Genomic characterization of the first insectivoran papillomavirus reveals an unusually long, second non-coding region and indicates a close relationship to Betapapillomavirus

Eric Schulz,1 Marc Gottschling,1,2 Ignacio G. Bravo,3 Ullrich Wittstatt,4 Eggert Stockfleth1 and Ingo Nindl1,5

1Clinic for Dermatology, Venereology and Allergology, Skin Cancer Center Charité (HTCC), Charité – Universitätsmedizin Berlin, Berlin, Germany
2Department of Biology, Systematic Botany, Ludwig-Maximilians-Universität München, Munich, Germany
3Experimental Molecular Evolution, Institute for Evolution and Biodiversity, Westfälische Wilhems Universität Münster, Münster, Germany
4Institute für Lebensmittel, Arzneimittel und Tierseuchen, Zentrum für Infektionsdiagnostik (ZID), Berlin, Germany
5DKFZ – Charité, Viral Skin Carcinogenesis, Viral Transformation Mechanisms, German Cancer Research Center (DKFZ), Heidelberg, Germany

Knowledge about biological diversity is the prerequisite to reliably reconstruct the evolution of pathogens such as papillomaviruses (PV). However, complete genomes of non-human PV have only been cloned and sequenced from 8 out of 18 orders within the Placentalia, although the host-specific variety of PV is considered much larger. We isolated and sequenced the complete genome of the first insectivoran PV type from hair follicle cells of the European hedgehog (Erinaceus europaeus), designated EHPV. We conducted phylogenetic analyses (maximum-likelihood criterion and Bayesian inference) with the genomic information of a systematically representative set of 67 PV types including EHPV. As inferred from amino acid sequence data of the separate genes E1, E2 and L1 as well as of the gene combination E6–E7–E1–E2–L1, EHPV clustered within the β-γ-π-ζ-PV supertaxon and constituted the closest relative of genus Betapapillomavirus infecting primates. Beside the typical organization of the PV genome, EHPV exhibited a 1172 bp, non-coding region between the E2 and the L2 open reading frames. This trait has been previously described for the only distantly related Lambdapapillomavirus, but a common evolutionary origin of both non-coding regions is unlikely. Our results underscore the modular organization of the PV genome and the complex natural history of PV.

INTRODUCTION

The diversity of papillomaviruses (PV) has not been sufficiently studied. Viruses from approximately 40 different host species have been isolated and (partly) sequenced, but it is assumed that all members of the Amniota potentially harbour specific PV (Van Ranst et al., 1995; de Villiers et al., 2004; Rector et al., 2007). The currently known hosts do not cover more than 8 of the 18 orders of the Placentalia, but PV from taxa such as Proboscidea and “Insectivora” have not been isolated or completely sequenced at this moment in time.

The dsDNA PV genome comprises approximately 8000 bp. It is structurally conserved and organized in up to eight well-defined open reading frames (ORF), including the four major genes E1, E2, L2 and L1 and additional, smaller genes. A non-coding region (NCR1) involved in regulation of gene expression is located upstream of the early genes. The genomic region between the genes E2 and L2 is an evolutionary hotspot. In genus Betapapillomavirus (β-PV, infecting primates) and other PV, this transitional region is short and non-coding, while it encodes the E5 oncogene(s) in genera Alphapapillomavirus (α-PV, infecting primates) and Deltapapillomavirus (δ-PV, infecting ungulates; Bravo &
In Lambda papillomavirus (λ-PV, infecting Carnivora), this region spans more than 1000 bp and does not appear to encode any protein (Terai & Burk, 2002; Rector et al., 2007). It has been therefore described as the second non-coding region (NCR2).

