Assessing Host-Virus Codivergence for Close Relatives of Merkel Cell Polyomavirus Infecting African Great Apes

http://researchonline.ljmu.ac.uk/9246/

Article
For more information please contact researchonline@ljmu.ac.uk

http://researchonline.ljmu.ac.uk/
Assessing Host-Virus Codivergence for Close Relatives of Merkel Cell Polyomavirus Infecting African Great Apes


Epidemiology of Highly Pathogenic Microorganisms, Robert Koch Institute, Berlin, Germany; Department of Primatology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany; Institut de Recherches en Ecologie Tropicale, Libreville, Gabon; FG12, Measles, Mumps, Rubella and Viruses Affecting Immunocompromised Patients, Robert Koch Institute, Berlin, Germany; Department of Medicine, University of California, San Diego, California, USA; Centre de Recherche pour le Développement, Université Alaisane Ouattara de Bouake, Bouake, Côte d’Ivoire; North Carolina Zoological Park, Asheboro, North Carolina, USA; Institut Congolais pour la Conservation de la Nature, Kinshasa, Democratic Republic of Congo; Diane Fossey Gorilla Fund International, Atlanta, Georgia, USA; Department of Environmental Sciences and Program in Population Biology, Ecology and Evolution, Emory University, Druid Hills, Georgia, USA; Division of Neurobiology, Ludwig Maximilians University, Munich, Germany; Centre for Research and Conservation, Royal Zoological Society of Antwerp, Antwerp, Belgium; Department of Environmental Health, Rollins School of Public Health, Emory University, Druid Hills, Georgia, USA; International Gorilla Conservation Program, Kigali, Rwanda; Institut National de Recherche Biomédicale, Kinshasa, Democratic Republic of Congo; Wildlife Conservation Society, Grauer’s Gorilla Project, Kinshasa-Combe, Democratic Republic of Congo; School of Human Evolution and Social Change, Arizona State University, Tempe, Arizona, USA; Institute of Vertebrate Biology, Academy of Sciences, Brno, Czech Republic; Department of Pathology and Parasitology, University of Veterinary and Pharmaceutical Sciences, Brno, Czech Republic; Biology Centre, Institute of Parasitology, Academy of Sciences of the Czech Republic, Ceske Budejovice, Czech Republic; Liberec Zoo, Liberec, Czech Republic; World Wildlife Foundation, Dzanga Sangha Protected Areas, Bangui, Central African Republic; Tai Chimpanzee Project, Centre Suisse de Recherches Scientifiques, Abidjan, Côte d’Ivoire; Institute of Biology, University of Neuchatel, Neuchatel, Switzerland; Budongo Conservation Field Station, Masindi, Uganda; School of Psychology, University of St. Andrews, St. Andrews, Scotland, United Kingdom; Unité Mixte Internationale 233, Institut de Recherche pour le Développement, INSERM U1175, and University of Montpellier, Montpellier, France; Computational Biology Institute, Montpellier, France

ABSTRACT
It has long been hypothesized that polyomaviruses (PyV; family Polyomaviridae) codiverged with their animal hosts. In contrast, recent analyses suggested that codivergence may only marginally influence the evolution of PyV. We reassess this question by focusing on a single lineage of PyV infecting hominine hosts, the Merkel cell polyomavirus (MCPyV) lineage. By characterizing the genetic diversity of these viruses in seven African great ape taxa, we show that they exhibit very strong host specificity. Reconciliation analyses identify more codivergence than noncodivergence events. In addition, we find that a number of host and PyV divergence events are synchronous. Collectively, our results support codivergence as the dominant process at play during the evolution of the MCPyV lineage. More generally, our results add to the growing body of evidence suggesting an ancient and stable association of PyV and their animal hosts.

IMPORTANCE
The processes involved in viral evolution and the interaction of viruses with their hosts are of great scientific interest and public health relevance. It has long been thought that the genetic diversity of double-stranded DNA viruses was generated over long periods of time, similar to typical host evolutionary timescales. This was also hypothesized for polyomaviruses (family Polyomaviridae), a group comprising several human pathogens, but this remains a point of controversy. Here, we investigate this question by focusing on a single lineage of polyomaviruses that infect both humans and their closest relatives, the African great apes. We show that these viruses exhibit considerable host specificity and that their evolution largely mirrors that of their hosts, suggesting that codivergence with their hosts played a major role in their diversification. Our results provide statistical evidence in favor of an association of polyomaviruses and their hosts over millions of years.

