPV39	A7	305
HPV45	A7	68
HPV59	A7	189
HPV68	A7	287
HPV70	A7	371
HPV40	A8	2843
HPV7	A8	2469
HPV43	<none>	
HPV16	A9	72
HPV31	A9	99
HPV33	A9	201
HPV35h	A9	104
HPV52	A9	28
HPV58	A9	254
RhPV1	A9	17
HPV11	A10	587
HPV13	A10	2248
HPV44	A10	2677
HPV55	A10	2607
HPV6b	A10	767
PCPV1	A10	1694
HPV34	A9	37
HPV54	<none>	
HPV61	<none>	
CgPV1	<none>	

Table I-A (cont.)

Type	Group	Windows
HPV5	B1	2534
HPV36	B1	2838
HPV8	B1	2391
HPV12	B1	2622
HPV47	B1	2528
HPV14d	B1	2963
HPV20F	B1	2802
HPV21	B1	2988
HPV19	B1	3203
HPV25	B1	2904
HPV22	B1	218
HPV23	B1	260
HPV9	B1	521
HPV15	B1	792
HPV17	B1	670
HPV37	B1	659
HPV38	B1	390
HPV24	B1	1434
HPV49	B1	319
HPV4	B2	506
HPV48	B2	20
HPV50	B2	21
HPV65	B2	428
HPV60	<none>	
BPV1	C1	3957
BPV2	C1	3999
DPV	C2	794
EEPV	C2	955
RPV	C2	98
BPV3	D	201
BPV4	D	123
BPV6	D	195
COPV	E	11
HPV1a	E	80
HPV63	E	73
HPV41	<none>	
CRPV	<none>	
ROPV	<none>	
MmPV	<none>	
MnPV	<none>	
FPV1	<none>	

Table I-B

Type	Group	Windows	Group	Windows
HPV16_6b	A9	348	A10	667
HPV44_6b	A10	2710		

Table I-C

Type	Group	Windows	Group	Windows	
HPV16_6b	A9	6	A10	243	(window size = 100)
	A9	105	A10	571	(window size = 200)
	A9	264	A10	796	(window size = 300)
HPV44_6b	A10	2035			(window size = 100)

Table I. Intergroup Analyses. The table presents the group to which each query sequence was most similar, and the number of windows in which confidence in the assignment reached 99%, for comparison of nucleotide sequences with a window size of 100, including invariant positions, unless otherwise noted. **A.** Known types for which sequence of at least one complete ORF is available, ordered by phylogenetic groups as defined elsewhere in this compendium. **B.** Artificial chimeric sequences, original sequences retained in background. **C.** Artificial chimeric sequences, original sequences removed from background.

Table II-A

Type	Group	Windows
HPV3	HPV28	249
HPV28	HPV3	259
HPV10	**HPV3**	**11**
	HPV28	**16**
HPV29	<none>	
HPV2a	HPV27	707
HPV27	HPV2a	602
HPV57	<none>	
HPV30	HPV53	1164
HPV53	HPV30	1003
HPV56	HPV53	33
HPV18	HPV45	207
HPV45	HPV18	392
HPV39	HPV68	211
HPV68	HPV39	280
HPV59	<none>	
HPV70	<none>	
HPV7	HPV40	155
HPV40	HPV7	198
HPV43	<none>	
HPV16	<none>	
HPV31	HPV35h	3
HPV35h	<none>	
HPV52	<none>	
HPV33	HPV58	1243
HPV58	HPV33	1444
RhPV1	<none>	
HPV6b	HPV11	738
HPV11	HPV6b	913
HPV44	HPV55	3511
HPV55	HPV44	3319
HPV13	PCPV1	132
PCPV1	HPV13	227
HPV5	HPV36	61
HPV36	HPV5	39
HPV12	HPV8	43
HPV8	HPV12	38
HPV47	<none>	
HPV14d	HPV21	1
HPV20F	HPV21	57
HPV21	HPV14d	18
HPV19	HPV25	42
HPV25	HPV19	8
HPV15	<none>	
HPV17	HPV37	1
HPV37	HPV17	60
HPV22	HPV23	135
HPV23	HPV22	215
HPV38	HPV23	1
HPV24	<none>	
HPV49	<none>	
HPV9	<none>	
HPV4	HPV65	5381
HPV65	HPV4	5426
HPV48	HPV50	309
HPV50	HPV48	420
HPV60	<none>	
BPV1	BPV2	416
BPV2	BPV1	438
DPV	RPV	3
EEPV	RPV	40
RPV	EEPV	2
BPV3	<none>	
BPV4	<none>	
BPV6	<none>	

Table II-B

Type	Group	Windows	Group	Windows
6b_44	HPV6b	3005	HPV44	1767

Table II-C

Type	Group	Windows	Group	Windows
6b_44	HPV11	551	HPV55	2157

Table II. Intragroup Analyses. The table presents the type or types to which each query sequence was most similar, and the number of windows in which confidence in the assignment reached 99%, for comparison of nucleotide sequences with a window size of 100, including invariant positions. **A.** Known types for which sequence of at least one complete ORF is available, ordered by phylogenetic groups as defined elsewhere in this compendium. **B.** Artificial chimeric sequence, original sequences in background. **C.**. Artificial chimeric sequence, original sequences not in background.