PV systematics traditionally relies on the sequence comparison of the L1 gene. More rigorous evaluation of PV evolution, applying computation of confidence values for internal nodes and appropriate outgroup choice, have become standard procedures for phylogenetic analyses during the past years (García-Vallvé et al., 2005; Narechania et al., 2005; Rector et al., 2007). Using these approaches, four PV supertaxa have been identified, namely alpha-omikron (α+0), beta-gamma-pi-xi (β+γ+π+ξ), delta-epsilon (δ+ε), and kappa-lambda-mu-nu-sigma (κ+λ+μ+ν+σ)-PV. The interpretation of the corresponding molecular trees suggest multiple evolutionary mechanisms driving PV diversification (Gottschling et al., 2007b). They include (i) co-divergence between the viruses and their hosts (Van Ranst et al., 1995; Rector et al., 2007), (ii) infection across species borders (‘zoonosis’: Myers et al., 1996; Rector et al., 2005; Gottschling et al., 2007a), (iii) establishment of new ecological niches by adaptive radiation (García-Vallvé et al., 2005; Jackson, 2005), and/or (iv) recombination (Narechania et al., 2005; Varsani et al., 2006).

Well-resolved host trees are the necessary prerequisite to test the hypothesis of co-phylogeny between PV and their hosts (Page et al., 1996). During the past years, molecular and morphological studies have tremendously improved the phylogenetic systematics of the Placentalia (Murphy et al., 2001; Reyes et al., 2004; Springer et al., 2004; Bininda-Emonds et al., 2007). Within Laurasiatheria, Erinaceus europaeus Linnaeus, 1758 (Erinaceidae) belongs to the Eulipotyphla that are a part of the former Mammalian order “Insectivora” (hence recognized polyphyletic). The Eulipotyphla are the sister group of the remaining laurasiatherian taxa such as Carnivora, Cetartiodactyla, Chiroptera, and Perissodactyla. If co-divergence with hosts alone drove PV evolution, a virus isolated from E. europaeus should be closely related to those laurasiatherian PV. This hypothesis would be rejected by significantly different host and PV topologies.

In this study, we present the complete genome of the first insectivoran PV isolated from Erinaceus europaeus. In order to avoid violating the principle of priority, we have designated the new type as EHPV (‘European hedgehog’), because EEPV is given to a PV type from the European elk (Moreno-López et al., 1981; Ahola et al., 1986). We describe here the genome organization of EHPV and determine its phylogenetic relationships within the family Papillomaviridae by using the maximum-likelihood (ML) criterion and Bayesian inference.

**METHODS**

**DNA isolation and sequencing of the complete viral genome.**

Approximately 30 hairs including the follicle cells were collected under sterile conditions from the facial region of a free ranging European hedgehog (Erinaceus europaeus; male adult, field sample, Raul-Wallenberg-Straße, 126 79 Berlin-Marzahn-Hellersdorf, Germany; Supplementary Table S1, available in JGV Online) and were stored at –20 °C. Extraction of genomic DNA was performed as previously described (Gottschling et al., 2008), and DNA was stored at –20 °C further until further analysis. From this sample, a 574 bp L1 sequence fragment, generated and verified in two independent PCR experiments using degenerated FA-primers (Forslund et al., 1999; Antonnsson & Hansson, 2002), has been published previously (GenBank accession number EF396272; Gottschling et al., 2008).

Rolling circle amplification (RCA) of the viral, episomal DNA was performed with a TempliPhi 500 Amplification kit (Amersham Biosciences), following an optimized protocol using additional nucleotides to obtain higher folds of amplification. Genomic DNA (1 μl) was added to 5 μl ‘sample buffer’, denatured at 95 °C for 3 min, and was subsequently stored on ice. The denatured DNA was mixed with 6.2 μl of a solution containing 5 μl ‘reaction buffer’, 0.2 μl enzyme mix, and 1 μl 4.5 mM additional dNTPs. Incubation was performed at 30 °C overnight, and the reaction was stopped at 65 °C for 10 min.

Synthetic hexamers (2 nM) containing the EcoRI cleavage site were added to replenish single stranded regions of the concatemeric RCA products. They were digested using 10 units of EcoRI (New England Biolabs). Restriction fragments were separated by electrophoresis on a 0.8 % agarose gel, and a band of approximately 8000 bp was excised. DNA was purified with a QIAquick Gel Extraction kit (Qiagen), following the manufacturer’s instructions.