Viral diversification is notably shaped by processes that promote host specificity, for example, antagonistic coevolution (1), and opportunities to colonize new hosts, i.e., cross-species transmission events. Depending on their balance, host-virus coevolutionary patterns may arise and persist over the long term. Long-term coevolution may have played an important role in the diversification of some double-stranded DNA (dsDNA) viruses, e.g., herpesviruses and papillomaviruses (2–5). Polyomaviruses (PyV; family Polyomaviridae) are small non-enveloped viruses with a circular double-stranded DNA genome (ca. 5 kb in length) (6). They infect a broad range of animals, including arthropods and vertebrates (fish, birds, and mammals), and comprise at least 13 distinct viruses infecting humans (7, 8). In humans, infections occur in childhood, persist lifelong, and are usually asymptomatic (9). At least five PyV have been associated with disease in immunosuppressed individuals (10–12). Routes of transmission are poorly characterized but may involve respiratory droplets and/or environmental contamination. Putative codivergence events of hosts and their PyV have repeatedly been invoked in the literature to explain the structure of PyV diversity. Reconciliation analyses performed on the family scale sometimes supported a significant contribution of codivergence events (8, 13), but other studies have failed to detect any global codivergence signal (14, 15). Similarly, authors focusing on
more recent evolutionary events defended opposing views as to the potential codivergence of humans and JC polyomaviruses (JCV) (16, 17–19). An alternative scenario combining ancient noncodivergence events and subsequent lineage-specific codivergence with their hosts, as proposed for papillomaviruses (3), still remains to be tested. The disparate sampling of the PyV animal hosts as well as the lack of resolution of many internal branches of this viral family tree severely compromises the power to detect such patterns from currently available data.

To overcome these limitations, we designed a formal test to assess the influence of codivergence on the evolution of PyV and characterized the genetic diversity of a single lineage of PyV that infects a set of recently diverged host species with a well-resolved phylogeny. Specifically, we focused on viruses infecting African great apes (here, referred to simply as great apes) belonging to the lineage comprising the Merkel cell polyomavirus (MCPyV), an oncogenic human virus (Human polyomavirus 5, genus Alphapolyomavirus) (10, 20, 21, 22).

MATERIALS AND METHODS

Samples. We collected a total of 386 fecal samples in the wild from seven great ape taxa (Table 1). Great ape samples were collected opportunistically or from habituated animals and preserved in RNAlater (Qiagen, Hilden, Germany), in liquid nitrogen, or by drying over silica. We also collected 197 fecal samples from two human populations in Côte d’Ivoire and the Democratic Republic of the Congo (Table 1). Human samples were preserved in liquid nitrogen. For animal samples, authorization was obtained from responsible local authorities. For human samples, institutional authorization was received along with the written consent of all participants in the study.

Molecular biology. DNA extraction was performed using a Roboklon stool kit (Roboklon, Berlin, Germany), according to manufacturer’s instructions.

To identify Merkel cell polyomavirus (MCPyV)-related sequences in DNA extracts, a nested PCR assay was set up that made use of generic, degenerate primers targeting a ca. 700-bp VP1 fragment (Table 2, PCR1). These primers were designed on the basis of published MCPyV sequences and those of MCPyV-related PyV of nonhuman primates (NHP). First-round PCR mixes were set up so as to reduce the risk of carryover contamination with PCR products. The mixtures contained 0.2 μM each primer, 200 μM deoxynucleoside triphosphate (dNTP) mix (with dUTP replacing dTTP), 0.3 U of AmpErase uracil N-glycosylase (UNG; Invitrogen, Carlsbad, CA, USA), 4 mM MgCl2, 1× PCR buffer, and 1.25 U of Platinum Taq polymerase (Invitrogen). Second-round PCR mixes were prepared in the same way but did not include UNG. Cycling conditions were as follows: 7 min at 45°C (UNG activity) and 7 min at 95°C, followed by 47 cycles (first round) or 45 cycles (second round) of 30 s at 95°C, 30 s at 57°C (first round) or 58°C (second round), and 2 min at 72°C, with a final 10 min at 72°C.

Twenty-two positive samples were then selected based on the results of preliminary phylogenetic analyses to attempt additional nested long-distance (LD) amplification of partial genomes (approximately 2.5 kb) with generic, degenerate primers (Table 2, PCR2) using a TaKaRa-Ex kit (TaKaRa Bio, Inc., Otsu, Japan) according to the manufacturer’s instructions. Nondegenerate primers (sequences available from the authors upon request) were used for amplification of the remaining part (approximately 2.8 kb) of the genome with LD nested PCR. LD PCR cycling conditions followed those reported in Leendertz et al. (21).

A total of 174 human DNA extracts were also screened using a semi-nested PCR system targeting a ca. 200-bp VP1 fragment (Table 2, PCR3). This system was designed to be specific to members of lineage 1 (see below) and was validated on a selection of great ape DNA extracts of known status before being employed on human DNA extracts (data not shown). PCR mix preparation and cycling conditions followed those mentioned above.

Short PCR products were purified using ExoSAP-IT (Affymetrix, Santa Clara, CA, USA) whereas LD PCR products were purified using a column-based PCR purification kit (Qiagen, Venlo, Netherlands). All purified products were sequenced with a BigDye Terminator cycle sequencing kit on a 377 DNA automated sequencer (Applied Biosystems, Warrington, United Kingdom).

