JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 10, 2009

Pp. 1461–1473
DOI: 10.1089/cmb.2009.0091

Research Article

A Whole Genome Study and Identification of Specific Carcinogenic Regions of the Human Papilloma Viruses

*ABDOULAYE BANIRÉ DIALLO,[1,2] *DUNAREL BADESCU,[1]
MATHIEU BLANCHETTE,[2] and VLADIMIR MAKARENKOV[1]

ABSTRACT

In this article, we undertake a study of the evolution of human papillomaviruses (HPV), whose potential to cause cervical cancer is well known. First, we found that the existing HPV groups are monophyletic and that the high risk of carcinogenicity taxa are usually clustered together. Then, we present a new algorithm for analyzing the information content of multiple sequence alignments in relation to epidemiologic carcinogenicity data to identify regions that would warrant additional experimental analyses. The new algorithm is based on a sliding window procedure and a p-value computation to identify genomic regions that are specific to HPVs causing disease. Examination of the genomes of 83 HPVs allowed us to identify specific regions that might be influenced by insertions, by deletions, or simply by mutations, and that may be of interest for further analyses. Supplementary Material is provided (see online Supplementary Material at www.libertonline.com).

Key words: algorithm for carcinogenic region detection, evolutionary events, human papilloma viruses, phylogenetic trees.

1. INTRODUCTION

HUMAN PAPILLOMA VIRUSES (HPV) have a causal role in cervical cancer, with almost half a million new cases identified each year (Angulo and Carvajal Rodriguez, 2007; Bosch et al., 1995; Muñoz, 2000). The HPV genomic diversity is well known (Antonsson et al., 2000). About one hundred HPV types are identified, and the whole genomes of more than eighty of them are sequenced (see the latest Universal Virus Database report by International Committee on Taxonomy of Viruses [ICTV]). A typical HPV genome is a double-stranded, circular DNA genome of size close to 8 Kbp, with complex evolutionary relationships and a small set of genes. In general, the E5, E6, and E7 genes modulate the transformation process; the two regulatory proteins, E1 and E2, modulate transcription and replication; and the two structural proteins L1 and L2 compose the viral capsid. Protein E4 has an unclear function in the HPV life cycle; however, several studies indicate that it could facilitate the viral genome replication and the activation of viral late functions (Wilson et al., 2007), and it could also be responsible for virus assembly (Prétet et al., 2007). A HPV is considered to belong to a new HPV type if both its complete genome has been cloned and the DNA sequence

[1]Département d'informatique, Université du Québec à Montréal, Montréal Québec, Canada.
[2]McGill Centre for Bioinformatics and School of Computer Science, McGill University, Montréal, Québec, Canada.
*The two first authors contributed equally to the work and should be considered as joint first authors.

of the gene L1 differs by more than 10% from the closest known HPV type. The comparison of HPV genomes, conducted by ICTV, is based on nucleotide substitutions only (Muñoz et al., 2003; de Villiers et al., 2004). Older HPV classifications were built according to their higher or lower risk of cutaneous or mucosal diseases. Most of the HPV studies were based on single gene (usually E6 or E7) analyses. The latter genes are predominantly linked to cancer due to the binding of their products to the p53 tumor suppressor protein and the retinoblastoma gene product pRb (Van Ranst et al., 1992). To define carcinogenic types, we used epidemiologic data from a large international survey on HPVs in cervical cancer and from a multicenter case-control study conducted on 3,607 women with incident, histologically confirmed cervical cancer recruited in 25 countries (Muñoz et al., 2003, 2004). HPV DNA detection and typing in cervical cells or biopsies were centrally done using polymerase chain reaction (PCR) assays, which attests to the quality of the study (Muñoz et al., 2003). More than 89% of patients had squamous cell carcinoma, and about 5% had adenosquamous carcinoma (Muñoz et al., 2003) (Table 1). More than half of the infection cases are due to the types 16 and 18 of HPV, which are thus referred to as high-risk HPVs (Chan et al., 1995).

In this article, we studied a whole genome phylogenetic classification of the HPV and the insertion and deletion (indel) distribution among HPV lineages leading to the different types of cancer. First, we inferred a phylogenetic tree of 83 HPVs based on whole HPV genomes. We found that the evolution of the L1 gene, used by ICTV to establish the HPV classification, generally reflects the whole genome evolution. Second, we compared the gene trees built for the 8 most important HPV genes (E1, E2, E4, E5, E6, E7, L1, and L2) using the normalized Robinson and Foulds topological distance (Robinson and Foulds, 1981). Then, we described a new algorithm for analyzing the information content of multiple sequence alignments in order to identify regions that may be responsible for the carcinogenicity. This algorithm is based on a new formula taking into account the sequence similarity among carcinogenic taxa and the sequence dissimilarity between the carcinogenic and non-carcinogenic taxa, computed for a genomic region bounded by the position of the sliding window. To facilitate the identification of relevant regions, we compute p-values for the different regions according to their score obtained with our new formula. Using the new technique we developed, we examined all available genes in 83 HPV genomes and identified the specific genomic regions that would warrant interest for future biological studies.

Table 1. Distribution of Carcinogenic HPVs for the "Squam" and "Adeno" Types of Cancer

	Squamous cell carcinoma		*Adenocarcinoma and adenosquamous carcinoma*	
HPV types	*Number*	*% positive*	*Number*	*% positive*
HPV-16	1,452	54.38	77	41.62
HPV-18	301	11.27	69	37.30
HPV-45	139	5.21	11	5.95
HPV-31	102	3.82	2	1.08
HPV-52	60	2.25		
HPV-33	55	2.06	1	0.54
HPV-58	46	1.72	1	0.54
HPV-56	29	1.09		
HPV-59	28	1.05	4	2.16
HPV-39	22	0.82	1	0.54
HPV-51	20	0.75	1	0.54
HPV-73	13	0.49		
HPV-82	7	0.26		
HPV-26	6	0.22		
HPV-66	5	0.19		
HPV-6	2	0.07		
HPV-11	2	0.07		
HPV-53	1	0.04		
HPV-81	1	0.04		
HPV-55	1	0.04		
HPV-83	1	0.04		
Total	2,293	85.89	168	90.37

Complete genomic sequence data is not available yet for HPVs-35, HR, 68, and X.

