

Discovery of a polyomavirus in European badgers (*Meles meles*) and the evolution of host range in the family *Polyomaviridae*

Sarah C. Hill,^{1†} Aisling A. Murphy,^{2†} Matthew Cotten,³ Anne L. Palser,³ Phillip Benson,² Sandrine Lesellier,⁴ Eamonn Gormley,⁵ Céline Richomme,⁶ Sylvia Grierson,⁷ Deirdre Ni Bhuachalla,⁵ Mark Chambers,^{4,8} Paul Kellam,^{3,9} María-Laura Boschioli,¹⁰ Bernhard Ehlers,^{11‡} Michael A. Jarvis^{2‡} and Oliver G. Pybus^{1‡}

Correspondence

Bernhard Ehlers
ehlersb@rki.de
Michael A. Jarvis
michael.jarvis@plymouth.ac.uk
Oliver G. Pybus
oliver.pybus@zoo.ox.ac.uk

¹Department of Zoology, University of Oxford, UK

²School of Biomedical and Healthcare Sciences, Plymouth University, UK

³Wellcome Trust Sanger Institute, UK

⁴Bacteriology Department, Animal and Plant Health Agency, UK

⁵School of Veterinary Medicine, University College Dublin (UCD), Ireland

⁶ANSES, Nancy Laboratory for Rabies and Wildlife, France

⁷Department of Virology, Animal and Plant Health Agency, UK

⁸School of Veterinary Medicine, University of Surrey, UK

⁹MRC/UCL Centre for Medical Molecular Virology, University College London, UK

¹⁰University Paris-Est, ANSES, Laboratory for Animal Health, Bovine Tuberculosis National Reference Laboratory, France

¹¹Robert Koch Institute, Division 12 'Measles, Mumps, Rubella and Viruses Affecting Immunocompromised Patients', Germany

Polyomaviruses infect a diverse range of mammalian and avian hosts, and are associated with a variety of symptoms. However, it is unknown whether the viruses are found in all mammalian families and the evolutionary history of the polyomaviruses is still unclear. Here, we report the discovery of a novel polyomavirus in the European badger (*Meles meles*), which to our knowledge represents the first polyomavirus to be characterized in the family Mustelidae, and within a European carnivoran. Although the virus was discovered serendipitously in the supernatant of a cell culture inoculated with badger material, we subsequently confirmed its presence in wild badgers. The European badger polyomavirus was tentatively named *Meles meles* polyomavirus 1 (MmePyV1). The genome is 5187 bp long and encodes proteins typical of polyomaviruses. Phylogenetic analyses including all known polyomavirus genomes consistently group MmePyV1 with California sea lion polyomavirus 1 across all regions of the genome. Further evolutionary analyses revealed phylogenetic discordance amongst polyomavirus genome regions, possibly arising from evolutionary rate heterogeneity, and a complex association between polyomavirus phylogeny and host taxonomic groups.

Received 24 November 2014

Accepted 23 January 2015

†These authors contributed equally to this work.

‡These authors are co-senior authors of this paper.

The GenBank/EMBL/DDBJ accession numbers for the UK and French MmePyV1 isolates are KP644239 and KP644238, respectively. The raw reads for the UK isolate have been assigned ENA secondary sample accession number ERS155731.

Four supplementary tables are available with the online Supplementary Material.

INTRODUCTION

Viruses of the family *Polyomaviridae* are small, non-enveloped icosahedral viruses, comprising ~70 putative species from three proposed genera (Imperiale & Major, 2007; Johne *et al.*, 2011). Polyomaviruses have been found in many different avian and mammalian hosts, including rodents, birds, bats, humans, non-human primates, carnivorans, elephants,

dolphins, horses and Artiodactyla. In human populations, polyomavirus seroprevalence can reach up to 90% and most infections are asymptomatic (Van Ghelue *et al.*, 2012). However, in immunocompromised individuals infection can cause an array of symptoms, including Merkel cell carcinoma, kidney disease and progressive multifocal leukoencephalopathy (Imperiale & Major, 2007). Although renal and respiratory diseases, tumours, and wasting have been reported in some other animals, the consequences of polyomavirus infection in many non-human species remain to be determined.

All identified polyomaviruses share a similar genome organization and virion structure. Polyomaviruses typically exhibit a 40–45 nm diameter non-enveloped icosahedral capsid, composed of 72 monomers (Baker *et al.*, 1989; Imperiale & Major, 2007). The capsid encloses a single, circular dsDNA genome of ~5000 bp (Johne *et al.*, 2011). All polyomavirus genomes are composed of early and late regions, and regulatory regions called the non-coding control region (NCCR) (Imperiale & Major, 2007). The NCCR contains the origin of replication, transcription factor binding sites, promoters and enhancers. Transcription from the NCCR is bidirectional, producing mRNA encoding either early proteins or late proteins (Johne *et al.*, 2011). All polyomaviruses produce at least two early mRNA products encoding large and small tumour antigens (LT-Ag and St-Ag, respectively) (Imperiale & Major, 2007). These proteins are involved in viral genome replication and modulation of host cell conditions required for virus replication (Stevens *et al.*, 2013). The mouse and hamster polyomaviruses also encode a middle tumour antigen (MT-Ag) (Imperiale & Major, 2007). Recently, an alternate reading frame gene (called ALTO) overlapping the LT-Ag gene was identified in Merkel cell polyomavirus (MCPyV) and predicted to occur in polyomaviruses that are phylogenetically related to MCPyV (Carter *et al.*, 2013). Late polyomavirus mRNA transcripts encode capsid proteins (typically VP1, VP2 and VP3) (Imperiale & Major, 2007; Johnne *et al.*, 2011). Several mammalian polyomaviruses (Fig. 1) also encode an agnoprotein or other short proteins between the NCCR and the VP2 ORF (Johne *et al.*, 2011). Agnoproteins are involved in the control of viral protein expression and have effects throughout the viral life cycle (Gerits & Moens, 2012).

Virus cross-species zoonotic transmission is a frequent cause of emerging epidemics in humans (e.g. influenza A viruses, Lassa virus, Ebola virus, severe acute respiratory syndrome coronavirus and Middle East respiratory syndrome coronavirus) (Parrish *et al.*, 2008). Understanding the evolutionary history of viral families can help define the propensity of viruses to switch host species. However, little is known about the evolutionary history of polyomaviruses or their capacity for host switching. In an early phylogenetic study (Shadan & Villarreal, 1993), the similarity of mammalian and polyomavirus evolutionary trees led to the suggestion that polyomaviruses had co-evolved with their host species. This idea evolved to incorporate a combination of host-switching and virus–host co-divergence events

(notably, the basal split between avian and mammalian polyomaviruses) (Pérez-Losada *et al.*, 2006). However, recent statistical re-evaluation that incorporated novel polyomaviruses from a wider range of taxa has largely rejected the hypothesis of polyomavirus–host co-divergence (Krumbholz *et al.*, 2009; Warden & Lacey, 2012).

In the absence of evidence for co-divergence, alternative models have been proposed to explain the phylogenetic distribution of polyomavirus host species. In the first model, cross-species transmission of polyomaviruses is proposed to be a relatively common evolutionary event (Tao *et al.*, 2013). This hypothesis has consequences for human and animal health, and raises the question of what factors determine the rate of virus transmission between species (Streicker *et al.*, 2010). In the second model, the polyomavirus phylogeny is proposed to result from recombination amongst polyomaviruses, which results in the appearance of host species switching. A recent study found support for recombination of several polyomaviruses, yet was unable to identify ancestors of the putative recombinant lineages with certainty (Tao *et al.*, 2013). In a third model, heterogeneity in the evolutionary rates of different genes amongst the polyomavirus lineages is used to explain their complex evolutionary history (Tao *et al.*, 2013).