Overlapping partial sequences were used to reconstruct circular genomes using Geneious, version 7.1.4 (Biomatters, Ltd., Auckland, New Zealand) (23). Genomes were subsequently annotated with Geneious.

Phylogenetic analyses. Partial VP1 and complete genome data sets were assembled that comprised sequences generated in this study and a selection of (partial VP1) or all (complete genome) MCPyV sequences as well as any publicly available great ape MCPyV-related sequences. Both data sets were reduced to unique sequences and aligned using MUSCLE, as implemented in SeaView, version 4 (24). Conserved nucleotide blocks were selected from the alignments using Gblocks (in SeaView) (25) and used for recombination analyses using RDP4, version 4.46 (26). Two final alignments were generated: one with 74 sequences and 838 positions and another with 16 sequences and 5,150 positions. Further analyses were performed only on the partial VP1 alignment as this comprised more genetic diversity.

The best model of nucleotide substitution (gene time-reversible matrix with rate variation across sites [GTR+G4]) was selected with jModelTest, version 2.1.4 (27), using the Bayesian information criterion. Maximum likelihood (ML) analyses were performed using PhyML, version 3 (28), as implemented on the PhyML Web server (29). The best-fit root of the ML tree was identified using TempEst, version 1.5 (30; http://tree.bio.ed.ac.uk/software/tempest/). Bayesian Markov chain Monte Carlo (BMCMC) analyses were performed in BEAST, version 1.8.2, under a log-normal relaxed clock (uncorrelated) and three different models of diversification: a pure coalescent model assuming a constant population size, a multispecies coalescent model using the 14-species scheme suggested by species delineation analyses (see below), and a birth-death specification model (31, 32). Convergence of BMCMC runs (at least two runs per model) and appropriate sampling of the posterior were checked with Tracer, version 1.6 (http://tree.bio.ed.ac.uk/software/tracer/). Branch robustness was assessed through nonparametric bootstrapping (250 pseudoreplicates for ML analyses) or posterior probabilities (BMCMC analyses).

Host specificity analyses. Host specificity was assessed by running Bayesian tip-association significance testing (BaTS) on all posterior samples of trees (PST) generated by BMCMC analyses (33). BaTS allows for tests of the correlation of trait states with ancestry while accounting for phylogenetic uncertainty suggested by the PST. It compares observations to a null distribution generated under the assumption that trait values are not influenced by ancestry. Host species/subspecies was defined as the...
<table>
<thead>
<tr>
<th>Species and subspecies</th>
<th>Country</th>
<th>Site</th>
<th>No. of samples</th>
<th>No. of positives</th>
<th>Proportion (% [95% CI])</th>
<th>Minimum identity within host subspecies (%)</th>
<th>Maximum identity with a publicly available sequence (% [GenBank accession no. and host subspecies])</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Gorilla gorilla gorilla</em></td>
<td>Cameroon</td>
<td>Belgique</td>
<td>20</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Central African Republic</td>
<td>Dzanga-Sangha Special Reserve</td>
<td>23</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Democratic Republic of</td>
<td>Kahuzi-Biega National Park</td>
<td>34</td>
<td>1</td>
<td>20.6 (9.3–38.4)</td>
<td>97.4 (74.7, <em>Gorilla gorilla gorilla</em>)</td>
<td>99 (HQ385752, <em>Gorilla gorilla gorilla</em>)</td>
</tr>
<tr>
<td><em>Gorilla gorilla diehli</em></td>
<td>Cameroon</td>
<td>Greater Takamanda-Mone Landscape</td>
<td>22</td>
<td>1</td>
<td>1.9 (0.3–7.4)</td>
<td>98.5 (95.1, <em>Gorilla gorilla gorilla</em>)</td>
<td>99 (HQ385752, <em>Gorilla gorilla gorilla</em>)</td>
</tr>
<tr>
<td><em>Gorilla beringei beringei</em></td>
<td>Rwanda</td>
<td>Volcanoes National Park</td>
<td>51</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Gorilla beringei graueri</em></td>
<td>Democratic Republic of</td>
<td>Kahuzi-Biega National Park</td>
<td>34</td>
<td>1</td>
<td>1.2 (0–7.6)</td>
<td>97.4 (95.1, <em>Gorilla gorilla gorilla</em>)</td>
<td>99 (HQ385752, <em>Gorilla gorilla gorilla</em>)</td>
</tr>
<tr>
<td><em>Pan troglodytes troglodytes</em></td>
<td>Cameroon</td>
<td>Belgique</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
<td>83 (HQ385747, <em>Pan troglodytes verus</em>)</td>
</tr>
<tr>
<td></td>
<td>Cameroun</td>
<td>Cameroun</td>
<td>10</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Campo Ma’an National Park</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mambele</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gabon</td>
<td>Loango National Park</td>
<td>27</td>
<td>3</td>
<td>11.5 (4.8–24.1)</td>
<td>97.4 (95.1, <em>Pan troglodytes verus</em>)</td>
<td>95 (HQ385748, <em>Pan troglodytes verus</em>)</td>
</tr>
<tr>
<td><em>Pan troglodytes schweinfurthii</em></td>
<td>Uganda</td>
<td>Budongo Central Forest Reserve</td>
<td>33</td>
<td>9</td>
<td>30.3 (20–43)</td>
<td>97.4 (95.1, <em>Pan troglodytes verus</em>)</td>
<td>95 (HQ385747, <em>Pan troglodytes verus</em>)</td>
</tr>
<tr>
<td></td>
<td>Kibale Forest National Park</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Salonga National Park</td>
<td>26</td>
<td>14</td>
<td>53.8 (33.7–72.9)</td>
<td>77.4 (74.7, <em>Pan troglodytes verus</em>)</td>
<td>91 (HQ385746, <em>Pan troglodytes verus</em>)</td>
<td></td>
</tr>
<tr>
<td><em>Pan paniscus</em></td>
<td>Democratic Republic of</td>
<td>Salonga National Park</td>
<td>26</td>
<td>14</td>
<td>53.8 (33.7–72.9)</td>
<td>77.4 (74.7, <em>Pan troglodytes verus</em>)</td>
<td>91 (HQ385746, <em>Pan troglodytes verus</em>)</td>
</tr>
<tr>
<td></td>
<td>the Congo</td>
<td></td>
<td>101</td>
<td>16</td>
<td>15.2 (10.7–21.2)</td>
<td>99 (JF812999, <em>Homo sapiens</em>)</td>
<td></td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>Côte d’Ivoire</td>
<td>Tai National Park</td>
<td>96</td>
<td>16</td>
<td></td>
<td></td>
<td>100 (JF812999, <em>Homo sapiens</em>)</td>
</tr>
</tbody>
</table>