Table III

	Query	Matches	Parameters
A.	HPV10	HPV3, HPV28, HPV29	w=50,100; including
			w=50; excluding
	HPV3	HPV28, HPV10	w=50; including
			w=50; excluding
	HPV28	HPV3, HPV10	w=50; including
	HPV30	HPV53, HPV56	w=50; including
	HPV53	HPV30, HPV56	w=50; including
	HPV39	HPV68, HPV70	w=100; excluding
	HPV18	HPV45, HPV59	w=50; including
			w=50; excluding
		HPV18, HPV45	w=50; including
	HPV31	HPV16,HPV35h	w=50; including
			w=50; excluding
	HPV37	HPV15, HPV17	w=50; including
	HPV50	HPV48, HPV60	w=50; including
B.	HPV6b_44	HPV6, HPV44	w=50,100,200,300,500; incl.
			w=50,100,200,300,500; excl.
C.	HPV6b_44	HPV11, HPV55	w=50,100,200,300,500; incl.
			w=50,100,200,300,500; excl.

Table III. Positive intragroup comparisons. Parameter settings examined were: window sizes of 50, 100, 200, 300, and 500 including invariant positions and window sizes of 50, 100, and 200 excluding invariant positions.“Including” (or “excluding”) means including (excluding) invariant positions. All positive results are shown. **A.** Known types. **B.** Artificial HPV6b-HPV44 chimera, HPV6b and HPV44 in background. **C.** Artificial HPV6b-HPV44 chimera, HPV6b and HPV44 not in background.

```
             E7          E5              L2         L1               LCR
          66777 333333333344444444 44444444444 666666666 77777777777777777777777777
          14389 888889999900000000 11122233444 344445555 12234455667777777777788888
          67295 566780788911344578 18923801001 934584567 93358825681122456889923334
          ..... 767123817026919868 33473070698 035510876 21294801882389213058953681

HPV16R    AATTT TTGCCTaTTCCGTaTAAT TATTATGAGAA AaAATCCTA gAAGAGgGCCTTAATACTCGAGGAAG

CASKI           ---A--c------g----                               --a-------------------
GB10      ----- ------------------                               --aA------------------
GB13      -----                                                  --a---G---------------
SIHA            -----Ac--G---g----                               --a---------...-------
OR.4724                            ----------- -g-------     --a-------------------
OR.4997                            ----------- -g-------     --a-------------------
OR.5110                            ----------- ---------     ----------------------
OR.6311                            -------G--- -g-------     --a-------------------   Background E
OR.8329                            ----------- -g-------     --a-------------------   Sequences
OR.6170                            ---------G- ---------     ----------------------
OR.8987                            ----------- -g-------     --a-------------------
NM.T197                            ----------- -g-------     --a-------------------
OR.9237                            ----------- -g-------     --a-------------------
NM.T446                            ----------- -g-------     --a-------------------
OR.0198                            ----------- -g-------     --a-------------------
NM.T455                            ----------C ---------     ----------------------

SB13            ------cC----Cg--T-                               --a------------------A
SB7             ------c------g--T-                               --a------C-----------A   Background
OR.2087                            ----------- -g-------     --a------C-----------A   As (E-like)
OR.5428                            ----------- -g-------     --a------C-----------A   Sequences
OR.4716                            ----------- -g-------     --a------C-----------A
OR.7574                            ----------- -g-------     --a------C---C-------A

BT11      ----- ------------------ ---C------- --------- --------------------------
DT4       ----- ------------------ ---NN------ --------- t-------------------------
DT24      T---- ------------------ --GCC------ --------- t-------------------------
BT23      ----- ------c------g-N-- ---C------- -g------- t-------------------------
DT42      ----- ------c------g---- ---C------- -g------- t-----a-------------------   E-like isolates
DC255     ----- ------c------g---- ---C------- -g------- t-----a-------------------   of Smits et al
DC212     ----- ------c------g---- ---C------- -g------- t-----a-------------------
DC141     ----- ------c------g---- AT-C------- -g------- t-----a-------------------
DC207     ----- ------c------g---- AT-C------- -g------- t--C--a-------------------
DC269     ----- ----T-c------g---- ---C------- -g------- t-----a-------------------

BT7       ---CG C-----c--G---g-C-- ----------- -g--C---- t-C---a-------------------
BT8       --CCG C-----c--G-A-g---- ----------- -g------- t-C---a-------------------
BT15      --CCG C-----c--G-A-g---- ----------- -g------- t-C---a-------------------
BT9       ---CG C-A---c--T---T---- ----------- -g------- tG----a-------------------   Mosaic
BT12      ---CG C-A---c--T---T---- ---N------- -g------- tG----a-------------------   isolates
BT10      ---CG C-A---c--T---TG--- ---N------- -g------- tG----a-------------------   of Smits et al
BT20      ---CG C-A---c--T---T---- ----------- -gC------ tG----a-------------------
BT22      ---CG CNA---c--T---T-N-- ----------- -g-------    ---a-------------------

SB16            C-----c--G-A-g---C                               CAa-TA--C-G-T-T-------
NM.T529                            -----C----- -g---AT--     CAa-TA--C-G-T-T-------
NM.4094                            -----C----- -g----T--     CAa-TA--C-G-T-T-------   Background AA
OR.4541                            -----C----- -g----T--     CAa-TA--C-GCT-T-------

SB21A           C-----c--G---g-C-C                               CAa-T-------T-T--ATTG-
TB1       -G-CG C-----c--G---g-C-C                               CAa-TA------T-T--ATTG-   Background Af2
OR.3473                            -----C----- -g--C-T--     CAa-TA------T-T--ATCG-   Sequences
OR.3759                            -----C--A-- -g--C-T--     CAa-TA------T-T--ATCG-
OR.7145                            -----C----- -g--C-T--     CAa-T-------T-T--ATCG-

TB13            C-A---G-CTT--T---C                               -Aa----A----T-T---T---
TB15      -G-CG                                                  -Aa-TA------T-TAG-T---
TB16      ---CG CCA---c--T---T---C                               -Aa-TA------T-T---T---
TB4       ---CG C-A---c--T---T---C                               -Aa--A-A----T-T---T---   Background Af1
TB8       ---CG C-A---c--T---T---G                               -Aa--A------T-T---T---   Sequences
OR.7587                            -----CA---- -g----TAG     -Aa--A------T-T---T---
OR.1905                            -----CA---- -g----TA-     -Aa--A-A----T-T---T---
```

