Title: Prisoner of War dynamics explains the time-dependent pattern of substitution rates in viruses

Mahan Ghafari[1], Peter Simmonds[2], Oliver G Pybus[1], Aris Katzourakis[1,*]

1 – Department of Zoology, University of Oxford
2 – Nuffield Department of Medicine, University of Oxford

* E-mail: aris.katzourakis@zoo.ox.ac.uk

Abstract

Molecular clock dating is widely used to estimate timescales of phylogenetic histories and to infer rates at which species evolve. One of the major challenges with inferring rates of molecular evolution is the observation of a strong correlation between estimated rates and the timeframe of their measurements. Recent empirical analysis of virus evolutionary rates suggest that a power-law rate decay best explains the time-dependent pattern of substitution rates and that the same pattern is observed regardless of virus type (*e.g.* groups I-VII in the Baltimore classification). However there exists no explanation for this trend based on molecular evolutionary mechanisms. We provide a simple predictive mechanistic model of the time-dependent rate phenomenon, incorporating saturation and host constraints on the evolution of some sites. Our model recapitulates the ubiquitous power-law rate decay with a slope of -0.65 (95% HPD: -0.72, -0.52) and can satisfactorily account for the variation in inferred molecular evolutionary rates over a wide range of timeframes. We show that once the saturation of sites starts - typically after hundreds of years in RNA viruses and thousands of years in DNA viruses - standard substitution models fail to correctly estimate divergence times among species, while our model successfully re-creates the observed pattern of rate decay. We apply our model to re-date the diversification of genotypes of hepatitis C virus (HCV) to 396,000 (95% HPD: 326,000 - 425,000) years before present, a time preceding the dispersal of modern humans out of Africa, and also showed that the most recent common ancestor of sarbecoviruses dates back to 23,500 (95% HPD: 21,100 - 25,300) years ago, nearly thirty times older than previous estimates. This not only creates a radical new perspective for our understanding the origins of HCV but also suggests a substantial revision of evolutionary timescales of other viruses can be similarly achieved.

Main

The timescale over which viruses evolve and how this process is connected to host adaptation has been an area of considerable research and methodological progress in recent decades. Mammalian RNA viruses, in particular, exhibit extraordinarily rapid genomic change[1-3] and analyses of their genetic variation has enabled detailed reconstruction of the emergence of viruses such as HIV-1[4], hepatitis C virus[5] and influenza A virus[6,7]. RNA viruses display evolutionary change over short-timescales (weeks to months) and can re-model a substantial part of their genomes following a host switch[8-13]. Well characterised examples in both RNA and DNA viruses include the emergence of HIV-1 in humans from a chimpanzee reservoir[4,14,15], and the adaptation of myxomatosis in rabbits[16].

These rapid rates of virus sequence change stand in striking contrast with evidence for extreme conservation of virus genome sequences over longer periods of evolution and at higher taxonomic

levels[17,18]. Inferred short term rates of virus sequence change should create completely unrecognisable genome sequences if they were extrapolated over thousands or even hundreds of years, yet endogenous viral elements (EVEs) that integrated into host genomes throughout mammalian evolution are recognisably similar to contemporary genera and families of *Bornaviridae*, *Parvoviridae* and *Circoviridae* amongst many examples[19-24]. This observation is complemented by increasing evidence from studies of virus / host co-evolution[25-27], and more recently from analysis of viruses recovered from ancient DNA in archaeological remains[28-32], that together indicate a remarkable degree of conservatism in viral genome sequences and their inter-relationships at genus and family levels. This conundrum has been attributed to the time-dependent rate phenomenon (TDRP), which is the observation that apparent rates of evolution are dependent on timescale of measurement. The TDRP has been explained by processes such as sequence site saturation, short-sighted within-host evolution, short-term changes in selection pressure and potential errors in estimation of short-term substitution rates[17,18,33-35]. Empirically, substitution rates show a striking linear relationship between log-transformed rates and timescale of measurement, across RNA and DNA viruses, despite the large differences among viruses in their initial short term substitution rates[33]. The gradient of regressions of observation time to estimated evolutionary rates is consistently around -0.65, for all virus groups for which long term substitution rates can be calculated or inferred, implying a common underlying evolutionary process.

We recently developed a model of virus evolution in which the primary driver of sequence change over long evolutionary time-scales was host adaptation, in which virus sequence change is severely curtailed by stringent fitness constraints[36]. Viruses exist within a tightly-constraining host niche to which they rapidly adapt; paradoxically, their high mutation rates, large population sizes and consequent ability to adapt rapidly serve to restrict their long-term diversification and sustained sequence change, rendering them evolutionary "Prisoners of War" (PoW). Ultimately over longer timescales, rates of viral evolution will be bounded by the rate of evolution of their hosts [36].

In the current study, we develop the PoW model to explain the longer-term evolutionary trajectories of viruses and the time-dependence of their inferred substitution rates. The model accounts for genetic saturation of rapidly-evolving sites, and host constraints on site evolution, with the proportion of fast- to slow-evolving sites being exponentially distributed. These sites will saturate chronologically from the fastest evolving to those that evolve epistatically, to those that evolve at the host substitution rate. This model can reproduce effectively the empirically observed TDRP patterns and the inflection points where time-dependent rate changes become manifest. We demonstrate that the model predictions are robust to intrinsic and marked differences in substitution rates among different virus groups or assumptions about the relative proportion of sites evolving at different rates.

Power-law rate decay due to site-saturation

First, we show how a time-dependent rate effect emerges when estimating the sequence divergence using a standard evolutionary model. Suppose that a sequence has diverged from its ancestor t generations ago under a constant and uniform substitution rate, μ. The proportion of pairwise differences between the sequence and its ancestor, $p(t)$, reaches its maximum value, hereafter called the saturation frequency, α, at time $t^* \approx \alpha / \mu$ (see **Equation S1**). As the ancestral and derived sequence continue to evolve beyond the saturation point, t^*, their observed proportion of pairwise

differences, $\hat{p}$, is effectively unchanged. Thus, using any conventional substitution model, the inferred genetic distance, $\hat{d}$, remains constant and the estimated substitution rate, $\hat{\mu}$, follows a power-law drop as the observed divergence time, $\hat{t}$, increases , i.e. $\hat{\mu} \sim 1/\hat{t}$, with slope -1 on a log-log graph (see **Figure S1 a**) – this is assuming the substitution model correctly infers the saturation frequency, i.e. $\alpha_M \geq \alpha$, where α_M is the expected saturation frequency set by the model (see **Equation S2**).

Saturation under rate heterogeneity

In the presence of among-site rate heterogeneity, a fraction of sites may evolve at a rate that is much slower (or faster) than some other sites. Under such circumstances, if we apply a measure of genetic distance based on a constant and homogeneous substitution rate per site, the substitution model may reliably recover the expected (mean) rate up to and before the fastest-evolving sites reach saturation, beyond which point the time-dependent pattern of rate decay emerges while the remaining sites continue to accumulate changes until the slowest-evolving sites also reach saturation (i.e. the point at which the entire sequence space has been fully explored) and a power-law rate decay with slope -1 emerges (see **Figure S1 b**).

Although our focus so far has been on the saturation of observed pairwise differences and how it can create a time-dependent rate effect, the same holds true when tracking the evolutionary changes of a large number of sequences through time. Using a standard Jukes-Cantor substitution model on a set of simulated sequences (see Methods section), both in the absence and presence of rate heterogeneity, we can recreate similar patterns of time-dependent rate decay and show that, over longer timescales, i.e. when the divergence time between two populations is much longer than the typical coalescent times $(2N_e)$ of a neutral population of size N_e, the variation in inferred substitution rates is dominated by the saturation along the longest (internal) branch connecting the two populations (**Figure S2**). We also find that, over short timescales, systematic under-estimation of the Time to the Most Recent Common Ancestor (TMRCA) results in inflated substitution rate estimates (**Equation S4**).

Saturation under the Prisoner of War model

The PoW model of virus evolution is based on the principle that site saturation over long timescales dominates the time-dependency of virus substitution rates. Building upon a collection of virus evolutionary rate estimates from >130 publications[33], we use 389 nucleotide substitution rate estimates across six major viral groups to find the line of best fit between our predicted time-dependent substitution rate (the PoW model) and the evolutionary rate estimates for each viral group (data) using the geometric least squares method[37]. We then show that the PoW model captures all of the important properties of the TDRP and its variation among different groups of viruses.