* At the species/subspecies level. NA, not assessed; CI, confidence interval.

PyV sequences from Western chimpanzees (*Pan troglodytes verus*) were already available from a previous study (14).
trait of interest. Its association with ancestry was assessed at the host sub-
species level (8 states) and species level (5 states) independently by run-
ning separate Bats analyses during which 500 null replicates per tree were
generated. Global as well as state-specific statistics of association were
computed (global, association index [AI] and Fitch parsimony score [PS];
state-specific, maximum exclusive single-state clade size [MC]).

To investigate the association of host and PyV diversification pro-
cesses, we performed PyV species delineation analyses with the R package
SPLITS (34), using the maximum clade credibility tree derived from
BMCMC analyses performed under the (coalescent) constant popula-
tion size model. SPLITS implements general mixed Yule coalescent
(GMYC) models (34, 35) which are optimized and compared to the
null hypothesis that the tree was generated by pure coalescent pro-
cesses, i.e., that it reflects diversity within a single species. When the
GMYC model outperforms the null model, the parts of the tree most
likely to have been generated by between-species and within-species
processes can be identified, thereby delineating species (according to
the phylogenetic species concept).

Codivergence analyses. The degree of topological congruence and the
number of events necessary to explain (reconcile) incongruences were
assessed using Jane, version 4 (36). Jane implements a genetic algorithm to
quickly identify the most parsimonious scenarios of coevolution involv-
ing several types of events (codivergence, duplication, duplication with
host switch, loss, and failure to diverge). As input, it requires host and
parasite phylogenies and the respective tip mapping as well as an event
cost matrix. A simplified version of the PyV phylogeny was used as input,
whereby single-host clades were collapsed. Three sets of costs were tested:
(i) set 1 with the parameter codivergence set at 0, duplication at 1 (under
the assumption that duplication incurs costs related to within-host spe-
ciation, e.g., maintaining of distinct lineages in the face of within-host
competition or tropism change within the same host), duplication with
host switch at 1 (host switch incurs costs), loss at 1 (prevalence was always
high), and failure to diverge at 1 (given their respective evolutionary time-
scales, viruses are unlikely to fail to diverge when their hosts do so); (ii) set
2, with the same parameters as set 1 but with loss set at 0 (prevalence may
have been low at some point in the past); (iii) set 3, with codivergence set
at 1 and all noncodivergence events at 0. Set 3 is a variation of set 1 with
the same relative costs but where all costs are shifted to the left. This allows
costs and codivergence events to be equated. Jane was run using the ver-
tex-based cost mode, and the parameters of the genetic algorithm were
kept at their default values (population size, 100; number of generations,
100). To determine the probability of observing the inferred costs by
chance, costs were also calculated on a set of 500 samples for which tip
mapping was randomized. Settings of the genetic algorithm were kept at
default values.

Topology tests were performed to assess whether exceptions to a sce-
nario of perfect codivergence observed in the PyV phylogenetic tree were
better supported by the data than a perfect codivergence model. This was
done by using approximately unbiased (AU) tests, as implemented in
CONSEL, version 0.11 (37).