Well-resolved DNA was cloned into pBK-CMV plasmid (Stratagene) at the EcoRI restriction site using T4 Ligase (New England Biolabs). After ligation, One Shot Top 10 competent E. coli (Invitrogen) were transformed with the construct, and plated on kanamycin-containing selective agar plates. Five white colonies were picked and plasmid DNA was isolated using a QIAquick Spin Miniprep kit (Qiagen) and analysed after digestion with EcoRI. Flanking regions of DNA fragments with a size of approximately 8000 bp were sequenced using standard M13 primers. A bacterial clone harbouring the complete PV genome was transferred to 300 ml Luria–Bertani medium containing kanamycin (50 μg ml⁻¹) and incubated at 250 r.p.m. and 37 °C overnight. Two stock cultures containing 30 % glycerine as well as aliquots of the plasmid max prep were stored at –80 °C. Plasmid max prep were generated with a Qiagen Plasmid Maxi kit. The insert containing the complete genome of the novel PV was sequenced by primer walking (GENterprise).

**Protein prediction and codon usage.** Open reading frames (ORF) were predicted using ‘ORF Finder’ (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). They were confirmed by the manual alignment of nucleotide (nt) and amino acid (aa) sequences to homologous regions of most similar PV types in ‘Se-Al’ (v2.0a72; http://tree.bio.ed.ac.uk/software/seal/). Primary sequence analysis of the predicted proteins was performed with ProtParam (Gasteiger et al., 2005; wwww.expasy.ch/tools/protparam.html), Prosite (Hulo et al., 2008; http://www.expasy.ch/prosite.html), and PSORT II (Nakai & Horton, 1999; http://psort.im.s.u-tokyo.ac.jp/form2.html). The different codon usage of E4 with respect to the remaining ORF of EHPV (see below) was the basis to identify the putative start of the E4 gene, as previously described (Puigbó et al., 2008a).

Transcription factor binding sites (TFBS) were predicted with Match (http://www.gene-regulation.com/pub/programs.html#match ), using a library of mononucleotide weight matrices from TRANSFAC 9.3 (Matys et al., 2003). A positive match was considered true positive, if the score value was higher or equal to the core similarity cut-off. We chose a high cut-off value to minimize the number of false-positive
and false-negative results. The presence of the predicted TFBS in the NCR2 was confirmed with Cister (http://zlab.bu.edu/~mfrith/cister.shtml), using posterior decoding and based on a hidden Markov model (Frith et al., 2001). Only elements with a posterior probability higher than 0.8 were considered positive, and only results calculated by both algorithms were considered true positives.

The Codon Adaptation Index (CAI) is a measure of the similarity of synonymous codon usage between a gene and a reference set. An interval of expected CAI values were calculated using the E-CAI server, accounting for biases in nucleotide and/or amino acid composition (Puigbo et al., 2008b; http://genomes.urv.cat/CAIcalc/E-CAI/). We aimed to discern whether the differences in CAI between viruses and hosts were statistically significant, or whether they were merely artefacts arising from internal biases in the G+C content and/or aa composition of query sequences. Codon usage for E. europaeus was downloaded from the codon usage database (http://www.kazusa.or.jp/codon/).

Sequence alignment and phylogenetic analyses. The taxon sample for phylogenetic analyses covered the currently known diversity of PV and comprised 67 complete PV sequences (including 20 systematically representative human PV types; see Supplementary Table S1). After exclusion of the highly divergent E4 gene region, an aa alignment (comprising the genes E6–E7–E1–E2–L2–L1) was constituted using MAFFT (v6.523; Katoh et al., 2005; http://align.bmr.kyushu-u.ac.jp/mafft/software/), with the default settings. The final matrix is available at http://htcc.pt-dlr.de/dateien/SchulzEHPV.nex. A sequence identity matrix of the L1 was generated by BioEdit (v7.0.0; Hall, 1999). Phylogenetic analyses were performed for each of the four large genes separately, and a multi-gene analysis was conducted after exclusion of the L2 gene because it perturbs phylogenetic reconstructions (Gottschling et al., 2007b). All computations were run on a Linux cluster at the North German High Performance Computer.