The model categorises into M discrete rate classes that are equally spaced on a log-scale, with a common ratio ΔM between consecutive rate groups, ranging from those evolving the fastest, at rate $\mu_{\max}$ per site per year (SSY), to the ones evolving at the host substitution rate, $\mu_{\min}$. The fraction of sites, m_i, in each rate group i, with the corresponding substitution rate, μ_i, is exponentially distributed, $m_i = Ce^{\lambda i}$, where C is the normalisation factor, i.e. $C = 1/\sum_{j=1}^{M} e^{\lambda j}$, and the exponent

coefficient, λ, sets the tendency of sites to be either mostly slowly ($\lambda < 0$) or rapidly ($\lambda > 0$) evolving (**Figure 1 a**). Assuming a fixed incremental difference between any two consecutive rate groups, i.e. $\mu_{i+1} = \Delta M \mu_i$, the substitution rate at the fastest-evolving sites is determined by the total number of groups, M, which, in turn, sets the inflection point for when the time-dependent rate decay emerges. Once the fastest-evolving sites diverge to saturation, other rate groups that evolve more slowly (e.g. via epistatic and compensatory substitutions), saturate sequentially as the timespan of rate measurement, t, increases. This chronological saturation effect continues until the inferred rate decays to the host substitution rate, $\mu_{\min}$ (**Figure 1 b**). Therefore, the time-dependent rate curve, according to the PoW model is given by

$$\hat{\mu}(t) = -\alpha_M Ln\left(1 - \frac{1}{\alpha_M}\sum_{i=1}^{M} \alpha m_i \left(1 - e^{-\mu_i t/\alpha}\right)\right)/t \qquad (1)$$

where the observed genetic distance between the derived and ancestral sequences is assumed to increase with the timespan of rate measurement, t. While over short timescales, i.e. $t \ll 1/\mu_{\max}$, several methodological (e.g. internal node calibration errors) and biological (e.g. purifying selection) artefacts may inflate the substitution rate estimates in viruses (i.e. such that $\hat{\mu}(t)$ underestimates the inferred substitution rates), over longer time-scales (i.e. after a few years) the rate estimates are expected to converge to the mean substitution rate[38,39], $\langle \mu \rangle = \sum_{i=1}^{M} m_i \mu_i$.

Upon the saturation of the fastest-evolving sites (i.e. the inflection point), $\hat{\mu}(t)$ follows the empirically observed power-law decay with the slope -0.65 $(95\%\ \mathrm{HPD} = -0.72, -0.57)$ across all viral groups (**Figure 1c-h**). Finding the exact point at which the rate decay pattern emerges likely depends on the choice of substitution model and varies between different virus groups. For instance, a previous study[40] has shown that substitution models with Gamma-distributed rate heterogeneities across sites may perform better at estimating the mean substitution rate over longer timescales, thereby delaying the emergence of the inflection point in the power-law rate decay, compared to models with a strict molecular clock.

We find that the mean substitution rates and fastest-evolving rate groups in double-stranded DNA viruses (dsDNA) is the lowest among all viral groups and that, together with reverse-transcribing DNA (RT-DNA) and single-stranded DNA viruses (ssDNA), dsDNA viruses have typically 1-2 orders of magnitude slower rates than RNA viruses (see **Table 1**). Conversely, the estimated mean and fastest rates are very similar among the groups of positive-strand RNA viruses (+ssRNA), negative-strand RNA viruses (-ssRNA), and RNA retroviruses (RT-RNA). The estimated substitution rates for rapidly evolving sites in RNA viruses, i.e. $\mu_{\max} \sim 10^{-2}$ SSY, resembles their inferred mutation rates[41] which may also result in their rapid saturation after a few decades[42].

To ensure that our model predictions are not biased towards a particular virus family with more evolutionary rate estimates (i.e. more data points to fit to the PoW model in **Equation 1**), we remove all the short-term rate estimates (< 100 years) within each viral group except for one virus family or genus to re-calibrate the mean substitution rates (see **Table S1**). We find that despite the broad range of evolutionary rate estimates across all viral groups, the estimated parameters of the PoW model are robust to such changes and are not an artifact of systematic biases in selecting rate estimates from a

particular virus family. We note that, in group VI, the stark difference in evolutionary rates between *Lentivirus* and *Deltaretrovirus* families over short timescales results in a noticeably different pattern of rate decay (see **Figure S3 e**) and the long-term rates are more aligned with the predictions based on the *Deltaretrovirus* re-calibration. The predicted mean and maximum substitution rate of *Lentivirus* families are 1-2 orders of magnitude higher than the *Deltaretrovirus* family. The latter evolves at rates similar to RT-DNA viruses. We also see that a larger fraction of sites in DNA viruses tend to evolve at rates closer to the host substitution rates (i.e. have a sharper negative gradient, λ) compared to RNA viruses, which largely have an equal fraction of sites across all rate groups. It is worth noting that all rate groups, including the ones with lowest proportion of sites, $m_{\min}$, have representative proportions across the genome, i.e. $m_{\min} \gg 10^{-4} > 1/L$, where L is a typical genome size of an RNA virus. We also carried out a similar sensitivity analysis at the level of virus genera which further confirms that the rate curves predicted by the PoW model are still accurate at this level and are not an artefact of measured rates at the level of Baltimore groups (see **Figure S4**).

To illustrate the radical effect of applying the PoW model to virus evolutionary timescales, we analysed an alignment of complete hepatitis C virus (HCV) genome sequences that represent its component genotypes and subtypes (**Figure 2**). Using the trajectory of the PoW-transformed evolutionary rate decay for Group IV and the expected (short-term) substitution rate of 1.2 x 10^{-3} substitutions/site/year for HCV (**Figure 2 a, b**), the model demonstrates a clear separation of timescales for the diversification of variants within genotypes ($\sim 50-500$ years), among subtypes ($\sim 1,000-20,000$ years), and among genotypes ($\sim 40,000-200,000$ years) with an estimated root age for HCV of 396,000 (95% HDP: 326,000 – 425,000) years before present (BP). While the predicted divergence times for within-genotype variants using the PoW model is similar to those obtained using a standard substitution model, the latter models estimate the root age of HCV to be only 970 (95% HDP: 850 – 1100) years BP with no clear separation of timescales for among-genotype diversifications (**Figure 2 c**). These results contrast with estimates of 500 – 2,000 years of genotype diversification by simple extrapolation from short term rates[43], while among-subtype divergence times of 1,000 – 20,000 years are up to 50 times higher than the 300 – 500 years estimated in previous molecular epidemiological analysis[44-46]. The revised, very early evolutionary origin of HCV genotypes (326,000years, 425,000 years 95% HPD) predicted by our model is striking. While these early dates still fit currently proposed hypotheses for multiple and potentially relatively recent zoonotic sources of HCV in humans, associated with different genotypes[47,48], the existence of a common ancestor of HCV before human migration of Africa (150,000BP) support alternative scenarios where HCV diversified within anatomically modern humans. HCV genotypes may have arisen from geographical separation in Africa (genotypes 1, 2, 4, 5, 7) and migrational separation of human populations migrating out of Africa into Asia (genotypes 3, 6 and 8).

We also carried out a similar analysis to investigate the origins of the SARS-CoV-2 sarbecovirus lineage (**Figure 3**). While the PoW-transformed phylogeny recovers the previous estimates for SARS-CoV-1 and SARS-CoV-2 diversification from their most closely related bat virus over short timescales (i.e. less than hundreds of years BP), it extends the root age back to 23,000 (21,000 - 25,000) years BP, nearly 20 times older than previous estimates[49]. Our results indicate that humanity may have been exposed to these viruses since the Paleolithic if they had come into contact with their natural hosts. Also, our date estimates of the origin of the sarbecovirus lineage are in remarkable concordance with signatures

of selection on human genomic datasets that indicate an arms race with corona-like viruses dating back 25,000 years[50], providing an external comparator for our methodology.

The PoW model creates an over-arching evolutionary framework that can reconcile, and incorporate timescales derived from co-evolutionary and ancient DNA studies. Further radical re-evaluations of timescales of other RNA and DNA viruses using this approach will provide new insights into their origins and evolutionary dynamics[27,51]. Application of the PoW will place ancestors of more divergent virus sequences far further back into the past than conventional reconstructions. The good fit between modelled and observed substitution rates and the gradient of rate decay over time were based on a minimal number of assumptions about mutational fitness effects, proportion of sites evolving at a particular rate, and robust to substantial differences in substitution rates across different viral groups. By finding the short-term substitution rate (the flat part of the modelled rate decay) and the value of the fastest-evolving rate group (which sets the inflection point of the curve), the PoW model can reconstruct corrected substitution rates for virus genotypes with increasingly divergent nucleotide sequences.

References

1 Duffy, S., Shackelton, L. A. & Holmes, E. C. Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* **9**, 267-276, doi:10.1038/nrg2323 (2008).

2 Pybus, O. G. & Rambaut, A. Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet* **10**, 540-550, doi:10.1038/nrg2583 (2009).

