4800–4813 *Nucleic Acids Research, 2015, Vol. 43, No. 10* *Published online 22 April 2015*
doi: 10.1093/nar/gkv378

Interspecific adaptation by binary choice at *de novo* polyomavirus T antigen site through accelerated codon-constrained Val-Ala toggling within an intrinsically disordered region

Chris Lauber[1,2,†], Siamaque Kazem[1,†], Alexander A. Kravchenko[3,‡], Mariet C.W. Feltkamp[1] and Alexander E. Gorbalenya[1,3,4,*]

[1]Department of Medical Microbiology, Leiden University Medical Center, 2300-RC Leiden, The Netherlands, [2]Institute for Medical Informatics and Biometry, Technische Universität Dresden, 01307 Dresden, Germany, [3]Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, 119899 Moscow, Russia and [4]Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, 119899 Moscow, Russia

Received September 12, 2014; Revised March 22, 2015; Accepted April 9, 2015

ABSTRACT

It is common knowledge that conserved residues evolve slowly. We challenge generality of this central tenet of molecular biology by describing the fast evolution of a conserved nucleotide position that is located in the overlap of two open reading frames (ORFs) of polyomaviruses. The *de novo* ORF is expressed through either the ALTO protein or the Middle T antigen (MT/ALTO), while the ancestral ORF encodes the N-terminal domain of helicase-containing Large T (LT) antigen. In the latter domain the conserved Cys codon of the LXCXE pRB-binding motif constrains codon evolution in the overlapping MT/ALTO ORF to a binary choice between Val and Ala codons, termed here as codon-constrained Val-Ala (COCO-VA) toggling. We found the rate of COCO-VA toggling to approach the speciation rate and to be significantly accelerated compared to the baseline rate of chance substitution in a large monophyletic lineage including all viruses encoding MT/ALTO and three others. Importantly, the COCO-VA site is located in a short linear motif (SLiM) of an intrinsically disordered region, a typical characteristic of adaptive responders. These findings provide evidence that the COCO-VA toggling is under positive selection in many polyomaviruses, implying its critical role in interspecific adaptation, which is unprecedented for conserved residues.

INTRODUCTION

Intrinsically disordered regions (IDRs), which are either not structured or may become structured upon interaction with diverse partners (1), have been identified in many proteins and implicated in various biological processes as adaptive responders (2–5). They have a biased amino acid residue composition and evolve faster than structured proteins (6,7), with exception of very small islands of relative conservation, known as short linear motifs (SLiM), that mediate protein–protein interactions (8).

IDRs are frequently encoded by overlapping open reading frames (ORFs) that evolved *de novo* by overprinting the ancestral ORFs (9–13) and are common in viruses (9–13). This overlapping of ORFs is accompanied by suppression of synonymous substitution rate in the ancestral ORFs (negative or purifying selection) compared to that of non-overlapping ORFs, indicative of codon constraints in the *de novo* ORFs due to their expression. The observed phenomenon has been extensively used for *in silico* identification of functional *de novo* ORFs (12,14–15), which often led to the elucidation of non-canonical expression mechanisms of these ORFs (e.g. (16–19)). Suppression of synonymous substitution rate is also reciprocally imposed on *de novo* ORFs by the overlapping ancestral gene. These observations led to analysis of relative rate change of substitutions in the *de novo* genes compared to ancestral or non-overlapping genes (12,20–23).

This ORF-wide analysis has not been extended to individual codons of the *de novo* ORFs due to formidable technical challenges. A common approach to characterize site-specific evolution is to estimate deviation from the substitution rate under a model of neutral evolution for each

[*]To whom correspondence should be addressed. Tel: +31 71 5261652; Fax: +31 71 5266761; Email: a.e.gorbalenya@lumc.nl
[†]These authors contributed equally to the paper as first authors.
[‡]Deceased; of blessed memory.

codon of an ORF. Suppression and acceleration of the substitution rate is attributed to negative and positive selection, respectively, with positive selection being seen as the hallmark signature of adaptation during intra-species evolution (24). One particular pattern of variation under positive selection is the frequent exchange of residues with pervasive return to the wild-type state, dubbed residue toggling (25). Identification of codons under selection, either negative or positive, is part of the established evolutionary-based pipeline that informs functional characterization of proteins encoded in non-overlapping ORFs (26,27). However, the available techniques were not developed to untangle selection forces acting on the overlapping ORFs, which constrain evolution of each other. This may explain the lack of identification of *de novo* codon(s) under positive selection, despite broad recognition of a prominent role that the overlapping ORFs play in adaptation of viruses to host (12).

One of the largest and poorly characterized pairs of proteins encoded by overlapping ORFs is expressed by members of the fast growing *Polyomaviridae* family (Supplementary Table S1). These viruses cause latent infections in diverse mammals and birds, and in humans, some of these viruses have been responsible for different pathologies in immunocompromised individuals (28,29). Polyomaviruses employ multi-ORF double-stranded DNA (dsDNA) genomes of ~5 kb (30,31). Genomes of a large subset of polyomaviruses include two overlapping ORFs (15,32–33), designated here ORF2 and ORF5 (Figure 1A; for other designations see Text S1 and Supplementary Table S2). ORF2 encodes the second exon of the large T antigen (LT) that includes the functionally important LXCXE pRB-motif in the ORF5 overlapping part (34–36) and a helicase domain in the non-overlapping part (30,37). ORF5 is expressed as a separate protein (ALTO) in Merkel cell polyomavirus (MCPyV) (15); while it encodes the second exon of Middle T antigen (MT) in murine and hamster polyomaviruses (MPyV and HaPyV) (33,38–39). The ORF5-encoded part of MT antigen is implicated in control of cell transformation (33,38–39), enriched with Pro residues (40,41) and includes a C-terminal transmembrane domain (38) that is essential for the oncogenic function of MT (42). This function and interaction of MT with different cellular proteins may be modulated by phosphorylation at several Ser, Thr and Tyr residues in rodent polyomaviruses (39). We will use ORF5-plus and ORF5-less to refer to respective subsets of polyomaviruses; ORF5-plus viruses are also known as Almipolyomaviruses (15). Likewise, and purely for the sake of uniformity, hereafter we have designated the ORF5-encoding product as MT/ALTO for all ORF5-plus polyomaviruses. Because ORF5 is conserved in only ORF5-plus polyomaviruses, while the overlapping part of ORF2 is found in all mammalian polyomaviruses (15), these ORFs are defined as *de novo* (ORF5) and ancestral (ORF2), according to Sabath *et al.*, (12). ORF5-plus viruses form a large monophyletic cluster in one of the main branches of polyomavirus tree (15), dubbed Orthopolyomaviruses I (Ortho-I); with three other branches being Orthopolyomaviruses II, Malawipolyomaviruses and Wukipolyomaviruses (43), although branch delineation and designation may vary in different studies (15,44).

To understand the evolution of overlapping ORFs, we studied ORF2 and ORF5 at codon resolution. We found that one of the most conserved ORF5 codons, located in a SLiM of ORF5, experienced an accelerated evolutionary rate despite being strongly constrained to two amino acids by the overlapping ancestral ORF2. Using available and specially developed evolutionary-based approaches we revealed an unprecedented frequent toggling between these two residues during large-scale multi-species evolution in the Ortho-I clade of polyomaviruses. This analysis is, to our knowledge, the first to identify a conserved position of *de novo* protein under positive selection. Its results suggest a new IDR-mediated adaptation mechanism employed by many mammalian polyomaviruses with potential relevance to understanding adaptation of other viruses and organisms.

MATERIALS AND METHODS

Datasets: viruses, sequences and alignments

Full-length genome sequences of 55 polyomaviruses available in the Genbank/RefSeq database on February 2013 (Supplementary Table S1) were downloaded into the Viralis platform (45). When several genomes per species were available, the RefSeq sequence was chosen for presentation. The Muscle (46) and ClustalW (47) programs were used to generate family-wide multiple amino-acid alignments for viral capsid protein (VP)1 encoded in ORF3, VP2 (ORF4) and LT (ORF2), followed by manual curation. For each of the three protein alignments, strongly conserved blocks (48) were extracted using the Blocks Accepting Gaps Generator (BAGG) tool (www.genebee.msu.su/~antonov/bagg/cgi/bagg.cgi) to produce a concatenated multiple sequence alignment used for phylogenetic reconstruction and other analyses (see below). The ORF2-wide alignment was also mapped on the genome sequences, which were then translated in the alternative reading frame (RF −3) encoding ORF5 in 22 viruses of the ORF5-plus group to produce an ORF5 alignment. ORF5 size varies from 441 nucleotides (nt) to 846 nt and ORF5 sequence conservation was detectable only in some subsets of polyomaviruses (Supplementary Table S3; data not shown, see also (15)).

For analysis of site-specific evolutionary selection by Datamonkey programs, we used 10 alignments of selected positions of the ORF5 and ORF2 (datasets, D1–D10). These 10 alignments represented different groups of viruses, including all mammalian polyomaviruses (D1–D2), Ortho-I viruses (D3–D4), ORF5-plus viruses (D5–D6), ORF5-less viruses (D7) and three non-overlapping lineages of ORF5-plus viruses (D8–D10), each analysed separately (see Supplementary Table S4 for details). Using conservation considerations, some codons of ORF5 and ORF2 were selected, so all datasets included ORF5 codons while D2, D4, D6 and D7 included also ORF2 codons. For ORF5 of D1–D7, those codons were chosen whose overlapping codon in ORF2 (−1 frame) was aligned with no gaps across mammalian polyomaviruses. For ORF5 of D8–D10 and ORF2, most conserved codons in respective alignments were used after manual pruning of weakly aligned codons.

