OPEN ACCESS

2nd International Electronic Conference on Entropy and Its Applications
15-30 November 2015

www.sciforum.net/conference/ecea-2

Conference Proceedings Paper – Entropy

Entropy Based Computational Identification of Genomic Markers for Human Papillomavirus Detection and Genotyping

Gerlane Barros, Edilaine Araújo and Marcus Batista *

Department of Biology, Federal University of Sergipe. Av. Marechal Rondon, s/n, Jardim Rosa Elze, São Cristóvão, Sergipe, Brazil; gerlane15@live.com (G.B.); edilaine.doria@hotmail.com (E.A.); mvabatista@hotmail.com (M.B.)

* Author to whom correspondence should be addressed; E-Mail: mvabatista@hotmail.com (M.B.); Tel: +55-79-2105-6615

Published: 13 November 2015

Abstract: Papillomaviruses are circular double-stranded DNA viruses that specifically infect skin or mucocutaneous epithelium of mammals, reptiles and birds causing asymptomatic infections, benign and malignant lesions. The classification of papillomaviruses is based on the L1 gene sequence identity. However, several studies on Human papillomavirus (HPV) diversity make use of only 450 bp fragment in L1 in order to classify novel HPV types, subtypes, and variants. It has been observed that this fragment is not appropriated for detection and genotyping of HPV. The aim of this study was to develop and apply a novel computational tool based on entropy in order to identify phylogenetic informative genomic regions that could be used as markers for the detection and genotyping of HPV. To develop the method, a comparative analysis was performed to assess the genetic variability of L1 sequences from *Alphapapillomavirus*, *Betapapillomavirus* and *Gammapapillomavirus* genera. Shannon entropy was calculated. Informative sites were identified by using a cutoff of 1.0 bits of information. Phylogenetic trees were constructed based on those informative sites. The entropy measure presented itself as a good approach to identify phylogenetic informative genomic regions, which is important to correctly position novel HPV types in a phylogenetic tree, relevant to genotype these viruses.

Keywords: Human Papillomavirus; Entropy; Phylogenetic Analysis; Molecular Marker.

---

1. Introduction

Papillomaviruses (PVs) are circular double-stranded DNA viruses that specifically infect the skin or mucocutaneous epithelium of mammals, reptiles and birds causing asymptomatic infections, benign and malignant lesions [5]. The classification of these viruses is based on the sequence identity of L1 gene, which is the most conserved in the viral genome [1,19].

Among these viruses, five genera are related to human host, *Alphapapillomavirus*, *Betapapillomavirus*, *Gammapapillomavirus*, *Mupapillomavirus* and *Nupapillomavirus* [3]. Human papillomavirus (HPV) has a great oncogenic potential and so, it has major medical importance. HPV is the main etiological factor for cervical cancer development. Tumors associated with HPV in the cervical region are frequent and constitutes a serious public health problem, especially in developing countries [17]. More than 80% of cases of cervical cancer occur in developing countries, where the population has no easy access the preventive screening and appropriate treatment of the disease [9].

Invasive cervical cancer (ICC) is the second most common female cancer [8]. For development of ICC, the presence of carcinogenic HPV is necessary [12]. Alphapapillomaviruses are known for their association with lesions that can progress to cancer [13]. Among them, 71% of the ICC global burden is attributed to infections caused by HPV types 16 and 18 [16].

The diagnosis of HPV is based on the identification of cellular changes and the identification of viral DNA by molecular biology methods. Molecular methods based on polymerase chain reaction (PCR) have been widely used worldwide for the detection and genotyping of HPV using the primers MY09/11 and GP5+/6+ [7]. It is recommended to use a set of primers for best results HPV detection methods [4,10,14] and reducing uncertainties. However, the low sensitivity of MY09/11 primers may increase the number of false negatives [11]. In addition, GP5+/6+ primers are not appropriated for HPV genotyping because they do not have sufficient phylogenetic information so we can statistically confirm novel HPV types, subtypes and variants. In this context, novel genomic markers and primers that amplified these genomic regions are needed in order to improve molecular diagnostic methods for HPV detection and genotyping.