Here, we report *Meles meles* polyomavirus 1 (MmelPyV1) as a novel polyomavirus present in wild European badger (*Meles meles*) populations and attempt to clarify the evolutionary history of the mammalian polyomaviruses. To our knowledge, this represents the first description of a polyomavirus in the family Mustelidae, expanding the host range of these viruses for which full-length sequences exist to 18 distinct mammalian families. To explore the evolutionary history of this novel virus and the mammalian polyomaviruses as a whole, we performed comprehensive phylogenetic analyses and analyses of viral recombination using whole-genome sequences from all available mammalian polyomavirus species.

RESULTS

Virus discovery in cell culture

The badger polyomavirus was first discovered in a cell culture supernatant derived from tissues taken from a badger lung in Cornwall, England, in 1996 (Banks *et al.*, 2002). The original purpose of the culture was to maintain a badger herpesvirus, *mustelid herpesvirus 1* (MusHV-1), present in the original tissue sample (Banks *et al.*, 2002). The clarified supernatant of this cell culture had been stored at -80°C since 1997 (henceforth referred to as ‘1997 supernatant’). The 1997 supernatant was enriched using a sucrose cushion as part of an attempt to sequence the badger herpesvirus present. DNA was extracted from the sucrose-cushion-purified supernatant, amplified using random priming and sequenced using Illumina deep sequencing.

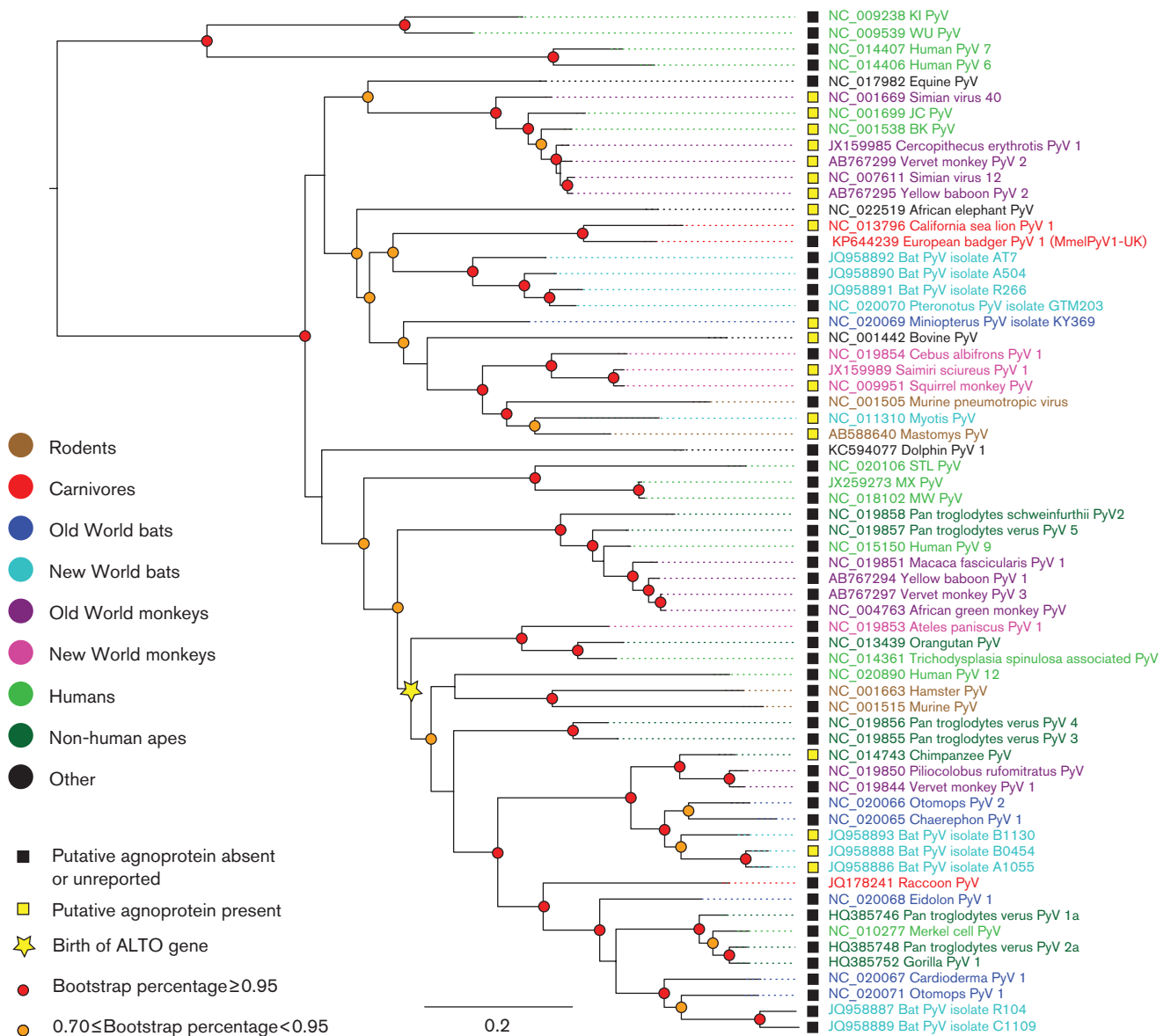


Fig. 1. Maximum-likelihood phylogeny of all known polyomavirus (PyV) species or putative species, including MmeIPyV1, estimated from the ‘genome-wide’ alignment. Viral taxa are coloured according to host species and bootstrap support scores are indicated using coloured circles (as indicated in the key). Scale bar represents expected number of substitutions per site (first and second codon positions only).

In addition to herpesvirus contigs (in preparation), two polyomavirus contigs of 4495 and 468 nt were identified, which appeared to encode a full-length circular genome (Fig. 2) (here denoted the MmeIPyV1 UK isolate; GenBank accession number KP644239). Overlapping identical sequences of 110 nt were found at both ends of the molecules. Furthermore, additional reads were present in the short read sequence data that spanned both ends of the sequences. When the original MiSeq reads were mapped to the genome, 1278 reads (of 6.3 million) were identified with perfect homology to the genome. In alignable regions, nucleotide

identity with any other polyomavirus genome was never more than 72% and the closest known viral genome was that of California sea lion polyomavirus 1 (CSLPyV1). Under the definition of <81–84% sequence identity suggested by John *et al.* (2011), the polyomavirus described here represents a new species and the species name ‘Meles meles polyomavirus 1’ (MmeIPyV1) is proposed.

