GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Entropy-based approach for selecting informative regions in the L1 gene of bovine papillomavirus for phylogenetic inference and primer design

M.V.A. Batista, A.C. Freitas and V.Q. Balbino

Departamento de Genética, Centro de Ciências Biológicas,
Universidade Federal de Pernambuco, Recife, PE, Brasil

Corresponding author: M.V.A. Batista
E-mail: mvabatista@hotmail.com

Genet. Mol. Res. 12 (1): 400-407 (2013)
Received April 26, 2012
Accepted August 2, 2012
Published February 8, 2013
DOI http://dx.doi.org/10.4238/2013.February.8.4

ABSTRACT. Bovine papillomaviruses (BPVs) cause many benign and malignant lesions in cattle and other animals. Twelve BPV types have been identified so far, and several putative novel BPV types have been detected based on the analysis of L1 gene fragments, generated by FAP59/64 and MY11/09 primers. Phylogenetic trees are important in studies that describe novel BPV types. However, topological mistakes could be a problem in such studies. Therefore, we made use of entropy to find phylogenetic informative regions in the BPV L1 gene sequences from all 12 BPVs. Six data sets were created and phylogenetically compared to each other using neighbor-joining and maximum likelihood methods of phylogenetic tree reconstruction. We found two major regions in the L1 gene, using an entropy-based approach, which selects regions with low information complexity. More robust phylogenetic trees were obtained with these regions, when compared to the ones obtained with FAP59/64 and MY11/09 primers. More robust phylogenetic trees are important to accurately position novel BPV types, subtypes and variants. We conclude that an entropy-based approach is

a good methodology for selecting regions of the L1 gene of BPVs that could be used to design more specific and sensitive degenerate primers, for the development of improved diagnostic methods.

Key words: Bovine papillomavirus; L1 gene; Phylogenetic analysis; Entropy

INTRODUCTION

Papillomaviruses (PVs) form a diverse group of non-enveloped viruses with a circular double-stranded DNA that infects a wide variety of hosts. The genome is approximately 8 kb in size and contains around eight genes. The L1 gene encodes the major capsid protein and has been used to classify PVs into genera, species, types, subtypes, and variants (de Villiers et al., 2004; Bernard et al., 2010). Among PVs, bovine papillomaviruses (BPVs) have a major role in veterinary medicine. They cause benign and malignant lesions in cattle and are associated with equine, zebra, and buffalo lesions (Lörh et al., 2005; Silvestre et al., 2009; van Dyk et al., 2009; Bogaert et al., 2010; Somvanshi, 2011).

To date, 12 BPV types have been identified and classified into three genera (Deltapapillomavirus, Epsilonpapillomavirus, and Xipapillomavirus); one of them remains unassigned (Campo, 2006; Ogawa et al., 2007; Tomita et al., 2007; Hatama et al., 2008, 2011; Zhu et al., 2012). Furthermore, several putative novel BPV types and subtypes have been detected based on analyses of L1 gene fragments of approximately 450 bp (Antonsson and Hansson, 2002; Ogawa et al., 2004; Claus et al., 2008; Carvalho et al., 2012). These fragments have been generated using FAP59/64 or MY11/09 primers designed to detect human PVs (Manos et al., 1989; Forslund et al., 1999). The DNA sequence variability between human PVs and BPVs is considerable; thus, the sensitivity of these primers could be compromised.

Every study describing the detection or characterization of novel BPVs uses a phylogenetic tree to classify isolates into genera and prove statistically that they are novel types. However, the observed trees are not as robust as they should be, which is of central concern because topological bias could be inserted in the analysis, causing possible interpretation/classification errors.

A successful entropy-based approach has recently been described for the selection of phylogenetic informative genomic regions in PVs (Batista et al., 2011). The aim of this study was to make use of entropy to find new regions in the L1 gene of BPVs that are more suitable for phylogenetic inferences.

MATERIAL AND METHODS

The analysis was carried out using L1 gene sequences of the 12 BPVs characterized thus far. Sequences were retrieved from GenBank database and aligned using the Muscle algorithm incorporated in Molecular Evolutionary Genetics Analysis version 5 (Tamura et al., 2011). The GenBank accession Nos. are BPV1 (X02346), BPV2 (M20219), BPV3 (NC_004197), BPV4 (X05817), BPV5 (NC_004195), BPV6 (AJ620208), BPV7 (DQ217793), BPV8 (NC_009752), BPV9 (NC_010192), BPV10 (NC_010193), BPV11 (AB543507), and BPV12 (JF834523).

