An Evolutionary Study of the Human Papillomavirus

UQAM, Février 2009, Montréal, Canada

Abdoulaye Baniré Diallo (UQAM, MCB)

UQÀM

McGill

McGill Centre for Bioinformatics
McGill University . Montreal . Quebec . Canada

Human Papilloma Viruses

1) Form a well known family of viruses

2) More than 100 types identified and more than 80 types are fully sequenced

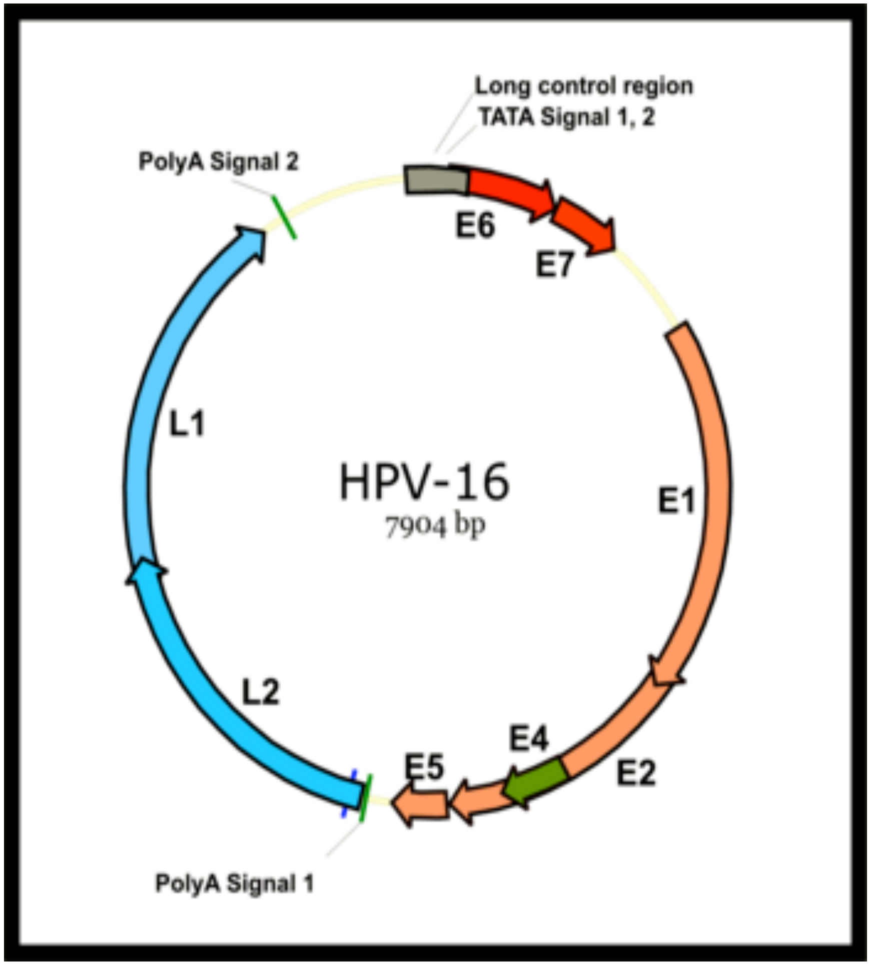
3) Cause Genital warts (condyloma), cervical and skin cancer) (see Dede, the tree-man)

4) They are double-stranded, circular DNA with sizes close to 8 Kbp

5) Complex evolutionary relationships and a small set of genes

Human Papilloma Viruses

- E5, E6, and E7 modulate the transformation process
- E1 and E2, regulatory proteins, modulate transcription and replication
- L1 and L2 are structural proteins and compose the viral capsid
- L1 gene is used to identify new type
- E4 has an unclear function although it could facilitate viral genome replication and assembly
- Genes E6 and E7 are important in cancer, due to the binding to *p53* tumor repressor



Wilson, R., Ryan, G.B., Knight, G.L., Laimins, L.A., Roberts, S., Virology 362(2) (2007) 453–460

Pr´etet, J.L., Charlot, J.F., Mougin, C., Bulletin Academic National de Medecine 191(3) (2007) 611–613

Human Papilloma Viruses and Cervical Cancer

1) Diagnostic data from 3607 women with cervical cancer from 25 countries

2) more than 89% of them have squamous cell carcinoma (SQUAM cancer)

3) 5% have adenosquamous carcinoma (ADENO cancer)

4) More than the half of the infections are due to type 16 and 18 the most infectious one

N.Munoz, et al., *N Engl J Med*, (2003), 384, pp.518 –527.