3 Holland, J. J. Transitions in understanding of RNA viruses: a historical perspective. *Current topics in microbiology and immunology* **299**, 371-401 (2006).

4 Sharp, P. M. *et al.* Origins and evolution of AIDS viruses: estimating the time-scale. *Biochem.Soc.Trans.* **28**, 275-282 (2000).

5 Nakano, T., Lu, L., Liu, P. & Pybus, O. G. Viral gene sequences reveal the variable history of hepatitis C virus infection among countries. *J Infect Dis* **190**, 1098-1108 (2004).

6 Nelson, M. I. & Holmes, E. C. The evolution of epidemic influenza. *Nat Rev Genet* **8**, 196-205, doi:10.1038/nrg2053 (2007).

7 Olsen, B. *et al.* Global patterns of influenza a virus in wild birds. *Science* **312**, 384-388, doi:10.1126/science.1122438 (2006).

8 Sawyer, S. L., Emerman, M. & Malik, H. S. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS biology* **2**, E275, doi:10.1371/journal.pbio.0020275 (2004).

9 Taubenberger, J. K. & Kash, J. C. Influenza virus evolution, host adaptation, and pandemic formation. *Cell host & microbe* **7**, 440-451, doi:10.1016/j.chom.2010.05.009 (2010).

10 Li, W. *et al.* Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. *Embo J* **24**, 1634-1643, doi:10.1038/sj.emboj.7600640 (2005).

11 Urbanowicz, R. A. *et al.* Human Adaptation of Ebola Virus during the West African Outbreak. *Cell* **167**, 1079-1087.e1075, doi:10.1016/j.cell.2016.10.013 (2016).

12 Allison, A. B. *et al.* Host-specific parvovirus evolution in nature is recapitulated by in vitro adaptation to different carnivore species. *PLoS pathogens* **10**, e1004475, doi:10.1371/journal.ppat.1004475 (2014).

13 Bhatt, S. *et al.* The evolutionary dynamics of influenza A virus adaptation to mammalian hosts. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **368**, 20120382, doi:10.1098/rstb.2012.0382 (2013).

14 Sauter, D. *et al.* Tetherin-driven adaptation of Vpu and Nef function and the evolution of pandemic and nonpandemic HIV-1 strains. *Cell host & microbe* **6**, 409-421, doi:10.1016/j.chom.2009.10.004 (2009).

15 Wain, L. V. *et al.* Adaptation of HIV-1 to its human host. *Molecular biology and evolution* **24**, 1853-1860, doi:10.1093/molbev/msm110 (2007).

16 Alves, J. M. *et al.* Parallel adaptation of rabbit populations to myxoma virus. *Science* **363**, 1319-1326, doi:10.1126/science.aau7285 (2019).

17 Duchene, S., Holmes, E. C. & Ho, S. Y. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proceedings. Biological sciences* **281**, doi:10.1098/rspb.2014.0732 (2014).

18 Ho, S. Y. *et al.* Time-dependent rates of molecular evolution. *Mol Ecol* **20**, 3087-3101, doi:10.1111/j.1365-294X.2011.05178.x (2011).

19 Katzourakis, A. & Gifford, R. J. Endogenous viral elements in animal genomes. *PLoS genetics* **6**, e1001191, doi:10.1371/journal.pgen.1001191 (2010).

20 Taylor, D. J., Leach, R. W. & Bruenn, J. Filoviruses are ancient and integrated into mammalian genomes. *BMC evolutionary biology* **10**, 193, doi:10.1186/1471-2148-10-193 (2010).

21 Katzourakis, A., Tristem, M., Pybus, O. G. & Gifford, R. J. Discovery and analysis of the first endogenous lentivirus. *Proc Natl Acad Sci U S A* **104**, 6261-6265, doi:10.1073/pnas.0700471104 (2007).

22 Han, G. Z. & Worobey, M. Endogenous lentiviral elements in the weasel family (Mustelidae). *Molecular biology and evolution* **29**, 2905-2908, doi:10.1093/molbev/mss126 (2012).

23 Gifford, R. J. *et al.* A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution. *Proc Natl Acad Sci U S A* **105**, 20362-20367, doi:10.1073/pnas.0807873105 (2008).
24 Hron, T., Farkasova, H., Padhi, A., Paces, J. & Elleder, D. Life History of the Oldest Lentivirus: Characterization of ELVgv Integrations in the Dermopteran Genome. *Molecular biology and evolution* **33**, 2659-2669, doi:10.1093/molbev/msw149 (2016).
25 Sharp, P. M. & Simmonds, P. Evaluating the evidence for virus/host co-evolution. *Curr Opin Virol* **1**, 436-441, doi:10.1016/j.coviro.2011.10.018 (2011).
26 Katzourakis, A., Gifford, R. J., Tristem, M., Gilbert, M. T. & Pybus, O. G. Macroevolution of complex retroviruses. *Science* **325**, 1512, doi:10.1126/science.1174149 (2009).
27 Aiewsakun, P. & Katzourakis, A. Marine origin of retroviruses in the early Palaeozoic Era. *Nature communications* **8**, 13954, doi:10.1038/ncomms13954 (2017).
28 Muhlemann, B. *et al.* Ancient human parvovirus B19 in Eurasia reveals its long-term association with humans. *Proc Natl Acad Sci U S A* **115**, 7557-7562, doi:10.1073/pnas.1804921115 (2018).
29 Muhlemann, B. *et al.* Ancient hepatitis B viruses from the Bronze Age to the Medieval period. *Nature*, doi:10.1038/s41586-018-0097-z (2018).
30 Krause-Kyora, B. *et al.* Neolithic and Medieval virus genomes reveal complex evolution of Hepatitis B. *Elife* **7**, doi:10.7554/eLife.36666 (2018).
31 Duggan, A. T. *et al.* 17(th) Century Variola Virus Reveals the Recent History of Smallpox. *Curr Biol* **26**, 3407-3412, doi:10.1016/j.cub.2016.10.061 (2016).
32 Patterson Ross, Z. *et al.* The paradox of HBV evolution as revealed from a 16th century mummy. *PLoS pathogens* **14**, e1006750, doi:10.1371/journal.ppat.1006750 (2018).
33 Aiewsakun, P. & Katzourakis, A. Time-Dependent Rate Phenomenon in Viruses. *J Virol* **90**, 7184-7195, doi:10.1128/Jvi.00593-16 (2016).
34 Lythgoe, K. A., Gardner, A., Pybus, O. G. & Grove, J. Short-Sighted Virus Evolution and a Germline Hypothesis for Chronic Viral Infections. *Trends in microbiology* **25**, 336-348, doi:10.1016/j.tim.2017.03.003 (2017).
35 Lythgoe, K. A., Pellis, L. & Fraser, C. Is Hiv Short-Sighted? Insights from a Multistrain Nested Model. *Evolution* **67**, 2769-2782, doi:10.1111/evo.12166 (2013).
36 Simmonds, P., Aiewsakun, P. & Katzourakis, A. Prisoners of war - host adaptation and its constraints on virus evolution. *Nat Rev Microbiol* **17**, 321-328, doi:10.1038/s41579-018-0120-2 (2019).
37 Crawford, G. & Williams, C. A Note on the Analysis of Subjective Judgment Matrices. *J Math Psychol* **29**, 387-405, doi:Doi 10.1016/0022-2496(85)90002-1 (1985).
38 Holmes, E. C., Dudas, G., Rambaut, A. & Andersen, K. G. The evolution of Ebola virus: Insights from the 2013-2016 epidemic. *Nature* **538**, 193-200, doi:10.1038/nature19790 (2016).
39 Meyer, A. G., Spielman, S. J., Bedford, T. & Wilke, C. O. Time dependence of evolutionary metrics during the 2009 pandemic influenza virus outbreak. *Virus Evol* **1**, doi:ARTN vev006
10.1093/ve/vev006 (2015).
40 Soubrier, J. *et al.* The Influence of Rate Heterogeneity among Sites on the Time Dependence of Molecular Rates. *Molecular biology and evolution* **29**, 3345-3358, doi:10.1093/molbev/mss140 (2012).
41 Sanjuan, R. From Molecular Genetics to Phylodynamics: Evolutionary Relevance of Mutation Rates Across Viruses. *PLoS pathogens* **8**, doi:ARTN e1002685
10.1371/journal.ppat.1002685 (2012).
42 Duchene, S., Ho, S. Y. W. & Holmes, E. C. Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models. *BMC evolutionary biology* **15**, doi:ARTN 36
10.1186/s12862-015-0312-6 (2015).