ML-based phylogenetic analyses were conducted using the parallel Message Passing Interface (MPI) version of RAxML (v7.0.3; Stamatakis, 2006; http://icwww.epfl.ch/~stamatak/index-Dateien/Page443.htm) applying the rtREV matrix (Dimmic et al., 2002), as previously described (Gottschling et al., 2007b). We executed 10-tree searches from distinct random stepwise addition sequence maximum-parsimony (MP) starting trees. Thereafter, we executed 1000 non-parametric bootstraps with ‘RAxML’ under the partition data mode, and the bootstrap support values were drawn on the best-scoring ML-tree.

Bayesian phylogenetic analyses were performed with BEAST (v1.4.7; Drummond & Rambaut, 2007; http://evolve.zoo.ox.ac.uk/beast/). For

![Fig. 1. A 1172 bp non-coding region between the genes E2 and L2 of EHPV. Organization of the EHPV genome and distribution of potential cis regulative elements (see legend). Gene lengths are to scale and are indicated by boxes, including information on start and end nucleotide positions. BS, Binding site.](image-url)
### Systematic position of EHPV within the family Papillomaviridae

The L1 sequence of EHPV was most similar to that of McPV-2 (63.6 % similarity), as inferred from a sequence similarity matrix (available as Supplementary Data, available in JGV Online). The corresponding region in the complete genome was identical with an L1 sequence fragment (GenBank accession number EF396272), generated by using the primer pair FAP59 and FAP64. The alignment for the multi-gene phylogenetic analysis (i.e. excluding the genes E4 and L2) was 2262 aa positions in length. Of these sites, 1574 (70 %) were parsimony-informative (23.5 per terminal taxon). Details of the different partitions are given in Table 1.

Fig. 2 shows the best-scoring ML-tree (–ln 5150 954.879), with the statistical support values for the two phylogenetic approaches used. Mammalian PV were monophyletic, irrespectively whether the data were analysed under the likelihood criterion (bootstrap support value from ML analysis: 100 LBS) or the Bayesian approach (Bayesian posterior probability: 1.00 BPP). They segregated into the four monophyletic supertaxa α-(50 LBS), β+γ+µ+σ-PV (99 LBS, 1.00 BPP). With in the β+γ+µ+ξ-PV, EHPV was the sister of genus Betapapillomavirus in the multi-gene alignment (87 LBS, 1.00 BPP) as well as in separate gene analyses of E1 (67 LBS, 1.00 BPP), E2 (0.99 BPP), and L1 (63 LBS, 1.00 BPP, see Supplementary Figures S1–S4, available in JGV Online). In the L2 gene analysis, EHPV was the closest relative of CpPV-2 (58 LBS), and together they constituted the sister group of β-PV + MnPV-1 + RaPV.

### Table 1. Primary sequence analysis of the EHPV genes and optimal model parameters of the different alignment partitions

<table>
<thead>
<tr>
<th>Gene(s)</th>
<th>All</th>
<th>E6</th>
<th>E7</th>
<th>E1</th>
<th>E2</th>
<th>E4</th>
<th>L2</th>
<th>L1</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of nucleotides (bp)</td>
<td>8256</td>
<td>414</td>
<td>252</td>
<td>1794</td>
<td>1161</td>
<td>153</td>
<td>1518</td>
<td>1548</td>
</tr>
<tr>
<td>No. of amino acids</td>
<td>–</td>
<td>138</td>
<td>84</td>
<td>597</td>
<td>386</td>
<td>116</td>
<td>505</td>
<td>515</td>
</tr>
<tr>
<td>Molecular mass (kDa)</td>
<td>15.580</td>
<td>9.228</td>
<td>6.7584</td>
<td>43.615</td>
<td>13.459</td>
<td>54.473</td>
<td>58.685</td>
<td></td>
</tr>
<tr>
<td>Isoelectric point</td>
<td>–</td>
<td>7.92</td>
<td>6.27</td>
<td>5.27</td>
<td>9.43</td>
<td>5.47</td>
<td>4.96</td>
<td>8.54</td>
</tr>
<tr>
<td>Alignment length (aa positions)</td>
<td>3140</td>
<td>313</td>
<td>201</td>
<td>832</td>
<td>310 (excl. E4)</td>
<td>Excluded</td>
<td>878</td>
<td>606</td>
</tr>
<tr>
<td>No. of parsimony-informative positions (percentage; per terminal taxon)</td>
<td>2160 (68 %; 32.2)</td>
<td>149 (48 %; 2.2)</td>
<td>109 (54 %; 1.6)</td>
<td>613 (74 %; 9.1)</td>
<td>262 (85 %; 3.9)</td>
<td>Excluded</td>
<td>586</td>
<td>441</td>
</tr>
<tr>
<td>Alpha (RAxML)</td>
<td>–</td>
<td>1.4393</td>
<td>1.7922</td>
<td>1.2560</td>
<td>1.1024</td>
<td>Excluded</td>
<td>Excluded</td>
<td>1.0498</td>
</tr>
<tr>
<td>Invar (RAxML)</td>
<td>–</td>
<td>0.0423</td>
<td>0.0261</td>
<td>0.0503</td>
<td>0.0116</td>
<td>Excluded</td>
<td>Excluded</td>
<td>0.0771</td>
</tr>
</tbody>
</table>