Human Papilloma Viruses and Cervical Cancer

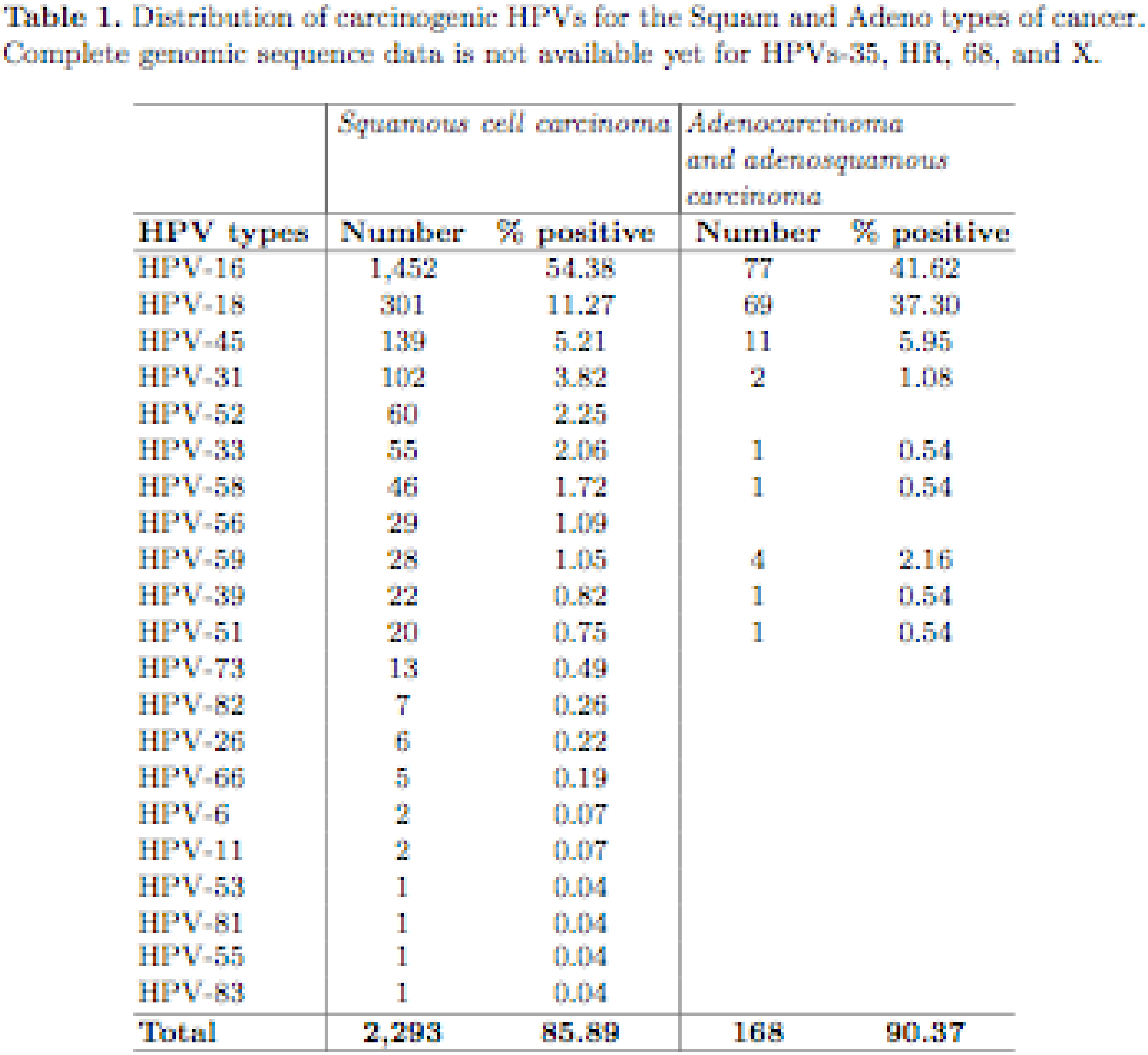


Table 1. Distribution of carcinogenic HPVs for the Squam and Adeno types of cancer. Complete genomic sequence data is not available yet for HPVs-35, HR, 68, and X.

	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**

N.Munoz, et al., *N Engl J Med*, (2003), 384, pp.518 –527.

Our goals

1) Study the whole genome phylogeny of the Human Papilloma Viruses

2) Identify the evolutionary patterns of the viral lineages

3) Define a new algorithm to identify regions that may be responsible for the carcinogenicity of the HPVs.

4) Identify relevant hit regions

N.Munoz, et al., *N Engl J Med*, (2003), 384, pp.518 –527.

Phylogeny of the Human Papilloma Viruses

1) 83 whole genomes have been obtained from the International Committee on Taxonomy of Viruses (ICTV)

2) Orthologous regions have been sorted

3) Sorted regions have been aligned using Clustal-W

4) The phylogenetic tree of 83 HPV was inferred using the PHYML with the HKY model of evolution

5) the bovine PV of type 1 was used as an outgroup to root the phylogenetic tree

6) 100 replicates for bootstrap have been chosen

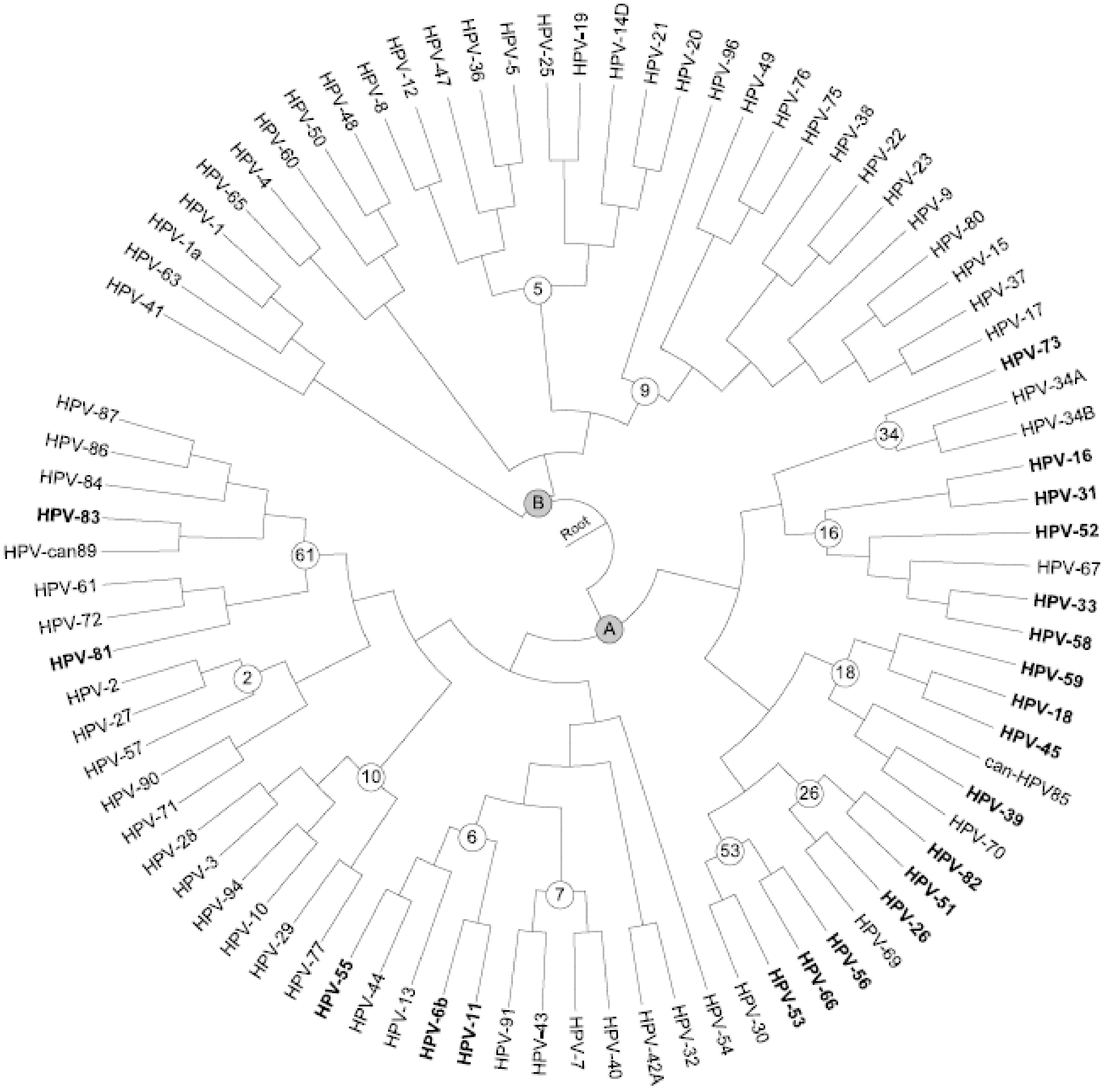
ICTVdB Management. *Edn. Büchen-Osmond, C.*, (Columbia U, New York, USA, 2006).