Therefore, this study makes use of a novel entropy-based method to identify HPV genomic regions that are phylogenetically more informative in order to improve HPV detection and genotyping. The method is based on Shannon's entropy and it was designed and prepared to identify conserved regions in HPV genomes. In this way, novel molecular diagnostic methods could be developed based on these novel genomic regions.

2. Results and Discussion

Phylogenetic inferences were made to the genera *Alpha*, *Beta* and *Gammapapillomavirus*, which presented respectively in 1602, 1587 and 1724 sites. We observed the number of conserved sites as 412 (25.72%), 550 (34.65%), and 206 (11.95%), respectively. The number of variables sites were 1157 (72.22%), 1034 (65.15%), 1477 (85.67%), respectively. The number of variable sites is high, which was expected since this is a very diverse group of viruses [1]. However, this gene is the most conserved region of the viral genome, thus it is used for the viral classification, which is important for diagnosis and genotyping [1,3,18]. Among the variable sites, 1090 (68.04%), 947 (59.67%), 1287 (74.65%) sites, respectively, were parsimony informative.

The region selected by the entropy measure shows a higher number of conserved sites for *Alpha*, *Beta* and *Gammapapillomavirus*, 191 (27.80%), 293 (37.42) and 99 (12.64%), respectively (Table 1). The low entropy (the region with entropy values under 1.0) region presents a higher percentage of conserved sites than the region comprehended by the primers MY09/11. The number of variable sites was also lower in the region specified by the entropy method. This result shows that these low-entropy regions could be used as a marker in order to develop novel primers to detect HPVs. In addition, the fact that the entropy approach is able to detect the most conserved regions in the HPV genome is relevant because we could accurately reconstruct a phylogenetic tree for HPV genotyping.

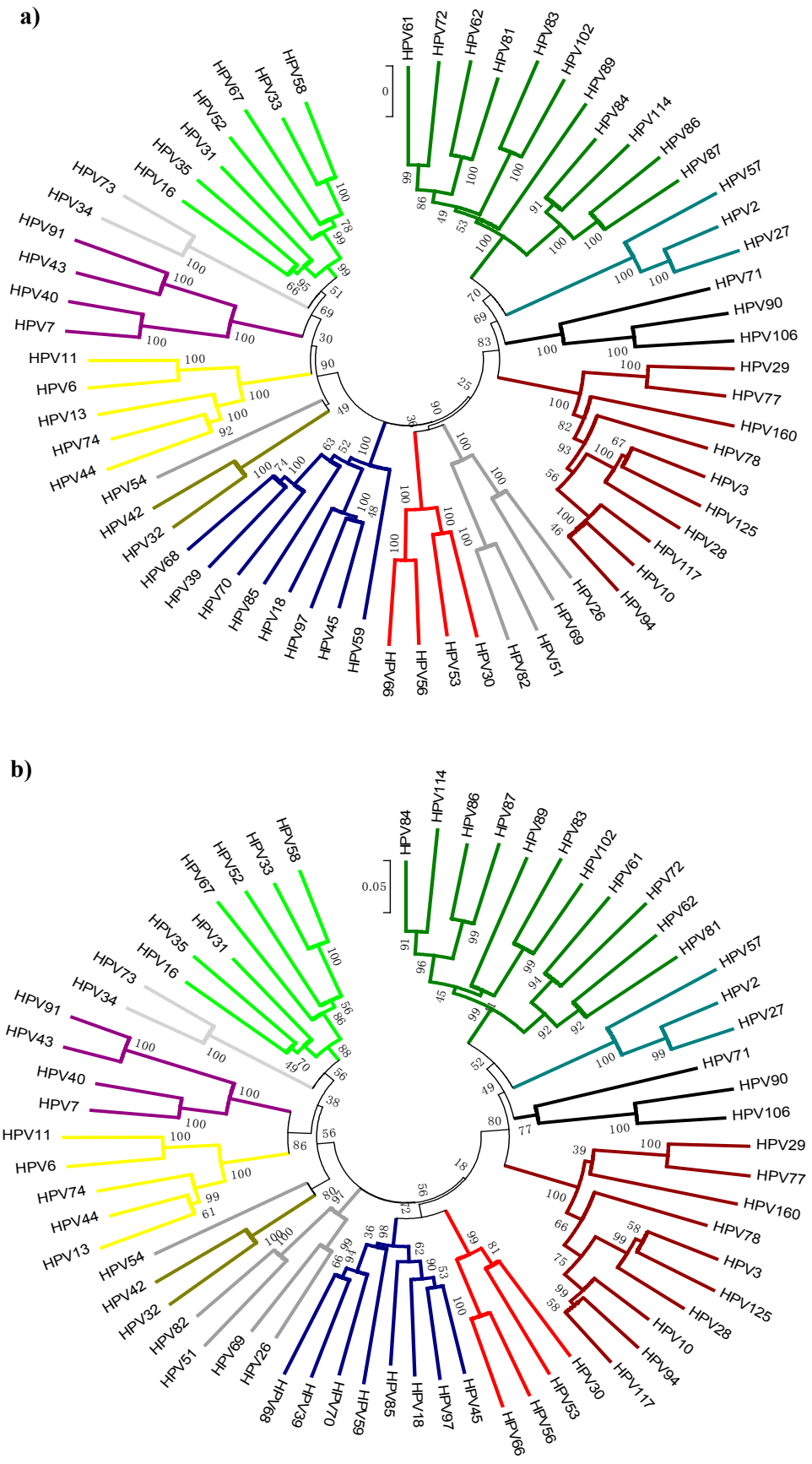
Table 1. Genetic diversity of diferent human papillomavirus genera.