### Table 2. Open reading frames in the EHPV genome and codon adaptation index (CAI) either to the codon usage of the virus (EHPV) or to the host (Erinaceus europaeus)

The expected CAI average was the average CAI value for 500 random sequences of the same length, same G+C content, and same amino acid composition as the corresponding viral sequence. The normalized CAI value was the quotient of CAI value and expected CAI average. The 95 % confidence was the upper one-side 95 % tolerance interval, representing the upper limit or estimated maximum value of the CAI of a random sequence. G+C content is given for the full-length gene as well as for each of the three possible coding positions. Reference values for the host genome are given in parentheses in the column head. The E4 gene is better adapted to the codon usage of the host than any other of the EHPV genes.

<table>
<thead>
<tr>
<th>Gene</th>
<th>CAI to EHPV codon usage (expected average)</th>
<th>CAI to E. europaeus codon usage (expected average)</th>
<th>Normalized expected CAI to E. europaeus codon usage (95 % confidence)</th>
<th>%G+C</th>
<th>%G+C(I)</th>
<th>%G+C(2)</th>
<th>%G+C(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E6</td>
<td>0.871 (0.808)</td>
<td>0.666 (0.643)</td>
<td>1.03 (1–1.10)</td>
<td>39.6</td>
<td>47.1</td>
<td>39.1</td>
<td>32.6</td>
</tr>
<tr>
<td>E7</td>
<td>0.773 (0.741)</td>
<td>0.647 (0.661)</td>
<td>0.97 (1–1.09)</td>
<td>47.9</td>
<td>58.4</td>
<td>40.4</td>
<td>44.9</td>
</tr>
<tr>
<td>E1</td>
<td>0.844 (0.775)</td>
<td>0.648 (0.629)</td>
<td>1.03 (1–1.10)</td>
<td>39.4</td>
<td>47.5</td>
<td>38.6</td>
<td>32.1</td>
</tr>
<tr>
<td>E2</td>
<td>0.731 (0.686)</td>
<td>0.678 (0.677)</td>
<td>1.00 (1–1.09)</td>
<td>50.9</td>
<td>53.0</td>
<td>51.4</td>
<td>48.3</td>
</tr>
<tr>
<td>E4</td>
<td>0.668 (0.605)</td>
<td>0.795 (0.706)</td>
<td>1.12 (1–1.09)</td>
<td>63.4</td>
<td>74.5</td>
<td>41.2</td>
<td>74.5</td>
</tr>
<tr>
<td>L2</td>
<td>0.793 (0.728)</td>
<td>0.661 (0.616)</td>
<td>1.07 (1–1.10)</td>
<td>44.3</td>
<td>54.9</td>
<td>44.3</td>
<td>33.8</td>
</tr>
<tr>
<td>L1</td>
<td>0.797 (0.742)</td>
<td>0.653 (0.624)</td>
<td>1.04 (1–1.08)</td>
<td>40.9</td>
<td>48.4</td>
<td>39.9</td>
<td>34.3</td>
</tr>
</tbody>
</table>
**DISCUSSION**