Figure 1. Summary of HPV16 nucleotide sequence variation in fragments of E7, E5, L2 and L1 ORFs and the LCR. Sequences reported by Smits and colleagues [18] compared with sequences reported elsewhere [1, 3, 20] ("background sequences"). Each column in the figure corresponds to a position in the nucleotide sequence at which at least one sample differs from the corrected reference sequence (HPV16R); numbers at the top of the figure (read vertically) refer to the nucleotide position relative to the reference. A dash ("-") represents a match to the reference sequence at that position. Uppercase letters represent changes at positions in which the sequence of the reference clone matches the sequence of the majority of variants which have been sequenced; lowercase letters are used in positions at which the reference sequence appears to contain an unusual (mutant) base. Blanks represent unsequenced regions. From top to bottom are given: class E and class As (closely related to E) background sequences, sequences from Smits *et al* which appear to be from unexceptional class E isolates, the mosaic sequences from Smits *et al* which appear to be derived from Ax isolates over E7 and E5 but E isolates from L2, L1 and LCR, and background sequences from AA, Af2, and Af1 isolates.

```
A. HPV10 ATTGGCACTTATTGCGTGTAGAAAATGCTTTGCTGTACAAAGCAAGAGAATGTGGACTGACACATATTGGCCATCAGGTGGTGCCACCTCTTAGTGTAAC
   HPV3  -------A--GA----GT----GC-A------T-------------G---------T-A-----C--------C--------------------------
   HPV28 -------A--GA----------------------------------G---------T-------------------------------------------
   HPV29 -----T-TC-TA----------G-G---G---TAT--T------------------A------G---A-----C------------A-A---------G-

B. HPV10 CACCCCAGCGTCCACCCAAGCCCAGGTGGGCGCGTCCGAGGGACCGGAACAAAAGCGACAGCGACTCGAGGCGGTC...GACGGACAGCACCAGCAGCAGCGA
   HPV3  -----A----C----------------------------A--------G---------------------A-----...T----GG----G--------A-AG
   HPV28 ----AAG-----G---G--A--TC----------A----A------C-G-----A---------------A-AC--...A--T-GG----G-------GA-AG
   HPV29 -----GCT--C------------C-T----GC-TGAG-------A-----GG--A----GCA-G--T--A--A---GGGCCT--G-CA--G----------AG
```

Figure 2. Sequence comparison of HPV10 to HPV3, HPV28 and HPV29. The two regions presented here are regions which resulted in classifying HPV10 as a possible mosaic genome. These are the regions which cross the 99% confidence level in Figure 3a. **A.** 100 nt region showing greatest similarity between HPV10 and HPV28 (start = nt 2872 of HPV10; E2 ORF). **B.** 100 nt region showing greatest similarity between HPV10 and HPV3 (start = nt 3437 of HPV10; E2/E4 ORF).