Finally, divergence dates were also estimated. Topological congruence
could emerge independently of codivergence, e.g., through preferential
host switching (38). Observing synchronicity in timing of divergence events
of hosts and their parasites reinforces the codivergence hypothesis.
When viral lineage duplication occurs, synchronicity of parasite diver-
gence events is also expected (provided the viral lineages maintain similar
degrees of association to their host). Divergence date estimates were ob-
tained using two methods: (i) as part of the aforementioned BMCMC
analyses or (ii) by reestimating branch lengths of the ML tree under codon
models using HyPhy, version 2.2.4 (39), and making the resulting tree
ultrametric using a relaxed clock model implemented in r8s (40).
The codon models used for this second set of analyses were a pure branch
model derived of MG94, in which the ratio of nonsynonymous substitu-
tions per nonsynonymous site to synonymous substitutions per synony-
mous site is estimated for each branch but assumed to be unchanged
across sites (41), and an adaptive branch site random effects model
(ABSEL), in which this ratio is estimated for each branch and allowed to
vary across sites (42). We detected marked saturation at synonymous sites
(data not shown); such strong saturation complicates analyses under both
nucleotide and codon models. For both BMCMC and ML-based analyses,
the relaxed clock was calibrated by setting a prior distribution (BEAST) or
enforcing a fixed age (r8s) for the time to the most recent common ances-
tor (tMRCA) of lineage 1 using a published estimate of the split date of all
hominine species (either 5.6 million years or a normal distribution with a
mean of 5.6 million years and standard deviation of 0.25 million years
[43]). Because we used the split date of all hominine species, estimates of
times to the most recent common ancestors for viruses should be regarded
as minimum bounds (viral coalescence times will necessarily predate the
effective ancestral host population/species split). It should also be noted
that divergence dates of the different hominine lineages are a point of
active debate; this stems from both a scarce paleontological record and
uncertainty in estimates of long-term mutation rates at genomic scales.
For example, the estimate we opted for here (5.6 million years) is drawn
from genomic analyses that proposed two estimates (5.6 or 11.2 million
years), depending on priors on the substitution rates ($1 \times 10^{-9}$ or $0.5 \times
10^{-9}$ mutations · bp$^{-1}$ · year$^{-1}$) (43). The focus of our synchronicity
analyses was, however, on relative internode lengths, not absolute dates.
Calendar years can thus be replaced with genetic distances and/or ratios of
interest (see Table 6).

Accession numbers. Partial VP1 and whole-genome sequences, re-
spectively, were deposited at the European Nucleotide Archive and

### Table 2: Primers used in this study

<table>
<thead>
<tr>
<th>PCR</th>
<th>Primer name</th>
<th>Primer sequence (5′–3′)</th>
<th>Annealing temp (°C)</th>
<th>Fragment size (kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCR1</td>
<td>PCR1.1-f</td>
<td>TGTGTCTCCTAAGCCBGGGATG</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PCR1.1-r</td>
<td>ACTACTGGGTATGGRRTTMYMCC</td>
<td>58</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>PCR1.2-f</td>
<td>CTGAATCCAGAAGGTTGGGAGT</td>
<td>58</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td>PCR1.2-r</td>
<td>CATGAAANGGCGATTTTTNCCACT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCR2</td>
<td>PCR2.1-f</td>
<td>CTGAAGYCTGAGGGACGTTGAG</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PCR2.1-r</td>
<td>GCAAACATRTGTGAATTGACCTC</td>
<td>58</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td>PCR2.2-f</td>
<td>TCAAGCWCAGTGCAGGAGG</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PCR2.2-r</td>
<td>GCAAATCYARRGGYTCTCCCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCR3</td>
<td>PCR3.1-f</td>
<td>TGATATGCAGCCAAATMWCCARC</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PCR3.1-r</td>
<td>AACATGTGATAATTGACCTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PCR3.1-f</td>
<td>TGATATGCAGCCAAATMWCCARC</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PCR3.2-r</td>
<td>AATTGACCTCCCTAATAGGATG</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
RESULTS

Detection of short MCPyV-related sequences. Using a specific PCR system designed to amplify a ca. 700-bp fragment of the VP1 gene, we screened 386 fecal great ape and 197 human samples (Table 1). We detected MCPyV-related sequences in 50 great ape DNA extracts representing all hosts but *Gorilla gorilla diehli*, with fecal detection rates between 1.2% (*Gorilla beringei beringei*) and 53.8% (*Pan paniscus*). Nearly all sequences were found only at one site; a single sequence was detected in five and two Eastern chimpanzees (*Pan troglodytes schweinfurthii*) at two distinct sites in Uganda. For species/subspecies from which more than two sequences were obtained, considerable sequence divergence was observed; e.g., maximum observed distances were over 20%, possibly reflecting the circulation of viruses belonging to different lineages (discussed in more detail below). Minimum observed distances to publicly available sequences were often relatively high, i.e., between 5 and 17%. Finally, we also detected MCPyV sequences with >99% identity to published MCPyV sequences in 30 human DNA extracts (fecal detection rate, 15.2%). Most human DNA extracts were also screened with a PCR system intended to be lineage 1 specific (see below); all assays were negative.