Knowledge about PV diversity is still sparse. In this study, we present the complete genome of the first PV type from the "Insectivora", EHPV, and have investigated its phylogenetic relationships within the family *Papillomaviridae*. We render the preliminary systematic position of this PV type more precisely than it has been possible in a previous L1 sequence fragment analysis (Gottschling et al., 2008). Despite the limited knowledge about PV diversity, the viruses have been generally regarded as co-diverging with their corresponding hosts. (Chan et al., 1995; Van Ranst et al., 1995; Bernard et al., 2001; Springer et al., 2004; Bininda-Emonds et al., 2007).

Two possible interpretations are conceivable for the systematic position of EHPV: (i) either the close relationship between EHPV and β-PV results from alternative evolutionary processes such as interspecies transmission, as it has been reconstructed for other regions of the PV tree (Gottschling et al., 2007a, b); or (ii) numerous PV lineages from various mammalian host taxa have not been identified so far (due to the limited investigation of non-human hosts) that would show a co-phylogenetic structure in a PV subtree if they were known. However, even the latter interpretation would lead to the rejection of the co-phylogenetic hypothesis in general because it necessarily postulates the origin of several (subsequently parallel and co-diverging, 'paralogous'; Jackson, 2005) lineages by another evolutionary mechanism than co-phylogeny.

The main core of the genome, comprising the four major genes E1, E2, L2 and L1, is highly conserved among the family *Papillomaviridae*. PV appear to have a modular organization of the genome, and some clades are characterized by apomorphic traits such as deletions or insertions of several smaller genes. One such trait is the presence of a long NCR2 between the genes E2 and L2, as described for the λ-PV infecting carnivores (Terai & Burk, 2002; Rector et al., 2007). EHPV infecting an insectivoran species similarly exhibits a long NCR2 between E2 and L2, but homology with the NCR2 from λ-PV is unlikely. This assumption is reinforced by the distant relationships between λ-PV and EHPV, as inferred from the phylogenetic analysis.

Some of the predicted TFBS present in the NCR1 of EHPV, such as E2, AP-1, NF-Y, NF-1, GATA, USF or the CCAAT-box as promoter element, have been previously reported from other PV types and may constitute the core of the regulation repertoire of PV gene expression (García-Vallvé et al., 2006). Additionally, other predicted TFBS might be present in the NCR2, but their physiological importance remains to be investigated experimentally.

All ORFs in EHPV are poorly adapted to the codon usage of the host, *E. europaeus*, as is generally the case for PV (Bravo & Müller, 2005). Since PV do not encode any molecular machinery involved in translation, systematic deviation from the codon usage of the host may lead to inefficient translation, decreased amount of translated protein, or decreased amount of properly folded protein (Bravo & Müller, 2005; Drummond & Wilke, 2008). The only exception to this de-adaptation is E4 (Zhao et al., 2003), usually the most highly expressed gene during PV infection (Doorman, 2005). The shift in codon usage in E4, nested within the E2 gene, can in fact be used to predict the E1^E4 annotation, despite the absence of a well-defined splice acceptor site (Puigbó et al., 2008a). However, an explanation integrating the evolution systematic codon usage bias and translation in PV is still wanting.

In conclusion, EHPV belongs to the β + γ + π + ζ-PV super-taxon, which thus comprises viruses infecting Eulipotyphla, Primates, Carnivora, Rodentia and "Artiodactyla" (the latter is a paraphyletic mammalian taxon). The new PV type constitutes the sister group of β-PV infecting primates. The intermingled phylogenetic relationships within the β + γ + π + ζ-PV supertaxon with respect to their hosts, and particularly the position of EHPV therein, underline the complex evolutionary history of the family *Papillomaviridae*.

**ACKNOWLEDGEMENTS**

We thank B. Kallies and T. Steinke (Konrad-Zuse-Zentrum für Informationstechnik, Berlin) for technical guidance during the...
establishment of our HPC access. I.G.B. was supported by the Volkswagen Foundation.

REFERENCES