S. Guindon , O. Gascuel, Syst. Biol., (2003) 52 pp. 696-704.

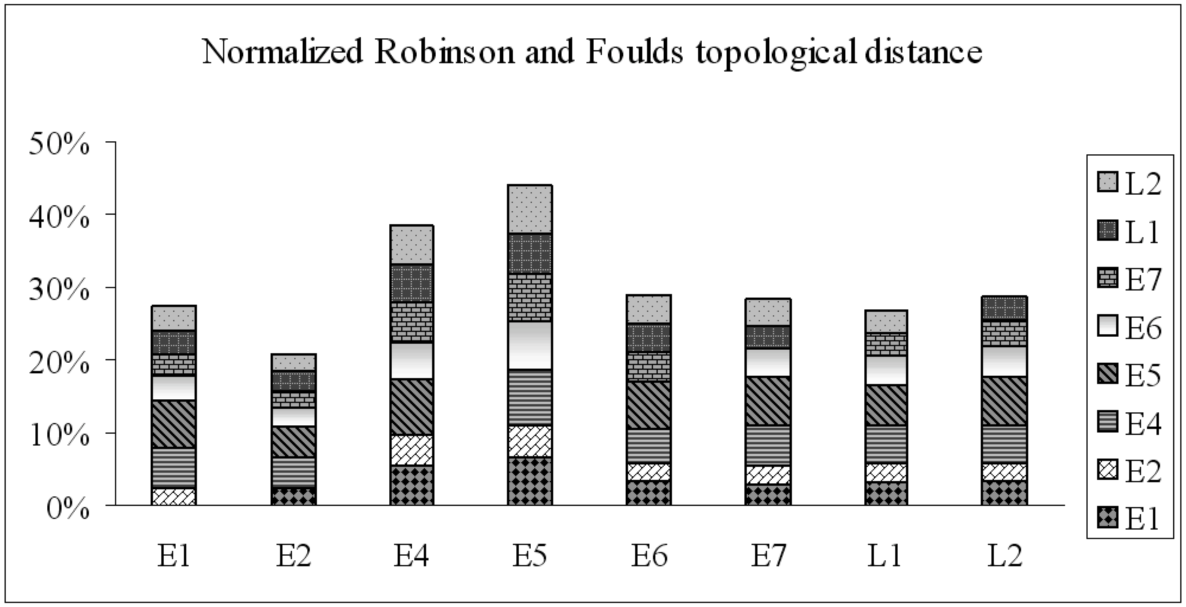
J.D. Thompson, D.G. Higgins, T.J. Gibson, *Nucl. Acids Res*. (1994), 22, pp. 4673-4680.

M. Van Ranst, J.B. Kaplanlt, R.D. Burk , *J. Gen. Virol*. (1992), 73, pp. 2653-2660.

Phylogeny of the Human Papilloma Viruses



Different Gene Evolutionary Histories



1) Different genes undergo different evolutionary histories, two HPV gene phylogenies differ from each other by about 5%, on average.

2) This could be explained by the hypothesis made in a number of HPV studies, that most HPV genes undergo frequent recombination events.

Narechania, A., Chen, Z., DeSalle, R., Burk, R.D., Journal of Virology 79 (2005) 15503-15510

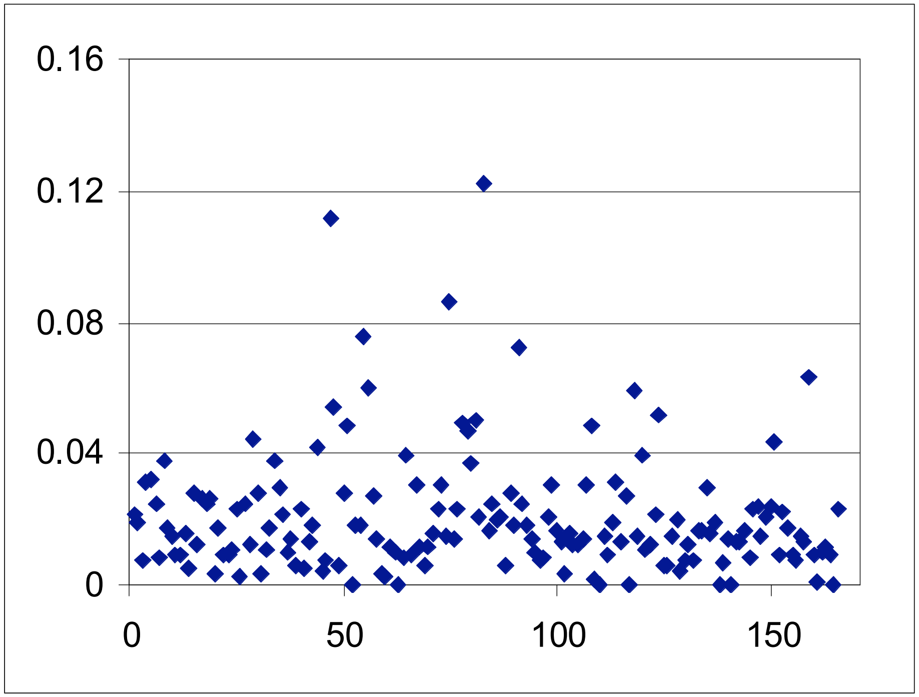
Varsani, A., Van der Walt, E., Heath, L., Rybicki, E.P., Williamson, A.L., Martin D.P.:,Journal of General Virology,87 (2006) 2527-2531

Phylogeny of the Human Papilloma Viruses and Indel Distribution

- Most likely indel Scenario has been built as well as the level of confidence in the prediction

- Most of the genes have more than 90% of the characters conserved throughout the evolution

- The indel frequencies are higher in the subtrees rooted by the node 61 where there are only low risks of carcinogenicity