43 Smith, D. B. & Simmonds, P. Characteristics of nucleotide substitution in the hepatitis C virus genome: Constraints on sequence change in coding regions at both ends of the genome. *J Mol Evol* **45**, 238-246, doi:Doi 10.1007/Pl00006226 (1997).
44 Simmonds, P. & Smith, D. B. Investigation of the pattern of diversity of hepatitis C virus in relation to times of transmission. *J Viral Hepatitis* **4**, 69-74, doi:DOI 10.1111/j.1365-2893.1997.tb00163.x (1997).
45 Markov, P. V. *et al.* Phylogeography and molecular epidemiology of hepatitis C virus genotype 2 in Africa. *J Gen Virol* **90**, 2086-2096, doi:10.1099/vir.0.011569-0 (2009).
46 Iles, J. C. *et al.* Phylogeography and epidemic history of hepatitis C virus genotype 4 in Africa. *Virology* **464**, 233-243, doi:10.1016/j.virol.2014.07.006 (2014).
47 Pybus, O. G. & Gray, R. R. VIROLOGY The virus whose family expanded. *Nature* **498**, 310-311, doi:DOI 10.1038/498310b (2013).
48 Pybus, O. G. & Theze, J. Hepacivirus cross-species transmission and the origins of the hepatitis C virus. *Curr Opin Virol* **16**, 1-7, doi:10.1016/j.coviro.2015.10.002 (2016).
49 Boni, M. F. *et al.* Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat Microbiol* **5**, 1408-+, doi:10.1038/s41564-020-0771-4 (2020).
50 Souilmi, Y. *et al.* An ancient coronavirus-like epidemic drove adaptation in East Asians from 25,000 to 5,000 years ago. *bioRxiv*, doi:10.1101/2020.11.16.385401 (2020).
51 Wertheim, J. O., Chu, D. K. W., Peiris, J. S. M., Pond, S. L. K. & Poon, L. L. M. A Case for the Ancient Origin of Coronaviruses. *J Virol* **87**, 7039-7045, doi:10.1128/Jvi.03273-12 (2013).
52 Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* **4**, doi:10.1093/ve/vey016 (2018).
53 Tajima, F. & Nei, M. Estimation of Evolutionary Distance between Nucleotide-Sequences. *Molecular biology and evolution* **1**, 269-285 (1984).
54 Felsenstein, J. Evolutionary Trees from DNA-Sequences - a Maximum-Likelihood Approach. *J Mol Evol* **17**, 368-376, doi:Doi 10.1007/Bf01734359 (1981).
55 Ho, S. Y. W., Phillips, M. J., Cooper, A. & Drummond, A. J. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Molecular biology and evolution* **22**, 1561-1568, doi:10.1093/molbev/msi145 (2005).

Figures

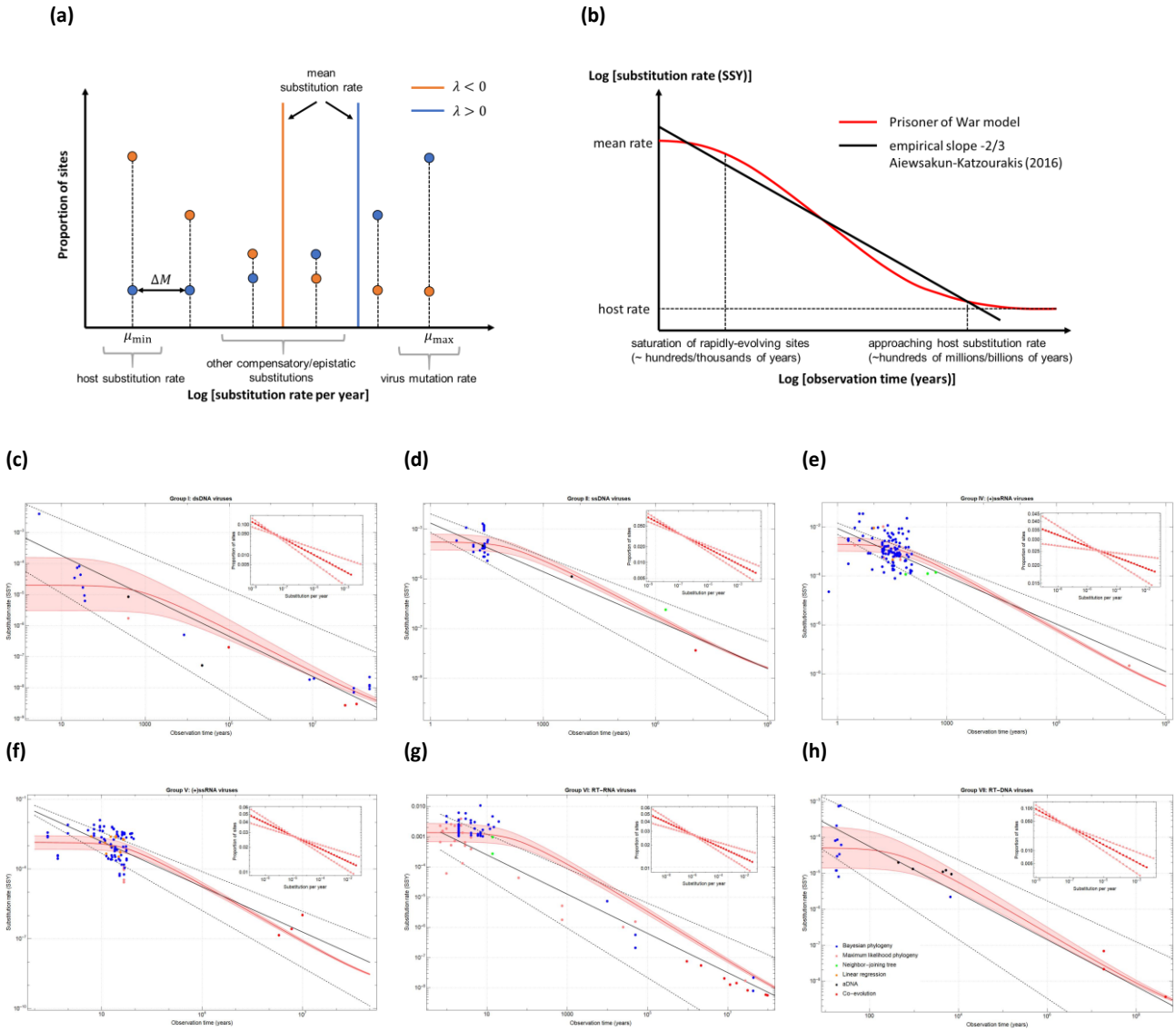


Figure 1: **(a)** Distribution of the fraction of sites per rate group according to the PoW model. A fraction of sites, m_i, belonging to rate group i (evolving at rate μ_i), is an exponentially distributed number with parameter λ. The rate groups are equally spaced on a log-scale from the slowest (with fixed value), $\mu_{\min} = 10^{-9}$, to the fastest (variable), $\mu_{\max}$, with a common ratio, ΔM. The finer the value of ΔM gets, the more accurate the predicted values become. However, typically, less than 50 rate groups (i.e. $M < 50$) is sufficient for all model predictions. For a fixed ΔM, the exponent coefficient, λ, together with the number of rate group, M, are the two free parameters of the PoW model which set the mean, $\langle \mu \rangle$, and maximum rates, $\mu_{\max}$, for any given data set. (**b**) Schematic plot of the time-dependent rate dynamics according to the PoW model (red) and the empirical observation made by Aiewsakun and Katzourakis[33] (black). (**c**)-(**h**) Estimated time-dependent rate curves for each viral group according to the PoW model. A total of 389 viral rate estimates (coloured circles representing various phylogenetic methods used for estimating rates) was collected from more than 130 publications, 23 estimates for group I, 32 for group II, 123 for group IV, 106 for group V, 85 for group VI, and 20 for group VII. The inset shows the distribution of rate groups in each virus group. The red line shows the best fit and shaded area the 95% confidence interval ($\Delta M = 1.58$ and $\alpha_M = \alpha = 3/4$).

