PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

rstb.royalsocietypublishing.org

Introduction

Cite this article: Katzourakis A. 2013 Paleovirology: inferring viral evolution from host genome sequence data. Phil Trans R Soc B 368: 20120493.
http://dx.doi.org/10.1098/rstb.2012.0493

One contribution of 13 to a Theme Issue 'Paleovirology: insights from the genomic fossil record'.

Subject Areas:
bioinformatics, genomics, paleontology, virology

Keywords:
virus evolution, genomics, endogenous viral element, endogenous retrovirus

Author for correspondence:
Aris Katzourakis
e-mail: aris.katzourakis@zoo.ox.ac.uk

Paleovirology: inferring viral evolution from host genome sequence data

Aris Katzourakis

Department of Zoology, University of Oxford, Oxford OX1 3PS, UK

1. Introduction

Paleovirology is the study of ancient viruses, typically over prehistoric or geological timescales. There is no physical 'fossil record' of viruses; virions persist for short time periods, and rapidly degrade leaving no direct trace of their existence. Many viruses can enter the genomes of their hosts—some, such as retroviruses, do so as an obligate step during their replication process, and others can occasionally do so, either by accident or as a latent part of their life cycle. When viral integrations occur in the germline of their host, they can be passed on to the next generation, potentially fixing in the host population. When this occurs, the integrated endogenous virus genomes evolve at host rates of mutation, and their sequence is relatively stably preserved. The study of this genomic 'fossil record' has led to the burgeoning field of paleovirology, which uses these endogenous viruses to disentangle the long-term evolutionary history of virus–host interactions.

Viruses are a pervasive feature of both eukaryotic and prokaryotic life, and offer some of the best-studied examples of evolution in action due to their rapid rates of evolution. We have observed the diversification of HIV into distinct subtypes over the past six decades, variously prevalent in different patient risk groups and geographical locations, as well as the selection of drug and immune escape mutations within the course of single infections [1]. Rates of viral evolution are so rapid that viruses are among the few organisms which can be used to validate the tools used to reconstruct molecular phylogenies [2]. However, at the same time, perhaps more so than for other biological entities, the long-term evolutionary history of viruses is far less well understood. This is, in part, because the rapid mutation rates that enable the study of their short-term evolution erode the signal of molecular evolution in nucleotide sequences [3], but also due to their physical fragility. Thus, the genomic fossil record is the only way to study viral evolutionary history on timescales spanning millions of years.

2. Scope

The presence of retroviral sequences in host genomes was noticed in the late 1960s, but these observations were initially greeted with scepticism. This special issue begins with a paper that charts the history of how endogenous retroviruses (ERVs) first came to light, and outlines the discoveries that led to the current understanding of viral sequences in host genomes [4]. Recent work has shown that all known viral genomic structures and replication strategies are represented in this fossil record, greatly expanding the scope of paleovirology beyond just the retroviruses [5]. These endogenous viral elements are known as EVEs, a broad term that encompasses both ERVs and non-retroviral integrations. The past few years have seen a number of high-profile reviews [6–9], and the work represented within them, alongside the papers included in this collection, mark the establishment of the field of paleovirology. This special issue brings together papers that span a range of approaches, including computational, theoretical and experimental, showcasing the varied approaches that are being undertaken and synthesizing the key insights within an evolutionary framework.

(a) Direct and indirect paleovirology

Retroviral sequences within the genomes of their hosts offer a direct way to date integration events, providing a chronological record of ancient infections. The

paired long terminal repeats at each end of a retrovirus are identical at the time of integration, and the divergence between them can be divided by the host neutral rate of evolution to provide an age estimate. This approach can be extended to any form of duplicated sequence within the host genome, and is therefore not limited to retroviruses. Furthermore, the identification of orthologous viral sequences between divergent host species provides a powerful means of obtaining minimum age estimates for very ancient viruses based on the known speciation dates of the hosts [10]. In this issue, Lee *et al*. [11] use this approach to date an ancient ERV orthologue and establish that it integrated prior to the divergence of placental mammals, approximately 100 Ma. Similar ancient dates have been inferred for foamy viruses [12], hepadnaviruses [13], circoviruses and bornaviruses [5].