HPV Genus	Length	Conserved sites	Variable sites	Parsimony informative sites
Alphapapillomavirus	687	191 (27,80%)	487 (70,89%)	468 (68,12%)
Betapapillomavirus	783	293 (37,42%)	487 (62,20%)	450 (57,47%)
Gammapapillomavirus	783	99 (12,64%)	667 (85,20%)	586 (74,84%)

Estimated conserved sites, variables and Parsim-informative of low-entropy region for the Alphapapillomavirus, Betapapillomavirus and Gammapapillomavirus genres.

Therefore, this method constitutes an effective tool in determining regions that have more probability to be homologous. In addition, the more conserved the HPV genomic region the more propitious to primer design. Another positive point in the entropy approach is to reduce the computational time spent on phylogenetic analysis [1].

HPV phylogeny was reconstructed using the Neighbor-Joining method, widely used in the reconstruction of PVs [18,20]. Our phylogenetic analyzes based on the entropy selected regions have shown that they are more informative than the one comprehended by the primers MY09/11 (Figure 1). In this study we adopted bootstrap greater than or equal to 70% to statistically evaluate the HPV types.


a)
0
HPV61
HPV72
HPV62
HPV81
HPV83
HPV102
HPV89
HPV84
HPV114
HPV86
HPV87
HPV57
HPV2
HPV27
HPV71
HPV90
HPV106
HPV29
HPV77
HPV160
HPV78
HPV3
HPV125
HPV28
HPV117
HPV10
HPV94
HPV26
HPV69
HPV51
HPV82
HPV30
HPV53
HPV56
HPV66
HPV59
HPV45
HPV97
HPV18
HPV85
HPV70
HPV39
HPV68
HPV32
HPV42
HPV54
HPV44
HPV74
HPV13
HPV6
HPV11
HPV7
HPV40
HPV43
HPV91
HPV34
HPV73
HPV16
HPV35
HPV31
HPV52
HPV67
HPV33
HPV58
b)
0.05
HPV84
HPV114
HPV86
HPV87
HPV89
HPV83
HPV102
HPV61
HPV72
HPV62
HPV81
HPV57
HPV2
HPV27
HPV71
HPV90
HPV106
HPV29
HPV77
HPV160
HPV78
HPV3
HPV125
HPV28
HPV10
HPV94
HPV117
HPV30
HPV53
HPV56
HPV66
HPV45
HPV97
HPV18
HPV85
HPV59
HPV70
HPV39
HPV68
HPV26
HPV69
HPV51
HPV82
HPV32
HPV42
HPV54
HPV44
HPV74
HPV13
HPV6
HPV11
HPV7
HPV40
HPV43
HPV91
HPV34
HPV73
HPV16
HPV35
HPV31
HPV52
HPV67
HPV33
HPV58