A total of six data sets were created for comparison. First, the complete L1 gene was used. Second, the phylogenetically most informative regions were selected using an entropy-

based approach described by Batista et al. (2011). A cutoff value of 1.0 was used, and every nucleotide site with an average entropy value under this cutoff was selected (total entropy). The third data set was the region defined by FAP59/64 primers. Fourth, the region defined by MY11/09 primers was used. The fifth data set was a 768-bp region obtained using the entropy-based approach (entropy region 1). The sixth data set was a 540-bp region also obtained using the entropy-based approach (entropy region 2).

For the phylogenetic analysis, the jModelTest 0.1.1 software (Posada, 2008) was used to select the model that best fit each data set. The models were selected under the Bayesian Information Criterion. The nucleotide substitution model selected for the complete L1 gene, total entropy, entropy region 1, and FAP59/64 region sequence alignments was GTR+G. For the MY11/09 region and entropy region 2, the substitution models were TPM3uf+G and TPM3uf+I+G, respectively.

The neighbor-joining method was used to reconstruct BPV trees in Molecular Evolutionary Genetics Analysis version 5 (Tamura et al., 2011). Maximum likelihood trees were created for each data set using the best fitting nucleotide substitution model in PhyML 3.0 (Guindon et al., 2010). Five substitution rate categories were used. The tree topology search was carried out with an algorithm developed with the best of the nearest-neighbor interchange and subtree pruning and regrafting methods. The quartet measures of Component 2.0 (Page, 1989) were used to compare the obtained topologies. The complete L1 gene tree was used as the template, and all other trees were compared to it. In addition, the confidence values of the nodes were compared.

RESULTS

The entropy-based approach uncovered five regions with low entropy values ($H \leq 1.0$), which are the phylogenetically most informative regions (Figure 1). Two regions with low entropy were selected to assess their phylogenetic potential. Compared with the FAP59/64 region, entropy region 1 was 278 bp longer. In addition, entropy region 2 was 70 bp longer than the MY11/09 region.

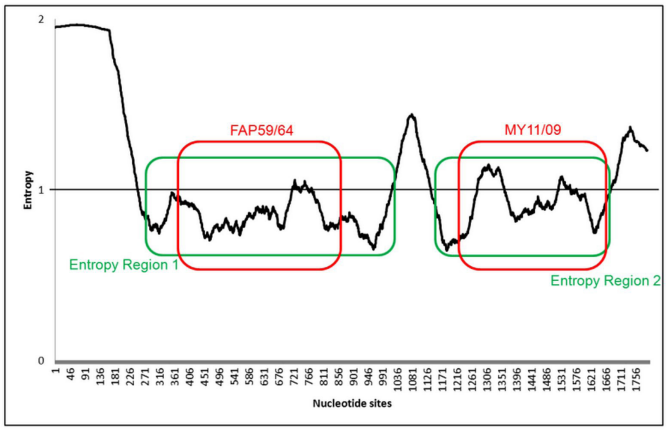

Figure 1. Entropy per site plot of the bovine papillomavirus L1 gene. Regions with entropy values equal to and below 1.0 bits were selected to the phylogenetic inference. Green boxes indicate two entropy regions analyzed in this study. Red boxes indicate regions comprised by FAP59/64 and MY11/09 primers.

Phylogenetic analysis using both methods for reconstructing trees showed that the trees using the complete L1 gene were more robust than the others (Figures 2 and 3). In general, clusters corresponding to genera could be identified but not without some changes inside these groups. However, the maximum likelihood tree of the complete L1 gene was more consistent, with higher confidence values. Although the trees constructed using the total entropy region were slightly less robust than those constructed with the complete L1 gene, their topology was very similar (see Figures 2 and 3). The quartet method showed that the smallest values of topological distance from the complete L1 gene for neighbor-joining and maximum likelihood trees were 0.032 and 0.051, respectively (Table 1).