Alignments of the conserved motifs in the N-terminal part of LT ORF2 and ORF5, partially described elsewhere

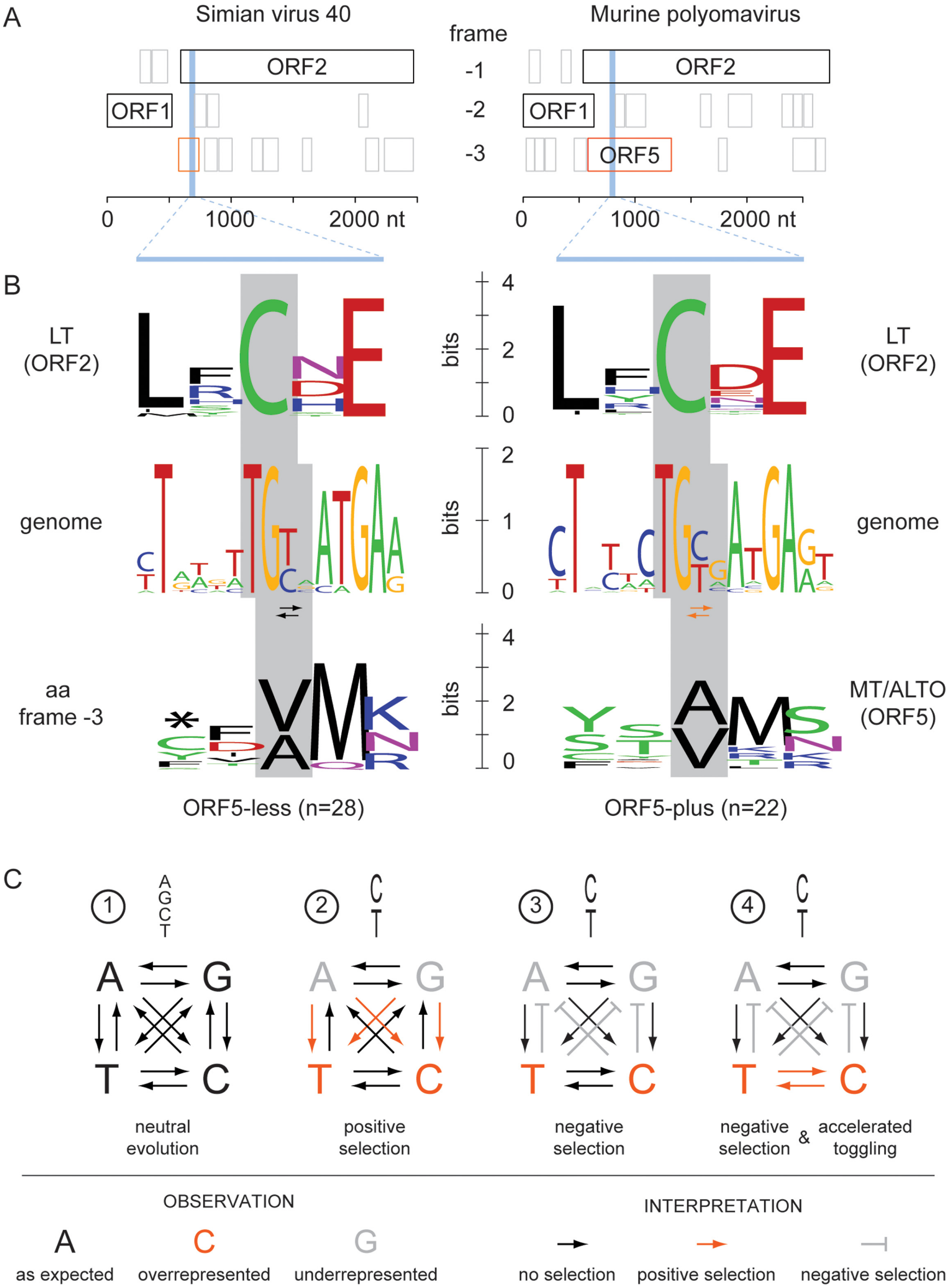


Figure 1. Toggling at the COCO-VA site in mammalian ORF5-plus and ORF5-less polyomaviruses. (**A**) ORF organization in three reading frames of the genomic region encoding the early genes is shown for Simian virus 40, SV-40 (left; NC_001669) and Murine polyomavirus, MPyV (right; Genbank accession NC_001515) representing ORF5-less and ORF5-plus polyomaviruses, respectively. The ORF2 frame was chosen as -1 frame for both viruses. ORF borders are defined here from stop to stop codon. Large expressed ORFs are boxed/outlined and named while other ORFs with a size of at least 75 nt are shown in grey. ORF5 of the ORF5-plus virus and one of its derivatives of the ORF5-less virus are highlighted in the 3 frame. The background highlighting indicates location of the LXCXE motif (an essential motif found in polyomaviruses and other viruses, and cellular proteins that mediates binding and inactivation of the cellular tumour-suppressor protein pRB (34–36)). (**B**) Shown are sequence logos of the LXCXE motif (top), the corresponding nucleotide sequence (middle) and the amino acid sequence translated from the ORF5 frame (bottom) for multiple alignment of the 28 ORF5-less (left) and 22 ORF5-plus viruses (right) analysed in this study using Viralis platform (45). The asterisk indicates stop codons in the −3 frame of some ORF5-less viruses. See 'Materials and Methods' section for other details. (**C**) Shown are four possible scenarios of evolution of a polynucleotide site under different selection regimes. In scenarios 2–4 different selection force(s) result in the same observed nucleotide diversity restricted to C or T. Scenarios 3 and 4 depict the COCO-VA toggling in ORF5-less and -plus viruses, respectively.

(15), were produced and converted into logos. To produce alignments as input for the RNAz program, we converted codon ORF5-based alignments of four subsets of ORF5-plus and two subsets of ORF5-less polyomaviruses, into the respective nucleotide alignments (Supplementary Table S3).

Phylogeny reconstruction

Phylogenetic analyses were performed by using a Bayesian approach implemented in BEAST version 1.7.4 (49) and the Whelan and Goldman (WAG) amino acid substitution matrix (50). Rate heterogeneity among sites was modelled using a gamma distribution with four categories, and a relaxed molecular-clock approach was tested against the strict molecular-clock approach (51) and was found to be superior. Markov chain Monte Carlo (MCMC) chains were run for 2 million steps and the first 10% were discarded as burn-in. Convergence of the runs was verified using the Tracer tool (http://beast.bio.ed.ac.uk/tracer).

Analysis of natural selection at codons

We have used Mixed Effects Model of Evolution (MEME) (52) and Fast, Unconstrained Bayesian AppRoximation (FUBAR) (27) at the Datamonkey website (http://www.datamonkey.org) (26) to test for natural selection at conserved ORF5 codons. In addition, we have screened for toggling at ORF5 residues using TOGGLE, an implementation of the residue toggling method developed for HIV-1 by Delport *et al.*, (25). We have analysed in total 10 different datasets, D1–D10 (see above), capturing different positions and virus diversities (see Supplementary Table S4). For each analysed dataset, selection of evolutionary model was performed automatically at the Datamonkey web site using default parameters prior to the analysis.

Analysis of COCO-VA toggling by BayesTraits

Evolution of non-synonymous replacements at the codon-constrained Val-Ala (COCO-VA) site of ORF5 was analysed by BayesTraits package using the Multistate model (http://www.evolution.rdg.ac.uk/BayesTraits.html) (53). This codon is constrained to encode either Ala or Val in all mammalian polyomaviruses due to the overlapping Cys codon of the LXCXE motif that is expressed in the LT ORF2 of these viruses. The analysed polyomaviruses were divided into two groups based on whether or not they express ORF5: ORF5-plus and ORF5-less viruses, respectively. The COCO-VA site is expressed as part of ORF5 in ORF5-plus, but not in ORF5-less viruses.

To test whether Ala-Val trait transitions are statistically more frequent in the ORF5-plus lineage compared to ORF5-less viruses, we applied the BayesTraits multistate model using a single trait (Ala/Val). We ran the analysis for three virus datasets: the combined set of mammalian polyomaviruses as well as separately for ORF5-plus and ORF5-less viruses, with respective posterior tree samples obtained through independent BEAST analyses. We then compared the estimated Ala-to-Val and Val-to-Ala transition rates between the three datasets, including an average Ala-Val exchange rate (corresponding to the toggling rate) by plotting the distributions. Statistical significance of differences in Ala-Val exchange rates was assessed using log Bayes Factors that was calculated with the R package Bayes Factor (http://bayesfactorpcl.r-forge.r-project.org/). As Ala-Val exchange is equivalent to T-C exchange at the second codon position of the COCO-VA codon (see Results and Discussion), we applied BayesTraits also to the third position of that codon as a control.

Statistical analyses of COCO-VA toggling using patristic distances

For each virus the smallest pairwise patristic distance (SPAT) to a virus encoding the same amino acid (monomorphic pairs: Ala↔Ala and Val↔Val; monoSPAT) and to that encoding the different amino acid (polymorphic pair: Ala↔Val, polySPAT) was calculated. Patristic distances were extracted from the polyomavirus phylogeny using the package Analyses of Phylogenetics and Evolution (APE) in R language (54).

We estimated the rate of COCO-VA toggling as the ratio of monoSPAT to the sum of polySPAT and monoSPAT values; designated SPAT ratio hereafter. Due to limited virus sampling at the intra-species level, we applied a sliding window approach to compare SPAT ratios between ORF5-plus and ORF5-less viruses. A window size of 0.15 and a shift of 0.05 at the monoSPAT scale were used. A two-sample non-parametric Mann–Whitney U test was utilized to test for statistically significant differences between the two virus groups within a particular window. A deviation of distributions of SPAT ratios from the average toggling rate of 0.5 was assessed using the Wilcoxon rank-sum test.

To independently assess the partitioning of mammalian polyomaviruses into ORF5-plus and ORF5-less virus groups, we determined the ranking of a predefined two-set partitioning among all possible two-set partitioning of the same type for the 30 viruses with monoSPAT values smaller than the derived threshold of 0.35. These 30 viruses comprise 14 ORF5-plus and 16 ORF5-less viruses or 16 Ortho-I and 14 non-Ortho-I viruses. We calculated the difference of mean toggling rate values between the two groups in each of these partitionings and determined its ranking among the differences of mean toggling rates obtained for all other 14–16 or 16–14 partitioning of the 30 viruses, whose total was 145,422,675 possible partitionings (e.g. combinations).