To determine that the polyomavirus was not a contaminant in the mink cell line used to culture the badger herpesvirus isolate, the mink cell line was tested for the presence of the virus. Primers designed from the

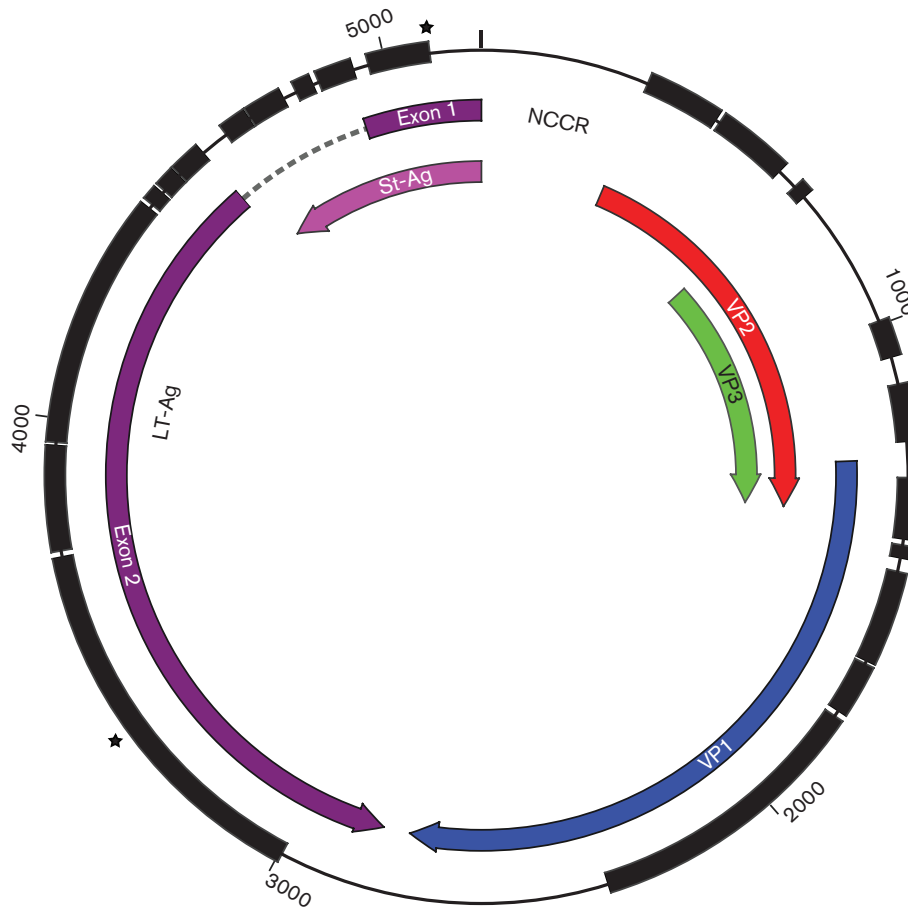


Fig. 2. Genome map of Mm1PyV1. Thick black blocks (outer circle) represent genomic regions that could be aligned across all mammalian polyomaviruses and were therefore included in the 'genome-wide' alignment. Coloured arrows represent ORFs. Early proteins are in purple; late proteins are in blue, red and green. Breakpoints used in recombination analysis are marked with two stars.

Mm1PyV1 genome generated strong PCR bands in a mink cell line sample that had been inoculated with material generated from the 1997 supernatant (Fig. 3) (see Methods). Indistinct or no bands were present in the negative control mock-infected mink cell line sample. This finding was consistent with the polyomavirus being derived from the original badger lung tissue and not the mink cell line or cell culture reagents.

Virus discovery in wild badgers

Fetal calf serum, which was used here throughout cell culturing, is the source of at least one species of polyomavirus and other species have been isolated from bovine tissue (Parry *et al.*, 1983; Peretti *et al.*, 2015; Schuurman *et al.*, 1991; Zhang *et al.*, 2014). To establish conclusively that Mm1PyV1 occurs in badgers, uncultured samples from wild badgers in France were tested in an independent laboratory for the presence of the virus. Diagnostic PCRs targeting VP1 and the NCCR were performed on faecal samples and seven different

tissues from 11 different badgers to determine prevalence. Eight badgers (73%) showed evidence of polyomavirus infection. PCR results were positive at least once in all sample types, except faecal samples which were all negative (Tables S1 and S2, available in the online Supplementary Material). Overall, 21 of 87 (24%) samples were positive for Mm1PyV1. Samples from liver and lymph nodes exhibited the highest rate of positivity. Neither VP1 nor the NCCR showed any sequence variability across the wild badger samples. Long-distance nested PCRs were used on one sample to generate an amplified genome, which was then sequenced. This generated a further polyomavirus isolate, referred to here as the Mm1PyV1 French isolate (GenBank accession number KP644238).

Genome characterization

Full-length comparison of the French and UK isolates showed that they differ by 34 single nucleotide polymorphisms (SNPs) and one indel. Thirty of the 34 SNPs

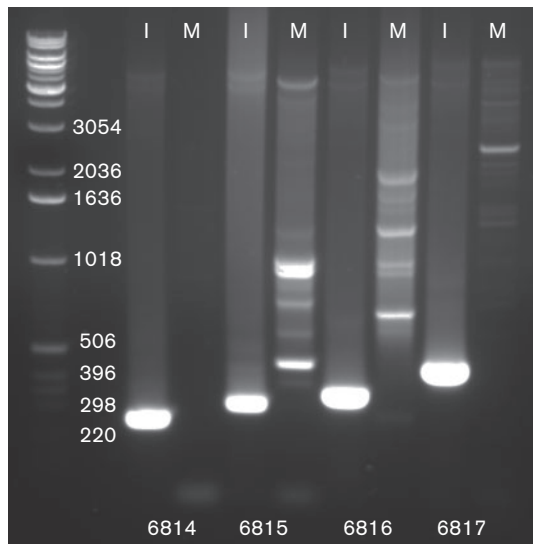


Fig. 3. Gel electrophoresis of mink cell line NBL-7. Lanes: I, infected with UK isolate cell line; M, mock infected. Numbers below the lanes represent the primers used. Numbers on the left represent the ladder fragment size (bp). Controls not shown.

fell in coding regions, four of which were non-synonymous (three in VP1 and one in LT-Ag) (Ser1770Thr, Ile2520Val, Ala2557Gly and Ala3044Pro; nucleotide positions given relative to the MmelPyV1 UK isolate genome). (Comparison to the type species SV40 is inappropriate here due to sequence length variation between MmelPyV1 and SV40 in VP1.) The single indel occurred in the NCCR.

The divergence time of the UK and French isolates was estimated based on previously published rates of polyomavirus nucleotide substitution. Assuming a substitution rate in coding regions of 4.34×10^{-5} substitutions per site per year [95% confidence interval (CI) $2.42\text{--}6.41 \times 10^{-5}$ substitutions per site per year] (Firth *et al.*, 2010), our observation of 30 SNPs in 4791 bp of coding sequence suggests that the French and UK isolates diverged within the past few hundred years (mean 144 years ago; 95% CI 97–260 years ago). This estimate should be considered tentative as it is based on an evolutionary rate for the BK polyomavirus (Firth *et al.*, 2010) and the rate for MmelPyV1 may be different.

The MmelPyV1 genome is a 5186 or 5187 bp circular molecule with a G+C content of 42.8%. These values fall within the range for known mammalian polyomaviruses (length 4697–5722 bp; G+C content 36.4–47.3%). The genome organization of MmelPyV1 was typical of the family *Polyomaviridae*, encoding ORFs from both strands of the genome separated by the NCCR (Fig. 2). ORFs encoding known viral proteins were conserved between the MmelPyV1 isolates and their closest relative, CSLPyV1 (Fig. 4). Proteins homologous to polyomavirus VP1, VP2, VP3, LT-Ag and St-Ag are present.

The LT-Ag sequence shares many features with other polyomaviruses, including a DnaJ domain (An *et al.*, 2012; Pipas, 1992) (including HPDKGG at aa 42–47; amino acid positions given relative to SV40), a Rb binding motif (An *et al.*, 2012; Pipas, 1992) (LRCDE at aa 103–107), ATPase motifs (Pipas, 1992) (GPINSGKT at aa 426–433 and GCVKVNLE at aa 503–510), a zinc finger motif (An *et al.*, 2012; Pipas, 1992) (CMDCLLEEQIITHYKYH at aa 302–317), a TPPK motif (DeCaprio & Garcea, 2013) (TPPK at aa 124–127), a Bub-1 motif (An *et al.*, 2012) (WERWW at aa 91–95) and a Cr1 domain (Pipas, 1992) (LMQLL at aa 13–17). Other motifs identified in SV40 appear to be absent (Cul-7 FNXEX; Ehlers & Moens, 2014). The St-Ag has motifs similar to other polyomaviruses, including two conserved PP2A binding motifs (Ehlers & Moens, 2014) (CQRNVNPKRCRCLMCRKRKH at aa 103–122 and WGMCYCYSCYQW at aa 135–148).