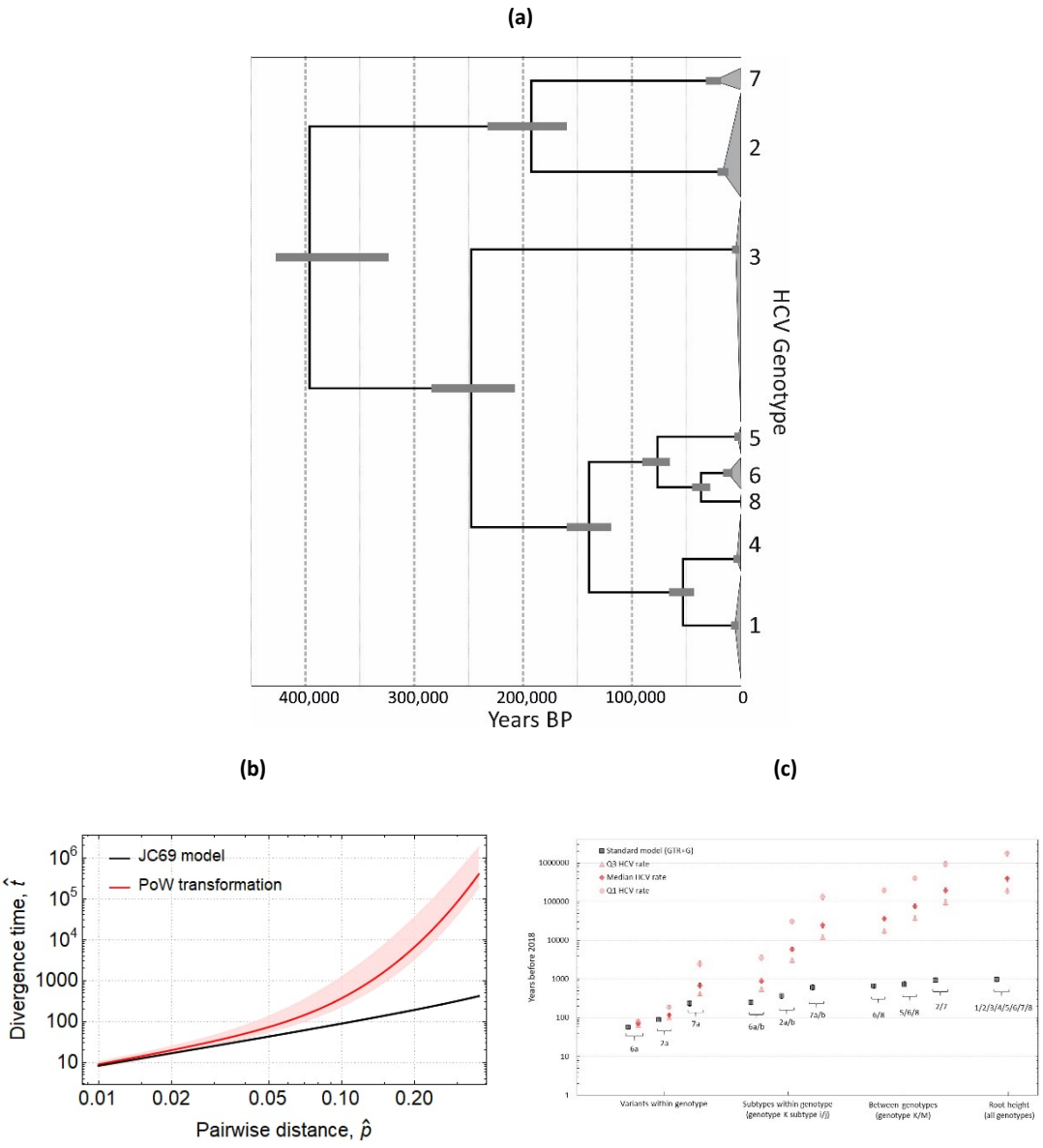
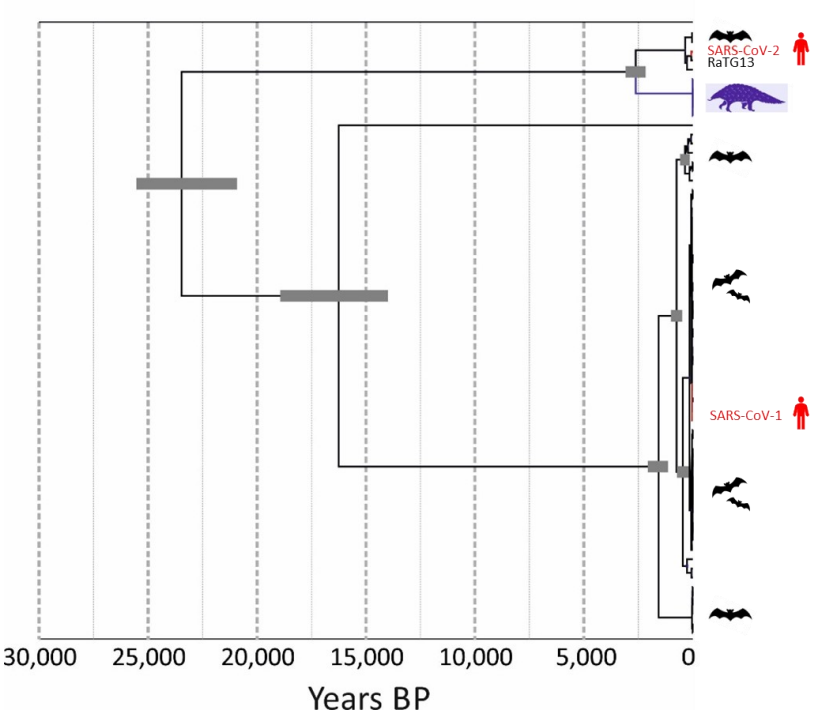


Figure 2: (a) The PoW-transformed time-calibrated phylogeny represents a maximum clade credibility tree inferred for HCV (including all its 8 genotypes). An ultrametric tree was first estimated in the Bayesian phylogenetic framework under a strict clock assumption and a JC69 substitution model assuming a fixed substitution rate equal to 1 (branch lengths are in units of substitutions per site) using BEAST software platform (v.1.10)[52]. Then the branch lengths (along with their corresponding 95% HPD) are rescaled according to the PoW transformation. The two insets show magnified parts of the tree where genotypes 5, 6, 7, and 8 along with their subtypes (i.e. a and b) and nearest within-genotype variants (e.g. MH940291/-/2015 is a variant within genotype 7a) are highlighted. **(b)** Compares the estimated divergence time for a pair of HCV sequences as a function of their inferred genetic distance (ranging from 1% to 37%) using a standard JC69 substitution model with an estimated distance $d_{\mathrm{JC}} = -\alpha_M Ln(1-\hat{p}/\alpha_M)/\langle\mu\rangle$ and a PoW-transformed measure of distance. The expected (short-term) HCV substitution rate (taken out of 9 estimates from the literature[33]) is $\langle\mu\rangle = 1.2(\mathrm{IQR}: 1.1-1.4)\times 10^{-3}$ SSY and $\mu_{\max}$ is assumed to be the same as the one inferred for Group IV (see **Table 1**). (**c**) Shows the estimated divergence times within and between various HCV genotypes and subtypes using a general time-reversible substitution model with a four-bin gamma rate distribution (GTR+G) and PoW-transformed tree. Error bars represent the 95%HDP for the inferred maximum clade credibility tree.

(a)



(b)

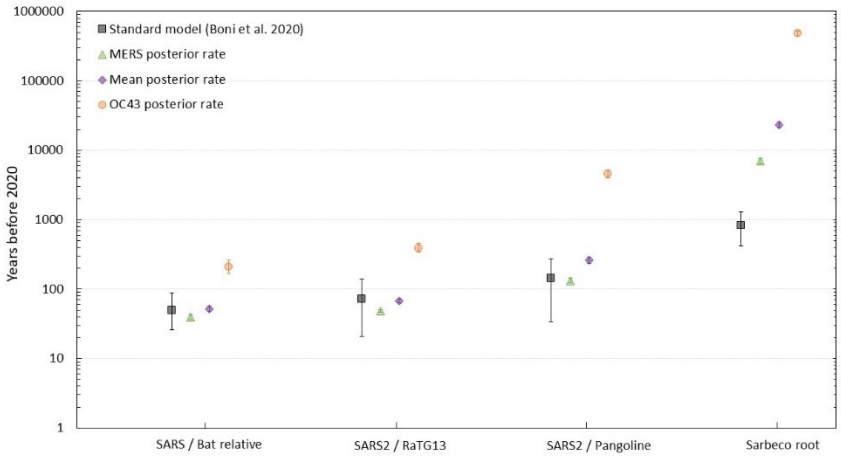


Figure 3: (a) The PoW-transformed time-calibrated phylogeny represents a maximum clade credibility tree inferred for the SARS-CoV-2 sarbecovirus lineage based on the non-recombinant alignment 3 (NRA3)[49] which includes SARS-CoV-1 and SARS-CoV-2 viruses in humans among other closely related bat and pangolin viruses. An ultrametric tree was first estimated in the Bayesian phylogenetic framework under a strict clock assumption and a JC69 substitution model assuming a fixed substitution rate equal to 1 (branch lengths are in units of substitutions per site) using BEAST software platform (v.1.10)[52]. Then the branch lengths (along with their corresponding 95% HPD) are rescaled according to the PoW transformation. The two insets show magnified parts of the tree where SARS-CoV-1 (blue) and SARS-CoV-2 (red) are located. **(b)** Shows the divergence time estimates for SARS-CoV-2 and 2002-2003 SARS-CoV from their most closely related viruses according to a standard substitution model[49] (black) and the PoW-transformed phylogeny using the consensus posterior rate centred around $\langle \mu \rangle = 5.5 \times 10^{-4}$ SSY (purple), HCoV-OC43 rate prior (orange), 2.4×10^{-4} SSY, and MERS-CoV rate prior (green), 7.8×10^{-4} SSY (see Extended data figure 3 in ref[49]).