Many early studies of ERV diversity focused on the evolutionary history of the reverse-transcriptase domain of the retroviral polymerase gene. Retroviruses are, however, known to recombine, and Henzy & Johnson [14] review the literature describing recombination events between the retroviral polymerase gene and the envelope gene, which have surprisingly distinctive evolutionary histories. The envelope protein plays a key role in defining host range, and the authors contrast the gammaretroviruses, that have a broad host range, with the betaretroviruses, that have a much narrower host range. On a number of occasions, betaretroviruses have acquired gammaretrovirus-derived *env* genes, allowing them to colonize a novel host. The resulting change in host range could drive viral diversification and open up new niches [14], and the swapping of envelopes can also lead to reactivation of non-functional ERVs [15].

Approaches that rely on the identification of viral sequences in host genomes have been termed 'direct paleovirology'. However, the existence of ancient viruses can also be inferred by investigation of their effects on the host genes that have evolved to control them. This is particularly true of innate immune genes such as the APOBEC family, TRIM5α TRIMCyp and SAMHD1, all of which are potent inhibitors of viral activity, their evolution shaped by conflict with viruses (see [16,17] for reviews). We can infer the presence of ancient paleoviruses by looking at differences in the activity of antiviral genes with a known target across host species, and reconstructing the history of diversifying selection that is characteristic of conflict with a pathogen. This approach has been termed 'indirect paleovirology' and is a powerful technique to trace both the ancient history of this conflict and how it has shaped susceptibility to modern viruses. Furthermore, this indirect approach can be the only way to study ancient infections in cases where EVEs cannot be identified [6]. Most viruses will not leave a convenient trace of their passage in the form of EVEs; thus this approach is an extremely valuable addition to the paleovirological toolkit.

Indirect paleovirology is most powerful where the interaction between a viral protein and host protein are known. It can be limited in its ability to formally prove the existence of a paleovirus and to rule out alternative scenarios; for example, in cases where the viruses that have shaped the evolution of a particular gene are unknown or extinct. The most robust studies of viral–host interaction will come from examples of conflicts that have involved viruses with at least a partial EVE record, combining direct and indirect approaches. Compton *et al*. [18] investigate the evolution of antiviral genes known to restrict lentiviruses in primates, and conclude that the ancestors of modern simian immunodeficiency virus (SIV) existed in simian primates more than 10 Ma, and may, in fact, have episodically infected simians for most of their evolutionary history. While EVEs of the simian lineage of lentiviruses have not been identified, endogenous lentiviruses in rabbits, lemurs and ferrets suggest that the lentiviruses are at least 12 Myr old [10,19–21]. This age is consistent with the indirect paleovirological estimate and can be seen as an independent validation of the approach.

The genomic record formed by EVEs is essentially a fossil record by analogy only, but there are other forms of viral fossil—not preserved in sediment, but rather, viral particles or viral genomic sequences that have serendipitously been preserved in substrates such as ice, or biopsy samples stored in paraffin. The oldest preserved viral samples are no more than a few hundred years old [22], but have proved instrumental in studies of viral history, for example pushing back the origins of the HIV pandemic to near the beginning of the twentieth century [23]. These samples are more akin to historical or archaeological records than to paleontological records, and cannot be used to elucidate viral history over timescales spanning millions of years. The oldest authenticated ancient DNA samples are in the order of tens to hundreds of thousands of years old, and it is almost inconceivable that DNA that is millions of years old could be sequenced from an animal, let alone DNA from a much smaller and more fragile viral genome. In this issue, Gray *et al*. [24] investigate hepatitis C virus sequences obtained from sera sampled and frozen in 1953 to clarify the spread of this pathogen in the first half of the twentieth century. This study highlights some of the difficulties of studying viral evolution in the absence of EVE data, which have not been identified for hepatitis C, even with temporally sampled viral sequences.