A. B. Diallo, V. Makarenkov, M. Blanchette., *J. Comp. Biol.*doi:10.1089/cmb.2007.A006.

Phylogeny of the Human Papilloma Viruses and Indel Distribution

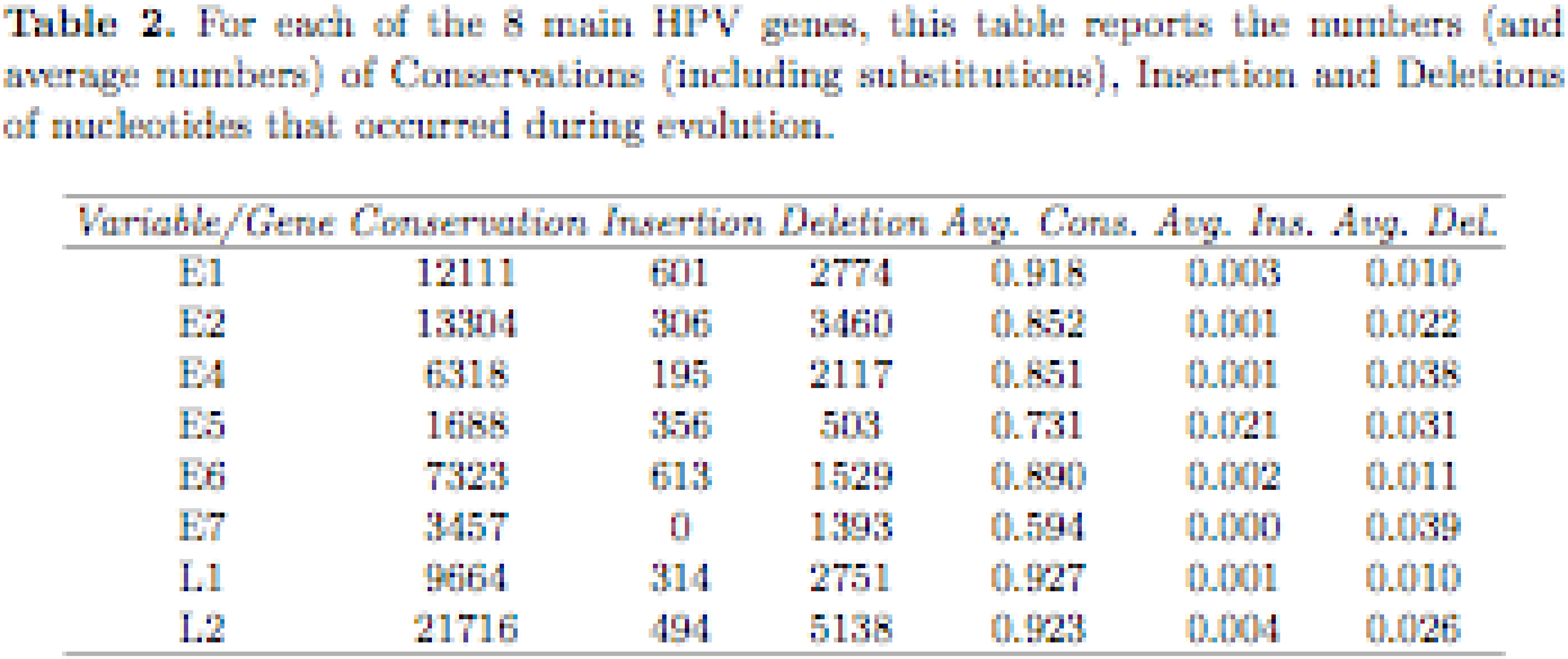


Table 2. For each of the 8 main HPV genes, this table reports the numbers (and average numbers) of Conservations (including substitutions), Insertion and Deletions of nucleotides that occurred during evolution.

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

A. B. Diallo, V. Makarenkov, M. Blanchette., *J. Comp. Biol.*doi:10.1089/cmb.2007.A006.

Relation between Squamous and Adeno Carcinoma and Evolutionary events

- linear and polynomial regressions to check for the presence of relationships between the explanatory variables and response variables.
- Explanatory variables are conservations, insertions and deletions
- Response variables are cancer/no cancer outcomes for the SQUAM and ADENO
- Eight HPV genes for the group of 83 HPV viruses have been considered

Relation between Squamous and Adeno Carcinoma and Evolutionary events

Statistics /Genes	*% of var iance for Lin. Regr.*	*% of variance for Pol. Regr.*	*Lin. Regr. p-value*	*Pol. Regr. p-value*	*Difference p-value*
E1 (81)	24.89	41.02	0.01	0.01	0.03
E2 (81)	24.49	41.70	0.01	0.01	0.02
E4 (57)	**32.12**	**58.47**	**0.01**	**0.01**	**0.01**
E5 (20)	39.84	64.98	0.49	0.72	0.71
E6 (81)	31.80	43.42	0.01	0.01	0.08
E7(81)	30.89	38.36	0.01	0.01	0.17
L1(83)	24.74	33.38	0.01	0.01	0.30
L2(83)	**42.55**	**47.54**	**0.01**	**0.01**	**0.64**
All genes	27.57	36.15	0.02	0.03	0.65

1. V. Makarenkov, P. Legendre, *Ecology*, (2002) 83(4), pp. 1146-1161.

Relation between Squamous and Adeno Carcinoma and Evolutionary events

- The results shows that when considering the HPV genes *E4* and *L2*, the presence and absence of the SQUAM and ADENO cancers correlate the best with the considered evolutionary event

- These two genes should be further analysed by virologists interested by studying the carcinogenic human papilloma viruses and evolutionary events

Our goals

1) Study the whole genome phylogeny of the Human Papilloma Viruses

2) Identify the evolutionary structures of the viral lineages

3) Define a new algorithm to identify regions that may be responsible for the carcinogenicity of the HPVs.

4) Identify relevant hit regions

N.Munoz, et al., *N Engl J Med*, (2003), 384, pp.518 –527.

Datasets for the identification of putatively carcinogenic regions

1) 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.