The NCCR contained three likely LT-Ag binding sites (An *et al.*, 2012; Pipas, 1992): one GAGGC at nt 275–279 and two reverse complement GCCTC at nt 77–81 and 84–88 (nucleotide positions relative to the UK isolate). Analyses using EMBOSS (Rice *et al.*, 2000) showed no evidence of tandem repeats or palindromes in the NCCR.

Neither ALTO nor agnoprotein genes were identified in MmelPyV1. A 53 aa ORF was identified overprinting the LT-Ag exon 2 in a frameshift position, in a similar genome position to the previously proposed ALTO gene (Carter *et al.*, 2013). However, MmelPyV1 is phylogenetically distinct from the clade of ALTO-containing polyomaviruses (Fig. 1) and the ORF is considerably shorter than identified ALTO proteins (53 aa compared with 248–250 aa; Carter *et al.*, 2013). Consequently, this ORF is unlikely to represent an ALTO gene. After searching for ORFs that were (i) conserved between the two MmelPyV1 isolate genomes, (ii) >30 aa long (corresponding to the shortest proposed current agnoprotein; Misra *et al.*, 2009) and (iii) located in the forward direction upstream of the VP2 start codon, we found no evidence of an agnoprotein ORF. We discovered one ORF >100 aa long in MmelPyV1 (location indicated by grey box in Fig. 4). This ORF does not have sequence identity with any known polyomavirus proteins, and is not conserved between MmelPyV1 and the closely related CSLPyV1, so is unlikely to represent a protein-encoding ORF.

It is unknown whether MmelPyV1 causes disease. The badger from which the UK isolate was derived was malnourished, harboured MusHV-1 (a virus found in most badgers; King *et al.*, 2004), and exhibited non-specific symptoms of inflammation and lesions in the kidneys, liver and lungs (Banks *et al.*, 2002). The badgers from which the French isolate was obtained showed no obvious disease symptoms when examined by a veterinarian.

Phylogenetic analysis of the family *Polyomaviridae*

To investigate the evolutionary history of the mammalian polyomaviruses, we constructed maximum-likelihood

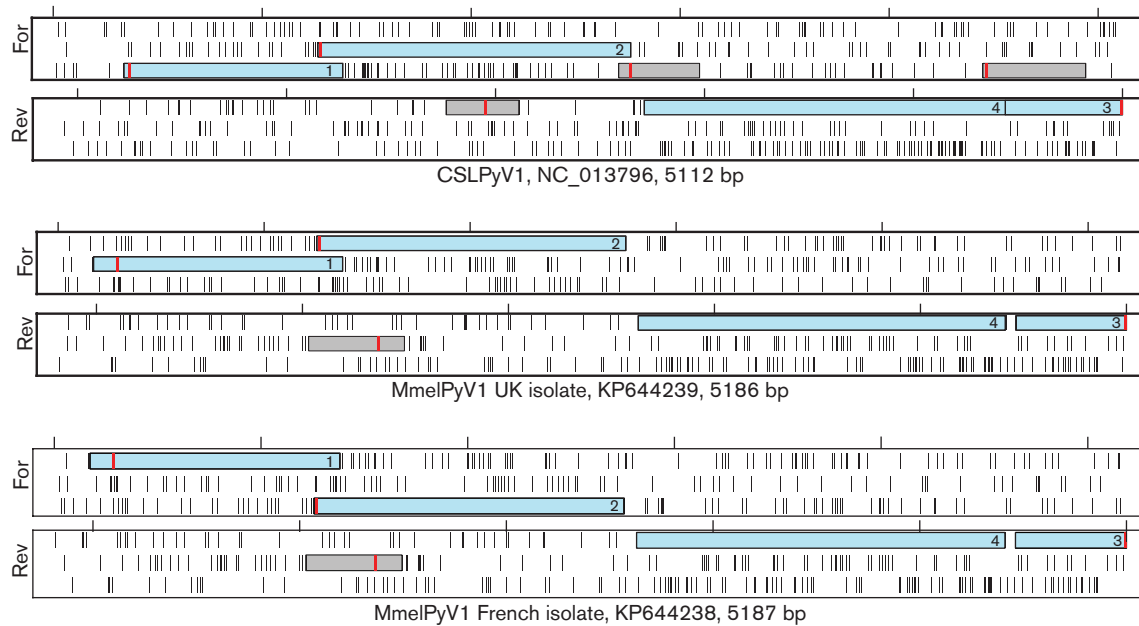


Fig. 4. Conserved ORFs of the two MmelPyV1 isolates and their nearest relative (CSLPyV1). All six translation frames are shown; the three forward (For) and reverse (Rev) reading frames are shown in separate boxes. All ORFs >100 aa are shown as boxes, with the first methionine marked in red. Solid vertical lines indicate stop codons and tick marks indicate 1000 bp markers. Blue coloured ORFs represent probable protein-coding ORFs identifiable by numbering: 1, VP2 and VP3; 2, VP1; 3, St-Ag and exon 1 of LT-Ag; 4, exon 2 of LT-Ag. Details of GenBank accession numbers and genome length are also included.

phylogenetic trees based on an alignment of 63 whole-genome polyomavirus sequences (obtained from GenBank in August 2014). In a maximum-likelihood phylogeny of polyomavirus genomes (Fig. 1), MmelPyV1 clustered with CSLPyV1 (Wellehan *et al.*, 2011). This result was supported by 100% bootstrap support in all phylogenies we estimated. In the genome regions retained in the genome-wide alignment (but including all three codon positions) (Fig. 2), MmelPyV1 and CSLPyV1 shared 72% nucleotide sequence identity. Based on a longer alignment of only the badger isolates and CSLPyV1 genomes across all coding regions, the estimated genetic *p* distance between the French and UK badger isolates was 0.006 substitutions per site, and between each MmelPyV1 isolate and the CSLPyV1 was 0.334 and 0.335 substitutions per site, respectively.

The maximum-likelihood phylogeny based on all known mammalian polyomavirus genomes comprises two well-supported clades: a smaller clade containing the KI, WU, HPyV6 and HPyV7 human polyomaviruses, and a larger clade containing a broader range of species (Fig. 1). MmelPyV1 and CSLPyV1 are clustered with monkey, bat, ape, rodent and dolphin viruses, amongst other hosts. This cluster is distinct from the Almipolyomavirus group that is defined by the presence of the ALTO gene (Carter *et al.*, 2013) (indicated by a yellow star in Fig. 1).

The division into these two clades is not preserved when different parts of the genome are considered separately. Phylogenies estimated from overlapping 600 bp partitions

of the genome-wide alignment (not shown) show that the WU, KI, HPyV6 and HPyV7 polyomaviruses, which form a strongly supported monophyletic clade in Fig. 1, are not consistently placed together when different sub-genomic regions are analysed. Such phylogenetic incongruence may indicate that these taxa have recombined in their evolutionary history. Analysis using Recco indicated statistical support for recombination of HPyV6, HPyV7, WU and KI, and also for the *Miniopertus* polyomavirus (small black star in Fig. 5) (alignment *P* values <0.001). Recco also reported recombination in the equine polyomavirus lineage, although this was less well supported (*P*=0.026).