(b) Functional approaches to paleovirology

EVEs have been studied using genomic data and evolutionary analysis, but it is possible to go beyond *in silico* approaches in several ways. Pioneering studies have shown that it is technically possible to re-create extinct viruses and study their biological properties by reconstructing the functional ancestral sequence of EVEs that have been inactivated by neutral mutations, and synthesizing the viral proteins [25–27]. This offers an opportunity to study the interactions between ancient viruses and their hosts in an *in vitro* setting. For safety reasons, not all the proteins that constitute a virus are synthesized, and important biological insights can be gained by substituting parts of viral proteins in well-established laboratory viral vectors. Using this approach, Goldstone *et al*. [27] have shown that the endogenous lentiviruses RELIK and pSIV encode a capsid with a functional cyclophilin-binding loop, proving conservation of this lentiviral phenotype for at least 12 Myr. In this issue, Yap & Stoye [28] take this approach further to test the hypothesis that the evolution of the lagomorph antiviral TRIM5α protein has been shaped by infection with ancient RELIK-like lentiviruses. They describe the variable restrictive ability of TRIM5α in different lagomorphs (rabbits, hares and pikas), offering evidence in favour of the idea that ancient lentiviruses have shaped the evolution of the restriction factors of their hosts. Their work exemplifies the complementarity of results from functional, direct and indirect paleovirology research.

EVEs can also be co-opted by their hosts to perform new functional roles, provided that the ancient integrations conferred

a selective advantage, a process termed exaptation. The *syncytin* gene in mammals represents one of the best-understood examples of viral genes that have been exapted by their hosts. Knockout experiments in mice have demonstrated that these are essential genes for placental development and embryonic survival, because cell–cell fusion at the fetal–maternal interface leads to formation of the syncytiotrophoblast [29]. Intriguingly, *syncytin* acquisition from distinct viruses has occurred independently at least seven times, each event happening after the divergence of the mammalian orders in which they are found. This raises an important evolutionary paradox—how is it possible that an essential function of the placenta, which itself has an ancient origin, is facilitated by these relatively recently acquired viral genes? Lavialle *et al*. [29] propose a model to explain this remarkable case of convergent evolution, where a founding retroviral *env* gene was subsequently replaced by independent subsequent acquisitions in diverse mammalian lineages. Evidence in favour of this model comes from 'lost *syncytins*', which have maintained fusogenic activity in Old World monkeys but have lost functionality in apes [30].

The question of the initial role of the founding *syncytin*-like gene remains, as its current essential role implies an intermediate state that offered a direct fitness advantage to its host. One possibility is that endogenized *env* genes, many of which have immunomodulatory domains, initially provided an immunosuppressive role that conferred tolerance to fetal antigens [29,31]. Exaptation after an initial co-option for a more general function would explain the scale of convergence and at the same time the diversity of placental form, a phenomenon that has also been ascribed to maternal–offspring conflict [32,33]. Alternatively, the immunosuppressive role could have played an antiviral function, and it has been suggested that co-option for antiviral properties facilitates gene transfer from viruses to their hosts [9].

In this issue, Herniou *et al*. [34] discuss an entirely different example of viral gene domestication, the integration of functionally active polydnavirus genomes within the genomes of multiple parasitoid wasps that have also occurred in a convergent manner. *Bracovirus* and *Ichnovirus* originated from the integration of distinct viruses into two separate lineages of parasitoid wasp. The parasitoid wasps inject viral particles into their lepidopteran hosts during oviposition in a process that is essential to their life cycle (reviewed by [34]). Bézier *et al*. [35] show that the 'macrolocus' that is responsible for this process in *Cotesia congregata* comprises two proviral loci that are joined by a region containing host wasp genes, resulting from a complex series of genomic rearrangements.

A potentially co-opted EVE has been identified in our own genomes. Horie *et al*. [36] review work that was performed on the analysis of EVE sequences derived from bornaviruses. Intriguingly, humans encode a bornaviral-derived protein from an open reading frame of the *N* gene, which has led to suggestions that it may be functional in humans, perhaps to combat other viral infections. Horie *et al*. [36] also present results from another bornavirus EVE in the 13-lined ground squirrel that has endogenized within the past 6 Myr. Furthermore, it could still be polymorphic within the host population, offering a potential model system to study the evolutionary dynamics of co-opted EVEs during the early stages of their formation. There are only a small number of cases of EVEs that can be observed during their endogenization, most notably the case of KoRV, a gammaretrovirus infecting koalas [37]. Given the rarity of endogenization, attempts to study them empirically are problematic. In this issue, Kanda *et al*. [38] outline a mathematical model that explores the effects of immunity and life history on the evolution of an ERV occupying a single locus within the population of its host.