General bioinformatics analyses

For selected phylogenetic lineages, alignments of ORF5 were converted HMM profiles and compared to each other using HHsearch (55) in both local and global alignment modes.

Sequence logos of selected alignments were produced using the WebLogo server (56,57).

Secondary structure and disorder prediction of protein sequences were generated using the Disorder Prediction MetaServer, which reports consensus results of eight protein disorder predictor tools: DISEMBL (58), DISOPRED (59), DISpro (60), FoldIndex (61), GlobPlot2 (62), IUPred (63), RONN (64) and VSL2 (65), and two protein secondary structure predictor tools: PROFsec

(66) and PSIPred (67) (http://www-nmr.cabm.rutgers.edu/bioinformatics/disorder/). The prediction of disorder was considered significant if at least four predictors gave a hit.

Secondary RNA structures in ORF2/ORF5 overlapping region were predicted with the program RNAz in a region of about 300–900 bp flanking the region encoding LXCXE motif sequence (68). The server uses an algorithm that detects thermodynamically stable and evolutionarily conserved RNA secondary structures in multiple RNA-sequence alignments on both RNA-strands, with number of sequences in alignments not exceeding six. If subsets were larger than six, they were reduced to a combination of six virus sequences. For structure prediction the default RNAz parameters of 'Standard Analysis' were utilized, which scored in the overlapping windows of 120 alignment columns with step-size of 40 nt (http://rna.tbi.univie.ac.at/cgi-bin/RNAz.cgi?PAGE=1&TYPE=S).

Proline enrichment in putative ORF5-encoded protein sequences was analysed by use of a custom R script that counts Proline residues and visualizes the counts with respect to location in the protein sequence and premature stop codons in the case of ORF5-less viruses (www.R-project.org) (69).

RESULTS AND DISCUSSION

Discovery of Codon-Constrained Val-Ala (COCO-VA) toggling in MT/ALTO

We were interested in understanding the evolution and function of the *de novo* ORF5. Only four short conserved motifs, designated ORF5m1 to ORF5m4, were evident in the ORF5-wide alignment (Supplementary Figure S1) due to an extremely high residue and two-fold size variation (see also below and (15)). They are counterparts of four motifs of LT antigen in the overlapping part of ORF2. Remarkably, the most conserved third aa residue of ORF5m2, identified in this study, has a restricted binary residue variation (Val/Ala) in both ORF5-plus and ORF5-less polyomaviruses (Figure 1AB). Val and Ala are encoded by eight **G**(**C**/**T**)(A/G/C/T) triplets which are the only codons compatible with the two **TG**(**C**/**T**) codons for conserved Cys of the LT LXCXE motif in the ancestral ORF2 (the 2-nt overlap between the Val/Ala and Cys codons is highlighted in bold). In other words, only variation at the second codon position of the COCO-VA codon (**C** or **T**) determines the encoded amino acid (Ala or Val) (Figure 1B). We named the observed phenomenon Codon-Constrained Val-Ala (COCO-VA) toggling.

The C/T variation represents only half of the full four-nucleotide variation possible at a polynucleotide position (Figure 1C). When each kind of nucleotide is equally frequent at a given position, it is likely to evolve at no selection (neutral evolution) (Figure 1C1), which may be found in the third codon positions of non-overlapping ORFs. In contrast, a restricted nucleotide variation, like C/T, may emerge as a result of selection, either positive (Figure 1C2) or negative (Figure 1C3), which is typically observed at the first and second positions of codons of non-overlapping ORFs. Evolutionary interpretation of the nucleotide variation is more complex in the overlapping ORFs, which may be subject to several evolutionary forces acting on each ORF. For instance, there is no doubt that the restricted C/T variation at the second codon position of the COCO-VA site is due to negative selection in the alternative ORF2 to maintain the Cys residue. On the other hand, this restricted variation would be equally compatible with no selection or positive selection in ORF5, with the latter scenario leading to accelerated toggling between C and T (compare Figure 1C3 and C4). Therefore, we asked whether selection is involved in the COCO-VA toggling.

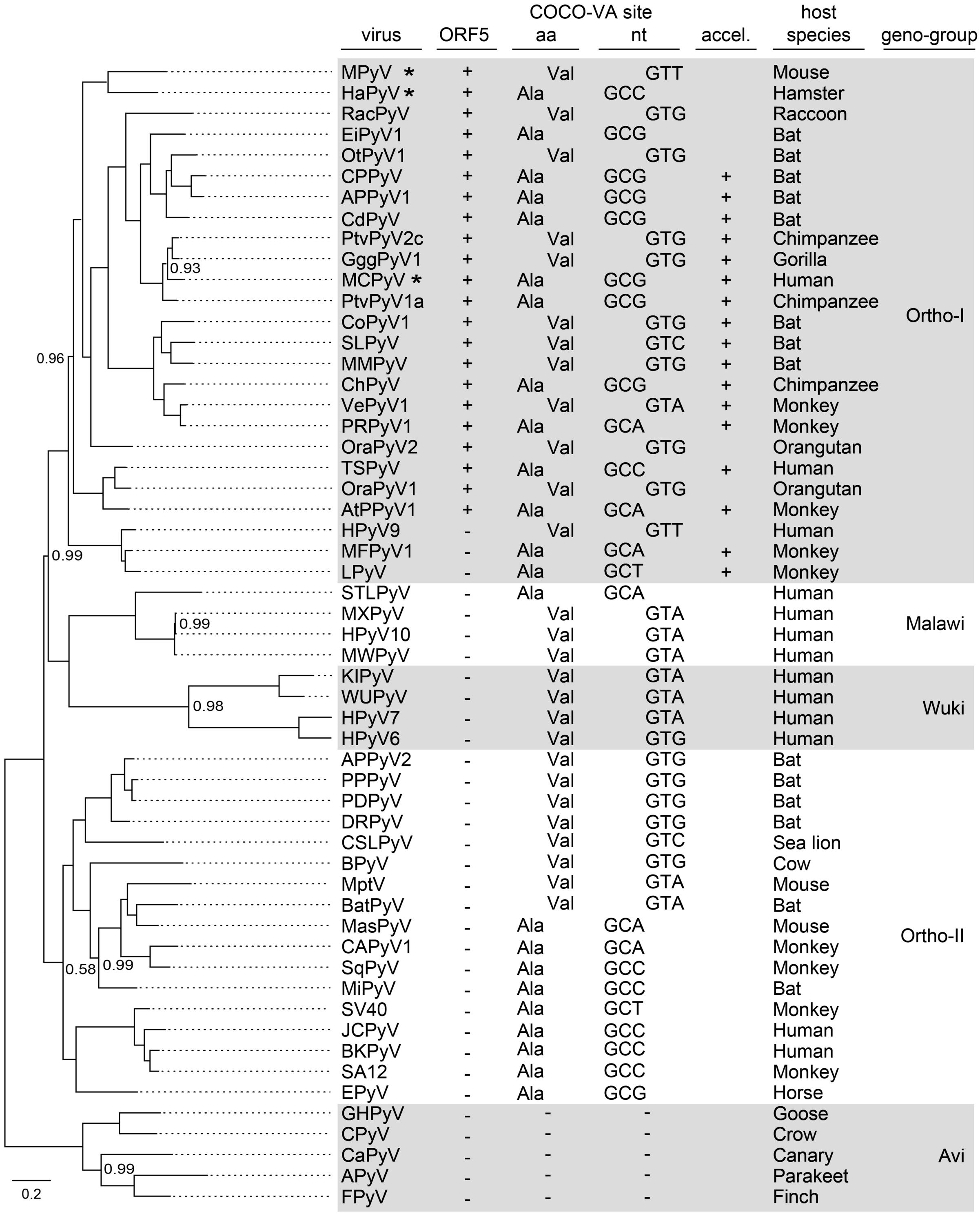
Phylogeny suggests accelerated COCO-VA toggling in ORF5-encoding polyomaviruses

First and in line with general reasoning (70), we note that conservation of the LXCXE Cys residue may not constrain the COCO-VA toggling. Second, if C and T nucleotides at the third position of the Cys codon are utilized unevenly, additional *non-ORF2* selection pressure(s), for instance on RNA, must be taken into account when analysing the toggling. Third, the ORF2 LXCXE conservation in *both* ORF5-plus and -less polyomaviruses provided us with two contrasting virus groups that differ in relation to the COCO-VA site expression through ORF5. Consequently, the COCO-VA site is not expected to be under selection pressure in ORF5-less viruses (Figure 1C3 scenario), while its evolution in ORF5-plus viruses may or may not be driven by selection depending on the functional importance of these residues (either Figure 1C3 or C4 scenario). Fourth, the restricted binary choice of aa residues at the COCO-VA site compared to the full 20 amino acid (aa) residue variation simplifies the evolutionary analysis of its residue variation.

Taking all these considerations into account, we reasoned that the relative *abundance of either Ala or Val* in ORF5-plus compared to ORF5-less polyomaviruses would be indicative of selection on residue *type.* Since Ala and Val are similarly and evenly abundant at the COCO-VA site in the known ORF5-plus and ORF5-less mammalian polyomaviruses: 11 versus 11 and 12 versus 14 (Figure 2), respectively, no indication for selection is apparent. This observation indicates also that the COCO-VA site may not have experienced other, non-ORF2-related selection favouring one of the two nucleotides. Accordingly, we have not found conserved RNA secondary structure elements in this region (see Text S2 and Supplementary Table S3, Supplementary Figures S2 and S3), which, potentially, could have been an alternative source of constraint on the non-synonymous substitution in ORF5.

Next, we investigated the *frequency* of COCO-VA toggling among polyomaviruses. In this and subsequent analyses, switching between Ala and Val residues was accounted with no regard to its direction: from Ala to Val or from Val to Ala. The analysis was limited to the interspecies comparisons. The rate of the COCO-VA toggling in the ORF5-less polyomaviruses provided a baseline rate of COCO-VA toggling that can be expected by chance mutation (neutral evolution). Comparison of this rate with that of the ORF5-plus viruses informed us about directional selection at the COCO-VA codon in the latter viruses.