2. INDEL ANALYSIS OF HPV GENOMES AND RECONCILIATION OF HPV GENE TREES

The 83 completely sequenced HPV genomes (all identified by the ICTV) were downloaded and aligned using ClustalW (Thompson et al., 1994), producing an alignment with 10426 columns. The phylogenetic tree of 83 HPVs (Fig. 1) was inferred using the PHYML program (Guindon and Gascuel, 2003) with the HKY substitution model. Bootstrap scores were computed to assess the robustness of the edges using 100 replicates. Most branches obtain support above 80%, but for a better readability, they are not represented in Figure 1. However, they are given in the Supplementary Material (see online Supplementary Material at www.liebertonline.com). As suggested in Van Ranst et al. (1992), the bovine PV of type 1 was used as outgroup to root this phylogeny. To the best of our knowledge, the constructed phylogenetic tree is the first whole genome phylogenetic tree of HPVs.

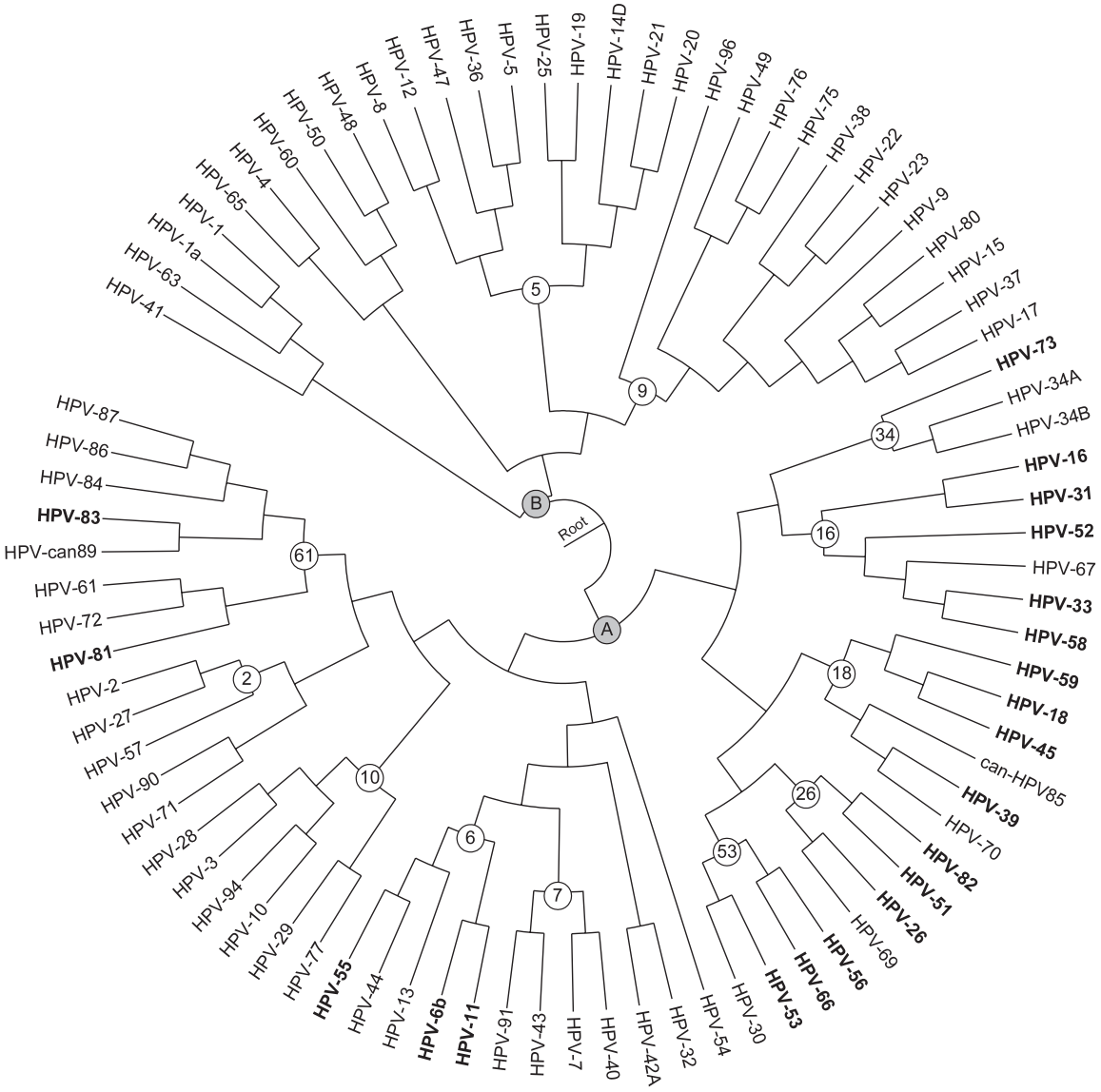

FIG. 1. Phylogenetic tree of 83 HPVs obtained with PHYML. The 21 carcinogenic HPV are shown in bold. The white nodes identify the existing HPV groups according to the ICTV and NCBI taxonomic classifications; the shaded nodes (**A**, **B**) distinguish between the non-carcinogenic and carcinogenic families. Bootstrap scores are above 80% for most of the branches; for a better readability, they are not represented. The HPVs 1 and 34 are present in two copies—(1 and 1a) and (34A and 34B)—respectively.

Our analysis revealed the presence of 12 known monophyletic HPV groups that are denoted by numerated nodes, labeled according to the ICTV annotation (Fig. 1). The other monophyletic groups obtained were not depicted by numbers. The whole-genome phylogeny obtained usually corresponds to the HPV classification provided by ICTV on the basis of the L1 gene. Most of the dangerous HPVs (Table 1) can be found in the sister subtrees rooted by the nodes 16 and 18.

As carcinogenicity may be introduced into a HPV by an insertion or deletion (indel) of a group of nucleotides, we first addressed the problem of indel distribution in the evolution of HPV. Thus, the most likely indel scenario was inferred using a heuristic method described in Diallo et al. (2006, 2007). Such a scenario includes the distribution of the predicted indel and base conservation events for all HPV genes. Table 2 reports, for each of the 8 main genes of HPV, the total number of conservations, insertions, and deletions of nucleotides that occurred during their evolution. Genes E1, L1, and L2 show more than 90% conservation at the nucleotide level; E2, E4, and E6 80–90%; and E5 and E7, respectively, 73% and 59%.

The highest indel frequencies are in the subtrees rooted by the node 61 where there are only low risks of carcinogenicity (Fig. 1). The groups included in the subtree *A* have low percentage of indels on in each branch. It is likely that the organisms of this subtree inherited their carcinogenicity from their closest common ancestor.