The approaches discussed thus far may also be useful in exploring the deleterious consequences of EVEs. Retroviruses with both endogenous and exogenous forms that are associated with disease are known from a range of mammals (e.g. mouse mammary tumour virus, Jaagsiekte, avian leukosis virus, murine leukemia virus), but the role of ERVs in human disease is far more controversial. This is partly due to the difficulties of studying their effects in humans but also to the chequered history of putative associations of ERVs with disease that have either proved impossible to replicate or represent laboratory contamination [31,39]. Biotechnological landmarks of the past decade, including the sequencing of the human genome and also the prospect of multiple genomes from different individuals and tissue types, open up new avenues of research into the association of ERVs with disease in humans. Magiorkinis *et al*. [31] suggest that the time is right to revisit and definitively address the association of ERVs with disease in the post-genomic era within a robust and carefully structured framework that makes full use of genomic data.

3. Concluding remarks

This special issue includes contributions that showcase the use of computational, experimental and theoretical approaches, to address a range of questions revealing the themes that are particularly reflective of recent literature. The authors have used direct, indirect and functional paleovirology approaches to highlight the antiquity of viruses themselves, as well as their interactions with their hosts. In addition to describing the natural history of ancient viral infections, the contributions highlight the consequences of these ancient interactions to their hosts, as exemplified by the presence of multiple exapted viral sequences. Interestingly, an unplanned emergent theme across this special issue is the emphasis that has been placed on the retroviral envelope gene [14,29,31]. It is fitting that the gene that is at the front line of the battle between viruses and their hosts appears to be at the forefront of recent research—pushing the endogenous envelope, as Henzy & Johnson [14] put it.

Paleovirological studies are in a phase of exponential growth. Data gathering and hypothesis generation are regularly interlinked, with unexpected fascinating insights emerging from data collected for different purposes. Paleovirology will come to encompass increasingly quantitative approaches as the number of described EVEs increases. The initial draft of the human genome was completed in 2001, giving rise to early papers that focused on the complement of ERVs within it, alongside studies that sequenced relatively limited genomic regions from multiple taxa (reviewed here by [4,14,31]). The ensuing decade has seen studies of multiple animal, plant and fungal genomes, including systematic surveys in some cases (reviewed by [6–9,36]). The genomic data that have been produced reveal a wealth of ancient viral diversity, a record of viral forms stretching back to the Cretaceous. The increasing availability of genomic and transcriptomic data from many individuals within a species

and multiple tissue types, including comparisons of germline and somatic genomes, will enable us to better understand the consequences of the ancient relationship between viruses and their hosts. These data can be properly understood within the framework made possible by our current understanding that both viruses and the genomes of their hosts have been shaped by a conflict that has spanned geological timescales. While paleovirology has emerged as a field in its own right, it has also become an integral part of modern virology and has redefined our conception of what a genome is.