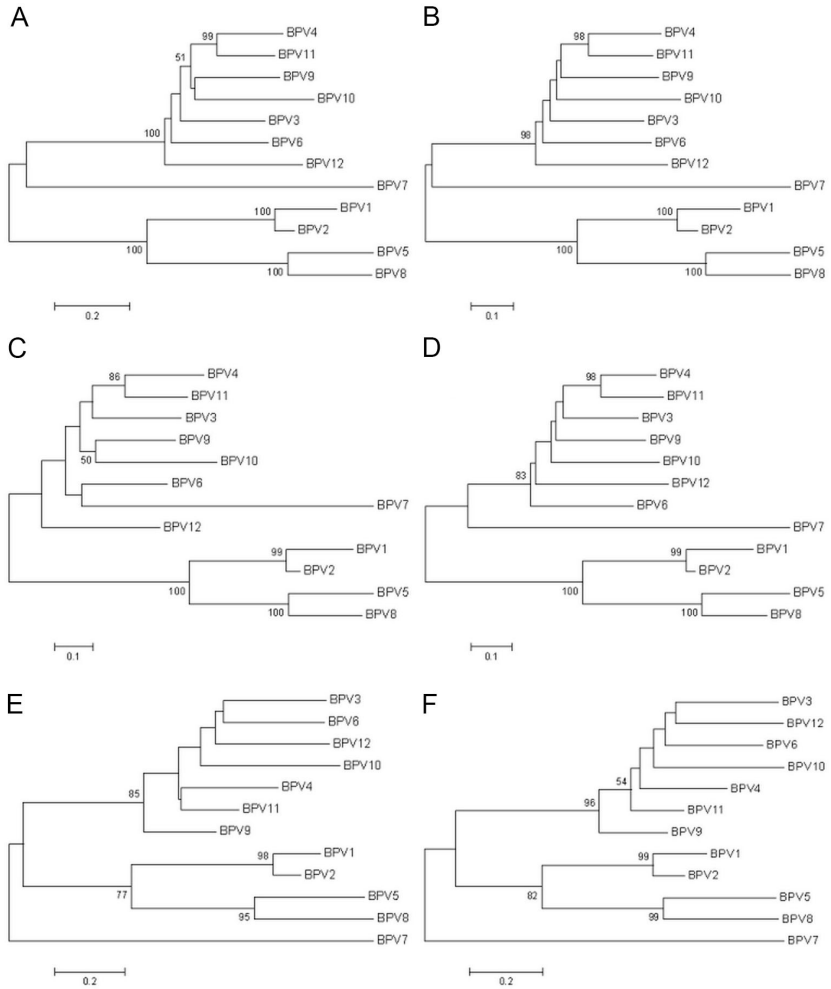

Figure 2. Phylogenetic trees of 12 bovine papillomaviruses inferred by the neighbor-joining method. Different data sets were used: **A.** the complete L1 gene; **B.** total entropy; **C.** FAP59/64 region; **D.** entropy region 1; **E.** MY11/09 region; **F.** entropy region 2. Bootstrap values under 50% are not shown. Tree branch length is in scale.