We also carried out an analysis intended to compare the topologies of the gene phylogenies built for the 8 main HPV genes. Thus, we first aligned, using ClustalW (Thompson et al., 1994), the HPV gene sequences, separately for each gene, and inferred 8 gene phylogenies using the PHYML program (Guindon and Gascuel, 2003) with the HKY model. In order to measure their degree of difference, we computed the Robinson and Foulds (RF) topological distances between each pair of gene trees (Robinson and Foulds, 1981). As the number of tree leaves varied from 70 to 83 (due to the non-availability of some gene sequences for a few HPVs), we reduced the size of some trees prior to this pairwise topological comparison and normalized all distances by the largest possible value of the RF distance, which is $2n - 6$ for two binary trees with n leaves. Figure 2 shows the results obtained, with RF distances are depicted as stacked rectangles. The results suggest that the trees representing the evolution of the E4 and E5 genes differ the most, on average, from the other gene phylogenies, whereas the phylogeny of E2 reconciles the most the topological differences of this group of gene trees. Two HPV gene phylogenies differ from each other by about 32%, on average. In the future, it might also be interesting to compare the gene trees we obtained using Maximum Likelihood tests such as Shimodaira-Hasegawa (Shimodaira and Hasegawa, 1999) or Kishino-Hasegawa (Kishino and Hasegawa, 1989) and to assess the confidence of phylogenetic tree selection using program such as CONSEL (Shimodaira and Hasegawa, 2001).

These results confirm the hypothesis made in a number of HPV studies (Narechania et al., 2005; Varsani et al., 2006), that most HPV genes undergo frequent recombination events. Uncritical phylogenetic analyses performed on recombinant sequences could lead to the impression of novel, relatively isolated branches. Recently, Angulo and Carvajal-Rodriguez (2007) have provided new support to the recent evidence of recombination in HPV. They found that the gene with recombination in most of the groups is L2 but the highest recombination rates were detected in L1 and E6. Gene E7 was recombinant only within the HPV16 type. The authors concluded that this topic deserves further study because recombination is an impor-

Table 2. For Each of the Eight Main HPV Genes: Numbers (and Average Numbers) of Conservations (including Substitutions), Insertion, and Deletions of Nucleotides that Occurred during Evolution

Variable/gene	*Conservation*	*Insertion*	*Deletion*	*Avg. cons.*	*Avg. ins.*	*Avg. del.*
E1	12111	601	2774	0.918	0.003	0.010
E2	13304	306	3460	0.852	0.001	0.022
E4	6318	195	2117	0.851	0.001	0.038
E5	1688	356	503	0.731	0.021	0.031
E6	7323	613	1529	0.890	0.002	0.011
E7	3457	0	1393	0.594	0.000	0.039
L1	9664	314	2751	0.927	0.001	0.010
L2	21716	494	5138	0.923	0.004	0.026

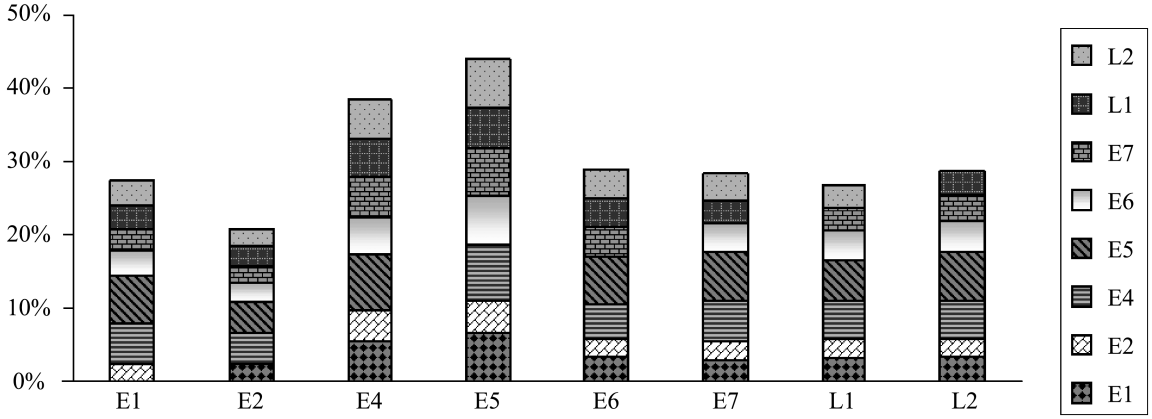

FIG. 2. Average normalized Robinson and Foulds topological distance for each of the 8 main HPV genes. Each column of the diagram represents a gene and consists of the stacked rectangles whose heights are proportional to the values of the normalized Robinson and Foulds topological distances between the phylogeny of this gene and those represented by the stacked rectangles. The column heights depicts the total average distance. For the sake of presentation, the percentage values on the ordinate axis were divided by 7 (which is the number of pairwise comparisons made for each gene tree).

tant evolutionary mechanism that could have a high impact both in pharmacogenomics and for vaccine development.

3. ALGORITHM FOR THE IDENTIFICATION OF PUTATIVELY CARCINOGENIC REGIONS

This section describes a new algorithm intended for finding genomic regions that may be responsible for HPV carcinogenicity. The algorithm is based on the hypothesis that sequence regions responsible for cancer are likely to be more similar among carcinogenic HPVs than between carcinogenic and non-carcinogenic HPVs. The following procedure was adopted. First, 83 available HPV genomes were downloaded and inserted into a relational database along with the clinical information regarding identified HPV types and histological type of cancer occurrences (Muñoz et al., 2003, 2004). We constructed three HPV types datasets: "High-Risk," containing HPVs16 and 18; "Squamous," containing HPV types responsible for squamous cell carcinoma (HPV-6, 11, 16, 18, 26, 31, 33, 39, 45, 51, 52, 53, 55, 56, 58, 59, 66, 73, 81, 82, 83); and "Adeno," with types responsible for adenocarcinoma (HPV-16, 18, 31, 33, 35, 39, 45, 51, 58, 59; Table 1). HPV types with incomplete genome information or without annotations were excluded from the dataset. As previously, we used the gene sequences aligned separately for each gene.

Then, we scanned all gene sequence alignments using a sliding window of a fixed width (in our experiments, the window width ranged from 3 to 20 nucleotides; Fig. 3). First, a detailed scan of each gene with increments of 1 nucleotide was performed to identifying the regions with a potential for causing carcinogenicity (the main results are reported in Table 3), and called here hit regions. Second, a non-overlapping windows of width 20 nucleotides was carried out for plotting Figures 4–8. Three separate analyses were made for the three above-described carcinogenic families: High-Risk, Squamous, and Adeno HPVs.