Table 1: Estimated mean and maximum substitution rate SSY according to the PoW model across 6 viral groups. Parentheses correspond to 95% confidence intervals.

Viral group	Type of virus	Mean substitution rate, $\langle \mu \rangle$	Fastest rate group, μ_{max}
Group I	dsDNA virus	$2(0.3-16)\times10^{-5}$	$3(0.6-10)\times10^{-3}$
Group II	ssDNA virus	$3(1-6)\times10^{-4}$	$2(1-3)\times10^{-2}$
Group IV	(+)ssRNA virus	$2(1-4)\times10^{-3}$	$4(3-6)\times10^{-2}$
Group V	(-)ssRNA virus	$1(0.7-3)\times10^{-3}$	$4(3-6)\times10^{-2}$
Group VI	RT-RNA virus	$1(0.7-3)\times10^{-3}$	$4(3-6)\times10^{-2}$
Group VII	RT-DNA virus	$5(1-20)\times10^{-5}$	$4(2-10)\times10^{-3}$

Methods

Time-dependent rate decay under a uniform substitution rate

Suppose that a sequence has diverged from its ancestor *t* generations ago under a substitution rate μ that is constant over time and equal across all sites. In this case, the proportion of pairwise differences, $p(t)$, is given by[53]

$$p(t) = \alpha(1 - e^{-\mu t/\alpha}) \quad \text{(S1)}$$

such that α is the maximum proportion of pairwise differences and is determined by $\alpha = 1 - \sum_i \pi_i^2$ where π_i is the base frequency of the *i*th nucleotide or amino acid. Assuming that d is the 'true' genetic distance between a pair of homologous sequences, i.e. $d = \mu t$, we can estimate the observed genetic distance, $\hat{d}$, with an observed proportion of pairwise differences, $\hat{p}$, using the Felseinstein's 1981 substitution model[54]

$$\hat{d} = \mu t = -\alpha_M Ln\{1 - \hat{p}/\alpha_M\} \quad \text{(S2)}$$

where α_M is the expected saturation frequency set by the model. If the model correctly estimates the saturation frequency, i.e. $\alpha_M = \alpha$, **Equation S2** accurately predicts the true genetic distance, i.e. $\hat{d} = d$, as long as the divergence time $t \ll \alpha/\mu$. As $p(t)$ approaches saturation frequency at $t \approx \alpha/\mu$, the observed proportion of pairwise differences will be bound by the number of evolving sites. For instance, if the saturation frequency is $\alpha = 3/4$, i.e. a standard Jukes-Cantor substitution model, to distinguish between an observed pairwise difference of $\hat{p}^* = 0.74$ and $\hat{p}^* = 0.741$ would require a sample of approximately one thousand evolving site at rate μ. Thus, for any small sample size, the estimated distance, $\hat{d}$, will remain effectively unchanged for $t \gtrsim \alpha/\mu$. In other words, if the pair have evolved beyond their saturation point, for a given divergence time, $\hat{t}$, the inferred rate follows a power-law rate drop with a slope -1 on a log-log plot, i.e. $\mu \approx \alpha_M/\hat{t}$ (see grey curve in **Figure S1 a**). The same power-law bias can be observed when estimating the divergence time, $\hat{t}$, assuming there is prior knowledge on the inferred substitution rate, $\hat{\mu}$.

In the presence of purifying selection and/or amino acid and nucleotide biases, the substitution model in **Equation S2** may incorrectly assume a maximum proportion of differences, α_M, that is higher than the true value, i.e. $\alpha_M > \alpha$. For instance, if we apply a Jukes-Cantor measure of nucleotide distance to a pair of sequences with a particular site preference that equally favours only two (out of the four) nucleotides, i.e. $\alpha = 1/2$, the true proportion of differences reaches saturation at earlier divergence times than what the selected substitution model would predict. As a result, similar to the previous scenario, the estimated rate drops as a power-law with slope -1 (see orange curve in **Figure S1 a**) after hitting the saturation point, i.e. $\hat{\mu} \approx \alpha_M Ln\{\alpha_M/(\alpha_M - \alpha)\}/t$.

The reverse side of this is when $\alpha_M < \alpha$ in which case the model underestimates the true saturation level. In this case, the observed proportion of pairwise differences passes the substitution model's expected point, i.e. $\hat{p}/\alpha_M \gtrsim 1$, at which point the estimated substitution rate, $\hat{\mu} \approx \alpha Ln\{\alpha/(\alpha - \alpha_M)\}/t$, goes to infinity (see blue curve in **Figure S1 a**).

Time-dependent rate decay in the presence of rate heterogeneity

In the previous scenario, **Equation S1**, we assumed there is no rate heterogeneity across sites. However, if the actual evolutionary process involves *M* group of sites with each group *i* evolving at rate μ_i and occupying a fraction m_i of sites, the proportion of pairwise differences would be given by

$$p(t) = \sum_{i=1}^{M} m_i \alpha_i (1 - e^{-\mu_i t/\alpha_i}) \qquad \text{(S3)}$$

such that $\sum_i m_i = 1$ and the expected substitution rate $\langle \mu \rangle = \sum_i m_i \mu_i$. If we apply a measure of distance based on **Equation S2** to an evolutionary process with rate heterogeneity, **Equation S3**, the model can reliably infer the expected substitution rate up to when the first group of sites with the fastest substitution rate, $\mu_{\max}$, reach saturation at $t \sim \mu_{\max}^{-1}$ at which point the pattern of time-dependent rate decay emerges until it plateaus at a rate corresponding to the slowest-evolving sites, $\mu_{\min}$. Once all the rate groups reach saturation at time $t \sim \mu_{\min}^{-1}$, the power-law rate decay with slope -1 emerges.

Suppose that a sequence has a fraction of sites m_1 evolving at rate $\mu_1 = \mu$ and the remaining sites $(1 - m_1)$ evolving epistatically such that a pair of sites need to mutate simultaneously for the offspring to recover wild-type fitness, i.e. $\mu_2 = \mu^2$. Assuming the saturation frequency across all sites are equal and that the model correctly identifies their frequency, i.e. $\alpha_i = \alpha = \alpha_M$, we can use **Equation S2** to recover the expected substitution rate $\langle \mu \rangle = m_1 \mu + (1 - m_1)\mu^2$. As the fast-evolving sites approach the saturation point at $t_1 \approx \alpha/\mu$, the inflection point of rate decay emerges and a sharp decline in estimated substitution rate follows while the remaining fraction of sites, $(1 - m_1)$, keep accumulating new substitutions at rate μ^2, slowing down the slope of rate decay until those sites also reach saturation at $t_2 \approx \alpha/\mu^2$ beyond which point the entire genome reaches saturation and the power-law rate decay with slope -1 emerges. We can also see from **Figure S1 b** that as the proportion of slow-evolving sites increases, the mean substitution rate goes down and the slope of the time-dependent rate decay becomes less steep.

Simulating a Wright-Fisher population to infer substitution rates

We simulate a neutral haploid Wright-Fisher population of size N_e with L evolving sites under a constant mutation rate μ per site such that every nucleotide (A, C, G, and T) can mutate to any other nucleotide at the same rate $\mu/3$ – mutation rate is equal to substitution rate under neutrality. We then sample from the entire population at two time points with an increasingly wider time gap, t^*. Initially, we allow the population to evolve for $10N_e$ generations before taking the first sample to ensure that neutral coalescent events reach their steady state distribution and that the population, on average, coalesce every $2N_e$ generations. We then take the second sample t^* generations later and repeat this process 100 times to generate replicate sequences at both time points and run each set of simulations in BEAST 1.10 to estimate the substitution rate (**Figure S2**). We load the simulated sequences (along with their sampling times) on BEAST and use a strict molecular clock with a continuous-time Markov chain reference prior on substitution rates, a constant population coalescent prior, and a Jukes-Cantor substitution model. For every simulated set, the Markov chain Monte Carlo was run for 10,000,000 steps and parameter convergence was inspected visually.