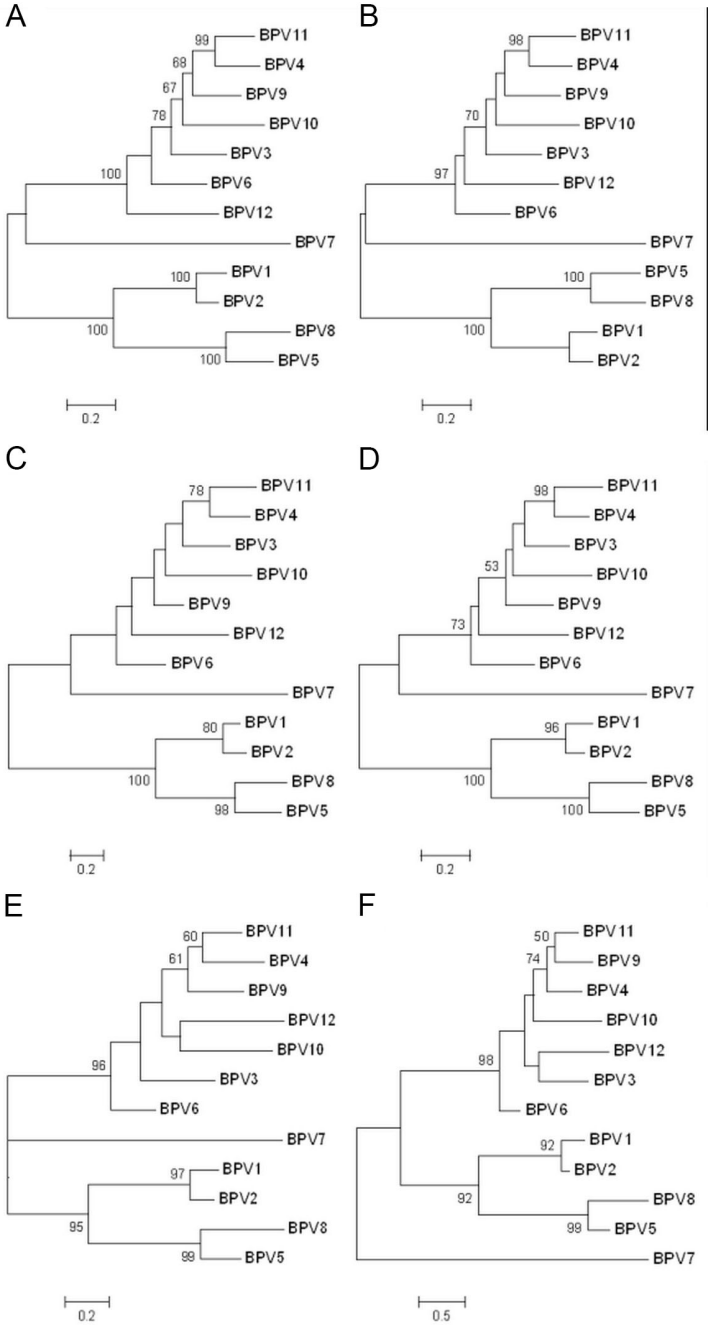

Figure 3. Phylogenetic trees of 12 bovine papillomaviruses inferred by the maximum likelihood method. Different data sets were used: **A.** the complete L1 gene; **B.** total entropy; **C.** FAP59/64 region; **D.** entropy region 1; **E.** MY11/09 region; **F.** entropy region 2. Bootstrap values under 50% are not shown. Tree branch length is in scale.

Table 1. Topological comparisons among the obtained trees.

Comparison		SD-NJ	SD-ML	s-NJ	d-NJ	s-ML	d-ML
Entropy	Complete L1	0.032	0.051	479	16	470	25
Entropy region 1	Complete L1	0.149	0.154	421	74	419	76
	Entropy	0.125	0.103	433	62	444	51
Entropy region 2	Complete L1	0.372	0.117	311	184	437	58
	Entropy	0.360	0.067	317	178	462	33
	Entropy region 1	0.354	0.170	320	175	411	84
FAP	Complete L1	0.156	0.154	418	77	419	76
	Entropy	0.188	0.103	402	93	444	51
	Entropy region 1	0.172	0.000	410	85	495	0
	Entropy region 2	0.455	0.170	270	225	411	84
MY	Complete L1	0.313	0.141	340	155	425	70
	Entropy	0.301	0.091	346	149	450	45
	Entropy region 1	0.315	0.184	339	156	404	91
	Entropy region 2	0.067	0.109	462	33	441	54
	FAP	0.400	0.184	297	198	404	91

Dissimilarity values: SD = symmetric difference; NJ = neighbor-joining tree; ML = maximum likelihood tree; s = resolved and same; d = resolved and different.

Comparison of the neighbor-joining trees from the FAP59/64 region and entropy region 1 showed that entropy region 1 presented a more robust topology, which was confirmed with the quartet method (see Figure 2 and Table 1). Neighbor-joining trees from the MY11/09 region and entropy region 2 were also compared and were similar (see Figure 2). This result was confirmed with the quartet measure that showed a topology distance of 0.067 (see Table 1). However, the tree for entropy region 2 showed bootstrap values slightly higher than those of the MY11/09 region.

We observed that the FAP59/64 region and entropy region 1 presented the same topology when the maximum likelihood trees were analyzed (see Figure 3). The confirmation was obtained with the quartet measure that showed no topology distance between them (see Table 1). However, the tree for entropy region 1 had higher confidence values, indicating a more robust phylogenetic tree. The tree for entropy region 2 had a topology more similar to that of the complete L1 gene than to that of the MY11/09 region (see Figure 3). The quartet measure showed that entropy region 2 had a topology distance value of 0.117 to the complete L1 gene tree, whereas the MY11/09 region had a value of 0.141 (see Table 1).

DISCUSSION

In this study we assessed and proposed novel regions in the L1 gene of BPVs to make phylogenetic inferences based on partial sequences. An entropy-based approach was used to select those regions, and they were identified as phylogenetically informative based on a previous study (Batista et al., 2011). The analysis supports the idea that new L1 gene regions should be taken into account in studies that aim to detect novel BPV types using degenerate primers.