Once the window position is fixed and the taxa are assigned to the sets X (carcinogenic HPVs) and Y (non-carcinogenic HPVs), the hit region identification function, denoted here as Q, can be computed. This function is defined as a difference between the means of the squared distances computed among the sequence fragments (bounded by the sliding window position) of the taxa from the set X and those computed only between the sequence fragments from the distinct sets X and Y. The mean of the squared distances computed among the sequence fragments of the carcinogenic taxa from the set X, and denoted here $V(X)$, is computed as follows:

$$V(X) = \frac{1}{(N(X)(N(X)-1)/2)} \sum_{\{x_1, x_2 \in X | x_1 \neq x_2\}} dist_h^2(x_1, x_2), \tag{1}$$

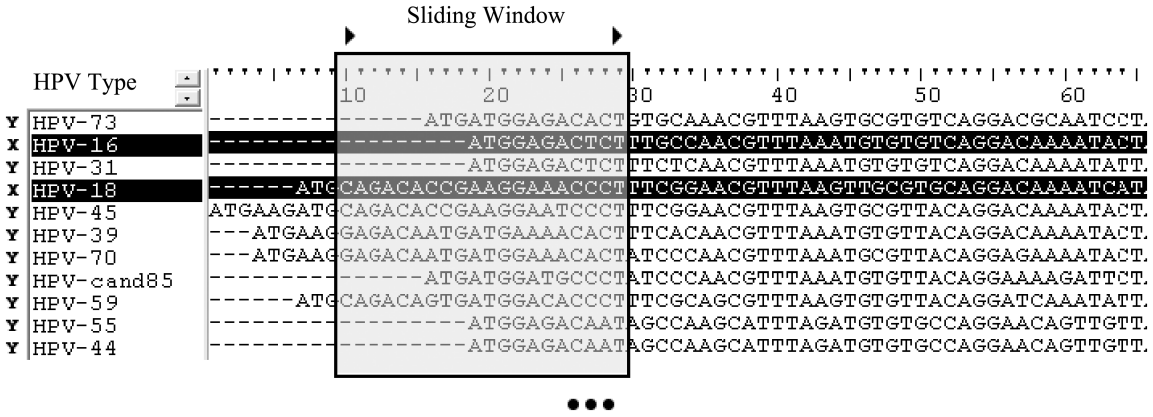

FIG. 3. A sliding window of a fixed width was used to scan each HPV gene separately. The sequences in black belong to the set X (carcinogenic HPVs; in this example, HPVs 16 and 18); all other sequences belong to the set Y (non-carcinogenic HPVs). The organism is indicated in the column on the extreme left.

and the mean of the squared distances computed only between the sequence fragments from the distinct sets X and Y, and denoted here as $D(X, Y)$, is computed as follows:

$$D(X,Y) = \frac{1}{N(X)N(Y)} \sum_{\{x \in X, y \in Y\}} dist_h^2(x,y), \tag{2}$$

where $N(X)$ and $N(Y)$ are the cardinalities of the sets X and Y, respectively, and $dist_h(x_1, x_2)$ is the Hamming distance between the sequence fragments corresponding to the taxa x_1 to x_2.

TABLE 3. SELECTED HIGH-SCORING REGIONS WITH RESPECT TO THE VALUES OF THE HIT REGION IDENTIFICATION FUNCTION Q

Dataset	*Gene*	Q	*Index*	*Window width*	$D(X, Y)$	$V(X)$
High-Risk	E1	0.417	695	16	0.74	0.22
Squam	E1	0.345	575	14	0.50	0.08
Adeno	E1	0.353	307	20	0.52	0.09
High-Risk	**E2**	**0.553**	**1289**	**13**	**0.76**	**0.02**
Squam	E2	0.385	613	16	0.47	0.00
Adeno	E2	0.415	1265	20	0.66	0.14
High-Risk	E4	0.480	606	17	0.62	0.00
Squam	E4	0.373	1035	15	0.46	0.01
Adeno	E4	0.395	549	15	0.49	0.00
High-Risk	E5	0.339	88	13	0.41	0.01
Squam	E5	0.401	72	16	0.50	0.00
Adeno	E5	0.363	72	16	0.44	0.00
High-Risk	E6	0.496	725	17	0.69	0.05
Squam	**E6**	**0.531**	**725**	**17**	**0.76**	**0.06**
Adeno	**E6**	**0.521**	**725**	**17**	**0.75**	**0.06**
High-Risk	E7	0.258	206	13	0.34	0.05
Squam	E7	0.263	445	16	0.38	0.08
Adeno	E7	0.262	110	16	0.40	0.10
High-Risk	**L1**	**0.574**	**241**	**14**	**0.79**	**0.02**
Squam	L1	0.294	1159	15	0.34	0.00
Adeno	L1	0.302	1181	17	0.56	0.20
High-Risk	L2	0.310	1751	14	0.65	0.28
Squam	L2	0.320	1916	15	0.38	0.00
Adeno	L2	0.313	1914	17	0.37	0.00

The best results for the contiguous regions of size 13–20 are reported. The best entry by HPV type (High-Risk, Squam, Adeno) and by gene is presented. The largest values of Q are in bold.

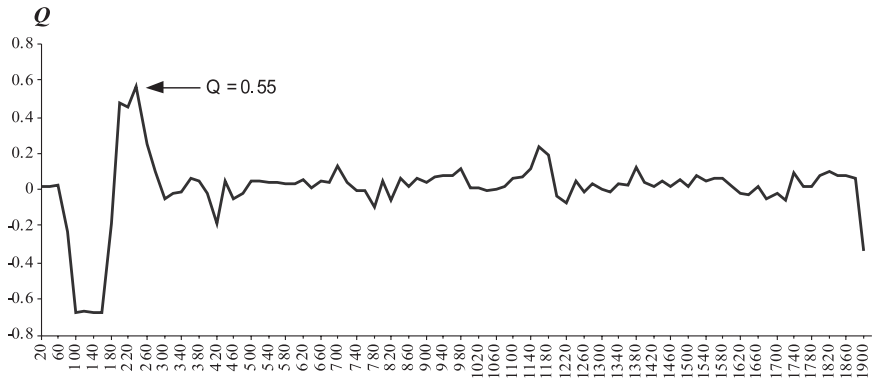

FIG. 4. The variation of the hit identification function Q for the High-Risk HPVs (HPVs 16 and 18) obtained with the non-overlapping sliding widows of width 20 during the scan of the L1 gene. The abscissa axis represents the window position.