In the framework of this comparison, we have first mapped COCO-VA toggling on a Bayesian phylogenetic



		COCO-VA site			host	
virus	ORF5	aa	nt	accel.	species	geno-group
MPyV *	+	Val	GTT		Mouse	Ortho-I
HaPyV*	+	Ala	GCC		Hamster	
RacPyV	+	Val	GTG		Raccoon	
EiPyV1	+	Ala	GCG		Bat	
OtPyV1	+	Val	GTG		Bat	
CPPyV	+	Ala	GCG	+	Bat	
APPyV1	+	Ala	GCG	+	Bat	
CdPyV	+	Ala	GCG	+	Bat	
PtvPyV2c	+	Val	GTG	+	Chimpanzee	
GggPyV1	+	Val	GTG	+	Gorilla	
MCPyV *	+	Ala	GCG	+	Human	
PtvPyV1a	+	Ala	GCG	+	Chimpanzee	
CoPyV1	+	Val	GTG	+	Bat	
SLPyV	+	Val	GTC	+	Bat	
MMPyV	+	Val	GTG	+	Bat	
ChPyV	+	Ala	GCG	+	Chimpanzee	
VePyV1	+	Val	GTA	+	Monkey	
PRPyV1	+	Ala	GCA	+	Monkey	
OraPyV2	+	Val	GTG		Orangutan	
TSPyV	+	Ala	GCC	+	Human	
OraPyV1	+	Val	GTG		Orangutan	
AtPPyV1	+	Ala	GCA	+	Monkey	
HPyV9	-	Val	GTT		Human	
MFPyV1	-	Ala	GCA	+	Monkey	
LPyV	-	Ala	GCT	+	Monkey	
STLPyV	-	Ala	GCA		Human	Malawi
MXPyV	-	Val	GTA		Human	
HPyV10	-	Val	GTA		Human	
MWPyV	-	Val	GTA		Human	
KIPyV	-	Val	GTA		Human	Wuki
WUPyV	-	Val	GTA		Human	
HPyV7	-	Val	GTA		Human	
HPyV6	-	Val	GTG		Human	
APPyV2	-	Val	GTG		Bat	Ortho-II
PPPyV	-	Val	GTG		Bat	
PDPyV	-	Val	GTG		Bat	
DRPyV	-	Val	GTG		Bat	
CSLPyV	-	Val	GTC		Sea lion	
BPyV	-	Val	GTG		Cow	
MptV	-	Val	GTA		Mouse	
BatPyV	-	Val	GTA		Bat	
MasPyV	-	Ala	GCA		Mouse	
CAPyV1	-	Ala	GCA		Monkey	
SqPyV	-	Ala	GCC		Monkey	
MiPyV	-	Ala	GCC		Bat	
SV40	-	Ala	GCT		Monkey	
JCPyV	-	Ala	GCC		Human	
BKPyV	-	Ala	GCC		Human	
SA12	-	Ala	GCC		Monkey	
EPyV	-	Ala	GCG		Horse	
GHPyV	-	-	-		Goose	Avi
CPyV	-	-	-		Crow	
CaPyV	-	-	-		Canary	
APyV	-	-	-		Parakeet	
FPyV	-	-	-		Finch	

Figure 2. Polyomavirus phylogeny and ORF5 characteristics. Shown is a Bayesian phylogeny using BEAST version 1.7.4 (49) for 55 polyomaviruses (listed in Supplementary Table S1) based on conserved regions in the LT, VP1 and VP2 proteins (see 'Materials and Methods' section for details). The numbers plotted in the tree show posterior probability support values for internal branching events <1. The scale bar is in average number of amino acid substitutions. Asterisks in the virus column indicate viruses for which the ORF5 expression has been demonstrated experimentally. The ORF5 column indicates the presence (+) or absence (−) of ORF5 in polyomaviruses genomes. The COCO-VA site column depicts the residue (Ala or Val) and corresponding codon at the COCO-VA site that is constrained by the Cys codon of the LT LXCXE motif in mammalian polyomaviruses (see Figure 1 and Supplementary Figure S1). The acceleration column (accel.) labels viruses that experienced selection-driven acceleration at the COCO-VA site. The geno-group column depicts the phylogenetic distribution of polyomaviruses according to Feltkamp *et al.*, (43). Please note that Carter *etal.* 2013 (15) divided all mammalian polyomaviruses into two groups, monophyletic Almipolyomaviruses and paraphyletic non-Almipolyomaviruses, which correspond to ORF5-plus and ORF5-less polyomaviruses, respectively. The tree was pseudorooted at the branch connecting mammalian and avian (Avi) polyomaviruses. See 'Materials and Methods' section for other details.

tree of polyomaviruses (Figure 2). Due to extreme sequence divergence of the ORF2/ORF5 overlap region in mammalian polyomaviruses (see above; (15)), reliable alignment of this region is limited to four motifs of only ∼30 residues in total (Supplementary Figure S1), which may not be sufficient for reliable phylogeny reconstruction. Therefore we choose to use a concatenated alignment of other conserved domains representing LT, VP1 and VP2 proteins and accounting for ∼50% of genome for phylogeny inference. Large monophyletic groups on this tree were formed by viruses, which were recognized as similar in the ORF2/ORF5 overlapping region. Additionally, we have observed good agreement between topologies of separate branches of this tree, each representing closely related polyomaviruses, with trees of these same viruses using alignments of the ORF2/ORF5 overlap region (Supplementary Figure S4). These observations showed that the ORF2/ORF5 overlap region is likely to have co-evolved with the LT, VP1 and VP2/3 proteins, whose tree was thus considered suitable for analysis of the COCO-VA toggling.

Subsequently, visual inspection of the tree revealed contrasting patterns of phylogenetic grouping for Ala- and Val-specific viruses in ORF5-plus and ORF5-less subsets of mammalian polyomaviruses, respectively (Figure 2). While Ala- and Val-specific viruses were largely intertwined in the first subset, they predominantly formed large residue-specific monophyletic groups in the second subset. This result was indicative of acceleration of the COCO-VA toggling in ORF5-plus viruses. To verify and extend this observation further, we have conducted additional evolutionary-based analyses using available and specially designed approaches.

No evidence for positive selection at the COCO-VA site by conventional evolutionary analyses

We started with employing two most advanced and widely used programs, MEME (52) and FUBAR (27), developed for evolutionary analysis of residue variation. Also included was TOGGLE (25), which was specifically developed for analysis of residue toggling. These three programs are available through the Datamonkey website (26). The employed programs differ in how they accommodate lineage- and site-specific variation in the analysed dataset to infer patterns of evolution and deduce selection forces acting on individual codons. Since these tools were developed for the analysis of non-overlapping ORFs, only a single evolutionary force, if identified, is reported. Consequently, we did not expect that these programs could infer both purifying selection (due to Cys conservation of LXCXE) and positive selection (accelerated Val-Ala toggling) at the COCO-VA site of the ORF5-plus viruses as depicted in Figure 1C4. Rather, we asked whether the programs could provide evidence for either negative or positive selection at this site of ORF5-plus viruses and negative selection at this site of ORF5-less viruses. This type of inferences depends on the number and diversity of alignment positions under analysis. Due to the high sequence divergence of the ORF2/ORF5 overlapping region, we thus analysed different subsets of mammalian polyomaviruses, in order to facilitate identification of selection forces. Specifically, the programs were applied to 10 different alignments of ORF5, D1–D10, representing selected ORF5 codons, which may or may not be merged with ORF2 codons for different subsets (see 'Materials and Methods' section and Supplementary Table S4). In none of the 30 conducted analyses, the COCO-VA codon was identified to be under positive/diversifying selection, including toggling. Also the COCO-VA codon was not found to be negatively selected in analyses that included only ORF5-plus viruses, neither in the entire set nor its D5, D6, D8, D9 and D10 subsets. However, upon analysis of the other five virus sets by FUBAR, either including all ORF5-plus viruses along with other viruses (D1–D4) or including only ORF5-less viruses (D7), the COCO-VA codon was identified to be under purifying selection. In contrast, various other ORF5 codons were identified as being positively or negatively selected or be involved in toggling, in many of these analyses (Supplementary Table S4).

The lack of evidence for positive selection/toggling at the COCO-VA codon in ORF5-plus viruses according to these analyses could be either a true negative result (lack of the phenomenon) or a false negative result (failure to detect a signal due to systematic technical deficiency). As detailed below, we believe that the latter explanation is most likely. Indeed, the employed three programs operate under the assumption that the entire codon table of 61 varieties is available for evolution at every site in the analysed alignments. Consequently, the eight different codons (four for Val and four for Ala) observed at the COCO-VA site were seen as severely restricted rather than representing the full spectrum allowed at this site (imposed by Cys conservation in the overlapping ORF2 codon). This misreading of the observed residue variation has profound implications for its evolutionary interpretation, since high diversity tends to be interpreted as a sign of positive selection, while restricted diversity is commonly associated with purifying selection during evolution of non-overlapping ORFs. It is thus not surprising that the COCO-VA site evolution was qualified to be under negative selection in several tests by FUBAR. This result could be seen as evidence for the dominance of purifying selection at the COCO-VA site according to FUBAR.