The putative breakpoints in the circular genomes of WU, KI, HPyV6 and HPyV7 were identified as occurring in the NCCR and towards the 3' end of the LT-Ag exon 2 (breakpoints marked on Fig. 2 with black stars). These breakpoint positions were used to define two partitions of the genome-wide alignment: (i) partition A, comprising the late region (~1000 nt) plus a short region of the LT-Ag exon 2 (~360 nt), and (ii) partition B, comprising the majority of the early region (~1000 nt of the genome-wide alignment). The phylogeny estimated from partition A (Fig. 5a) shows the WU, KI, HPyV6 and HPyV7 polyomaviruses to be a well-supported monophyletic cluster, placed as a sister group to the remainder of the mammalian polyomaviruses. The phylogeny estimated from partition B (Fig. 5b) does not support this cluster. Instead, WU and KI group together in a separate clade (red in Fig. 5b), and

Fig. 5. Maximum-likelihood phylogenies produced from different regions of the genome. (a) Partition A, comprising the late region plus a short region of LT-Ag exon 2. (b) Partition B, comprising the majority of the early region. Red and blue coloured branches indicate the major two clades as identified in the genome-wide alignment (Fig. 1). Green branches indicate human polyomaviruses within inconsistent phylogenetic locations. The *Miniopterus* polyomavirus is marked with a star to indicate identification in Recco as a further possible recombinant. Scale bar represents expected number of substitutions per site (first and second codon positions only).

HPyV6 and HPyV7 group together in a separate clade (blue in Fig. 5b), in both instances with a bootstrap support of 93%.

Visual inspection of the mammalian polyomavirus phylogenies indicates that viruses isolated from the same host family or order are not always clustered together, yet closely related virus species are often from similar hosts (Fig. 1). The program BaTS (Parker *et al.*, 2008) was used to quantify the statistical support for an association between the virus phylogeny and host species taxonomy. Host species were assigned to categories in three different ways (Table S3). In each case there was a significant correlation between host taxonomic category and viral phylogeny (Table 1). Thus, polyomaviruses from related hosts are grouped together in the viral phylogenetic tree more often than expected by chance (Table 1).

DISCUSSION

Under the species definition proposed by Johne *et al.* (2011), the polyomavirus reported here represents a novel species and we suggest the name ‘*Meles meles* polyomavirus 1’. To our knowledge, MmelPyV1 represents the first complete polyomavirus species found in the family Mustelidae. A metagenomic analysis of ferret faeces conducted by Smits *et al.* (2013) detected DNA fragments with sequence similarity to polyomaviruses in two ferrets, hinting that polyomaviruses may occur in mustelids (Smits *et al.*, 2013) – a hypothesis that has been confirmed here. Following the discoveries of the California sea lion and raccoon polyomaviruses (Dela Cruz *et al.*, 2013; Wellehan *et al.*, 2011), MmelPyV1 represents the third fully described polyomavirus species in carnivorans and the first to be described in a European carnivoran.

The closest relative of MmelPyV1 is CSLPyV1 (Wellehan *et al.*, 2011). Of all known polyomavirus hosts, the sea lion is the most closely related to the European badger. Analyses

using BaTS (Parker *et al.*, 2008) showed significant support for preferential clustering of polyomaviruses from more similar hosts (Table 1). Although previous analyses have ruled out strict viral-host co-divergence during polyomavirus evolution (Krumbholz *et al.*, 2009; Tao *et al.*, 2013; Warden & Lacey, 2012), this is not inconsistent with the fact that more phylogenetically related polyomaviruses are isolated from more similar hosts than expected by chance. Instead, the evolutionary history of cross-species transmission in the polyomaviruses is likely to result from the interplay of multiple factors, such as preferential host switching amongst related hosts, viral–host co-divergence and host species sympatry (e.g. Charleston & Robertson, 2002; Streicker *et al.*, 2010). Understanding the evolutionary history is complicated by the fact that some polyomaviruses have been isolated from captive animals rather than from animals living within their natural host ranges (e.g. Scuda *et al.*, 2013). It is possible that animals living in artificially close proximity may share pathogens more frequently than in the wild. Furthermore, such cross-species transmissions may be more likely to cause disease symptoms that are subsequently investigated and identified as caused by novel viruses.

Using a previously published nucleotide substitution rate for BK polyomaviruses (Firth *et al.*, 2010), we estimate that the two MmelPyV1 isolates diverged within the past few hundred years. The polyomavirus NCCR is the most variable genome region, both within and across virus species (De Gascun & Carr, 2013). The recent divergence of the French and UK isolates is supported by a deficit of genetic differences in the NCCR. The substitution rate we employed is, to our knowledge, the only published rate derived from longitudinal sampling of polyomavirus genomes. There are caveats to the rate used here. First, Firth *et al.* (2010) noted that the BK polyomavirus dataset they analysed contained comparatively weak temporal structure, hence their estimated evolutionary rate may be poorly resolved. Second, BK polyomavirus is not closely

Table 1. Results of the BaTS tests for association of virus phylogeny with host species taxonomy

Host species grouping (no. of categories)	Mean association index	Mean parsimony score	P
Mammalian Order (8)	0.47	13.79	<0.01
Mammalian Family (19)	1.69	26.00	<0.01
Modified Order grouping* (10)	1.44	26.74	<0.01

*Host species are grouped according to mammalian order, except bats (split into Old/New World bats), primates (split into Old/New World monkeys and apes) and ungulates (grouped into a single category). See Table S3 for further details.

related to MmelPyV1 (especially in the late region; Fig. 5a) and the extrapolation of evolutionary rates across different viral species may lead to errors in estimates of divergence times.

Our phylogenetic analysis of all available genomes enabled us to explore the evolutionary relationships of the mammalian polyomaviruses. Four human polyomaviruses (HPyV6, HPyV7, WU and KI) were placed in different phylogenetic positions depending on the genome region analysed (Fig. 5). Although this result may be compatible with a recombinant origin for this clade, it is known that distinguishing between genetic signatures of recombination and lineage-specific variation in evolutionary rate is difficult. Tao *et al.* (2013) reported significant evolutionary rate heterogeneity between the early and late regions in the HPyV6, HPyV7, WU and KI polyomaviruses, but observed no such heterogeneity in other polyomaviruses (Tao *et al.*, 2013). The ability of such rate heterogeneity to generate apparent signatures of recombination has been noted for other viruses, including influenza virus (Worobey *et al.*, 2002). These observations, plus the observation that genetic distances and bootstrap support values are generally lower in the early region phylogeny (Fig. 5b), suggest that rate variation may be the more likely explanation for phylogenetic incongruence in the mammalian polyomaviruses.

Diagnostic PCRs to determine the presence of MmelPyV1 in different tissues suggested that the virus occurs at highest frequencies in the lymph nodes and the liver, and rarely in blood, faeces or kidney samples. In a study of primate polyomaviruses, viruses were also rarely present in these latter tissues (Scuda *et al.*, 2013). A previous metagenomic analysis of badger faeces by van den Brand *et al.* (2012) did not identify polyomavirus, supporting our suggestion that MmelPyV1 is rarely found in faeces. Future attempts to characterize MmelPyV1 or to discover new mammalian polyomaviruses should perhaps focus on solid-tissue samples rather than blood or faeces.

The known diversity of the *Polyomaviridae* has increased substantially due to increased virus discovery efforts, yet unanswered questions remain. The true host range and genomic diversity of mammalian polyomaviruses are unknown. Evolutionary rates for the family are unclear, and a molecular epidemiological investigation of polyomavirus transmission and diversity within a wild animal population has yet to be conducted. Furthermore, it is unknown why only some polyomaviruses are associated with disease and information about the aetiology of polyomaviruses is mostly limited to the viruses infecting humans. This problem is unlikely to be solved without systematic virus sampling in natural host populations, thereby avoiding sampling biases arising from the convenient sampling of easy-to-catch, high-profile or dead animals.