Then, the hit region identification function Q is defined as follows:

$$Q = ln(1 + D(X, Y) - V(X)). \tag{3}$$

The larger the value of this function for a certain genomic region, the more distinct are the carcinogenic taxa from the non-carcinogenic ones. The use of the Hamming distance instead of the well-adapted sequence to distance transformations—such as the Jukes-Cantor (1969), Kimura (1980) 2-parameter, or TamuraNei (1993) corrections—is justified by the two following facts: first, often the latter transformation formulae are not applicable to short sequences (remember that in our experiments the sequence lengths, equal to the sliding window width, varied from 3 to 20 nucleotides), and second, most of the well-known transformation models either ignore gaps or assign a certain penalty to them. As the carcinogenicity of HPVs can be related to an insertion or deletion of a group of nucleotides, the gaps should not be ignored but rather considered as valid characters, with the same weight as the other nucleotides, when computing the pairwise distances between the genomic regions.

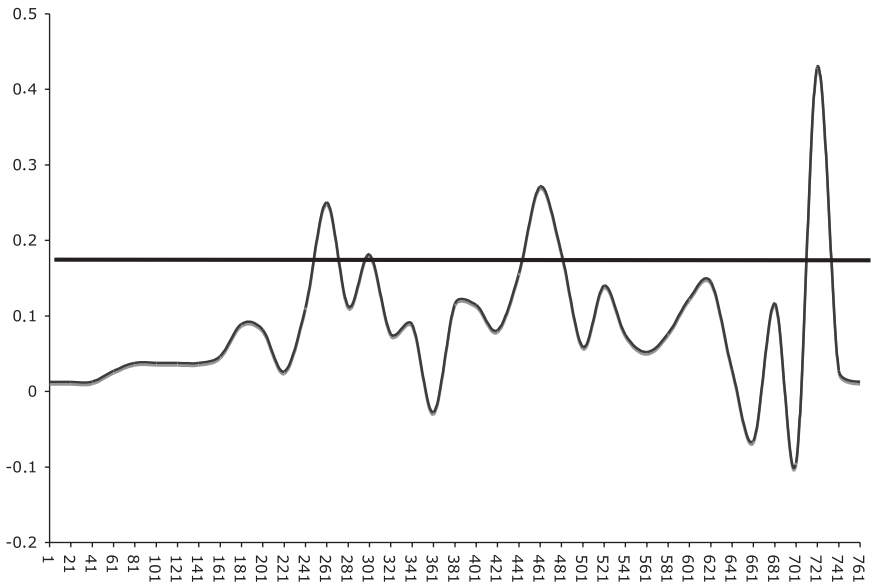

FIG. 5. The variation of the hit identification function Q for the High-Risk HPVs (HPVs 16 and 18) obtained with the non-overlapping sliding widows of width 20 during the scan of the E6 gene. The horizontal line cutting the graph represents the threshold of p-value less than 0.001. The abscissa axis represents the window position.

The time complexity of this algorithm executed with overlapping sliding windows of a fixed width, and advancing one alignment site by step, is $O(l \times n^2 \times w)$, where l is the length of the multiple sequence alignment, n the number of taxa, and w the window width. However, this complexity can be reduced to $O(n^2 \times l)$ if we avoid recomputing the Hamming distance for neighboring overlapping windows. This can be done by only removing the value of the left column of the sliding window while taking into account the value of added column in the Hamming distance of the sliding window. For a non-overlapping sliding window, the time complexity is $O(n^2 \times l)$. If the width of the sliding window varies, as was the case in our experiments, the time complexity should be obviously multiplied by the difference between the maximum and minimum window widths. The detailed algorithmic scheme is presented below (Algorithm 1).

Algorithm 1 Algorithmic scheme (MSA, *MSA_L*,*X*, *N(X)*, *Y*, *N(Y)*, *WIN_MIN*, *WIN_MAX*, *S*, *TH*)

Require: MSA: Multiple sequence alignment (considered as a matrix),
MSA_L: Length of MSA,
X: Set of carcinogenic taxa,
N(X): Cardinality of the set X,
Y: Set of non-carcinogenic taxa,
N(Y): Cardinality of the set Y,
WIN_MIN: Minimum sliding window width,
WIN_MAX: Maximum sliding window width,
S: Sliding window step,
TH: Minimum Q value for Hit (i.e., hit threshold).
Ensure: Set of Hit Regions: (*win_width*, *idx*, *Q*), where
win_width : Current sliding window width,
idx : Hit Index (i.e., its genomic position),
Q : Value of the hit region identification function.

for *win_width* **from** *WIN_MIN* **to** *WIN_MAX* **do**
  for *idx* **from** 0 **to** *MSA_L–win_width* **with step** *S* **do**
    MSA_X ← MSA[*X*][*idx..idx* + *win_width*]
    MSA_Y ← MSA[*Y*][*idx..idx* + *win_width*]
    $V(X) \leftarrow D(X, Y) \leftarrow 0$
    for all distinct $i, j \in X$ **do**
      $V(X) \leftarrow V(X) + dist_h^2(\text{MSA\_X}[i], \text{MSA\_X}[j])$
    end for
    $V(X) \leftarrow 2 \times V(X)/(N(X) \times (N(X) - 1))$
    for each $i \in X$ **and** $j \in Y$ **do**
      $D(X, Y) \leftarrow D(X, Y) + dist_h^2(\text{MSA\_X}[i], \text{MSA\_Y}[j])$
    end for
    $D(X, Y) \leftarrow D(X, Y)/(N(X) \times N(Y))$
    $Q \leftarrow ln(1 + D(X, Y) - V(X))$
    if $Q > TH$ **then**
      identify the current region (*win_width, idx, Q*) *as a hit region*
    end if
  end for
end for

To identifiy a region as a hit, one might use a measure to determine whether the given region has a value of Q higher than a given threshold. However, it is unclear what will be the best value of threshold, since the distribution of values of Q might be different in function of the alignment. One possibility could be to rank the Q values and choose a set of highest ones. Moreover, an approach involving the computation of p-values could be implemented to determine the regions that have a value of Q that is different from the normal Q values of the alignment. Here, we used the mentioned different approaches to choose the relevant regions according to their value of Q. To compute the p-value for each given region W_i with a Q value of Q_i, random sampling of the alignment columns according to the window size has been done. One million samples were generated and their Q values computed. For each given region, the number of times that Q from the sample is higher than Q_i is counted. It is worth noting that one would expect most of the regions with value of Q to have a p-value of less than 0.001.