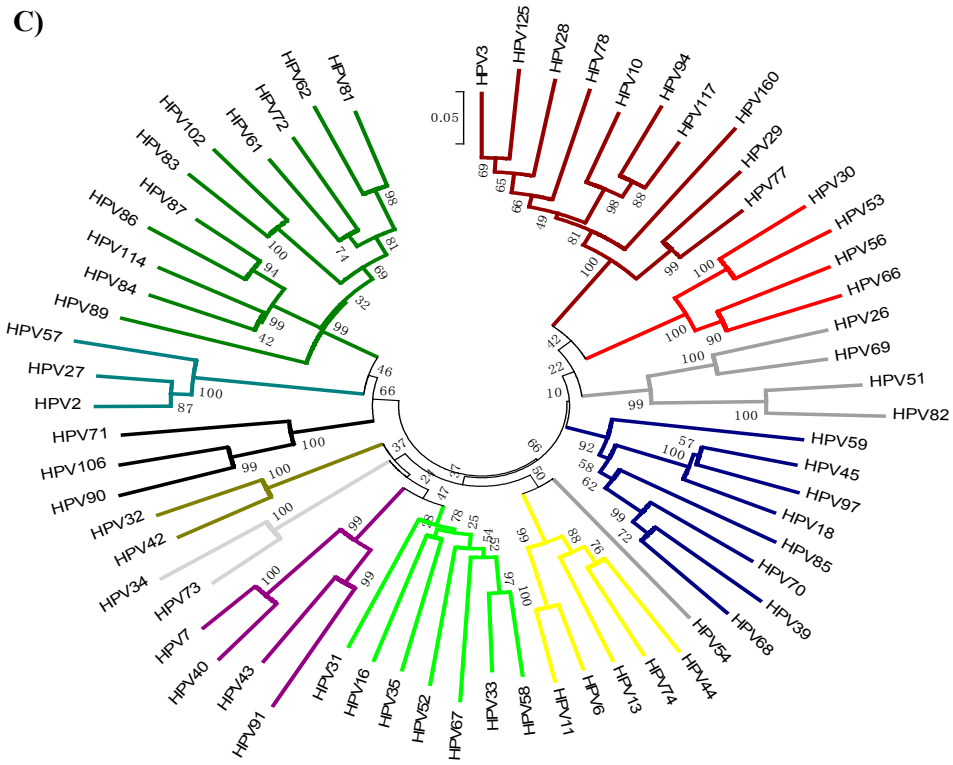


Figure 1. Neighbor-Joining phylogenetic tree of human papillomavirus L1 region based on amino acid sequences. **a)** Phylogenetic tree of *Alphapapillomavirus* using the complete L1 gene sequence. **b)** Phylogenetic tree of *Alphapapillomavirus* using the MY09/11 region. **c)** Phylogenetic tree of *Alphapapillomavirus* using the regions selected by the entropy approach.

The reconstructed trees based on entropy regions presented clusters similarly to the complete L1 gene and better statistical support than the MY09/11 region. The bootstrap values of internal nodes of the entropy region based tree of *Alphapapillomavirus* and *Betapapillomavirus* genera were above 70%, unlike the MY09/11 region, which presented internal nodes below this value. The internal nodes of the *Gammapapillomavirus* tree also showed better bootstrap values in the entropy based trees. However, not all groups presented bootstrap values above 70%, which was expected since the genetic variability in this group of viruses is very high.

A study carried out by Batista *et al.*, (2013) showed that the effectiveness of the use of entropy in selecting phylogenetic informative regions for proper reconstruction of bovine papillomavirus phylogeny [1–2]. In this study, we could show that this approach was also efficient and more accurate in order to select informative regions in HPV L1 gene, which is very important to identify novel markers to detect and genotype these viruses.

3. Experimental Section

3.1. Data collection and development of a local database

In order to develop a local database, 158 nucleotide sequences of the L1 gene of *Alphapapillomavirus*, *Betapapillomavirus* and *Gammapapillomavirus* were collected from the Papillomavirus Episteme (PAVE) database (http://pave.niaid.nih.gov). The nucleotide sequences were

stored in a local database along with DNA content information. These sequences were converted into amino acids prior to the alignment. Multiple Alignments were performed by using the MUSCLE algorithm [6], incorporated into the MEGA5 software [17]. The aligned amino acids were back translated into nucleotides. Sequence variability parameters were then estimated: total number of aligned sites, number of variable sites, and number of conserved sites.