METHODS

Viral discovery and sequencing of the UK isolate. MmelPyV1 was first discovered in a cell culture supernatant, derived from the

apical lobe of a badger lung in Cornwall, England, in 1996 (Banks *et al.*, 2002). The culture was established using the mink cell line NBL-7 (ATC CCL 64) by Banks *et al.* (2002). The clarified supernatant of this cell culture had been stored at -80°C since 1997 (the '1997 supernatant').

The 1997 supernatant was enriched using a sucrose cushion according to published methods (Wang *et al.*, 2006) and DNA was subsequently extracted. Randomly primed amplification (REPLI-g UltraFast reagents; Qiagen) was used to amplify all sample DNA. This method provides robust amplification of all DNA in preparation for deep sequencing, but also works particularly well at amplifying small circular templates such as the polyomavirus through rolling circle replication (reviewed by Johnne *et al.*, 2009). This material was sheared using a Covaris sonicator, and standard Illumina paired-end libraries were prepared and sequenced on an Illumina MiSeq. The short read sequence data (1×10^7 , 149 nt reads) were filtered to remove low-quality reads using QUASR (Watson *et al.*, 2013) and reads mapping to mink repetitive sequences.

De novo assembly with Velvet version 1.2.7 and VelvetOptimizer-2.2.0 was used to assemble larger sequence contigs (Zerbino, 2010; Zerbino & Birney, 2008). The resulting contigs were processed using SLIM, an iterative BLAST algorithm (Cotten *et al.*, 2014) to identify viral sequences.

The 1997 supernatant from which the MmelPyV UK isolate was discovered was recultured to increase virus stock using mink cell line NBL-7. Following lysis by freeze/thawing to release any remaining cell-associated virus, a new supernatant (henceforth called 'MusHV1') was produced by centrifugation. To establish whether the polyomavirus was of badger or mink cell line origin, cells from the mink cell line NBL-7 were cultured in minimum essential medium (10% FCS/1% penicillin/streptomycin/glutamine), and were either infected with MusHV1 or mock infected. Infection was allowed to proceed until extensive cytopathic effect of the MusHV1-infected sample. Cells and supernatant were harvested, centrifuged (1200 r.p.m., 5 min) and the cell pellet was resuspended. DNA was extracted using a Qiagen DNA Blood and Tissue kit, with lysates being passed through a QiaShredder column to reduce viscosity prior to loading onto the DNA binding columns. DNA was amplified using Invitrogen Accuprime *Taq* and primers listed in Table S4. The following reaction conditions were used: 94°C for 5 min; 35 cycles of denaturation (94°C , 30 s), annealing (58°C , 30 s) and elongation (68°C , 5 min); final elongation at 68°C for 10 min. DNA products were resolved by agarose gel electrophoresis and visualized using ethidium bromide staining.

Sequencing of the French isolate. Wild badger tissue samples were tested for the polyomavirus in an independent laboratory. Between March and May 2013, 11 badgers were captured in the Chavigny commune, France, as part of a study into a vaccine for *Mycobacterium bovis*. Following capture, badgers were maintained together in open-air pens. Ethical approval was given by ANSES/ENVA/UPEC, which is registered by the French National Working Group on Animal Ethics (CNREEA). Legal permissions were obtained from the French Ministry of Higher Education and Research (reference 00611.02) and from regional governmental committees [Direction Départementale des Territoires de Meurthe-et-Moselle, Direction Départementale de la Protection des Populations de Meurthe-et-Moselle (reference C54-431-1) and Direction Départementale de la Protection des Populations d'Indre-et-Loire (reference C 37-175-3)].

DNA was extracted from blood samples using QIAamp kits (Qiagen). Tissue samples were extracted according to published methods (Courcoul *et al.*, 2014; Hénault *et al.*, 2006). Diagnostic PCRs targeting VP1 and the NCCR were performed in eight sample types from 11 different badgers (Tables S1 and S2). A template of 250 ng

organ or whole-blood DNA, or 5 µl DNA extracted from faeces, was added to the PCR master mix (AmpliTaq Gold reagents; Applied Biosystems) (primers listed in Table S4). PCRs were completed under the following conditions: 95 °C for 12 min; 45 cycles of denaturation (95 °C, 30 s), annealing (60 °C, 30 s) and elongation (72 °C, 2 min); final elongation at 72 °C for 10 min. All PCR products of expected sizes were purified and sequenced using a BigDye Terminator Cycle Sequencing kit (Applied Biosystems) on a 377 Automated DNA Sequencer (Applied Biosystems). Bases were called using Sequencing Analysis version 5.4 (Applied Biosystems) software and the resulting files analysed using Geneious version 7.1.4 (Geneious).

One of the DNA samples positive for both the NCCR and VP1 (salivary gland of badger 6) was chosen for genome amplification. Long-distance nested PCR was performed using the TaKaRa-EX PCR system (TaKaRa Bio) and the primers listed in Table S4. The amplicons were sequenced and analysed as above, producing a full-length genome. To validate the genome, independent overlapping PCRs were completed and the new amplicons resequenced (Table S4).

EMBOSS was used to search for tandem repeats and palindromes in the NCCR (Rice *et al.*, 2000). Settings were: palindromes, minimum length 10 and 20 (one and three mismatches, respectively); tandem repeats, maximum repeat size 600, threshold score 12.

Phylogenetic analyses. To construct an alignment of polyomaviruses, 63 whole-genome polyomavirus sequences were obtained from GenBank (accessed August 2014), including reference sequences from all known mammalian polyomavirus species and putative novel mammalian polyomavirus species. Avian polyomaviruses were too divergent to be aligned at the whole-genome level and were excluded. The two MmelPyV1 isolates were nearly identical, varying at 35 sites (0.67% of the genome), so only the UK isolate was included in the phylogenetic study.

Nucleotide sequences were aligned using Muscle version 3.8.31 with default parameters (Edgar, 2004), which generated a comparatively poor-quality alignment that required manual editing using Se-AL (<http://tree.bio.ed.ac.uk/software/seal/>). Regions where the alignment was highly uncertain (including non-coding regions) were removed, resulting in a 'genome-wide' alignment of 3534 bp that covered ~70% of the MmelPyV1 polyomavirus genome length (regions retained are shown in Fig. 2). Two further alignments were derived from the genome-wide alignment: (i) an 'early-gene' alignment (2034 nt) and (ii) a 'late-gene' alignment (1500 nt). Third codon positions were found to be significantly saturated using the test of Xia *et al.* (2003) as implemented in DAMBE (Xia & Lemey, 2009; Xia & Xie, 2001; Xia *et al.*, 2003) and were removed.

MEGA6 (Tamura *et al.*, 2013) was used to test nucleotide substitution models. For both early and late regions, the best-fit model was a general time reversible (GTR) model of nucleotide substitution, with gamma (Γ)-distributed among-site rate heterogeneity and a category of invariant sites (GTR+Γ+I). Phylogenetic trees were estimated using maximum-likelihood as implemented in Garli-2.01 (Zwickl, 2006) from the genome-wide, early-gene and late-gene alignments outlined above. Ten heuristic searches for the maximum-likelihood tree were repeated for each alignment and the tree with highest likelihood retained. One thousand maximum-likelihood bootstrap replicates were performed and the support for each node annotated onto the maximum-likelihood tree using the SumTrees functionality of DendroPy version 3.12.0 (Sukumaran & Holder, 2010).