Characterization of full genomes. We attempted to determine full-genome sequences from a selection of DNA extracts (n = 22). This was possible for samples from *P. paniscus* (n = 2), *P. troglodytes schweinfurthii* (n = 3), *P. troglodytes diehli* (n = 1), and *Gorilla beringei graueri* (n = 1). Examination of putative open reading frames (ORFs) showed that all genomes displayed a typical PyV genome structure with an early region encoding regulatory proteins (small T and large T antigens) and a late region coding for structural proteins (VP1, VP2, and VP3) separated by a noncoding control region (NCCR). No open reading frame likely to encode a putative agnoprotein was identified. Overall, a ca. 80% sequence similarity to genomes of MCPyV and MCPyV-related nonhuman primate PyV was observed. Preliminary analyses revealed that the full genomes represented only a fraction of the overall genetic diversity detected in this study. To incorporate this broader diversity, we performed all of following phylogenetic analyses on an alignment of partial VP1 sequences (including sequences extracted from the novel full genomes).

Molecular phylogeny. We could not detect any signal indicative of recombination in the VP1 alignment (26). Phylogenetic analyses in both maximum likelihood (ML) (28) and Bayesian (31) frameworks supported the existence of a number of host-specific clades (Fig. 1 and 2). All clades seemed to derive from three ancient lineages: one that comprised only MCPyV sequences and two that included only viral sequences detected in gorillas, bonobos, and chimpanzees. Branching order partially recapitulated host divergence events in the two great ape lineages (here, referred to as lineages 1 and 2) (Fig. 1 and 2). We identified four exceptions: (i) the polyphyly of PyV infecting Western chimpanzees in lineage 1 and Eastern chimpanzees in lineage 2, (ii) the interspersion of PyV infecting Eastern lowland and mountain gorillas in lineage 1, and (iii) the basal position of MCPyV.

Host specificity. We estimated the statistical support for host specificity using BaTS (Table 3). We found that viral sequences found in a single host species were generally more likely to be closely related than expected by chance when both global and state-specific statistics were considered. The only exceptions corresponded to viral sequences identified in the sister subspecies *G. beringei beringei* and *G. beringei graueri*.

We also characterized the viral diversification process by running a species delineation analysis using general mixed Yule coalescent models (GMYC) (34, 35). The best GMYC model outperformed the null, full coalescent model (P = 0.0005) and identified 14 entities, among which 10 comprised several sequences. Nine multisequence entities comprised only sequences identified from a single host species/subspecies, indicating a close parallelism of PyV and host diversification processes (Fig. 1).

Codivergence. Taking the viral phylogeny presented in Fig. 1 as a given, we performed reconciliation analyses using Jane (Table 4). Under all tested cost sets, and whether the host species or subspecies phylogeny was considered, the number of codivergence events always exceeded the number of noncodivergence events. Randomization tests showed that, irrespective of the cost set, these results could not be explained by chance at the subspecies level. At the species level and using a P value threshold of 0.05, results obtained under two of the cost sets failed to reach statistical significance; it should, however, be noted that the species-level phylogeny comprises only five species, meaning that these tests had low power.

We also examined whether the viral topology presented in Fig. 1 was a better fit to our data than alternative topologies which enforced strict codivergence within lineages 1 and 2. The model forcing MCPyV to belong to lineage 1 was the only one that was rejected (AU test; P = 0.003). Monophyly of PyV infecting Western chimpanzees in lineage 1 and Eastern chimpanzees in lineage 2 as well as inclusion of MCPyV in lineage 2 could not be excluded (AU test; P = 0.52, 0.13, and 0.11, respectively). Given the very recent split of Eastern lowland and mountain gorillas (about 10,000 years ago [44]), the interspersion of PyV infecting these subspecies appeared biologically plausible, so we did not compare this scenario to a strict codivergence model.

Besides topological congruence, codivergence should result in synchronization of (i) viral and host divergence dates and (ii) viral divergence dates in the case of ancestral viral lineage duplication. We first estimated divergence dates using a relaxed clock model applied to nucleotide data in a Bayesian framework. For five of the six focal nodes of our analyses (nodes 1.2 to 4 and 2.1 to 3), these estimates were significantly older than host divergence events (Table 5). This pattern was compatible with the effects of the time dependency of molecular rates, i.e., the decay of molecular rates with increasing observation timescales, which can result in overestimating recent time to the most recent common ancestor (tMRCA) inferred from deep calibration points (19, 45–47). As this may arise through the effects of unaccounted-for purifying selection (among other possible mechanisms) (48, 49, 50), we reestimated all branch lengths using selection-aware models of codon evolution in an ML framework. A branch model of codon evolution resulted in divergence dates very close to those inferred by BRCMC analyses. Using an adaptive branch site random effect model of codon evolution, strong purifying selection was detected on a number of branches, including deep ones (data not shown). Most of the resulting increase in the overall tree length was supported by a single basal branch. This expansion prevented us from deriving any trustworthy tMRCA estimates.
Given the likely impact of strong purifying selection and our inability to properly account for it, we reexamined branch length/internode ratios by rescaling the results in Table 5, using the tMRCA of a young node, node 1.4 (divergence of lineage 1 PyV infecting *P. troglodytes troglodytes* and *P. troglodytes schweinfurthii*), as a new arbitrary unit (Table 6). This resulted in a good agreement of host and virus relative divergence dates for most nodes (nodes 1.3 and 2.3 and nodes 1.2 and 2.2). The tMRCA of lineage 2 PyV infecting all great apes was a large underestimate of the divergence date of their hominine hosts, as expected under the hypothesis that deep branch lengths are severely underestimated.