2) Three HPV Types Datasets:

- ”High-Risk”, (HPVs16 and 18)

- ”SQUAM”, 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)

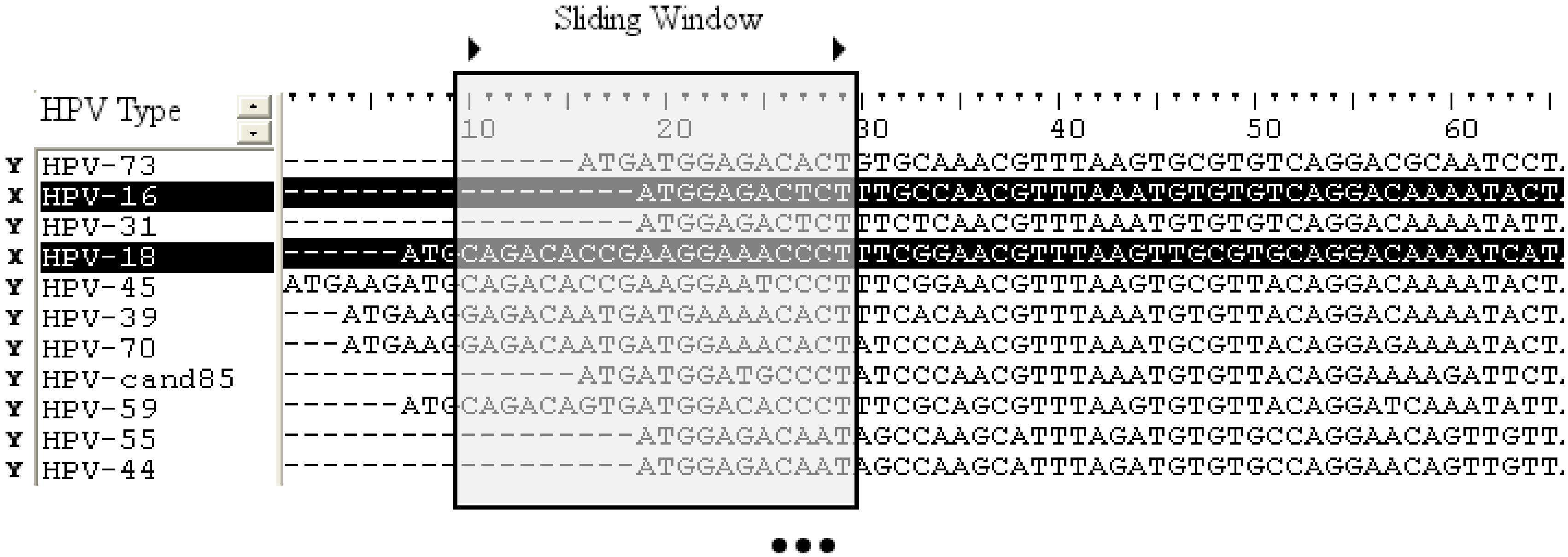
- “ADENO” with types responsible for Adenocarcinoma (HPV-16, 18, 31, 33, 35, 39, 45, 51, 58, 59)

3) Each gene was independently aligned using Clustal-W

Mu˜noz, N., Bosch, F.X., de Sanjos, S., Herrero, R., Castellsagu, X., Shah, K.V., Snijders, P.J.F., Meijer, C.J.L.M., New England Journal o
Medecine 384(2003) 518–527

Mu˜noz, N., Bosch, F.X., Castellsagu, X., Daz, M., de Sanjose, S., Hammouda, D., Shah, K.V., Meijer, C.J., International Journal of Cancer, 111(2004) 278–285

Hit Detection Problem



```
    HPV Type                10        20        30        40        50        60
Y   HPV-73        ---------------ATGATGGAGACACTGTGCAAACGTTTAAGTGCGTGTCAGGACGCAATCCT.
X   HPV-16        ------------------ATGGAGACTCTTTGCCAACGTTTAAATGTGTGTCAGGACAAAATACT.
Y   HPV-31        ------------------ATGGAGACTCTTTCTCAACGTTTAAATGTGTGTCAGGACAAAATATT.
X   HPV-18        ------ATGCAGACACCGAAGGAAACCCTTTCGGAACGTTTAAGTTGCGTGCAGGACAAAATCAT.
Y   HPV-45        ATGAAGATGCAGACACCGAAGGAATCCCTTTCGGAACGTTTAAGTGCGTTACAGGACAAAATACT.
Y   HPV-39        ---ATGAAGGAGACAATGATGAAAACACTTTCACAACGTTTAAATGTGTTACAGGACAAAATACT.
Y   HPV-70        ---ATGAAGGAGACAATGATGGAAACACTATCCCAACGTTTAAATGCGTTACAGGAGAAAATACT.
Y   HPV-cand85    ---------------ATGATGGATGCCCTATCCCAACGTTTAAATGTGTTACAGGAAAAGATTCT.
Y   HPV-59        ------ATGCAGACAGTGATGGACACCCTTTCGCAGCGTTTAAGTGTGTTACAGGATCAAATATT.
Y   HPV-55        ------------------ATGGAGACAATAGCCAAGCATTTAGATGTGTGCCAGGAACAGTTGTT.
Y   HPV-44        ------------------ATGGAGACAATAGCCAAGCATTTAGATGTGTGCCAGGAACAGTTGTT.
```

•••

- Label species as X (carcinogenic HPVs) and Y (non-carcinogenic HPVs)
- Compute the region identification function Q, for a genomic region bounded by the position of the sliding window based on:
  - Sequence similarity among carcinogenic taxa
  - Sequence dissimilarity between the carcinogenic and non-carcinogenic taxa.

Region Identification Function Q

- The mean of the squared distances computed between the sequence fragments from the distinct sets X and Y

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

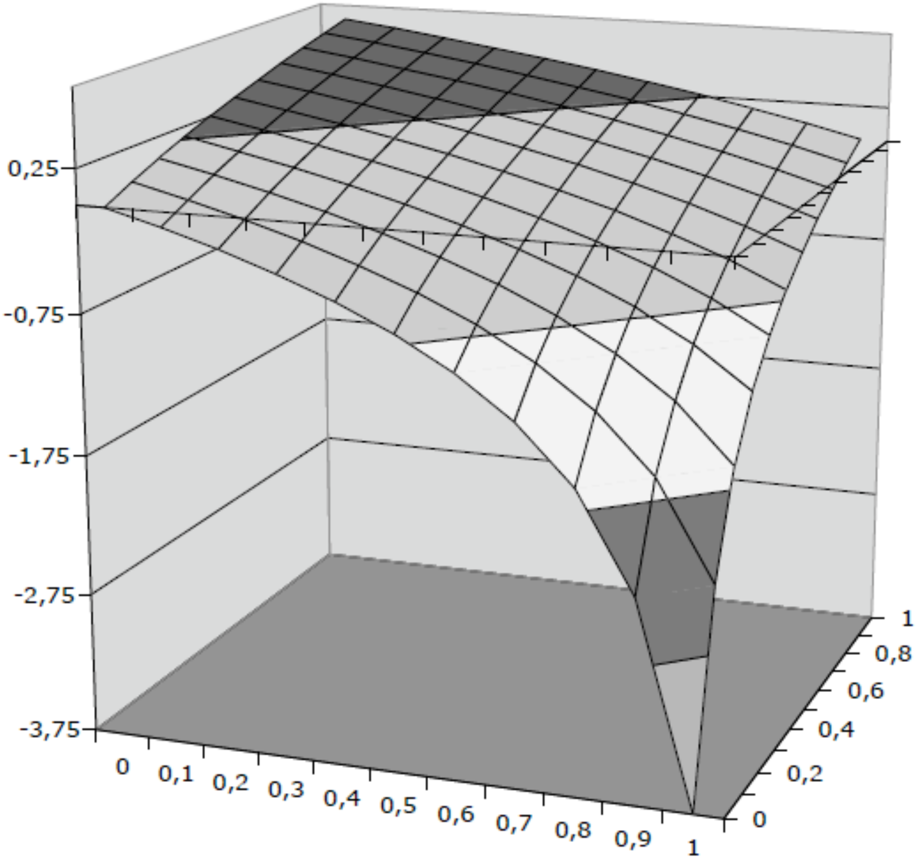
- The mean of the squared distances computed among only the sequence fragments of the carcinogenic taxa from the set X