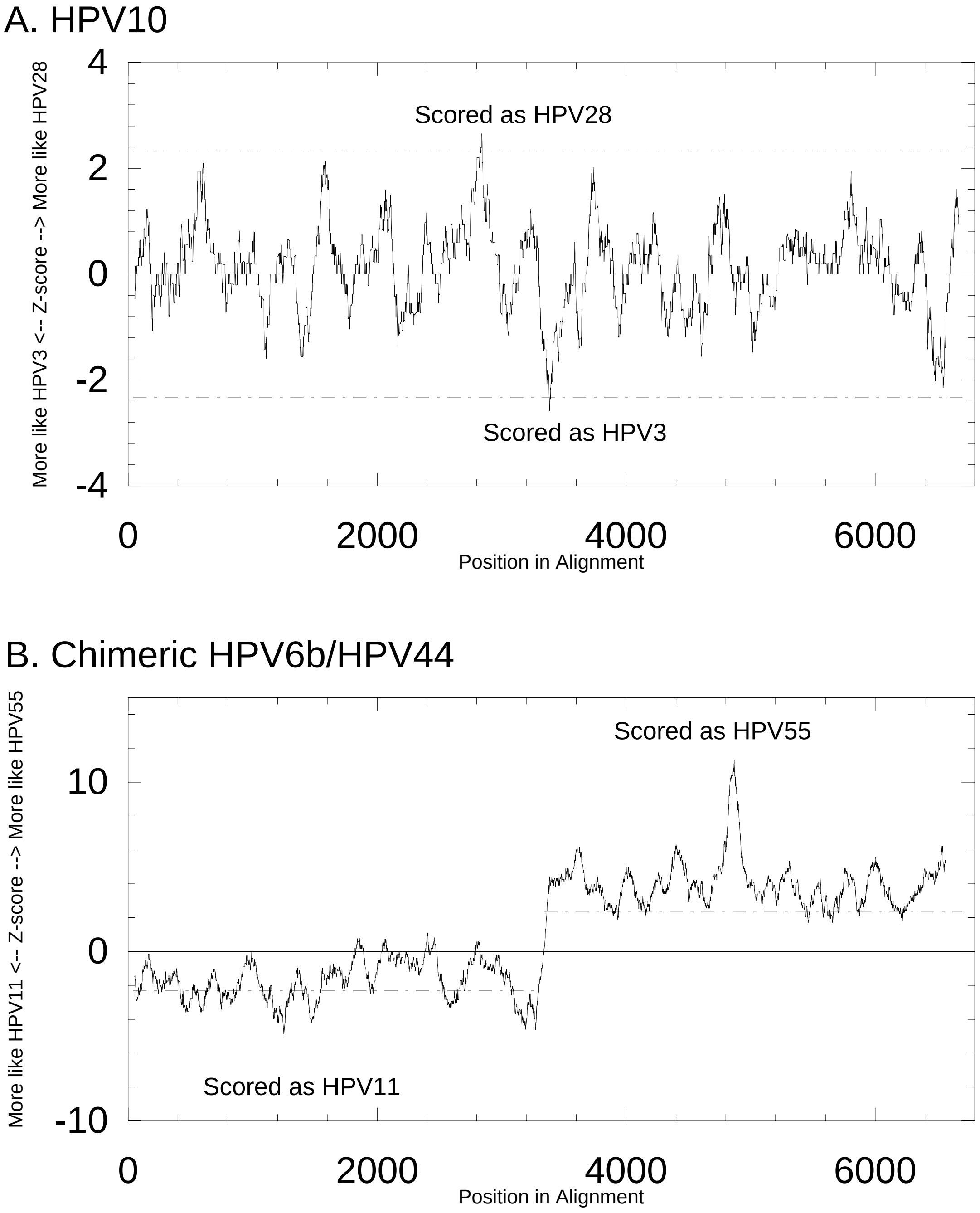

Figure 3. Patterns of similarity between query sequences and sequences to which they were scored as significantly most similar. The x-axis represents the position in the composite E6, E7, E1, E2, L2 and L1 nucleotide alignment. The y-axis represents the z-score for the difference in similarity between the query and the two related sequences. The dashed lines indicate the 99% confidence level. **A.** Comparison of HPV10 to HPV28 and HPV3. **B.** Comparison of the chimeric HPV6b/HPV44 sequence to HPV11 and HPV55, the closest relatives of HPV6b and HPV44.