References

1. Rambaut A, Posada D, Crandall KA, Holmes EC. 2004 The causes and consequences of HIV evolution. *Nat. Rev. Genet.* **5**, 52–61. (doi:10.1038/nrg1246)
2. Cunningham CW, Zhu H, Hillis DM. 1998 Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. *Evolution* **52**, 978–987. (doi:10.2307/2411230)
3. Holmes EC. 2003 Molecular clocks and the puzzle of RNA virus origins. *J. Virol.* **77**, 3893–3897. (doi:10.1128/JVI.77.7.3893-3897.2003)
4. Weiss RA. 2013 On the concept and elucidation of endogenous retroviruses. *Phil. Trans. R. Soc. B* **368**, 20120494. (doi:10.1098/rstb.2012.0494)
5. Katzourakis A, Gifford RJ. 2010 Endogenous viral elements in animal genomes. *PLoS Genet.* **6**, e1001191. (doi:10.1371/journal.pgen.1001191)
6. Patel MR, Emerman M, Malik HS. 2011 Paleovirology—ghosts and gifts of viruses past. *Curr. Opin. Virol.* **1**, 304–309. (doi:10.1016/j.coviro.2011.06.007)
7. Feschotte C, Gilbert C. 2012 Endogenous viruses: insights into viral evolution and impact on host biology. *Nat. Rev. Genet.* **13**, U283–U288. (doi:10.1038/nrg3199)
8. Holmes EC. 2011 The evolution of endogenous viral elements. *Cell Host Microbe* **10**, 368–377. (doi:10.1016/j.chom.2011.09.002)
9. Aswad A, Katzourakis A. 2012 Paleovirology and virally derived immunity. *Trends Ecol. Evol.* **27**, 627–636. (doi:10.1016/j.tree.2012.07.007)
10. Keckesova Z, Ylinen LMJ, Towers GJ, Gifford RJ, Katzourakis A. 2009 Identification of a RELIK orthologue in the European hare (*Lepus europaeus*) reveals a minimum age of 12 million years for the lagomorph lentiviruses. *Virology* **384**, 7–11. (doi:10.1016/j.virol.2008.10.045)
11. Lee A, Nolan A, Watson J, Tristem M. 2013 Identification of an ancient endogenous retrovirus, predating the divergence of the placental mammals. *Phil. Trans. R. Soc. B* **368**, 20120503. (doi:10.1098/rstb.2012.0503)
12. Katzourakis A, Gifford RJ, Tristem M, Gilbert MTP, Pybus OG. 2009 Macroevolution of complex retroviruses. *Science* **325**, 1512. (doi:10.1126/science.1174149)
13. Suh AB, Jurgen B, Schitz J, Kriegs JO. In press. The genome of a mesozoic paleovirus reveals the evolution of hepatitis B viruses. *Nat. Commun.* **4**, 1791. (doi:10.1038/ncomms2798)
14. Henzy JE, Johnson WE. 2013 Pushing the endogenous envelope. *Phil. Trans. R. Soc. B* **368**, 20120506. (doi:10.1098/rstb.2012.0506)
15. Young GR, Eksmond U, Salcedo R, Alexopoulou L, Stoye JP, Kassiotis G. 2012 Resurrection of endogenous retroviruses in antibody-deficient mice. *Nature* **491**, 774–778. (doi:10.1038/nature11599)
16. Daugherty MD, Malik HS. 2012 Rules of engagement: molecular insights from host-virus arms races. In: *Annu. Rev. Genet.* **46**, 677–700. (doi:10.1146/annurev-genet-110711-155522)
17. Duggal NK, Emerman M. 2012 Evolutionary conflicts between viruses and restriction factors shape immunity. *Nat. Rev. Immunol.* **12**, 687–695. (doi:10.1038/nri3295)
18. Compton AA, Malik HS, Emerman M. 2013 Host gene evolution traces the evolutionary history of ancient primate lentiviruses. *Phil. Trans. R. Soc. B* **368**, 20120496. (doi:10.1098/rstb.2012.0496)
19. Katzourakis A, Tristem M, Pybus OG, Gifford RJ. 2007 Discovery and analysis of the first endogenous lentivirus. *Proc. Natl Acad. Sci. USA* **104**, 6261–6265. (doi:10.1073/pnas.0700471104)
20. Gifford RJ, Katzourakis A, Tristem M, Pybus OG, Winters M, Shafer RW. 2008 A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution. *Proc. Natl Acad. Sci. USA* **105**, 20 362–20 367. (doi:10.1073/pnas.0807873105)
21. Han G-Z, Worobey M. 2012 Endogenous lentiviral elements in the weasel family (*Mustelidae*). *Mol. Biol. Evol.* **29**, 2905–2908. (doi:10.1093/molbev/mss126)