**DISCUSSION**

The lack of any physical viral fossil record considerably complicates the task of understanding the long-term association of viruses with their hosts. However, using their present-day distribution, their nucleic acid sequences, and (more rarely) other biological traits, we can try to infer how long and how closely viruses have been associated with their hosts. The aim of this study was to determine whether codivergence, i.e., viral diversification driven by host diversification, is an important driver of PyV evolution.

Measurable host specificity is an absolute prerequisite for characterizing historical codivergence events. Host specificity has often been assumed for PyV, with only a few well-identified exceptions, e.g., budgerigar fledgling disease virus and simian virus 40 (SV40). Over the last decade, this assumption has been repeatedly supported by the implementation of generic PyV detection tools which have not revealed any multihost PyV (20, 51). Here, we used a PCR assay designed to specifically target a single PyV lineage to generate a large sample of sequences from closely related PyV infecting wild African great apes. Statistical tests strongly supported marked host specificity, which was still detectable at the host subspecies level. Viral diversification/speciation—as revealed by a GMYC model, i.e., according to the phylogenetic species concept—appeared strongly influenced by host diversification.

Host specificity and a coupling of viral diversification/speciation with host diversity could also arise over much shorter timescales than those implied by codivergence events. If codivergence is a dominant evolutionary process, a key expectation is that virus and host phylogenies should often be congruent. Phylogenetic analyses of great ape MCPyV-like sequences highlighted the existence of two viral lineages within which viral divergence events were mostly in line with hominine divergence events. Exceptions to the expectation of perfect codivergence within these lineages were not statistically supported. In addition, reconciliation analyses identified more codivergence events than noncodivergence...
events, irrespective of the host taxonomic level and cost set, e.g., 10 codivergence events versus 5 non-codivergence events considering host subspecies and all cost sets. Codivergence may therefore be the dominant process at play, accompanied by less frequent non-codivergence events, e.g., the viral lineage duplication event that gave rise to lineages 1 and 2.

On short timescales, host relatedness may influence viral transmission in such a way that topological congruence ensues in the absence of real codivergence, e.g., if host jumps are facilitated by host phylogenetic proximity (the preferential host switch hypothesis) (38, 52). A further step in validating codivergence events consists of showing that host and virus divergence events are synchronized. This requires branch lengths to be properly estimated events, irrespective of the host taxonomic level and cost set, e.g., 10 codivergence events versus 5 non-codivergence events considering host subspecies and all cost sets. Codivergence may therefore be the dominant process at play, accompanied by less frequent non-codivergence events, e.g., the viral lineage duplication event that gave rise to lineages 1 and 2.

On short timescales, host relatedness may influence viral transmission in such a way that topological congruence ensues in the absence of real codivergence, e.g., if host jumps are facilitated by host phylogenetic proximity (the preferential host switch hypothesis) (38, 52). A further step in validating codivergence events consists of showing that host and virus divergence events are synchronized. This requires branch lengths to be properly estimated

### TABLE 3 Results of BaTS tests for host specificity

<table>
<thead>
<tr>
<th>Host species or subspecies (no. in group)</th>
<th>Mean association index</th>
<th>Mean parsimony score</th>
<th>Mean maximum exclusive single-state clade size</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species (5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G. beringei</td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.002</td>
</tr>
<tr>
<td>G. gorilla</td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.002</td>
</tr>
<tr>
<td>H. sapiens</td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.002</td>
</tr>
<tr>
<td>P. paniscus</td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.002</td>
</tr>
<tr>
<td>P. troglodytes</td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.002</td>
</tr>
<tr>
<td>Subspecies (8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G. beringei beringei</td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.002</td>
</tr>
<tr>
<td>G. beringei graueri</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>G. gorilla grauila</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>H. sapiens</td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.002</td>
</tr>
<tr>
<td>P. paniscus</td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.002</td>
</tr>
<tr>
<td>P. troglodytes</td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.002</td>
</tr>
<tr>
<td>P. troglodytes schweinfurthii</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P. troglodytes</td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.002</td>
</tr>
<tr>
<td>P. troglodytes verus</td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.002</td>
</tr>
</tbody>
</table>

*aThe values reported are derived from analyses performed on posterior sets of trees generated under the 14-species coalescent model. Values were very similar when posterior samples of trees obtained under a constant population size coalescent model or a birth-death speciation model were analyzed.