In **Figure S2**, we recreate the time-dependent pattern of rate decay both in the absence and presence of rate heterogeneity across sites, using a standard substitution model on simulated data. We find that while the inferred substitution rates exhibit a power-law rate decay with slope -1 over longer time intervals (see **Figure S2 (a)** and **(b)**), the inferred TMRCAs tend to be overestimated with a similar (inverse) power-law trend, i.e. $\hat{t} \sim 1/\hat{\mu}$ (see **Figure S2 (c)** and **(d)**). We also find an unexpected time-dependent rate effect over short timescales. This occurs when the observation gap, t^*, is much shorter than the expected coalescent time of the population, i.e. $t^* \ll 2N_e$. This also results in the underestimation of true TMRCAs which systematically makes worse predictions for higher substitution rates. The expected rate curves (dashed lines shown in **Figure S2 (a)** and **(b)**) can be approximated by replacing $p(t)$ from **Equation S1** into $\hat{p}$ from **Equation S2** which is given by

$$\hat{\mu}(t^*) \approx -\alpha_M Ln\left\{1-\frac{1}{L\alpha_M}\left\lfloor L\sum_{i=1}^{M} m_i\alpha(1-e^{-\mu_i(t^*+2N_e)/\alpha})\right\rfloor\right\}/(t^*+T) \qquad \text{(S4)}$$

such $\lfloor . \rfloor$ is the floor function which represents the finite size effect of having L evolving sites on saturation frequency. The mean divergence time between the two populations is approximately $t \approx t^* + 2N_e$ and the inferred divergence time is $\hat{t} \approx t^* + T$ – this resembles the mis-calibration effects reported elsewhere (see Equation 2 in ref[55]). The reason why **Equation S4** only works as an approximate is that the median inferred TMRCA from simulation results, T, also varies with respect to observation gap t^* (see **Figure S2 (c)** and **(d)**). However, for $t^* \gg 2N_e$, the variation in T becomes negligible compared to $t^*$and only has second-order effects on inferred substitution rates. **Figure S2 (e)** and **(f)** show the tree topology under the two extremes, $t^* \ll 2N_e$ and $t^* \gg 2N_e$, respectively. It indicates that, over long timescales, the time-dependent rate effects are dominated by the very long (and saturated) branch connecting the two populations that are t^* generations apart. As a result, the decay dynamics looks very similar to the analytical results in **Figure S1** where we estimate the substitution rate between a pair of sequences separated by a very long branch.

Supplementary Figures

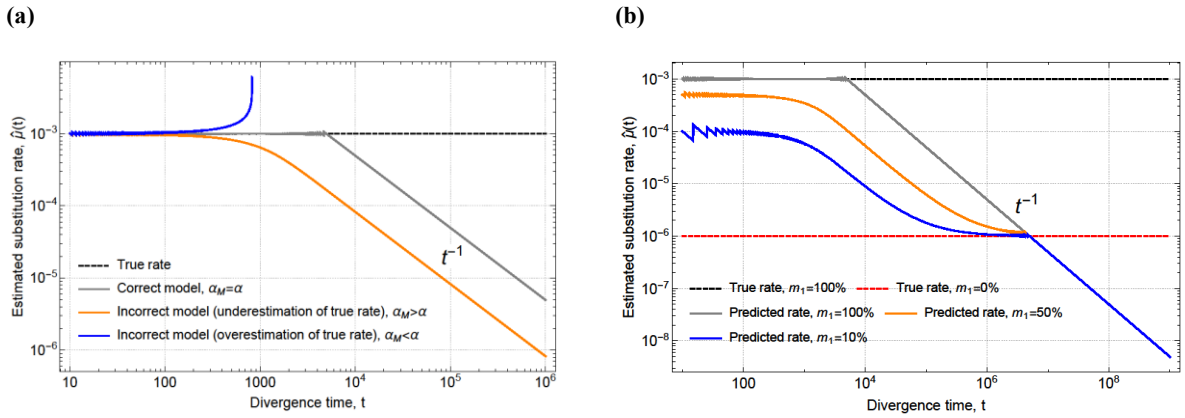


Figure S1: (a) Estimating the substitution rate $\mu = 10^{-3}$ (black dashed line) using **Equation S2** for a pair of sequences that have diverged from each other t generations ago: using the correct substitution model (gray), a model that over-estimates the true saturation frequency, α, (orange), and a model that under-estimates the true saturation frequency (blue). **(b)** Estimating the expected (or mean) substitution rate, $< \mu >$, when a fraction m_1 of sites evolve at rate $\mu_1 = 10^{-3}$ (black dashed line) and the remaining fraction, $1 - m_1$, at rate $\mu_2 = 10^{-6}$ (red dashed line). The expression t^{-1} in both plots shows the dominating term in rate decay with respect to divergence time, t, corresponding to slope -1 on the graphs.

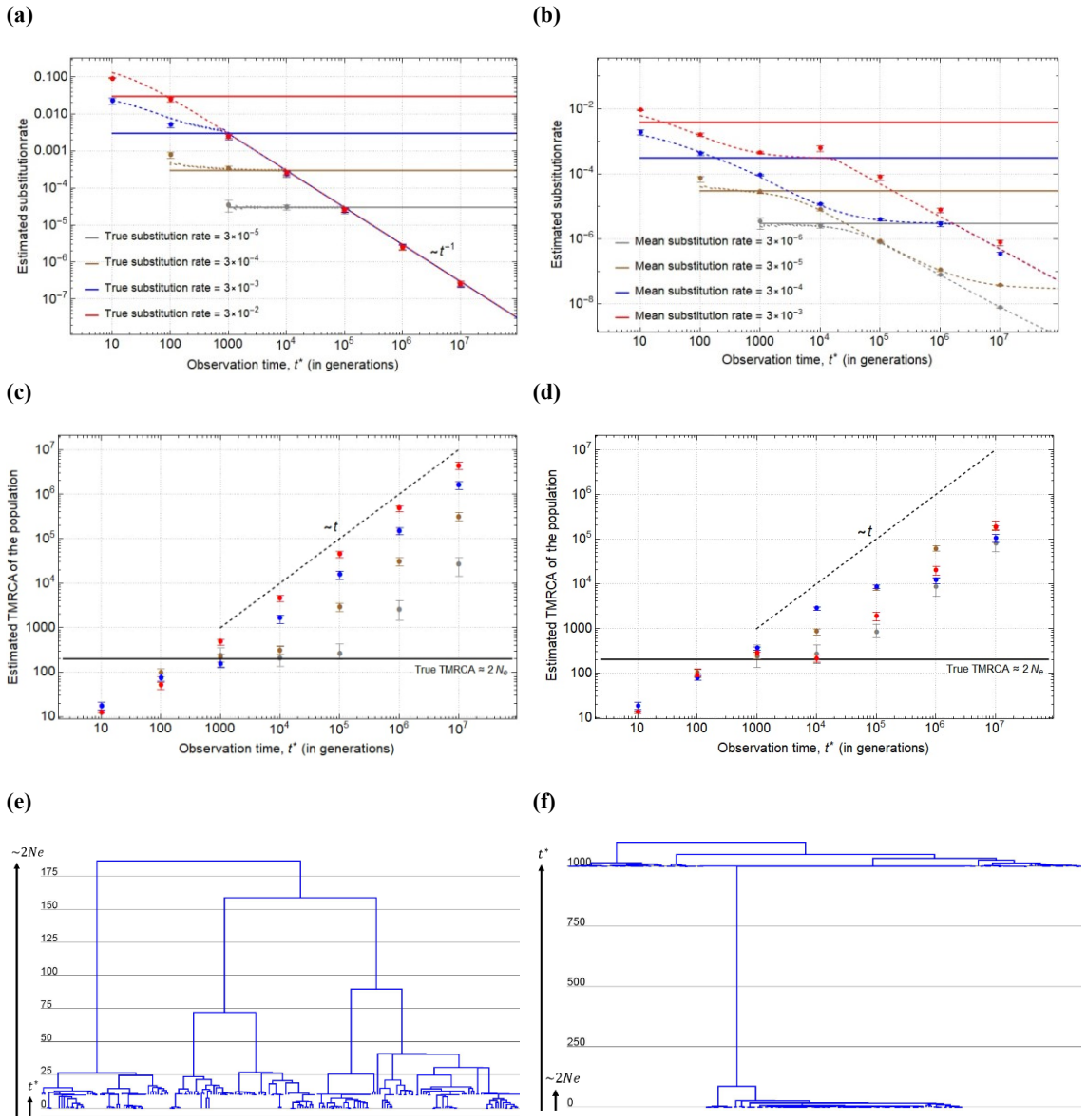


Figure S2: Estimated substitution rate for (**a**) a population of size N_e evolving according to a neutral model of substitution where $L_n = 100$ sites evolve at the same rate, μ, and (**b**) a model with rate heterogeneity across sites such that $L_1 = 100$ sites evolve at rate μ and the remaining $L_2 = 900$ sites at rate μ^2 measured as a function of observation gap, t^*, between when the first and second sample is taken from the population. Dashed lines show the theoretical prediction according to **Equation S4** and solid lines show the mean (expected) rates used for the simulations. (**c**) and (**d**) show the estimated TMRCA for the first group of sampled sequences, i.e. at $t^* = 0$, and solid lines show the mean TMRCA according to neutral theory. The rates and TMRCAs are estimated using BEAST 1.10 under a strict clock assumption (see Methods section). Dots represent the median value taken from 100 independent runs and error bars show the interquartile region. The maximum clade credibility trees for one simulation run corresponding to $\mu = 3 \times 10^{-5}$ is shown when the observation gap is (**e**) $t^* = 10$, $t^* \ll 2N_e$ and (**f**) $t^* = 1000$, $t^* \gg 2N_e$.