3.2. Entropy-based approach to identify phylogenetic informative genomic regions

The entropy approach to identify phylogenetic informative genomic regions in HPV sequences is based on the method described by Batista [1]. The entropy measures were calculated to assess the complexity and variability of each nucleotide site. The entropy was calculated for each position using the Shannon entropy formula:

$$\mathrm{H}\,i = -(\sum_{j=}^{4} 1\ \mathrm{P}ij\ \log 2\mathrm{P}ij)$$

where H*i* corresponds to the entropy of each site *i*; *j* is equal to 1, 2, 3 and 4, corresponding to the A, C, G and T nucleotides, respectively; and P*ij* is the proportion of the nucleotide *j* in the site *i*. A window size equal to 100 units was used in order to reduce information noise. Regions with low entropy values were selected as more informative for use in a molecular diagnostic system. Informative sites were defined as being less than or equal to 1.0 bit of information.

3.3. Phylogenetic reconstruction of HPV L1 sequences

These informative sites were selected in order to reconstruct the phylogenetic relationships of HPV L1 sequences. The neighbor-joining method was used to reconstruct HPV trees in MEGA5 software [15], using Kimura 2-parameter evolutionary model.

Ten datasets were then created for comparison. The complete L1 gene was used for each genus (one dataset for *Alphapapillomavirus*, one for *Betapapillomavirus*, and one for *Gammapapillomavirus*). Another dataset was created with the complete L1 gene using all three genera together. Datasets based on the region of MY09/11 primers for each genus were also created. Finally, datasets based on the entropy selected regions were created for each genus.

4. Conclusions

The results showed that it was possible to identify regions in HPV genome that provide robust phylogenetic topologies, and good statistical support. Therefore, the entropy measure presented itself as a good approach to identify phylogenetic informative genomic regions, which is important to correctly position novel HPV types in a phylogenetic tree, relevant to genotype these viruses. This entropy based approach could be used to design degenerate primers that are able to amplify phylogenetic informative regions, increasing the sensitivity and specificity of the HPV diagnosis.

Acknowledgments

The authors would like to acknowledge the Brazilian Institutes CNPq, CAPES, and FAPITEC/SE for financial support.

Author Contributions

All authors contributed equally for experimental design, data analysis and manuscript writing.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. BATISTA, M. V. A.; FERREIRA, T. A. E.; FREITAS, A. C.; BALBINO, V. Q. An entropy-based approach for the identification of phylogenetically informative genomic regions of Papillomavirus. *Infect. Genet. Evol.*, **2011**, *11*, 2026–2033.
2. BATISTA, M. V. A.; FREITAS, A. C.; BALBINO, V. Q. Entropy-based approach for selecting informative regions in the L1 gene of bovine papillomavirus for phylogenetic inference and primer design. *Genet. Mol. Res*., **2013**, *12*, 400–407.
3. BERNARD, H. U.; BURCK, R. D.; CHEN, Z.; DOORSLAER, K. V.; HAUSEN, H. Z.; VILLIERS, E. M. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology*, **2010**, *401*, 70–79.
4. CAMARGO, M.; LEON, S. S.; SANCHEZ, R.; MUNOZ, M.; VEJA, E.; BELTRAN, M.; PEREZ-PRADOS, A.; PATARROYO, M. E.; PATARROYO, M. A. Detection by PCR of human papillomavirus in Colombia: Comparison of GP5+/6+ and MY09/11 primer sets. *J. Virol. Methods*, **2011**, *178*, 68–74.
5. CAMPO, M. S. Animal models of papillomavirus pathogenesis. *Virus Res*., **2002**, *89*, 249–261.
6. EDGAR, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **2004**, *32*, 1792–1797.
7. EVANDER, M.; EDLUND, K.; BODEN, E.; GUSTAFSSON, A.; JONSSON, M.; KARLSSON, R.; RYLANDER, E.; WADELL, G. Comparison of a One-Step and a Two-Step Polymerase Chain Reaction with Degenerate General Primers in a Population-Based Study of Human Papillomavirus Infection in Young Swedish Women. *J Clin Microbiol* **1992**, *30*, 987–992.
8. FERLAY, J.; SOERJOMATARAM, I.; ERVIK, M.; DIKSHIT, R.; ESER, S.; MATHERS, C.; REBELO, M.; PARKIN, D. M.; FORMAN, D.; BRAY, F. Cancer Incidence and Mortality Worldwide. *IARC*, 2012, 1.0, Available online: http://globocan.iarc.fr. Accessed on 10/07/2015.
9. HERBERT, J.; COFFIN, J. Reducing Patient Risk for Human Papillomavirus Infection and Cervical Cancer. *JAOA* **2008**,*108*, 65–70.
10. KARLSEN, F.; KALANTARI, M.; JENKINS, A.; PETTERSEN, E.; KRISTENSEN, G.; HOLM, R.; JOHANSSON, B.; HAGMAR, B. Use of Multiple PCR Primer Sets for Optimal Detection of Human Papillomavirus. *J Clin Microbiol* **1996**, *34*, 2095–2100.
11. MAGALHÃES, J. M.; MOYSES, N.; AFONSO, L. A.; OLIVEIRA, L. H. S.; CAVALCANTI, S. M. B. Comparação de dois pares de oligonucleotídeos utilizados na reação em cadeia da polimerase para detecção de papilomavírus humanos em esfregaços cervicais. *J bras Doenças Sex Transm*, 2008, 19(2), 5–10.