Toggling at the COCO-VA site is significantly accelerated

Due to the above considerations, we decided to continue our testing of the toggling by applying an approach that could be free from the limitations of standard evolutionary based programs developed for non-overlapping ORFs. First, we sought to verify the elevated frequency of Ala-Val exchange in ORF5-plus compared to ORF5-less viruses that is apparent from polyomavirus phylogeny (Figure 2). To this end, we have applied the Multistate method of the BayesTraits package (53) to compare the COCO-VA site variation in mammalian ORF5-plus and ORF5-less polyomaviruses. The program employs continuous-time Markov models to estimate the transition rates between multiple states for a single trait (Ala/Val for COCO-VA site in this case) while it traverses a tree. The produced estimates take into account the uncertainty associated with tree reconstruction as it utilizes the full posterior tree sample. The estimated transition rate distribution was plotted for three virus datasets (Figure 3 left). From this plot it is evident that the estimated

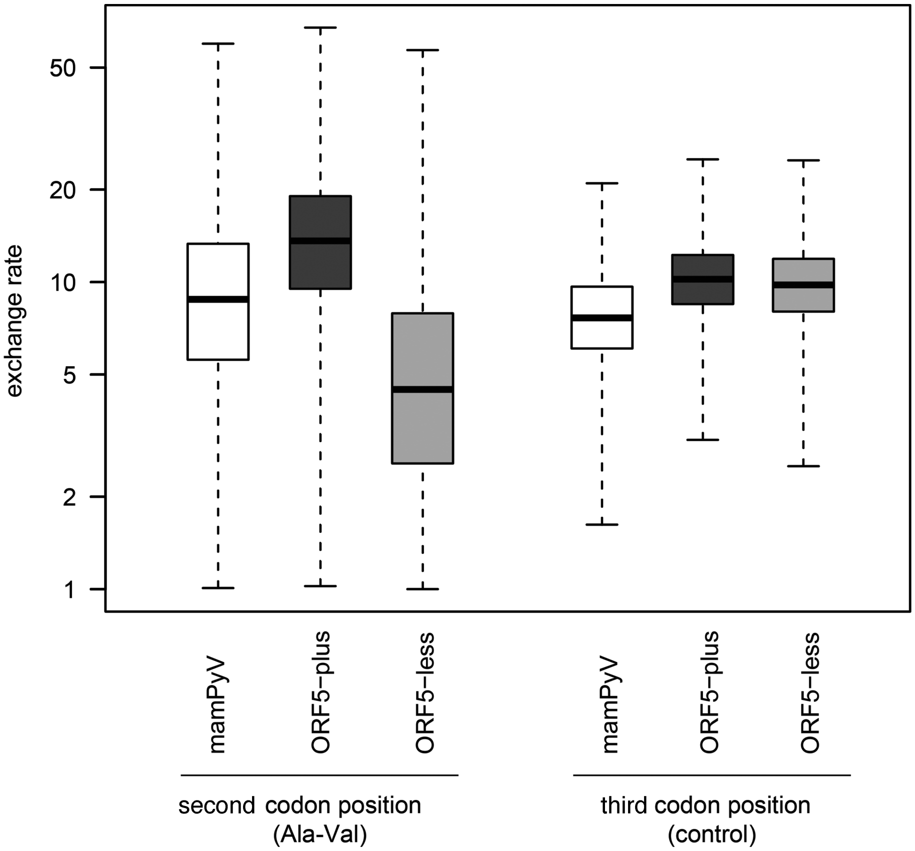


Figure 3. COCO-VA toggling is accelerated in ORF5-plus compared to ORF5-less viruses. Shown are results of BayesTraits multistate analysis of COCO-VA toggling rate in three groups of viruses. The distributions of estimated exchange rates at second (left side) and third (right side) codon position of the COCO-VA codon is shown. The exchange rate at the second codon position corresponds to the COCO-VA toggling while the rate at the third codon position serves as a control. The distributions are shown as Box-and-whisker graphs. The boxes span from the first to the third quartile and include the median (bold line), and the whiskers (dashed lines) extend to the extreme values.

Ala-Val exchange rate is more than three times higher (13.6 versus 4.3) for ORF5-plus viruses compared to ORF5-less viruses, with a 25–75% inter-quartile range of 9.5–19.1 and 2.5–7.7, respectively. This striking difference between the two datasets is strongly supported by a log Bayes Factor (logBF) of 3352.2, which is astronomically large and dwarfs the significance threshold of 2. As expected, the estimate for mammalian polyomaviruses was intermediate between those two with a 25–75% inter-quartile range of 5.6–13.3 (Figure 3 left). As Ala-Val exchange at the COCO-VA position is equivalent to C-T exchange at the second codon position (see Figure 2 and 'Materials and Methods' section) we have compared its exchange rate to that at the third codon position (Figure 3 right). This position accepts all four nucleotides and its variation is primarily driven by selection in the overlapping ORF2 in which it occupies the first codon position of the subsequent residue. As may be expected, the exchange rate at this position (now averaged over four instead of two nucleotides) is comparable for ORF5-plus and ORF5-less viruses (median and 25–75% inter-quartile range: 10.2 and 8.5–12.3 versus 9.8 and 8.0–11.9, respectively). Of notice, these numbers are still and consistently smaller than those of the Ala-Val exchange rate for ORF5-plus viruses.

Importantly, the observed difference at the second codon position (i.e. the Ala-Val exchange) may not be attributed to differences in virus diversity of the compared two datasets, whose distributions of smallest pairwise patristic distance (SPAT) values (median value and 25–75% inter-quartile range: 0.23 and 0.13–0.32 versus 0.28 and 0.12–0.46) were not different at a statistically significant level (Mann–Whitney U test; $P = 0.42$). Consequently, we concluded that the COCO-VA toggling rate is significantly and genuinely accelerated in the ORF5-plus compared to the ORF5-less polyomaviruses. Since the COCO-VA site is expressed in ORF5-plus but not ORF5-less viruses (although see below), this result implies positive selection on the COCO-VA site in ORF5-plus viruses.

Ratio approach to study accelerated COCO-VA toggling

To study the accelerated COCO-VA toggling further, we have developed a ratio approach remotely similar to that of comparing the ratio of non-synonymous to synonymous substitutions. We used the ratio of monoSPAT/(monoSPAT + polySPAT) values as a normalized measure of the COCO-VA toggling rate relative to Ala/Ala or Val/Val persistence, with polySPAT and monoSPAT resembling estimations of non-synonymous and synonymous substitutions, respec-

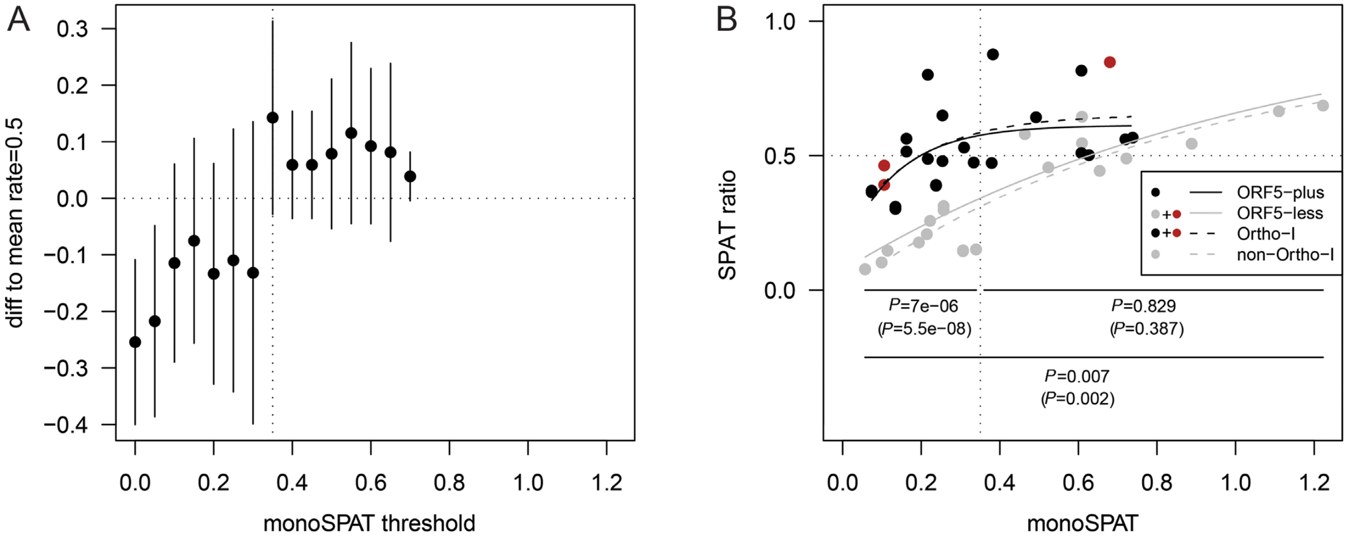


Figure 4. Accelerated toggling at the COCO-VA site in ORF5-plus and Ortho-I viruses. For the purpose of this analysis a representative set of 50 mammalian polyomaviruses (see Supplementary Table S1) was studied. Due to the lack of species demarcation criteria for polyomaviruses, we chose to consider viruses with different names as representing different species (dots in the plot). The only exception was made for MX polyomavirus, Human polyomavirus 10 and MW polyomavirus, which were represented only by the latter because of the very small distances that separate these three viruses. Two pairs of virus partitioning (subsets) of the mammalian polyomaviruses, based on the ORF5 presence and phylogeny, were considered. They and their codes are defined in the inset of panel (**B**). (**A**) The partitioning of the monoSPAT scale at 0.35 was derived based on the drop of the mean difference of SPAT ratios to the matching rate. Here, a sliding window (size 0.15, shift 0.05) starting at monoSPAT of 0.0 was moved along the monoSPAT range to calculate within-window mean differences (dots) and associated standard deviations (vertical lines). See also 'Materials and Methods' section and Supplementary Figure S5 for other details. (**B**) The curves show the fit of a 3-parameter logistic function to each of four different subsets. The numbers below show *P*-values of Mann–Whitney U tests comparing the SPAT ratio distributions between ORF5-plus and ORF5-less viruses (Orthopolyomavirus-I and non-Orthopolyomavirus-I) for two mono SPAT ranges (0–0.35, 0.35–1.25). A horizontal dotted line is drawn at the matching rate, whose evolutionary interpretation is defined in the text.

tively (for group designations see 'Materials and Methods' section). Only the C/T variation at the second codon position that controls Val-Ala exchange, rather than the entire codon for Ala/Val as it would have been the case upon analysis of a non-overlapping ORF by a conventional technique, was analysed in our test. We thus avoided complications to the analysis that would otherwise be caused by the unaccounted evolutionary pressure on the third position of Ala/Val codons by the ORF2 overlapping codon, where it occupies the first position of the subsequent residue (Figure 1B). An SPAT ratio of 0.5 indicates that the Ala/Val exchange rate matches that of Ala/Ala or Val/Val persistence during evolution of a particular lineage (hereafter, matching rate). Since amino acid residue persistence at the COCO-VA site in a pair of viruses may involve either no genetic change or synonymous substitution, the matching rate for toggling under the model of neutral evolution could be expected only at sufficiently large evolutionary distances when chance mutation, either synonymous or non-synonymous, is highly probable. Accordingly, persistence would dominate over toggling at smaller distances under this model, resulting in SPAT ratios smaller than 0.5. If positive selection is involved in toggling, increase of SPAT ratios compared to those expected under neutral evolution could be observed at sufficiently small evolutionary distances.