### TABLE 4 Results of reconciliation analyses with Jane

<table>
<thead>
<tr>
<th>Host phylogeny</th>
<th>Cost set</th>
<th>No. of events</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species level</td>
<td>1</td>
<td>5</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>5</td>
<td>0.066</td>
</tr>
<tr>
<td>Subspecies level</td>
<td>1</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>10</td>
<td>2</td>
</tr>
</tbody>
</table>

*a For the solution which was the most parsimonious in number of events.
TABLE 5 Absolute times to the most recent common ancestors of PyV in lineages 1 and 2

<table>
<thead>
<tr>
<th>Statistical framework</th>
<th>Diversification model or smoothing factor</th>
<th>Lineage 1</th>
<th>Lineage 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Node 1.1</td>
<td>Node 1.2</td>
</tr>
<tr>
<td></td>
<td>(all)</td>
<td>Node 1.2</td>
<td>Node 1.3</td>
</tr>
<tr>
<td></td>
<td>(panine)</td>
<td>Node 1.3</td>
<td>Node 1.4</td>
</tr>
<tr>
<td></td>
<td>(P. troglodytes)</td>
<td>Node 1.4</td>
<td>Node 2.1</td>
</tr>
<tr>
<td></td>
<td>(P. troglodytes + P. schweinfurthii)</td>
<td>Node 2.1</td>
<td>Node 2.2</td>
</tr>
<tr>
<td><strong>BMCMC</strong></td>
<td>Coalescent, constant population size</td>
<td>5.62</td>
<td>2.15</td>
</tr>
<tr>
<td></td>
<td>Multispecies coalescent</td>
<td>5.62</td>
<td>2.25</td>
</tr>
<tr>
<td></td>
<td>Speciation, birth-death</td>
<td>5.62</td>
<td>2.06</td>
</tr>
<tr>
<td><strong>ML</strong></td>
<td>1</td>
<td>5.62</td>
<td>2.29</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>5.62</td>
<td>2.24</td>
</tr>
<tr>
<td><strong>Prado-Martínez et al. (43)</strong></td>
<td></td>
<td>5.62</td>
<td>0.87</td>
</tr>
</tbody>
</table>

**a** Diversification models were used for Bayesian Markov chain Monte Carlo (BMCMC) analyses; smoothing factors were used for maximum likelihood (ML) analyses (under the MG94-like model of codon evolution).

**b** Estimates that are incompatible with those determined in Prado-Martínez et al. (43) appear in bold. The 95% highest posterior density (HPD) was used for BMCMC analyses; bootstrap (Bp) intervals were used for ML analyses.

**c** Bootstrap intervals were determined using 100 bootstrap pseudoreplicates of the codon data set from which branch lengths were reestimated on the ML topology; all trees were rooted using TempEst. tMRCA, time to the most recent common ancestor.

**d** Assuming a mutation rate of 1e−6 mutations/base pair/year.
The 95% highest posterior density (HPD) was used for BMCMC analyses; bootstrap (Bp) intervals were used for ML analyses. Diversification models were used for Bayesian Markov chain Monte Carlo (BMCMC) analyses; smoothing factors were used for maximum likelihood (ML) analyses (under the MG94-like model of codon evolution).

<table>
<thead>
<tr>
<th>Node</th>
<th>Branchlengths</th>
<th>Node 2.1</th>
<th>Node 2.2</th>
<th>Node 2.3</th>
<th>Node 1.1</th>
<th>Node 1.2</th>
<th>Node 1.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>493</td>
<td>4.98</td>
<td>2.39</td>
<td>1</td>
<td>32.11</td>
<td>4.98</td>
<td>2.39</td>
<td>1</td>
</tr>
</tbody>
</table>

Statistical framework or reference population size

TABLE 6 Relative times to the most recent common ancestors (tMRCA) of Py in lineages 1 and 2.
FUNDING INFORMATION
This work, including the efforts of Klara Judit Petzkelova, was funded by Academy of Sciences of the Czech Republic (RVO68081766). This work, including the efforts of Klara Judit Petzkelova, was funded by Grant Agency of the Czech Republic (206/09/0927). This work, including the efforts of Joel O. Wertheim, was funded by University of California Laboratory Fees Research Program (12-LR-236617). This work, including the efforts of Joel O. Wertheim, was funded by HHS | National Institutes of Health (NHI) (K01-AI110181). This work, including the efforts of Martine Peeters, was funded by Agence Nationale de Recherches sur le Sida et les Hepatites Virales (ANRS) (12125, 12128, 12555, and 12325).

REFERENCES

8540  jvi.asm.org  Journal of Virology  October 2016 Volume 90 Number 19