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

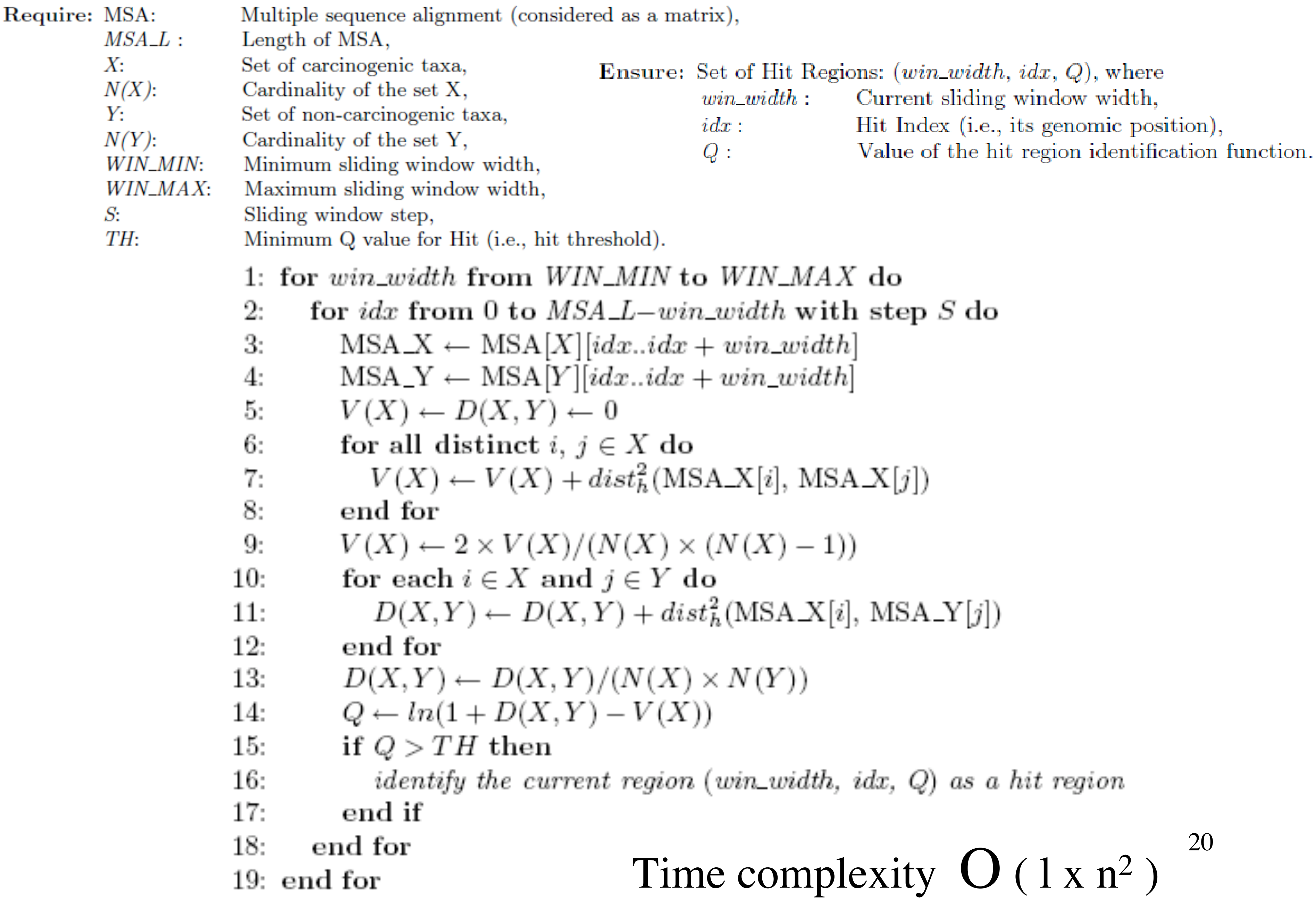
Region Identification Function Q

$$Q = ln(1 + D(X,Y) - V(X)).$$

V(X) < D(X,Y)	Q > 0
V(X) = D(X,Y)	Q = 0
V(X) > D(X,Y)	Q < 0



Algorithm



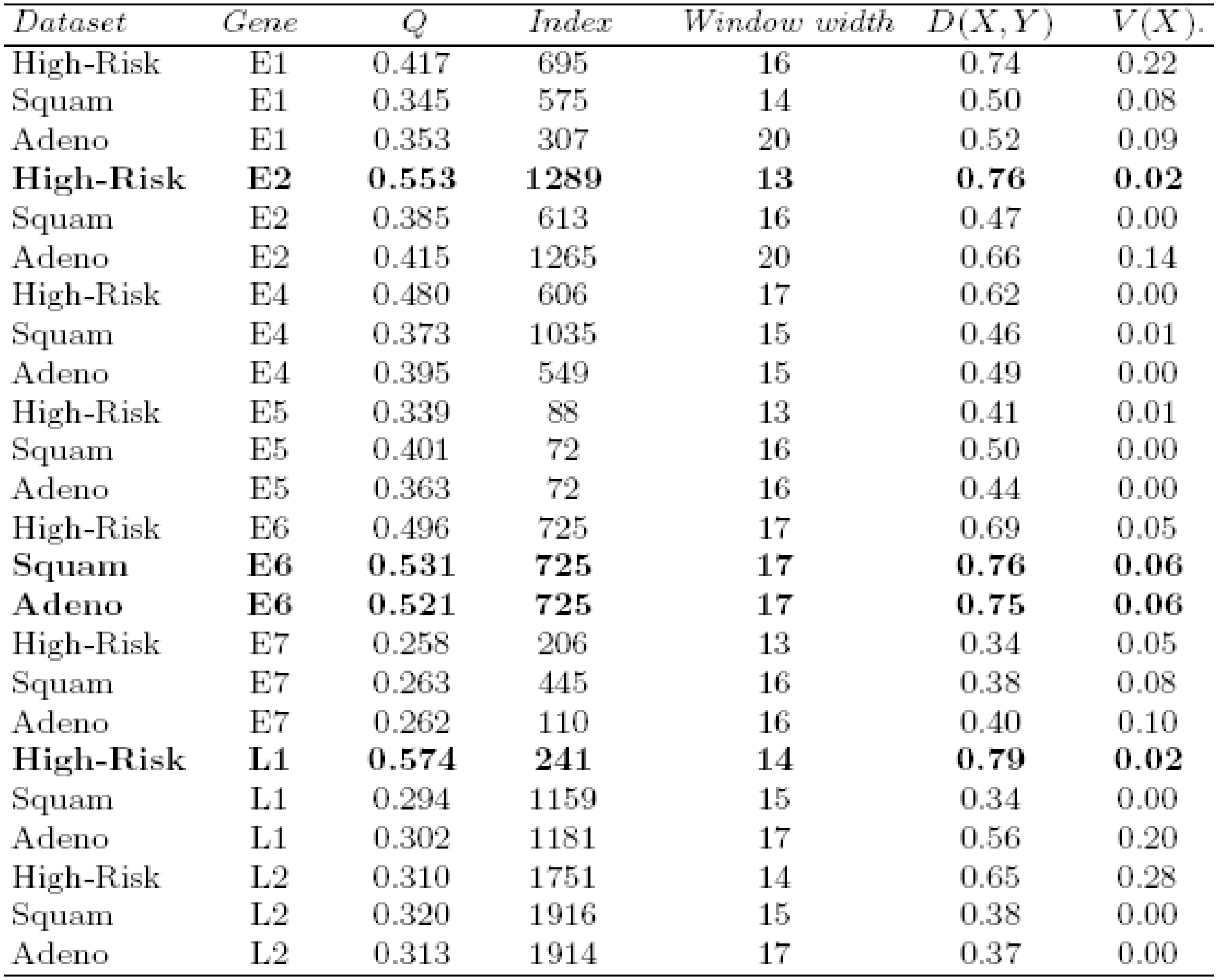
Require: MSA: Multiple sequence alignment (considered as a matrix),
MSA_L : Length of MSA,
X: Set of carcinogenic taxa,
N(X): Cardinality of the set X,
Y: Set of non-carcinogenic taxa,
N(Y): Cardinality of the set Y,
WIN_MIN: Minimum sliding window width,
WIN_MAX: Maximum sliding window width,
S: Sliding 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
        $\mathrm{MSA\_X} \leftarrow \mathrm{MSA}[X][idx..idx + win\_width]$
        $\mathrm{MSA\_Y} \leftarrow \mathrm{MSA}[Y][idx..idx + win\_width]$
        $V(X) \leftarrow D(X, Y) \leftarrow 0$
        for all distinct $i, j \in X$ do
            $V(X) \leftarrow V(X) + dist_h^2(\mathrm{MSA\_X}[i], \mathrm{MSA\_X}[j])$
        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
            $D(X,Y) \leftarrow D(X,Y) + dist_h^2(\mathrm{MSA\_X}[i], \mathrm{MSA\_Y}[j])$
        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
```

Time complexity $O(1 \times n^2)$

Selected High Scoring Regions - 1



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

Selected High Scoring Regions - 2