The above considerations indicate that under the model of neutral evolution we could expect different SPAT ratios at small and large evolutionary distances. To verify this and define ranges for small and large evolutionary distances separating pairs of monomorphic viruses, we analysed the difference between the matching rate and within-window distributions of SPAT ratios involving all mammalian polyomaviruses, which were plotted against monoSPAT values. Due to stochastic reasons, the estimated toggling rate may deviate from the actual rate for a virus. To address this limitation, we pooled SPAT ratios within a predefined window that was slid along the monoSPAT axis. Our analysis revealed that the Ala/Val exchange rate of polyomaviruses varies considerably, with very different median values being observed in two monoSPAT ranges (Figure 4A). In the monoSPAT range of 0–0.35, median SPAT ratios were consistently smaller compared to the matching rate, while in the monoSPAT range of 0.35–0.75, they were consistently larger than the matching rate. Accordingly, the entire monoSPAT range was split into two sub-ranges in our subsequent analyses. For evolutionary interpretations of the ratio test in subsequent analyses, we used the results obtained for ORF5-less viruses as a base-line, since Val-Ala toggling in these viruses is expected to experience no selection (Figure 1C3 scenario) over the entire evolutionary distance range.

Accelerated COCO-VA toggling is associated with inter-species diversification of ORF5-plus polyomaviruses

Is the accelerated toggling a characteristic of the entire ORF5-plus viruses or its subset? The difference between SPAT ratios in the distributions for ORF5-plus and ORF5-less viruses was statistically significant over the monoSPAT range of 0–0.35 (MWU test *P*-value = 7e-06), but not over the 0.35–1.2 range (MWU test *P*-value = 0.829) (Figure 4B). Importantly, this result may not be due to biases of the virus sampling which was comparable for ORF5-plus and -less viruses in the two distributions along the monoSPAT range (Supplementary Figure S5).

Consequently, the above observations indicate a selection-driven acceleration of COCO-VA toggling in the majority (15 out of 22) of ORF5-plus polyomaviruses, each of which is separated from another monomorphic virus by a monoSPAT of 0.35 or smaller (Figure 4). For the remaining seven ORF5-plus viruses, each of which is separated from another monomorphic virus by a monoSPAT larger than 0.35, no accelerated COCO-VA toggling was observed. This could be either due to specifics of evolution or the unavailability of close monomorphic relatives of these viruses in the current sampling. If the former is true, the viruses with accelerated COCO-VA toggling may be expected to cluster in the tree, while a random phyletic distribution is likely otherwise. Figure 2 shows that the 15 viruses with accelerated COCO-VA toggling are scattered across the entire branch of ORF5-plus viruses. This observation implies that the accelerated COCO-VA toggling may involve *all* ORF5-plus viruses (all terminal nodes in the respective tree branch) thus presenting an extreme case of convergent evolution. An improved, much larger virus sampling, which includes closely related viruses for each analysed virus species, will enable verification of this implication. Also, it may facilitate additional insights, including: (i) refining the estimate of the monoSPAT threshold at which the COCO-VA toggling acceleration can be observed and (ii) extending our analysis to poorly sampled intra-species diversity, in order to address the question whether accelerated COCO-VA toggling drives speciation or *vice versa*.

Could the observed difference between ORF5-plus and ORF5-less viruses in the monoSPAT range of 0–0.35 (Figure 4B) have emerged also under the evolutionary scenario that is alternative to that involving positive selection on ORF5-plus and no selection on ORF5-less viruses? If the COCO-VA site was under strong negative selection in ORF5-less viruses while being under either weak negative or no selection in ORF5-plus viruses, SPAT ratio of these viruses would differ. The following considerations make this scenario unlikely to be applicable to explain the data obtained in our study. First, this scenario implies that the COCO-VA site must be expressed in *all* ORF5-less viruses. These viruses include some of the most well characterized polyomaviruses, e.g. SV40, with no evidence for the expression of the COCO-VA site, although some of the poorly characterized ORF5-less viruses may indeed express this site (see below). We could also recall that ORF5-less viruses were defined as a group not having the property (ORF5) rather than having one, which would be required to link strong negative selection to the functional characteristic. Second, SPAT ratio of ORF5-plus viruses in the monoSPAT range of 0–0.35 is comparable to the matching rate (P = 0.390 in Wilcoxon rank sum test; see Supplementary Table S5). This result is in the excellent agreement with positive selection acting on the COCO-VA site in ORF5-plus viruses, while it may not be reconciled with the weak negative selection hypothesis. On the other hand, it would in principle be compatible with neutral evolution of the COCO-VA site in ORF5-plus viruses under the condition that polyomaviruses have very high mutation rate. The estimates of this rate vary greatly and generally this aspect has not been fully resolved (71). However, we note that the Val-Ala variation is already observed in several monophyletic subsets of ORF5-plus viruses which otherwise diverged little or modestly. This observation indicates that the Val-Ala variation may be among most frequent rather than average as would be expected under the neutral evolution scenario. This aspect could be studied most closely with the improved virus sampling. In conclusion, based on the available data the accelerated COCO-VA toggling due to positive selection is the most likely evolutionary scenario.

Accelerated COCO-VA toggling is most strongly associated with monophyletic Ortho-I viruses

The results described above provide evidence for accelerated COCO-VA toggling in the 0–0.35 monoSPAT range for ORF5-plus viruses. However, it is also evident that the distributions of ORF5-plus and less SPAT ratios overlap with two ORF5-less viruses deviating considerably from their group-mates and instead fitting into the other group rather well (Figure 4B, red dots in the 0–0.35 monoSPAT range). This grouping with ORF5-plus viruses received strong statistical support when analysing all of the 145,422,675 possible 16-by-14 combinations of the 30 viruses with monoSPAT values in the range of 0–0.35 (Text S3 and Supplementary Figure S6). Intriguingly, these two viruses (MFPyV1 and LPyV) along with another one (HPyV9) for which no closely related monomorphic virus is available in the current virus sampling (Figure 4B, red dot in the 0.35–1.20 monoSPAT range) form a sister lineage of ORF5-plus polyomaviruses at the root of the Ortho-I monophyletic group (Figure 2) (15,43). These results suggest that the accelerated toggling is most strongly associated with the Ortho-I group. Since the observed accelerated toggling is indicative of positive selection that may be realized only upon expression of the COCO-VA site in the two (three) poorly characterized basal Ortho-I viruses, such hypothesis must be considered. It could be achieved using a mechanism other than expression of the entire ORF5, for instance, through alternative splicing of mRNA(s) (72) that could fuse the COCO-VA site with other ORF(s). In the evolutionary framework, such expression of the COCO-VA site would be ancestral to those used by ORF5-plus viruses, implying that the COCO-VA site and the associated sequence motif could be a nucleation site for the subsequent ORF5 origin by ORF expansion (15).

COCO-VA toggling is located in a SLiM of an intrinsically disordered region

What could be the structural basis of COCO-VA toggling? Bioinformatics analyses indicate that MT/ALTO is a Pro residue rich IDR, whose motifs could form SLiMs (Supplementary Figures S1, S7, S8 and S9) (15). Thus, the interspecific toggling targets a SLiM, which is in line with the notion that IDRs evolving differently than structured protein regions (6,7). Since SLiMs promote protein folding in relatively flat energy landscapes (73) and mediate interactions with partners that are relatively weak (74,75), difference between physico-chemical properties of the just two possible COCO-VA residues, Val and Ala, could be of significance. For instance, these residues have contrasting structural propensities, favouring the formation of either α-helix (Ala) or β-sheet (Val) (76), which might be used to promote alternative folding of MT/ALTO upon interaction with partner(s). Unfortunately, this hypothesis may not be tested using the available computational approaches.

Concluding remarks

In complex protein networks, SLiMs are emerging as evolutionary adaptive transmitters of intracellular signals involving multiple interacting partners (2–3,13,77). Here we presented evidence for the evolutionary signature of adaptation in the otherwise uncharacterized SLiM of MT/ALTO. The effect of COCO-VA toggling on the SLiM may be similar to that of phosphorylation which could modulate SLiM activity considerably (78). MT antigen of rodent polyomaviruses has been shown to interact through its ORF5-encoded part with numerous cellular targets involved in signal transduction (33,38–39,42). The function of ALTO, identified just recently, has not been resolved yet (15). The described accelerated COCO-VA toggling is notable because of a unique combination of properties: it involves one of the just few conserved positions of the otherwise highly divergent MT/ALTO protein, and it may affect every species of Ortho-I polyomaviruses. These viruses are known to infect bats, rodents, monkeys, hominids and humans with apparently frequent host switching (Figure 2). Future studies should identify driving forces of the COCO-VA toggling to enable its comparison with intra-species residue toggling (25). The latter is likely driven by the cellular immune response and occurs at much smaller time and divergence scales and with the exchange of many residues. Practically, our study suggests that analysis of substitution rates can be applied to individual residues in overlapping ORFs. It extends the utility of the substitution rate analysis from mapping to dissecting functional elements in overlapping ORFs.