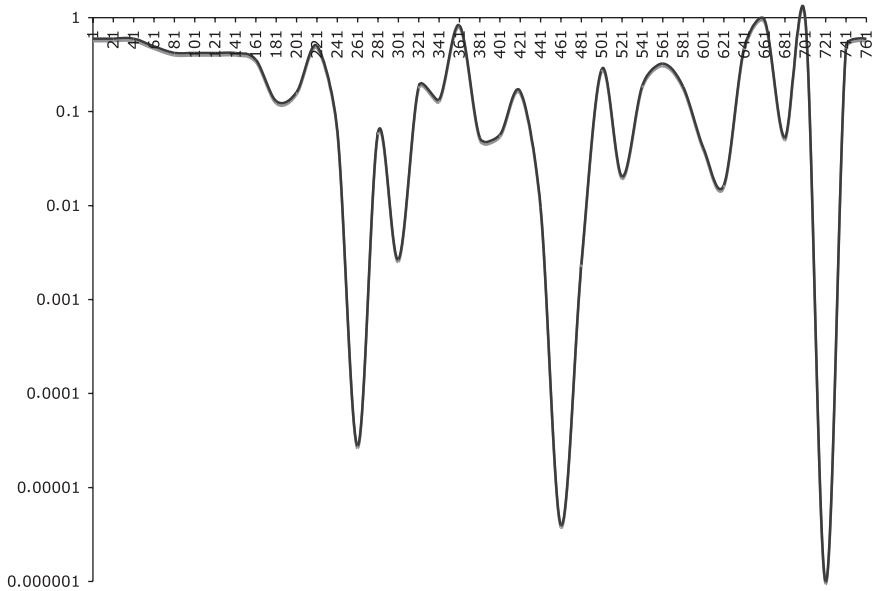

FIG. 6. The variation of the p-value in the different region of the alignment for the High-Risk HPVs (HPVs 16 and 18) obtained with the non-overlapping sliding widows of width 20 during the scan of the E6 gene. The abscissa axis represents the window position.

4. RESULTS AND DISCUSSION

The procedure for identifying hit regions in the 83 available HPV genomes was carried out twice: first, with overlapping windows of width w ($w = 3.20$), advancing one alignment site by step, and second, with non-overlapping windows of width 20. The 8 most important HPV genes (Table 3) were scanned in such a way. The scan based on the overlapping windows provided over 35,000 values of Q bigger than 0.25. From the best 100 results obtained for each gene, we manually selected (Table 3) the longest contiguous regions (up to 20 nucleotides) corresponding to the largest values of the hit region identification function Q. The values of Q were dependent on the window width, with better results usually associated with small windows.

For instance (Table 3), for larger window sizes, the largest values of Q were found during the scans of genes E2 and E6 for all types of HPVs, with the exception of the overall best score obtained during the scan of the gene L1 for the High-Risk HPV types (the value of 0.574 for a 14-nucleotide region starting with the index 241; Table 3). For windows of small width, the largest values of Q were observed during the scan of the gene E4 for the High-Risk HPV category, but in Table 3 we show only the best results for the longer contiguous regions of size 13 to 20 nucleotides. All the regions presented in Table 3 have a p-value of 0.

Figure 4 depicts the progressive results obtained during the scan of the L1 gene and the High-Risk HPVs (HPVs 16 and 18) with the non-overlapping windows of size 20 nucleotides. The highest score, for the non-overlapping windows of size 20 among all genes and all types of HPV-caused cancer, of the Q function ($Q = 0.55$) was obtained for this gene.

As most of the largest values of Q were obtained for the genes E2 and E6, we also present in Figures 7 and 8 the progressive results diagrams illustrating the scan of these genes with the non-overlapping windows of size 20. The largest values of the hit region identification function Q are usually found during the scan of the genes E2 and E6. Moreover, we found that in these two genes the number of regions obtaining p-values less than 0.001 is the largest. For instance, in gene E6, three large regions of size between 40 nucleotides and 60 nucleotides have a p-value less than 0.001 (Figs. 5 and 6). The last region of figure of E6 surprisingly corresponds to a PDZ domain-binding motif (-X-T-X-V) at the carboxy terminus of the protein, which is essential for targeting PDZ proteins for proteasomal degradation. Such proteins include hDlg, hScrib, MAGI-1, MAGI-2, MAGI-3, and MUPP1 (Choongho and Laimonis, 2004). The interaction between the E6 protein and hDLG or other PDZ domain-containing proteins could be an underlying mechanism in the development of HPV-associated cancers (Kiyono et al., 1997).