MrBayes version 3.2.2 (Ronquist & Huelsenbeck, 2003) was used to generate a posterior sample of trees from the genome-wide alignment, using a Markov chain Monte Carlo of 3×10^6 steps. The first 25% of trees were discarded as 'burn-in'. Trait values that represented the taxonomic group of the host species of each viral sequence were assigned to the phylogenetic tree tips (Table S3). The program BaTS

was used (with 200 randomizations of assigned trait values performed to define a null distribution) to statistically test whether viruses found in taxonomically related hosts clustered in the virus phylogeny (Parker *et al.*, 2008).

To assess the presence of phylogenetic signal for recombination amongst polyomaviruses, the genome-wide alignment was divided into 12 overlapping subgenomic partitions. Starting from position 1, each partition was 600 nt long with a neighbouring partition overlap of 200 nt. Phylogenetic trees were estimated from each partition using the maximum-likelihood approach detailed above (except that two heuristic searches and 200 bootstrap replicates were performed for each partition). Visual inspection of these trees revealed phylogenetic inconsistencies between genome regions. Subsequently, one species was chosen randomly from every monophyletic clade that was present in all 12 partition phylogenies. Sixteen clades that were well supported in the whole-genome phylogeny were identified (15 clades supported by bootstrap support values of 100% and one clade by 91%). Ten species that did not belong to any clade were also included in the analysis, resulting in a reduced alignment of 26 polyomavirus species. These 26 taxa were analysed using Recco (Maydt & Lengauer, 2006) to seek evidence of recombination. Recco attempts to find the minimum cost of reconstructing each sequence in an alignment from a combination of recombination and mutation of the other sequences, where recombination and mutation have a cost penalty (Maydt & Lengauer, 2006). Probable recombination breakpoints were generated based on the minimum-cost solutions. To further investigate potential recombinants, putative recombination breakpoints were used to define new genomic partitions and maximum-likelihood phylogenies were again estimated from each partition.

To determine the divergence of the MmelPyV1 isolate genomes (UK and French), we created a whole-genome alignment containing only the two isolates plus the closest related viral species (CSLPyV1) as an outgroup. All three genome positions and almost all regions except the NCCR could be retained (4812 sites). Pairwise genetic distances were calculated using MEGA6 (Tamura *et al.*, 2013). The date of divergence of the UK and French MmelPyV1 isolates was estimated using previous estimates of polyomavirus substitution rates (Firth *et al.*, 2010).

ACKNOWLEDGEMENTS

We thank Malcolm Banks for providing the Cornish supernatant sample. We also thank Pierre Sarradin, head of the experimental platform at INRA (Tours, France), where the badger post-mortem was conducted, Jacques Barrat, head of the experimental station at ANSES (Nancy, France), where badgers were initially housed in captivity, Michel Munier and Nicolas Penel, ANSES (Nancy, France), for animal husbandry, and Jean-Baptiste Schwyer, naturalist, and the staff from ANSES (Nancy, France) and ELIZ (Malzeville, France) for trapping the badgers. We thank Matthew Emery for his initial help in preparation of the DNA library used for next-generation sequencing, the Plymouth University Systems Biology Centre, and Sonja Liebmann and Cornelia Walter for excellent technical assistance. S.C.H. was supported by the Wellcome Trust (grant 102427). P.K., M.C. and A.L.P. were supported by the Wellcome Trust (grant 098051). The support of the European Community's Seventh Framework Programme (FP7/2007–2013) under the project EMPERIE (European Community grant agreement 223498) is acknowledged. These studies were partially supported by funds from the School of Biomedical and Healthcare Sciences, Plymouth University and by the Seale-Hayne Educational Trust. O.G.P. received support for this work from the John Fell OUP Fund.