The described phenomenon also challenges common perception of conservation of proteins, which is believed to be inversely and universally correlated with the rate of evolution. Accordingly, sites accepting relatively few residues are classified conserved and evolving slowly under negative selection. Typically, such residues are critical for maintaining protein core and/or playing an essential role in the active site of structured proteins. Besides the sites that are strictly invariant, those that accept only two residues during large-scale evolution are among the most conserved. Exchange of these residues could happen due to either rare fixation of non-synonymous mutation that is driven by episodic positive selection or residue drift. As a result, each of the two residues is likely to be associated with a large monophyletic clade in the tree (Figure 5 left panel), the pattern that can be recognized by available programs (e.g. (79)). Examples of this type of evolution are plenty in many protein families. For instance, rare exchange of the catalytic nucleophile Cys and Ser residues in virus proteases with chymotrypsin-like fold (80) or phosphate-binding Ser and Thr residues in the Walker-box GKS/T motif of nucleotide-binding proteins (81), are notable.

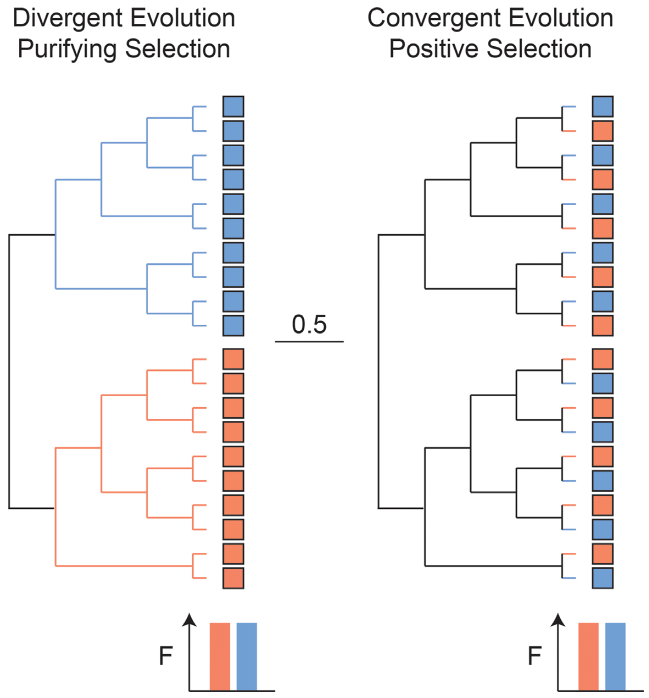


Figure 5. Contrasting modes of evolution at a conserved protein site accepting two residues. Shown are fictional examples of evolution of a conserved protein site with two-residue variation in families of structured (left panel) and unstructured (right) proteins, whose evolutionary scale of replacement at all sites was considerable (bar 0.5) and whose phylogeny is described by identical trees. In both cases, two residues are evenly distributed, each occupying 50% of terminal nodes. Residue type either clusters into two monophyletic groups (left) or is intertwined (right). The left panel depicts divergent evolution driven mostly by purifying selection as seen in many characterized structured proteins. The right panel depicts convergent evolution driven by positive selection as discovered at the COCO-VA site in the presented paper and may be experienced at other sites in unstructured proteins.

The above considerations indicate that in structured proteins, limited residue variation may largely be imposed by the molecular environment in which these proteins operate. In contrast, constraints on the genetic level is the chief factor determining residue variation in proteins encoded in overlapping ORFs. Consequently, this restricted residue variation in overlapping ORFs may not be linked to residue function in the manner described for structured proteins. Accordingly, overlapping ORFs predominantly encode unstructured proteins with their most conserved SLiMs mediating adaptation, a function that is commonly facilitated by the least conserved elements in structured proteins. Along the same line, we now provide evidence for the phylogenetic intertwining of viruses that employ, respectively, Val and Ala at the conserved COCO-VA site in the IDR of MT/ALTO. When depicted in a simplified form, this phylogenetic pattern can be contrasted with the clade-specific association of residues in a tree of structured proteins (compare right and left panels of Figure 5). This contrast is particularly striking since it is not evident in the cumulative frequency of each of these residues at terminal nodes (bottom panels underneath of trees in Figure 5). Thus, this logos-style representation of residue conservation, which is very popular in functional studies, may not capture residue change and its role in adaptation. Only analysis in the context of phylogeny could do it, as demonstrated in this study.

Since Cys is one of the least frequent amino acid residues and none of the other residues can constrain evolution in the overlapping −3 RF (or +1 RF) to only two residues, the described codon-constrained accelerated toggling might be viewed as an extremely exotic phenomenon limited to polyomaviruses. We believe that this perception is biased for several reasons. First of all, the (unknown) diversity of the Virus Universe is expected to be many orders of magnitude larger than the number of currently recognized few thousand virus species (82,83). This implies a good chance of discovering accelerated COCO-VA toggling in other viruses in the future. Furthermore, accelerated toggling might involve more than two residues at a site that could still be considered conservative relative to many other sites. Such constraint could be imposed by conserved amino acids other than Cys in the overlapping ORF or, if non-overlapping ORF is involved, by a different genetic mechanism, e.g. RNA structure, or even by a partner or partners interacting with an IDR site. Thus, the described accelerated COCO-VA toggling may represent an extreme case of common evolution of individual residues in IDRs of proteins, making it potentially relevant to understanding biology and pathology of adaptation of many organisms.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Jelle J. Goeman, Sander Kooijman, Andrey M. Leontovich and Els van der Meijden for useful discussions, and Igor Sidorov and Dmitry Samborskiy for help with Viralis. A.E.G. is member of the Netherlands Bioinformatics Center (NBIC) Faculty.

FUNDING

Leiden University Medical Center, The Netherlands, through the Collaborative Agreement in Bioinformatics between Leiden University Medical Center and Moscow State University (MoBiLe Program) (in part); Leiden University Fund. Funding for open access charge: Leiden University Medical Center.

Conflict of interest statement. None declared.

REFERENCES

1. Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, **293**, 321–331.
2. Dyson,H.J. and Wright,P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell. Biol.*, **6**, 197–208.
3. Dunker,A.K., Silman,I., Uversky,V.N. and Sussman,J.L. (2008) Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.*, **18**, 756–764.
4. Tompa,P. (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.*, **579**, 3346–3354.
5. Xue,B., Dunker,A.K. and Uversky,V.N. (2012) Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn.*, **30**, 137–149.
6. Brown,C.J., Johnson,A.K. and Daughdrill,G.W. (2010) Comparing models of evolution for ordered and disordered proteins. *Mol. Biol. Evol.*, **27**, 609–621.
7. Brown,C.J., Johnson,A.K., Dunker,A.K. and Daughdrill,G.W. (2011) Evolution and disorder. *Curr. Opin. Struct. Biol.*, **21**, 441–446.
8. Davey,N.E., Van Roey,K., Weatheritt,R.J., Toedt,G., Uyar,B., Altenberg,B., Budd,A., Diella,F., Dinkel,H. and Gibson,T.J. (2012) Attributes of short linear motifs. *Mol. Biosyst.*, **8**, 268–281.
9. Keese,P.K. and Gibbs,A. (1992) Origins of genes: 'big bang' or continuous creation? *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 9489–9493.
10. Firth,A.E. and Brown,C.M. (2006) Detecting overlapping coding sequences in virus genomes. *BMC Bioinformatics*, **7**, 75.
11. Belshaw,R., Pybus,O.G. and Rambaut,A. (2007) The evolution of genome compression and genomic novelty in RNA viruses. *Genome Res.*, **17**, 1496–1504.
12. Sabath,N., Wagner,A. and Karlin,D. (2012) Evolution of viral proteins originated de novo by overprinting. *Mol. Biol. Evol.*, **29**, 3767–3780.
13. Pushker,R., Mooney,C., Davey,N.E., Jacque,J.M. and Shields,D.C. (2013) Marked variability in the extent of protein disorder within and between viral families. *PLoS One*, **8**, e60724.
14. Ling,R., Pate,A.E., Carr,J.P. and Firth,A.E. (2013) An essential fifth coding ORF in the sobemoviruses. *Virology*, **446**, 397–408.
15. Carter,J.J., Daugherty,M.D., Qi,X., Bheda-Malge,A., Wipf,G.C., Robinson,K., Roman,A., Malik,H.S. and 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.
16. Firth,A.E. and Atkins,J.F. (2009) Evidence for a novel coding sequence overlapping the 5′-terminal approximately 90 codons of the gill-associated and yellow head okavirus envelope glycoprotein gene. *Virol. J.*, **6**, 222.
17. Firth,A.E. and Atkins,J.F. (2009) A case for a CUG-initiated coding sequence overlapping torovirus ORF1a and encoding a novel 30 kDa product. *Virol. J.*, **6**, 136.
18. Loughran,G., Firth,A.E. and Atkins,J.F. (2011) Ribosomal frameshifting into an overlapping gene in the 2B-encoding region of the cardiovirus genome. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, E1111–E1119.
19. Fang,Y., Treffers,E.E., Li,Y., Tas,A., Sun,Z., van der Meer,Y., de Ru,A.H., van Veelen,P.A., Atkins,J.F., Snijder,E.J. *et al.* (2012) Efficient -2 frameshifting by mammalian ribosomes to synthesize an additional arterivirus protein. . *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E2920–E2928.
20. Mizokami,M., Orito,E., Ohba,K., Ikeo,K., Lau,J.Y. and Gojobori,T. (1997) Constrained evolution with respect to gene overlap of hepatitis B virus. *J. Mol. Evol.*, **44**(Suppl. 1), S83–S90.
21. Hughes,A.L. and Hughes,M.A. (2005) Patterns of nucleotide difference in overlapping and non-overlapping reading frames of papillomavirus genomes. *Virus Res.*, **113**, 81–88.
22. De Groot,S., Mailund,T. and Hein,J. (2007) Comparative annotation of viral genomes with non-conserved gene structure. *Bioinformatics*, **23**, 1080–1089.
23. Sabath,N., Landan,G. and Graur,D. (2008) A method for the simultaneous estimation of selection intensities in overlapping genes. *PLoS One*, **3**, e3996.
24. Kryazhimskiy,S. and Plotkin,J.B. (2008) The population genetics of dN/dS. *PLoS Genet.*, **4**, e1000304.
25. Delport,W., Scheffler,K. and Seoighe,C. (2008) Frequent toggling between alternative amino acids is driven by selection in HIV-1. *PLoS Pathog.*, **4**, e1000242.
26. Kosakovsky Pond,S.L. and Frost,S.D. (2005) Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*, **21**, 2531–2533.
27. Murrell,B., Moola,S., Mabona,A., Weighill,T., Sheward,D., Kosakovsky Pond,S.L. and Scheffler,K. (2013) FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Mol. Biol. Evol.*, **30**, 1196–1205.
28. Dalianis,T. and Hirsch,H.H. (2013) Human polyomaviruses in disease and cancer. *Virology*, **437**, 63–72.
29. Kazem,S., van der Meijden,E. and Feltkamp,M.C. (2013) The trichodysplasia spinulosa-associated polyomavirus: virological background and clinical implications. *APMIS*, **121**, 770–782.