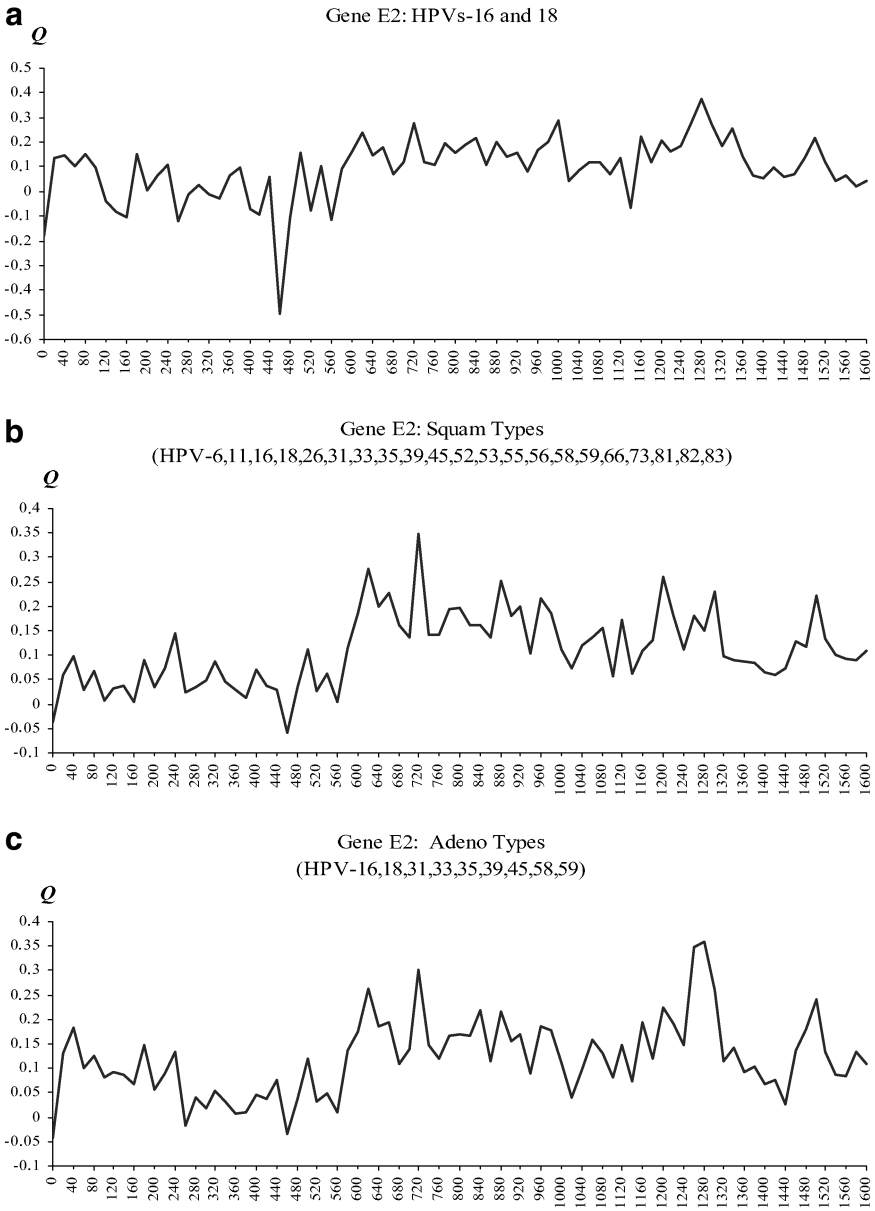

FIG. 7. The variation of the hit identification function Q for the following: (**a**) High-Risk HPVs (HPV-16 and 18). (**b**) Squam cancer causing HPVs. (**c**) Adeno cancer causing HPVs obtained with the non-overlapping sliding widows of width 20 during the gene E2 scan.

It is worth noting that according to recent findings the high expression of E6 and disruption of E2 might play an important role in the development of HPV-induced cervical cancer (Wang et al., 2007). As result of E6 high expression, the immune system is potentially evaded (Cordano et al., 2008). Disruption of the gene E2 was observed in invasive carcinomas (Chan et al., 2007) and in high-grade lesions (Graham and Herrington, 2000). Surprisingly, the overall largest value of Q was obtained for a specific region of the L1 gene. This underlines the possible use of our method for investigating particular regions of capsidal proteins in relation with vaccine design. It has been shown that linear epitopes within the protein L1 that induce neutralizing antibodies exist (Combita et al., 2002).

We observed that the results obtained depend on the window width. As substitutions affect individual sites whereas indels often involve several consecutive nucleotides, small window sizes will tend to favor the former. However, the use of the Hamming distance, which does not ignore gaps in calculation, and variable window width allows us to account for both substitution and indel events. In the future, it would be interesting

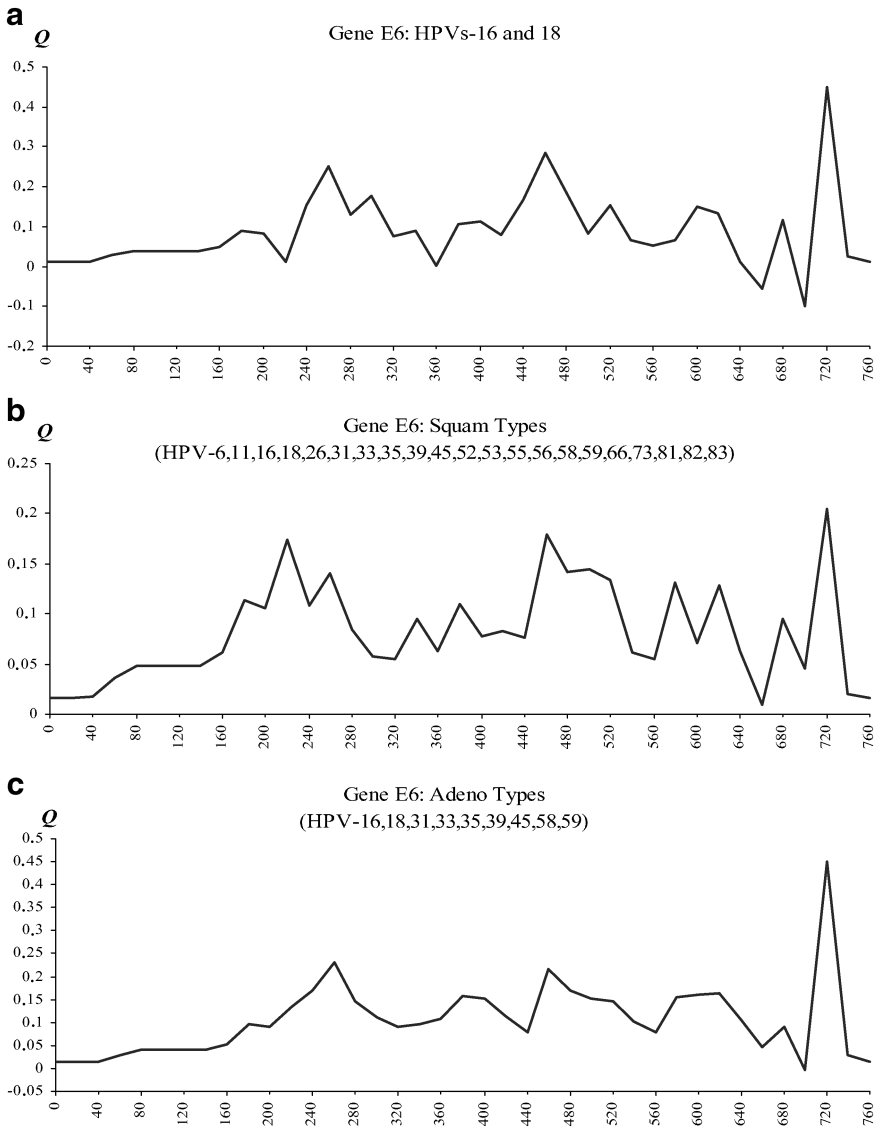

FIG. 8. The variation of the hit identification function Q for the following: (**a**) High-Risk HPVs (HPV-16 and 18). (**b**) Squam cancer causing HPVs. (**c**) Adeno cancer causing HPVs obtained with the non-overlapping sliding widows of width 20 during the gene E6 scan.

to study in more detail, in collaboration with virologists, all genomic regions providing the highest scores of the hit region identification function Q (particular attention should be paid to the E2, E6 and L1 genes), and to determine, for each selected region, the evolutionary events (substitutions or indels) responsible for the observed differences in the carcinogenic and non-carcinogenic HPVs, and then establish at which level (i.e., on which branch) of the associated gene phylogeny this event has occurred. It may also be interesting to consider merging our results to those given by methods for detecting sequences under lineage-specific selection such as DLESS (Siepel et al., 2006). Next, we plan to compare this work with other approaches on computational virology, which used some simpler methods, such as signatures, to analyze other viruses. Another interesting development would be to design more sophisticated statistical tests allowing one to measure the statistical significance of the obtained results.