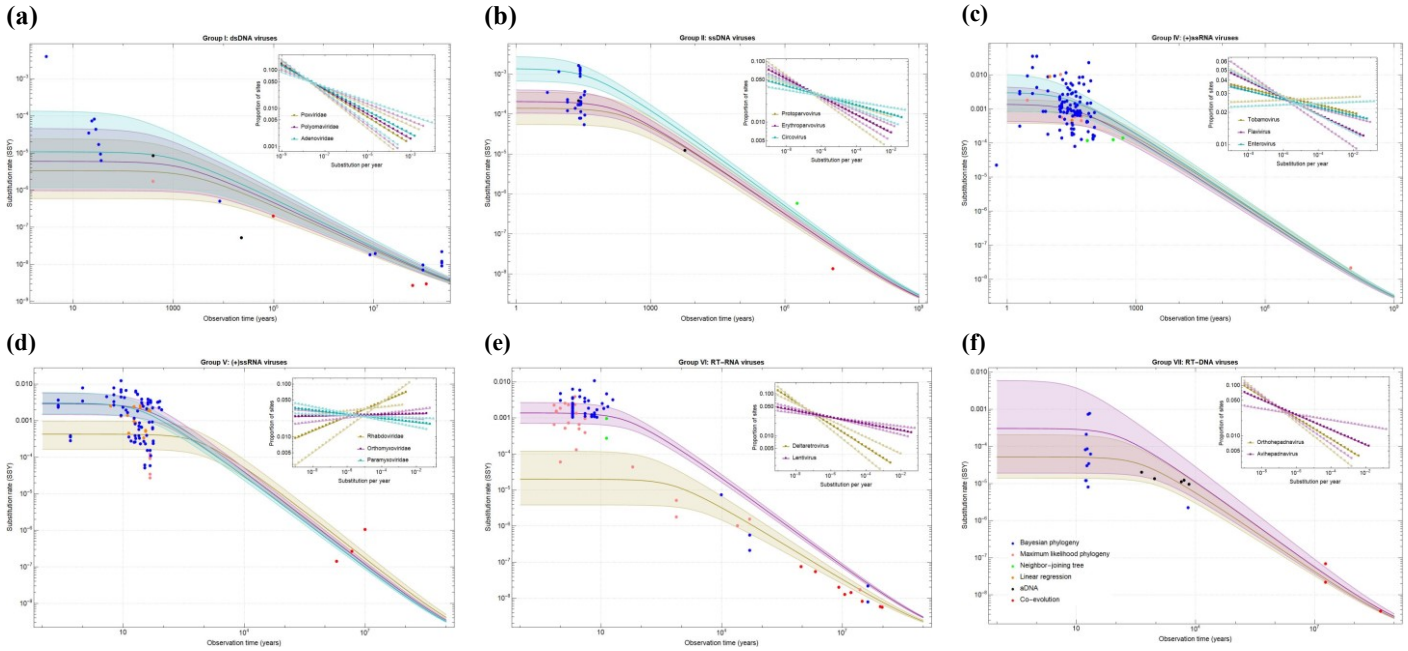


Figure S3: (**a**)-(**f**) Estimated time-dependent rate curves for each viral group according to the PoW model and their corresponding distribution of rate groups (inset). Two or three distinct mean substitution rates (coloured in gold, purple, and blue) are selected for each virus family to estimate rate curves according to the PoW model. The solid lines show the best fit and shaded areas the 95% confidence interval ($\Delta M = 1.58$ and $\alpha_M = \alpha = 3/4$).

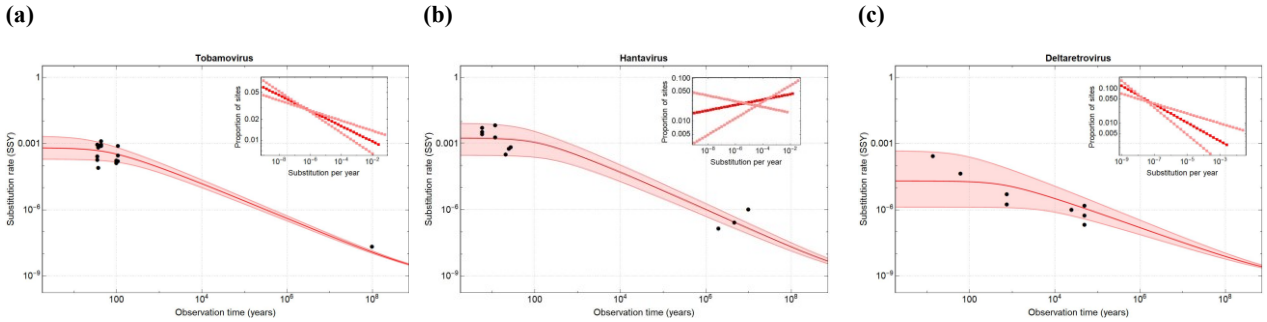


Figure S4: Estimated time-dependent rate curves for three selected genera. (**a**) Tobamoviruses with $\langle \mu \rangle = 0.6(0.2-2)\times 10^{-3}$ SSY and $\mu_{\max} = 3(1-6)\times 10^{-2}$ SSY, (**b**) Hantaviruses with $\langle \mu \rangle = 2(0.3-8)\times 10^{-3}$ SSY and $\mu_{\max} = 2(0.6-4)\times 10^{-2}$ SSY, and (**c**) Deltaretroviruses with $\langle \mu \rangle = 2(0.1-50)\times 10^{-5}$ SSY and $\mu_{\max} = 3(2-30)\times 10^{-3}$ SSY. These genera are selected as they have the largest timespan of rate measurements.

Table S1: Estimated mean and maximum substitution rate according to the PoW model using different virus families or genera to calibrate the mean rate across 6 viral groups. Parentheses correspond to 95% confidence intervals.

Viral group	Virus family/genus used for calibration*	Mean substitution rate, $\langle \mu \rangle$	Fastest rate group, μ_{max}
Group I	Poxviridae	$0.3(0.06-2)\times10^{-5}$	$0.6(0.2-3)\times10^{-3}$
	Polyomaviridae	$0.6(0.1-5)\times10^{-5}$	$1(0.3-4)\times10^{-3}$
	Adenoviridae	$1(0.1-10)\times10^{-5}$	$2(0.3-10)\times10^{-3}$
Group II	Protoparvovirus	$1(0.6-4)\times10^{-4}$	$0.6(0.4-1)\times10^{-2}$
	Erythroparvovirus	$2(1-4)\times10^{-4}$	$1(0.6-2)\times10^{-2}$
	Circovirus	$10(7-300)\times10^{-4}$	$4(3-6)\times10^{-2}$
Group IV	Tobamovirus	$0.4(0.9-2)\times10^{-3}$	$2(1-3)\times10^{-2}$
	Flavivirus	$1(0.4-5)\times10^{-3}$	$4(2-10)\times10^{-2}$
	Enterovirus	$3(0.9-10)\times10^{-3}$	$6(3-10)\times10^{-2}$
Group V	Rhabdoviridae	$0.4(0.2-1)\times10^{-3}$	$0.3(0.2-0.4)\times10^{-2}$
	Orthomyxoviridae	$3(1-6)\times10^{-3}$	$4(3-6)\times10^{-2}$
	Paramyxoviridae	$3(2-6)\times10^{-3}$	$6(4-10)\times10^{-2}$
Group VI	Deltaretrovirus	$0.02(0.004-0.1)\times10^{-3}$	$0.3(0.06-1)\times10^{-2}$
	Lentivirus	$1(0.7-3)\times10^{-3}$	$4(3-6)\times10^{-2}$
Group VII	Orthohepadnavirus	$5(1-20)\times10^{-5}$	$4(2-10)\times10^{-3}$
	Avihepadnavirus	$30(2-600)\times10^{-5}$	$20(2-100)\times10^{-3}$

*Only short-term rate estimates (measured over time scales of <100 years) – along with the long-term rate estimates (>100 years) from the entire data set – from this particular virus family or genus (rather than the entire viral group) is used for the calibration.