30. Decaprio,J.A. and Garcea,R.L. (2013) A cornucopia of human polyomaviruses. *Nat. Rev. Microbiol.*, **11**, 264–276.
31. DeCaprio,J.A., Imperiale,M.J. and Major,E.O. (2013) Polyomaviruses. In: Knipe,DM and Howley,PM (eds). *Fields VIROLOGY*. Wolters Kluwer/Lippincott Williams & Wilkins, Philadelphia, pp. 1633–1661.
32. Kazem,S., Lauber,C., van der Meijden,E., Kooijman,S., Bialasiewicz,S., Wang,R.C., Gorbalenya,A.E. and Feltkamp,M.C. (2013) Global circulation of slowly evolving trichodysplasia spinulosa-associated polyomavirus and its adaptation to the human population through alternative T antigens. *J. Neurovirol.*, **19**, 298–299.
33. Hutchinson,M.A., Hunter,T. and Eckhart,W. (1978) Characterization of T antigens in polyoma-infected and transformed cells. *Cell*, **15**, 65–77.
34. Kazem,S., van der Meijden,E., Wang,R.C., Rosenberg,A.S., Pope,E., Benoit,T., Fleckman,P. and Feltkamp,M.C. (2014) Polyomavirus-associated trichodysplasia spinulosa involves hyperproliferation, pRB phosphorylation and upregulation of p16 and p21. *PLoS One*, **9**, e108947.
35. Decaprio,J.A. (2009) How the Rb tumor suppressor structure and function was revealed by the study of Adenovirus and SV40. *Virology.*, **384**, 274–284.
36. De Souza,R.F., Iyer,L.M. and Aravind,L. (2010) Diversity and evolution of chromatin proteins encoded by DNA viruses. *Biochim. Biophys. Acta*, **1799**, 302–318.
37. Topalis,D., Andrei,G. and Snoeck,R. (2013) The large tumor antigen: a 'Swiss Army knife' protein possessing the functions required for the polyomavirus life cycle. *Antiviral Res.*, **97**, 122–136.
38. Courtneidge,S.A., Goutebroze,L., Cartwright,A., Heber,A., Scherneck,S. and Feunteun,J. (1991) Identification and characterization of the hamster polyomavirus middle T antigen. *J. Virol.*, **65**, 3301–3308.
39. Fluck,M.M. and Schaffhausen,B.S. (2009) Lessons in signaling and tumorigenesis from polyomavirus middle T antigen. *Microbiol. Mol. Biol. Rev.*, **73**, 542–563.
40. Magnusson,G., Nilsson,M.G., Dilworth,S.M. and Smolar,N. (1981) Characterization of polyoma mutants with altered middle and large T-antigens. *J. Virol.*, **39**, 673–683.
41. Yi,X. and Freund,R. (1998) Deletion of proline-rich domain in polyomavirus T antigens results in virus partially defective in transformation and tumorigenesis. *Virology*, **248**, 420–431.
42. Cheng,J., DeCaprio,J.A., Fluck,M.M. and Schaffhausen,B.S. (2009) Cellular transformation by Simian Virus 40 and Murine Polyoma Virus T antigens. *Semin. Cancer Biol.*, **19**, 218–228.
43. Feltkamp,M.C., Kazem,S., van der Meijden,E., Lauber,C. and Gorbalenya,A.E. (2013) From Stockholm to Malawi: recent developments in studying human polyomaviruses. *J. Gen. Virol.*, **94**, 482–496.
44. Johne,R., Buck,C.B., Allander,T., Atwood,W.J., Garcea,R.L., Imperiale,M.J., Major,E.O., Ramqvist,T. and Norkin,L.C. (2011) Taxonomical developments in the family Polyomaviridae. *Arch. Virol.*, **156**, 1627–1634.
45. Gorbalenya,A.E., Lieutaud,P., Harris,M.R., Coutard,B., Canard,B., Kleywegt,G.J., Kravchenko,A.A., Samborskiy,D.V., Sidorov,I.A., Leontovich,A.M. *et al.* (2010) Practical application of bioinformatics by the multidisciplinary VIZIER consortium. *Antiviral Res.*, **87**, 95–110.
46. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
47. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
48. Castresana,J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.
49. Drummond,A.J., Suchard,M.A., Xie,D. and Rambaut,A. (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.*, **29**, 1969–1973.
50. Whelan,S. and Goldman,N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, **18**, 691–699.
51. Drummond,A.J., Ho,S.Y., Phillips,M.J. and Rambaut,A. (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol.*, **4**, e88.
52. Murrell,B., Wertheim,J.O., Moola,S., Weighill,T., Scheffler,K. and Kosakovsky Pond,S.L. (2012) Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.*, **8**, e1002764.
53. Pagel,M., Meade,A. and Barker,D. (2004) Bayesian estimation of ancestral character states on phylogenies. *Syst. Biol.*, **53**, 673–684.
54. Paradis,E., Claude,J. and Strimmer,K. (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20**, 289–290.
55. Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
56. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
57. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
58. Linding,R., Jensen,L.J., Diella,F., Bork,P., Gibson,T.J. and Russell,R.B. (2003) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.
59. Ward,J.J., Sodhi,J.S., McGuffin,L.J., Buxton,B.F. and Jones,D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.
60. Deng,X., Eickholt,J. and Cheng,J. (2009) PreDisorder: ab initio sequence-based prediction of protein disordered regions. *BMC Bioinformatics*, **10**, 436.
61. Prilusky,J., Felder,C.E., Zeev-Ben-Mordehai,T., Rydberg,E.H., Man,O., Beckmann,J.S., Silman,I. and Sussman,J.L. (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, **21**, 3435–3438.
62. Linding,R., Russell,R.B., Neduva,V. and Gibson,T.J. (2003) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.
63. Dosztanyi,Z., Csizmok,V., Tompa,P. and Simon,I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.
64. Yang,Z.R., Thomson,R., McNeil,P. and Esnouf,R.M. (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, **21**, 3369–3376.
65. Vucetic,S., Brown,C.J., Dunker,A.K. and Obradovic,Z. (2003) Flavors of protein disorder. *Proteins*, **52**, 573–584.
66. Rost,B., Yachdav,G. and Liu,J. (2004) The PredictProtein server. *Nucleic Acids Res.*, **32**, W321–W326.
67. Buchan,D.W., Ward,S.M., Lobley,A.E., Nugent,T.C., Bryson,K. and Jones,D.T. (2010) Protein annotation and modelling servers at University College London. *Nucleic Acids Res.*, **38**, W563–W568.
68. Gruber,A.R., Neubock,R., Hofacker,I.L. and Washietl,S. (2007) The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. *Nucleic Acids Res.*, **35**, W335–W338.
69. Goodman,S.N. (1999) Toward evidence-based medical statistics. 2: The Bayes factor. *Ann. Intern. Med.*, **130**, 1005–1013.
70. Rogozin,I.B., Spiridonov,A.N., Sorokin,A.V., Wolf,Y.I., Jordan,I.K., Tatusov,R.L. and Koonin,E.V. (2002) Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet.*, **18**, 228–232.
71. Firth,C., Kitchen,A., Shapiro,B., Suchard,M.A., Holmes,E.C. and Rambaut,A. (2010) Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. *Mol. Biol. Evol.*, **27**, 2038–2051.
72. Zheng,Z.M. (2010) Viral oncogenes, noncoding RNAs, and RNA splicing in human tumor viruses. *Int. J. Biol. Sci.*, **6**, 730–755.
73. Fisher,C.K. and Stultz,C.M. (2011) Constructing ensembles for intrinsically disordered proteins. *Curr. Opin. Struct. Biol.*, **21**, 426–431.
74. Van Roey,K., Gibson,T.J. and Davey,N.E. (2012) Motif switches: decision-making in cell regulation. *Curr. Opin. Struct. Biol.*, **22**, 378–385.
75. Perkins,J.R., Diboun,I., Dessailly,B.H., Lees,J.G. and Orengo,C. (2010) Transient protein-protein interactions: structural, functional, and network properties. *Structure*, **18**, 1233–1243.
76. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

77. Kovacs,E., Tompa,P., Liliom,K. and Kalmar,L. (2010) Dual coding in alternative reading frames correlates with intrinsic protein disorder. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 5429–5434.
78. Deribe,Y.L., Pawson,T. and Dikic,I. (2010) Post-translational modifications in signal integration. *Nat. Struct. Mol. Biol.*, **17**, 666–672.
79. Gu,X., Zou,Y., Su,Z., Huang,W., Zhou,Z., Arendsee,Z. and Zeng,Y. (2013) An update of DIVERGE software for functional divergence analysis of protein family. *Mol. Biol. Evol.*, **30**, 1713–1719.
80. Gorbalenya,A.E. and Snijder,E.J. (1996) Viral cysteine proteinases. *Persp. Drug Discov. Design*, **6**, 64–86.
81. Koonin,E.V. and Gorbalenya,A.E. (1989) Tale of two serines. *Nature*, **338**, 467–468.
82. King,A.M.Q., Adams,M.J., Carstens,E.B. and Lefkowitz,E.J. (2012) *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*. Academic Press, London, http://www.sciencedirect.com/science/book/9780123846846.
83. Breitbart,M. and Rohwer,F. (2005) Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.*, **13**, 278–284.