12. MONSONEGO, J.; BOSCH, F. X.; COURSAGET, P.; COX, J. T.; FRANCO, E.; FRAZER, I.; SANKARANARAYANAN, R.; SCHILLER, J.; SINGER, A.; WRIGHT, T.; KINNEY, W.; MEIJER, C.; LINDER, J. Cervical cancer control, priorities and new directions. *Int. J. Cancer*, **2004**, *108*, 329 –333.
13. POLJAK, M.; CUZICK, J.; KOCJAN, B. J.; IFTNER, T.; DILLNER, J.; ARBYN, M. Nucleic Acid Tests for the Detection of Alpha Human Papillomaviruses. *Vaccine*, **2012**, *30*, 100–106.
14. REMMERBACH, T.; BRINCKMANN, U. G.; HEMPRICH, A.; CHEKOL, M.; KUHNDEL, K.; LIEBERT, U. G. PCR detection of human papillomavirus of the mucosa: comparison between MY09/11 and GP5+/6+ primer sets. *Virology*, **2004**, *30*, 302–308.
15. SAITOU, N.; NEI, M. The neighbor-joining method: a new method to reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **1978**, *4*, 406–425.
16. SANJOSE, S.; QUINT, W. G. V.; ALEMANY, L.; GERAETS, D. T.; KLAUSTERMEIER, J. E.; LLOVERAS,B.; *et al.* Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study. *Lancet Oncol*, **2010**, *11*, 1048–56**.**
17. SERRANO, B.; ALEMANY, L.; RUIZ, P. A.; TOUSA, S.; LIMA, M. A.; BRUNI, L.; JAIN, A.; CLIFFORD, G. M.; QIAO, Y. L.; WEISS, T.; BOSCH, X. F.; SANJOSE, S. Potential impact of a 9-valent HPV vaccine in HPV-related cervical disease in 4 emerging countries (Brazil, Mexico, India and China). *Cancer Epidemiology*, **2014**, *38*, 748–756**.**
18. TAMURA, K.; PETERSON, D.; PETERSON, N.; STECHER, G.; NEI, M.; KUMAR, S. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol. Biol. Evol.*, **2011**, *28*, 2731–2739.
19. VILLIERS, E. M.; FAUQUET, C.; BROKER, T. R.; BERNARD, H. U.; HAUSEN, H. Z. Classification of papillomaviruses. *Virology*, **2004**, *324*, 17–27.
20. ZHU. W.; DONG, J.; SHIMIZU, E.; HATAMA, S.; KADOTA, K.; GOTO, Y.; HAG, T. Characterization of novel bovine papillomavirus type 12 (BPV-12) causing epithelial papilloma. *Arch Virol*, **2012**, *157*, 85–91.