•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.

•As result of E6 high expression, the immune system is potentially evaded.

•Disruption of the gene E2 was observed in invasive carcinomas and in high-grade lesions.

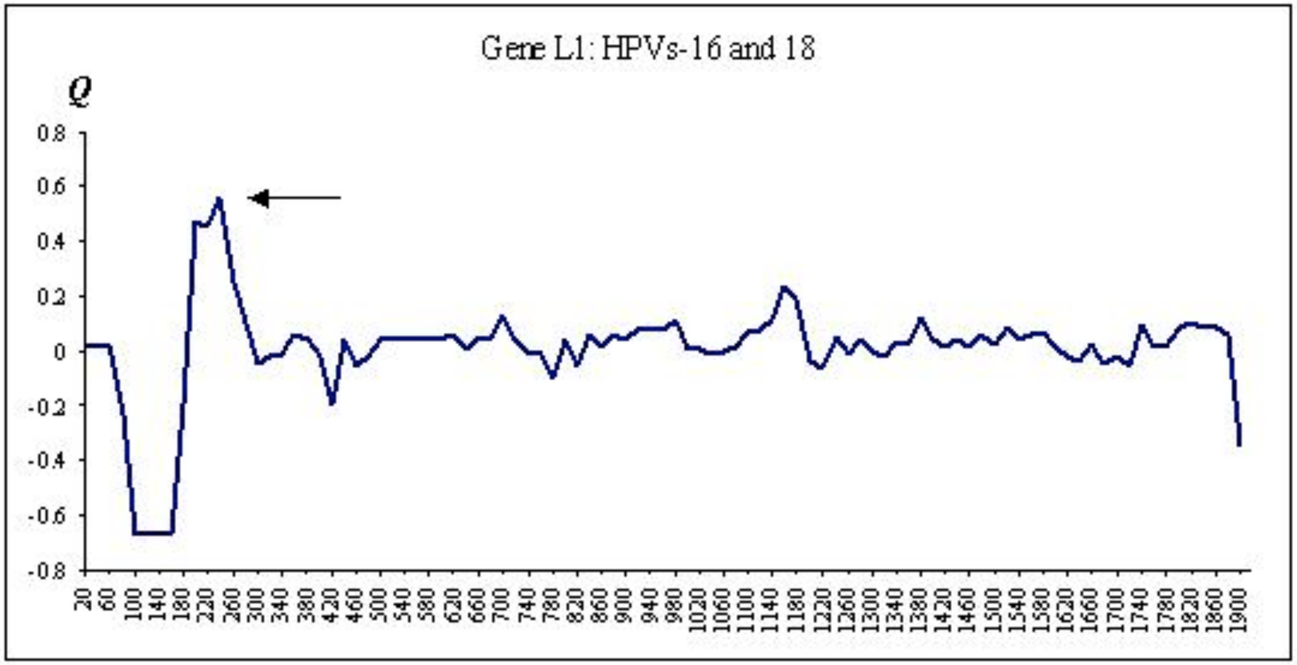
Wang, J.T., Ding, L., Gao, E.S., Cheng, Y.Y., Zhonghua Liu Xing Bing Xue Za Zhi 28(10) (2007) 968–971

Cordano P., Gillan, V., Bratlie, S., Bouvard, V., Banks, L., Tommasino, M., Cam
M.S., Virology 377(2) (2008) 408–418

Chan, P.K., Cheung, J.L., Cheung, T.H., Lo, K.W., Yim, S.F., Siu, S.S., Tang, J.W. , Journal of Infectious Diseases 196(6) (2007) 868–875

Graham, D.A., Herrington, C.S., Molecular Pathology 53 (2000) 201–206

Variation of Hit identification function Q



- Scanning with non-overlapping windows of size 20 nucleotides
- Contiguous high score regions in structural proteins could provide hints for drug design (e.g. for linear epitopes detection).

Combita, A.-L., Touz, A., Bousarghin, L., Christensen, N.D., Coursaget, P., Journal of Virology 76(13) (2002) 6480-6486

Our goals

1) Study the whole genome phylogeny of the Human Papilloma Viruses

2) Identify the evolutionary structures of the viral lineages

3) Define a new algorithm to identify regions that may be responsible for the carcinogenicity of the HPVs.

4) Identify relevant hit regions

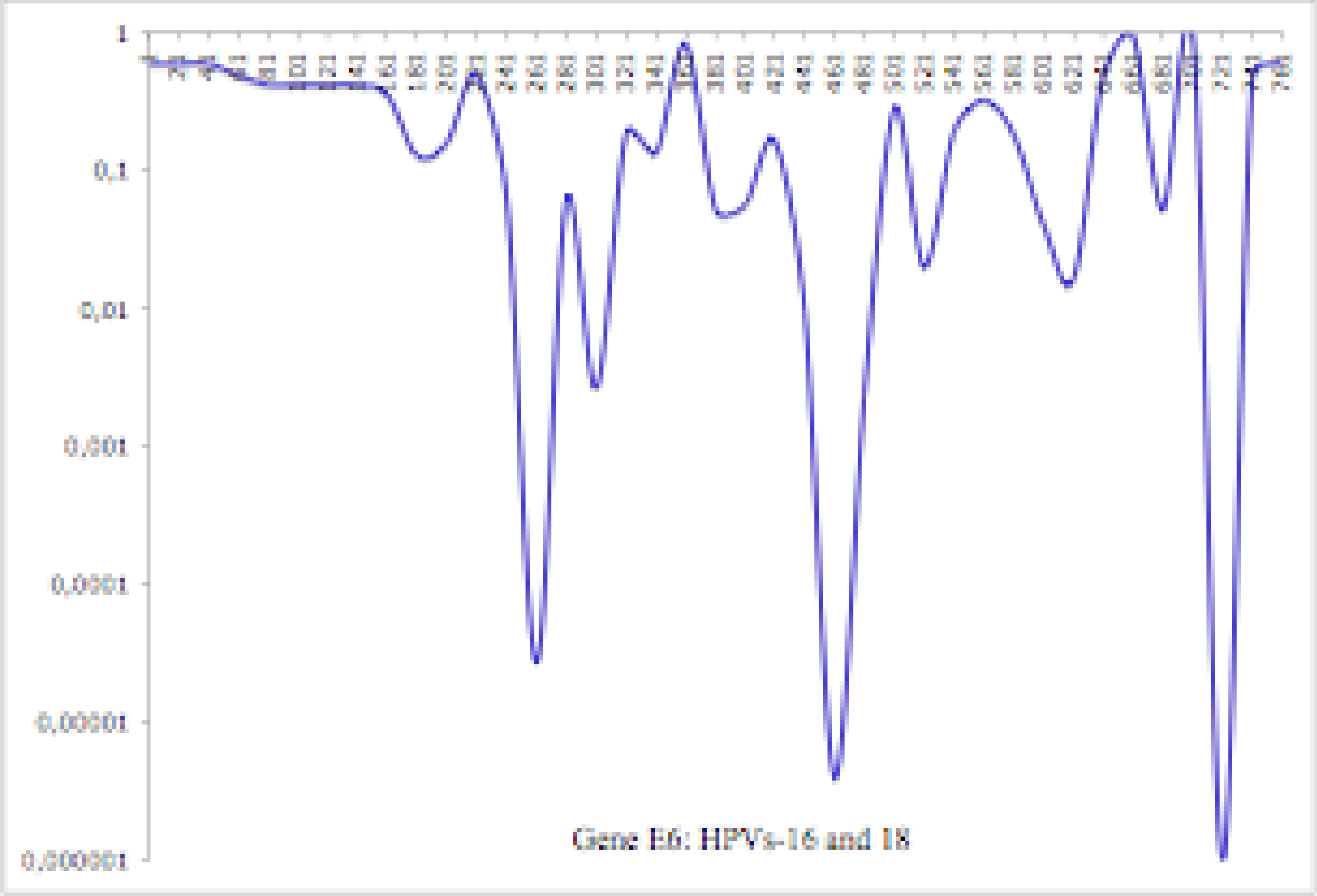