REFERENCES

- Am, P., Sáenz Robles, M. T. & Pipas, J. M. (2012). Large T antigens of polyomaviruses: amazing molecular machines. *Annu Rev Microbiol* **66**, 213–236.
- Baker, T. S., Drak, J. & Bina, M. (1989). The capsid of small papova viruses contains 72 pentameric capsomeres: direct evidence from cryo-electron-microscopy of simian virus 40. *Biophys J* **55**, 243–253.
- Banks, M., King, D. P., Daniells, C., Stagg, D. A. & Gavier-Widen, D. (2002). Partial characterization of a novel gammaherpesvirus isolated from a European badger (*Meles meles*). *J Gen Virol* **83**, 1325–1330.
- Carter, J. J., Daugherty, M. D., Qi, X., Bheda-Malge, A., Wipf, G. C., Robinson, K., Roman, A., Malik, H. S. & Galloway, D. A. (2013). Identification of an overprinting gene in Merkel cell polyomavirus provides evolutionary insight into the birth of viral genes. *Proc Natl Acad Sci U S A* **110**, 12744–12749.
- Charleston, M. A. & Robertson, D. L. (2002). Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. *Syst Biol* **51**, 528–535.
- Cotten, M., Oude Munnink, B., Canuti, M., Deijis, M., Watson, S. J., Kellam, P. & van der Hoek, L. (2014). Full genome virus detection in fecal samples using sensitive nucleic acid preparation, deep sequencing, and a novel iterative sequence classification algorithm. *PLoS One* **9**, e93269.
- Courcou, A., Moyen, J.-L., Brugère, L., Faye, S., Hénault, S., Gares, H. & Boschioli, M.-L. (2014). Estimation of sensitivity and specificity of bacteriology, histopathology and PCR for the confirmatory diagnosis of bovine tuberculosis using latent class analysis. *PLoS One* **9**, e90334.
- De Gascun, C. F. & Carr, M. J. (2013). Human polyomavirus reactivation: disease pathogenesis and treatment approaches. *Clin Dev Immunol* **2013**, 373579.
- DeCaprio, J. A. & Garcea, R. L. (2013). A cornucopia of human polyomaviruses. *Nat Rev Microbiol* **11**, 264–276.
- Dela Cruz, F. N., Jr, Giannitti, F., Li, L., Woods, L. W., Del Valle, L., Delwart, E. & Pesavento, P. A. (2013). Novel polyomavirus associated with brain tumors in free-ranging raccoons, western United States. *Emerg Infect Dis* **19**, 77–84.
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797.
- Ehlers, B. & Moens, U. (2014). Genome analysis of non-human primate polyomaviruses. *Infect Genet Evol* **26**, 283–294.
- Firth, C., Kitchen, A., Shapiro, B., Suchard, M. A., Holmes, E. C. & Rambaut, A. (2010). Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. *Mol Biol Evol* **27**, 2038–2051.
- Gerits, N. & Moens, U. (2012). Agnoprotein of mammalian polyomaviruses. *Virology* **432**, 316–326.
- Hénault, S., Karoui, C. & Boschioli, M. L. (2006). A PCR-based method for tuberculosis detection in wildlife. *Dev Biol (Basel)* **126**, 123–132.
- Imperiale, M. & Major, E. (2007). Polyomaviruses. In *Fields Virology*, 4th edn, pp. 2263–2298. Edited by D. Knipe, P. Howley, D. Griffen, R. Lamb, R. Martin, B. Roizman & S. Straus. Philadelphia, PA: Lippincott Williams & Wilkins.
- Johne, R., Müller, H., Rector, A., van Ranst, M. & Stevens, H. (2009). Rolling-circle amplification of viral DNA genomes using phi29 polymerase. *Trends Microbiol* **17**, 205–211.
- Johne, R., Buck, C. B., Allander, T., Atwood, W. J., Garcea, R. L., Imperiale, M. J., Major, E. O., Ramqvist, T. & Norkin, L. C. (2011). Taxonomical developments in the family *Polyomaviridae*. *Arch Virol* **156**, 1627–1634.
- King, D. R., Mutukwa, N., Lesellier, S., Cheeseman, C., Chambers, M. A. & Banks, M. (2004). Detection of mustelid herpesvirus-1 infected European badgers (*Meles meles*) in the British Isles. *J Wildl Dis* **40**, 99–102.
- Krumbholz, A., Bininda-Emonds, O. R. P., Wutzler, P. & Zell, R. (2009). Phylogenetics, evolution, and medical importance of polyomaviruses. *Infect Genet Evol* **9**, 784–799.
- Maydt, J. & Lengauer, T. (2006). Recco: recombination analysis using cost optimization. *Bioinformatics* **22**, 1064–1071.
- Misra, V., Dumonceaux, T., Dubois, J., Willis, C., Nadin-Davis, S., Severini, A., Wandeler, A., Lindsay, R. & Artsob, H. (2009). Detection of polyoma and corona viruses in bats of Canada. *J Gen Virol* **90**, 2015–2022.
- Parker, J., Rambaut, A. & Pybus, O. G. (2008). Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect Genet Evol* **8**, 239–246.
- Parrish, C. R., Holmes, E. C., Morens, D. M., Park, E.-C., Burke, D. S., Calisher, C. H., Laughlin, C. A., Saif, L. J. & Daszak, P. (2008). Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiol Mol Biol Rev* **72**, 457–470.
- Parry, J. V., Lucas, M. H., Richmond, J. E. & Gardner, S. D. (1983). Evidence for a bovine origin of the polyomavirus detected in foetal rhesus monkey kidney cells, FRhK-4 and -6. *Arch Virol* **78**, 151–165.
- Peretti, A., FitzGerald, P. C., Bliskovsky, V., Buck, C. B. & Pastrana, D. V. (2015). Hamburger polyomaviruses. *J Gen Virol*.
- Pérez-Losada, M., Christensen, R. G., McClellan, D. A., Adams, B. J., Viscidi, R. P., Demma, J. C. & Crandall, K. A. (2006). Comparing phylogenetic codivergence between polyomaviruses and their hosts. *J Virol* **80**, 5663–5669.
- Pipas, J. M. (1992). Common and unique features of T antigens encoded by the polyomavirus group. *J Virol* **66**, 3979–3985.
- Rice, P., Longden, I. & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276–277.
- Ronquist, F. & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574.
- Schuurman, R., van Steenis, B., van Strien, A., van der Noordaa, J. & Sol, C. (1991). Frequent detection of bovine polyomavirus in commercial batches of calf serum by using the polymerase chain reaction. *J Gen Virol* **72**, 2739–2745.
- Scuda, N., Madinda, N. F., Akoua-Koffi, C., Adjogoua, E. V., Wevers, D., Hofmann, J., Cameron, K. N., Leendertz, S. A. J., Couacy-Hymann, E. & other authors (2013). Novel polyomaviruses of nonhuman primates: genetic and serological predictors for the existence of multiple unknown polyomaviruses within the human population. *PLoS Pathog* **9**, e1003429.
- Shadan, F. F. & Villarreal, L. P. (1993). Coevolution of persistently infecting small DNA viruses and their hosts linked to host-interactive regulatory domains. *Proc Natl Acad Sci U S A* **90**, 4117–4121.
- Smits, S. L., Raj, V. S., Oduber, M. D., Schapendonk, C. M. E., Bodewes, R., Provacía, L., Stittelaar, K. J., Osterhaus, A. D. M. E. & Haagmans, B. L. (2013). Metagenomic analysis of the ferret fecal viral flora. *PLoS ONE* **8**, e71595.
- Stevens, H., Bertelsen, M. F., Sijmons, S., Van Ranst, M. & Maes, P. (2013). Characterization of a novel polyomavirus isolated from a fibroma on the trunk of an African elephant (*Loxodonta africana*). *PLoS One* **8**, e77884.
- Streicker, D. G., Turmelle, A. S., Vonhof, M. J., Kuzmin, I. V., McCracken, G. F. & Rupprecht, C. E. (2010). Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. *Science* **329**, 676–679.

- Sukumaran, J. & Holder, M. T. (2010).** DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**, 1569–1571.
- Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. (2013).** MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **30**, 2725–2729.
- Tao, Y., Shi, M., Conrardy, C., Kuzmin, I. V., Recuenco, S., Agwanda, B., Alvarez, D. A., Ellison, J. A., Gilbert, A. T. & other authors (2013).** Discovery of diverse polyomaviruses in bats and the evolutionary history of the *Polyomaviridae*. *J Gen Virol* **94**, 738–748.
- van den Brand, J. M. A., van Leeuwen, M., Schapendonk, C. M., Simon, J. H., Haagmans, B. L., Osterhaus, A. D. M. E. & Smits, S. L. (2012).** Metagenomic analysis of the viral flora of pine marten and European badger feces. *J Virol* **86**, 2360–2365.
- Van Ghelue, M., Khan, M. T. H., Ehlers, B. & Moens, U. (2012).** Genome analysis of the new human polyomaviruses. *Rev Med Virol* **22**, 354–377.
- Wang, J., Horner, G. W. & O’Keefe, J. S. (2006).** Genetic characterisation of bovine herpesvirus 1 in New Zealand. *N Z Vet J* **54**, 61–66.
- Warden, C. D. & Lacey, S. F. (2012).** Updated phylogenetic analysis of polyomavirus–host co-evolution. *J Bioinformatics Res* **1**, 46–49.
- Watson, S. J., Welkers, M. R. A., Depledge, D. P., Coulter, E., Breuer, J. M., de Jong, M. D. & Kellam, P. (2013).** Viral population analysis and minority-variant detection using short read next-generation sequencing. *Philos Trans R Soc Lond B Biol Sci* **368**, 20120205.
- Wellehan, J. F. X., Jr, Rivera, R., Archer, L. L., Benham, C., Muller, J. K., Colegrove, K. M., Gulland, F. M. D., St Leger, J. A., Venn-Watson, S. K. & Nollens, H. H. (2011).** Characterization of California sea lion polyomavirus 1: expansion of the known host range of the Polyomaviridae to Carnivora. *Infect Genet Evol* **11**, 987–996.
- Worobey, M., Rambaut, A., Pybus, O. G. & Robertson, D. L. (2002).** Questioning the evidence for genetic recombination in the 1918 “Spanish flu” virus. *Science* **296**, 211.
- Xia, X. & Lemey, P. (2009).** Assessing substitution saturation with DAMBE. In *Phylogenetic Handbook*, 2nd edn, pp. 615–630. Edited by P. Lemey, M. Salemi & A. M. Vandamme. Cambridge: Cambridge University Press.
- Xia, X. & Xie, Z. (2001).** DAMBE: software package for data analysis in molecular biology and evolution. *J Hered* **92**, 371–373.
- Xia, X., Xie, Z., Salemi, M., Chen, L. & Wang, Y. (2003).** An index of substitution saturation and its application. *Mol Phylogenet Evol* **26**, 1–7.
- Zerbino, D. R. (2010).** Using the Velvet *de novo* assembler for short-read sequencing technologies. *Curr Protoc Bioinformatics* **Chapter 11**, Unit 11. 5.
- Zerbino, D. R. & Birney, E. (2008).** Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* **18**, 821–829.
- Zhang, W., Li, L., Deng, X., Kapusinszky, B. & Delwart, E. (2014).** What is for dinner? Viral metagenomics of US store bought beef, pork, and chicken. *Virology* **468–470**, 303–310.
- Zwickl, J. (2006).** *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. Thesis, University of Texas at Austin, Austin, TX, USA.