ACKNOWLEDGMENTS

We thank Alix Boc and Emmanuel Mongin for their useful comments. B.D. is an NSERC fellow.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Angulo, M., and Carvajal Rodriguez, A. 2007. Evidence of recombination within human alpha-papillomavirus. *Virol. J.* 4, 33.

Antonsson, A., Forslund, O., Ekberg, H., et al. 2000. The ubiquity and impressive genomic diversity of human skin papillomaviruses suggest a commensalic nature of these viruses. *J. Virol.* 74, 11636–1164.

Bosch, F., Manos, M., Muñoz, N., et al. 1995. Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. International Biological Study on Cervical Cancer (IBSCC) study group. *J. Nat. Cancer Instit.* 87, 796–802.

Chan, P., Cheung, J., Cheung, T., et al. 2007. Profile of viral load, integration, and e2 gene disruption of hpv58 in normal cervix and cervical neoplasia. *J. Infect. Dis.* 196, 868–875.

Chan, S., Delius, H., Halpern, A., et al. 1995. Analysis of genomic sequences of 95 papillomavirus types: uniting typing, phylogeny, and taxonomy. *J. Virol.* 69, 3074–3083.

Choongho, L., and Laimonis, A. 2004. Role of the pdz domain-binding motif of the oncoprotein e6 in the pathogenesis of human papillomavirus type 31. *J. Virol.* 78, 12366–12377.

Combita, A.-L., Touzé, A., Bousarghin, L., et al. 2002. Identification of two cross-neutralizing linear epitopes within the l1 major capsid protein of human papillomaviruses. *J. Virol.* 76, 6480–6486.

Cordano, P., Gillan, V., Bratlie, S., et al. 2008. The e6e7 oncoproteins of cutaneous human papillomavirus type 38 interfere with the interferon pathway. *Virology* 377, 408–418.

de Villiers, E., Fauquet, C., Broker, T., et al. 2004. Classification of papillomaviruses. *Virology* 324, 17–27.

Diallo, A., Makarenkov, V., and Blanchette, M. 2007. Exact and heuristics methods to indel maximum likelihood problem. *J. Comput. Biol.* 14, 446–461.

Diallo, A., Makarenkov, V., and Blanchette, M. 2006. Finding maximum likelihood indel *Lect. Notes Comput. Sci.* 4205, 171–185.

Graham, D., and Herrington, C. 2000. Hpv-16 e2 gene disruption and sequence variation in cin 3 lesions and invasive squamous cell carcinomas of the cervix: relation to numerical chromosome abnormalities. *Mol. Pathol.* 53, 201–206.

Guindon, S., and Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *System. Biol.* 52, 696–704.

Kishino, H., and Hasegawa, M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from dna sequence data, and the branching order in hominoidea. *J. Mol. Evol.* 29, 170–179.

Kiyono, T., Hiraiwa, A., Fujita, M., et al. 1997. Binding of high-risk human papillomavirus E6 oncoproteins to the human homologue of the *Drosophila* discs large tumor suppressor protein. *Proc. Natl. Acad. Sci. USA* 94, 11612–11616.

Muñoz, N. 2000. Human papillomavirus and cancer: the epidemiological evidence. *J. Clin. Virol.* 19, 1–5.

Muñoz, N., Bosch, F., Castellsagué, X., et al. 2004. Against which human papillomavirus types shall we vaccinate and screen? The international perspective. *Int. J. Cancer* 111, 278–285.

Muñoz, N., Bosch, F., de Sanjosé, S., et al. 2003. Epidemiologic classification of human papillomavirus types associated with cervical cancer. *N. Engl. J. Med.* 384, 518–527.

Narechania, A., Chen, Z., DeSalle, R., et al. 2005. Phylogenetic incongruence among oncogenic genital alpha human papillomaviruses. *J. Virol.* 79, 15503–15510.

Prétet, J., Charlot, J., and Mougin, C. 2007. Virological and carcinogenic aspects of hpv. *Bull. Acad. Nat. Med.* 191, 611–613.

Robinson, D., and Foulds, L. 1981. Comparison of phylogenetic trees. *Math. Biosci.* 53, 131–147.

Shimodaira, H., and Hasegawa, M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* 16, 1114–1116.

Shimodaira, H., and Hasegawa, M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17, 1246–1247.

Siepel, A., Pollard, K., and Haussler, D. 2006. New methods for detecting lineage-specific selection. *Proc. RECOMB 2006* 190–205.

Thompson, J., Higgins, D., and Gibson, T. 1994. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680.

Van Ranst, M., Kaplanlt, J., and Burk, R. 1992. Phylogenetic classification of human papillomaviruses: correlation with clinical manifestations. *J. Gen. Virol.* 73, 2653–2660.

Varsani, A., Van der Walt, E., Heath, L., et al. 2006. Evidence of ancient papillomavirus recombination. *J. Gen. Virol.* 87, 2527–2531.

Wang, J., Ding, L., Gao, E., et al. 2007. Analysis on the expression of human papillomavirus type 16 E2 and E6 oncogenes and disruption of E2 in cervical cancer. *Zhonghua Liu Xing Bing Xue Za Zhi* 28, 968–971.

Wilson, R., Ryan, G., Knight, G., et al. 2007. The full-length E1–E4 protein of human papillomavirus type 18 modulates differentiation-dependent viral dna amplification and late gene expression. *Virology* 362, 453–460.

Address correspondence to:
Dr. Abdoulaye Baniré Diallo
Départment d' informatique
Université du Québec à Montréal
C.P. 8888
Succursale Centre-Ville
Montréal, Québec, H3C 3P8, Canada

E-mail: diallo.abdoulaye@uqam.ca