Phylogenetic analysis based on those regions showed that they are more informative than the regions determined using FAP59/64 and MY11/09 primers. These primers are widely used in studies that describe the detection or characterization of novel BPV types and subtypes (Antonsson and Hanson, 2002; Ogawa et al., 2004; Claus et al., 2008; Hatama et al., 2008; Carvalho et al., 2012). Because those studies use a phylogenetic tree to classify isolates into a genus and statistically prove that they are indeed novel types, the use of phylogenetically more

informative regions in the L1 gene of BPVs is critical for increased accuracy.

The two regions assessed in this study (entropy regions 1 and 2) comprised the regions of the primers FAP59/64 and MY11/09. However, the results showed that the increase in length of these regions, as suggested by the entropy approach, improved the phylogenetic signal. In addition, the final size of the two entropy regions is appropriate for any polymerase chain reaction and sequencing reactions, which makes them suitable in a BPV detection system.

Robust phylogenetic trees were obtained from the low entropy regions of the BPV L1 gene. These trees had very small topological distances to the complete L1 gene trees. This result showed that the approach proposed by Batista et al. (2011) is suitable for the selection of phylogenetically informative regions of the BPV L1 gene. Although the complete L1 gene trees were more robust than the entropy-based ones, we obtained very similar topologies with fewer data using this approach. Another interesting discovery was that the trees obtained using the maximum likelihood method was more robust than those constructed with the neighbor-joining method. Even though probabilistic methods are known to be more accurate than distance-based methods, many BPV detection or characterization studies still use the neighbor-joining method to reconstruct BPV phylogeny (Antonsson and Hanson, 2002; Ogawa et al., 2004, 2007; Tomita et al., 2007; Claus et al., 2008; Hatama et al., 2008, 2011; Lunardi et al., 2010; Zhu et al., 2012).

The fact that entropy region 1 was 278 bp longer than the FAP59/64 region and entropy region 2 was 70 bp longer than the MY11/09 region explains the improved topologies obtained. This result is in accordance with the idea that increasing sequence length is a good way to improve the support, resolution, and accuracy of phylogenetic inference, as suggested by Wortley et al. (2005). Entropy region 1 had 174 more parsimony informative sites than FAP59/64 does, and entropy region 2 had 37 more parsimony informative sites than MY11/09 has. These results show that the entropy approach not only extended the regions but also increased the information that was available for phylogenetic inference. However, the contribution of those sites to the phylogenetic trees was more evident in the neighbor-joining trees.

The entropy-based approach used in this study selected phylogenetically more informative regions in the L1 gene of BPVs. Two entropy regions were analyzed and compared to regions of established degenerate primers. In general, the two entropy regions were associated with more robust phylogenetic trees, which are important for positioning novel BPV types, subtypes, and variants accurately. This issue is central to BPV detection and characterization studies. The results of this study point to a solid methodology for the selection of regions in the L1 gene of BPVs, which could be applied to the design of novel degenerate primers with greater specificity and sensitivity for detecting those viruses. This finding is significant for the development of improved diagnostic methods and, consequently, the establishment of a surveillance program to identify the incidence and distribution of BPVs.

ACKNOWLEDGMENTS

We thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (Brazil) for providing a research fellowship for the doctoral degree of M.V.A. Batista (Proc. #142216/2010-0).

REFERENCES

Antonsson A and Hansson BG (2002). Healthy skin of many animal species harbors papillomaviruses which are closely related to their human counterparts. *J. Virol.* 76: 12537-12542.

Batista MV, Ferreira TA, Freitas AC and Balbino VQ (2011). An entropy-based approach for the identification of phylogenetically informative genomic regions of Papillomavirus. *Infect. Genet. Evol.* 11: 2026-2033.

Bernard HU, Burk RD, Chen Z, van Doorslaer K, et al. (2010). Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology* 401: 70-79.

Bogaert L, Martens A, Kast WM, Van Marck E, et al. (2010). Bovine papillomavirus DNA can be detected in keratinocytes of equine sarcoid tumors. *Vet. Microbiol.* 146: 269-275.

Campo MS (2006). Bovine Papillomavirus: Old System, New Lessons? In: Papillomavirus Research: From Natural History to Vaccine and Beyond (Campo M, ed.). Caister Academic Press, Wymondham.