22. Biagini P *et al*. 2012 Variola virus in a 300-year-old Siberian mummy. *N. Engl. J. Med.* **367**, 2057–2059. (doi:10.1056/NEJMc1208124)
23. Worobey M *et al*. 2008 Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* **455**, 661–664. (doi:10.1038/nature07390)
24. Gray RR, Tanaka Y, Takebe Y, Magiorkinis G, Buskell Z, Seeff L, Alter HJ, Pybus OG. 2013 Evolutionary analysis of hepatitis C virus gene sequences from 1953. *Phil. Trans. R. Soc. B* **368**, 20130168. (doi:10.1098/rstb.2013.0168)
25. Dewannieux M, Harper F, Richaud A, Letzelter C, Ribet D, Pierron G, Heidmann T. 2006 Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. *Genome Res.* **16**, 1548–1556. (doi:10.1101/gr.5565706)
26. Lee YN, Bieniasz PD. 2007 Reconstitution of an infectious human endogenous retrovirus. *PLoS Pathog.* **3**, 119–130. (doi:10.1371/journal.ppat.0030119)
27. Goldstone DC, Yap MW, Robertson LE, Haire LF, Taylor WR, Katzourakis A, Stoye JP, Taylor IA. 2010 Structural and functional analysis of prehistoric lentiviruses uncovers an ancient molecular interface. *Cell Host Microbe* **8**, 248–259. (doi:10.1016/j.chom.2010.08.006)
28. Yap MW, Stoye JP. 2013 Apparent effect of rabbit endogenous lentivirus type K acquisition on retrovirus restriction by lagomorph TRIM5αs. *Phil. Trans. R. Soc. B* **368**, 20120498. (doi:10.1098/rstb.2012.0498)
29. Lavialle C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, Vernochet C, Heidmann T. 2013 Paleovirology of '*syncytins*', retroviral *env* genes exapted for a role in placentation. *Phil. Trans. R. Soc. B* **368**, 20120507. (doi:10.1098/rstb.2012.0507)
30. Esnault C, Cornelis G, Heidmann O, Heidmann T. 2013 Differential evolutionary fate of an ancestral primate endogenous retrovirus envelope gene, the EnvV syncytin, captured for a function in placentation. *PLoS Genet.* **9**. (doi:10.1371/journal.pgen.1003400)
31. Magiorkinis G, Belshaw R, Katzourakis A. 2013 'There and back again': revisiting the pathophysiological roles of human endogenous retroviruses in the post-genomic era. *Phil. Trans. R. Soc. B* **368**, 20120504. (doi:10.1098/rstb.2012.0504)
32. Malik HS. 2012 Retroviruses push the envelope for mammalian placentation. *Proc. Acad. Natl Sci. USA* **109**, 2184–2185. (doi:10.1073/pnas.1121365109)
33. Haig D. 2012 Retroviruses and the placenta. *Curr. Biol.* **22**, R609–R613. (doi:10.1016/j.cub.2012.06.002)
34. Herniou EA, Huguet E, Thézé J, Bézier A, Periquet G, Drezen J-M. 2013 When parasitic wasps hijacked viruses: genomic and functional evolution of polydnaviruses. *Phil. Trans. R. Soc. B* **368**, 20130051. (doi:10.1098/rstb.2013.0051)
35. Bézier A *et al*. 2013 Functional endogenous viral elements in the genome of the parasitoid wasp *Cotesia congregata*: insights into the evolutionary dynamics of bracoviruses. *Phil. Trans. R. Soc. B* **368**, 20130047. (doi:10.1098/rstb.2013.0047)
36. Horie M, Kobayashi Y, Suzuki Y, Tomonaga K. 2013 Comprehensive analysis of endogenous bornavirus-like elements in eukaryote genomes. *Phil. Trans. R. Soc. B* **368**, 20120499. (doi:10.1098/rstb.2012.0499)
37. Tarlinton RE, Meers J, Young PR. 2006 Retroviral invasion of the koala genome. *Nature* **442**, 79–81. (doi:10.1038/nature04841)
38. Kanda RK, Tristem M, Coulson T. 2013 Exploring the effects of immunity and life history on the dynamics of an endogenous retrovirus. *Phil. Trans. R. Soc. B* **368**, 20120505. (doi:10.1098/rstb.2012.0505)
39. Voisset C, Weiss RA, Griffiths DJ. 2008 Human RNA 'rumor' viruses: the search for novel human retroviruses in chronic disease. *Microbiol. Mol. Biol. Rev.* **72**, 157–196. (doi:10.1128/MMBR.00033-07)