N.Munoz, et al., *N Engl J Med*, (2003), 384, pp.518 –527.

Computing of regions p-value

- Monte Carlo sampling was perfomed, to estimate the distribution of the Q values for a subset of W randomly chosen columns.
- million samples were generated and their Q values computed.
- The p-value of Q_i is then the fraction of samples that obtain a Q value larger or equal to Q_i .
- It is worth noting that one would expect most of the region with value of Q to have a p-value above 0.001.

N.Munoz, et al., *N Engl J Med*, (2003), 384, pp.518 –527.

Variation of Hit identification function Q



- 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.

What is next?

1) Filter hits according to proteomic alignments

2) Further study this genomic regions of E2, E6 and L1 in laboratory.

3) Identify the specific important evolutionnary events

4) Merge our results with existing methods (e.g. DLESS, signatures)

5) In-vitro analysis of the identified candidate

Acknowledgements

- UQAM
  - Alix Boc
  - Alpha Boubacar Diallo
  - Cherif Mballo
  - Mehdi Lay
  - Jin Xin Xie

- McGill
  - Mathieu Lavallée
  - Emmanuel Mongin
  - Michael Mayhew
  - Pablo Cingolani
  - Pierre-Étienne Jacques

CRSNG NSERC

Fonds de recherche
sur la nature
et les technologies
Québec

UQÀM

LACIM UQAM

McGill

McGill Centre for Bioinformatics
McGill University . Montreal . Quebec . Canada