Carvalho CC, Batista MV, Silva MA, Balbino VQ, et al. (2012). Detection of bovine papillomavirus types, co-infection and a putative new BPV11 subtype in cattle. *Transbound. Emerg. Dis*. DOI: 10.1111/j.1865-1682.2011.01296.x.

Claus MP, Lunardi M, Alfieri AF, Ferracin LM, et al. (2008). Identification of unreported putative new bovine papillomavirus types in Brazilian cattle herds. *Vet. Microbiol.* 132: 396-401.

de Villiers EM, Fauquet C, Broker TR, Bernard HU, et al. (2004). Classification of papillomaviruses. *Virology* 324: 17-27.

Forslund O, Antonsson A, Nordin P, Stenquist B, et al. (1999). A broad range of human papillomavirus types detected with a general PCR method suitable for analysis of cutaneous tumours and normal skin. *J. Gen. Virol.* 80: 2437-2443.

Guindon S, Dufayard JF, Lefort V, Anisimova M, et al. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59: 307-321.

Hatama S, Nobumoto K and Kanno T (2008). Genomic and phylogenetic analysis of two novel bovine papillomaviruses, BPV-9 and BPV-10. *J. Gen. Virol.* 89: 158-163.

Hatama S, Ishihara R, Ueda Y, Kanno T, et al. (2011). Detection of a novel bovine papillomavirus type 11 (BPV-11) using xipapillomavirus consensus polymerase chain reaction primers. *Arch. Virol.* 156: 1281-1285.

Löhr CV, Juan-Sallés C, Rosas-Rosas A, Paras GA, et al. (2005). Sarcoids in captive zebras (*Equus burchellii*): association with bovine papillomavirus type 1 infection. *J. Zoo. Wildl. Med.* 36: 74-81.

Lunardi M, Claus MP, Alfieri AA, Fungaro MH, et al. (2010). Phylogenetic position of an uncharacterized Brazilian strain of bovine papillomavirus in the genus Xipapillomavirus based on sequencing of the L1 open reading frame. *Genet. Mol. Biol.* 33: 745-749.

Manos MM, Ting Y, Wright DK, Lewis AJ, et al. (1989). The use of polymerase chain reaction amplification for the detection of genital human papillomaviruses. *Cancer Cells* 7: 209-214.

Ogawa T, Tomita Y, Okada M, Shinozaki K, et al. (2004). Broad-spectrum detection of papillomaviruses in bovine teat papillomas and healthy teat skin. *J. Gen. Virol.* 85: 2191-2197.

Ogawa T, Tomita Y, Okada M and Shirasawa H (2007). Complete genome and phylogenetic position of bovine papillomavirus type 7. *J. Gen. Virol.* 88: 1934-1938.

Page RDM (1989). COMPONENT User's Manual (Release 1.5). University of Auckland, Auckland.

Posada D (2008). jModelTest: phylogenetic model averaging. *Mol. Biol. Evol.* 25: 1253-1256.

Silvestre O, Borzacchiello G, Nava D, Iovane G, et al. (2009). Bovine papillomavirus type 1 DNA and E5 oncoprotein expression in water buffalo fibropapillomas. *Vet. Pathol.* 46: 636-641.

Somvanshi R (2011). Papillomatosis in buffaloes: a less-known disease. *Transbound. Emerg. Dis.* 58: 327-332.

Tamura K, Peterson D, Peterson N, Stecher G, et al. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28: 2731-2739.

Tomita Y, Literak I, Ogawa T, Jin Z, et al. (2007). Complete genomes and phylogenetic positions of bovine papillomavirus type 8 and a variant type from a European bison. *Virus Genes* 35: 243-249.

van Dyk E, Oosthuizen MC, Bosman AM, Nel PJ, et al. (2009). Detection of bovine papillomavirus DNA in sarcoid-affected and healthy free-roaming zebra (*Equus zebra*) populations in South Africa. *J. Virol. Methods* 158: 141-151.

Wortley AH, Rudall PJ, Harris DJ and Scotland RW (2005). How much data are needed to resolve a difficult phylogeny? Case study in Lamiales. *Syst. Biol.* 54: 697-709.

Zhu W, Dong J, Shimizu E, Hatama S, et al. (2012). Characterization of novel bovine papillomavirus type 12 (BPV-12) causing epithelial papilloma. *Arch. Virol.* 157: 85-91.