J Mol Evol (2002) 54:156–165
DOI: 10.1007/s00239-001-0064-3

JOURNAL OF MOLECULAR EVOLUTION

© Springer-Verlag New York Inc. 2002

Rates of Molecular Evolution in RNA Viruses: A Quantitative Phylogenetic Analysis

Gareth M. Jenkins, Andrew Rambaut, Oliver G. Pybus, Edward C. Holmes

Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

Received: 23 February 2001 / Accepted: 3 July 2001

Abstract. The study of rates of nucleotide substitution in RNA viruses is central to our understanding of their evolution. Herein we report a comprehensive analysis of substitution rates in 50 RNA viruses using a recently developed maximum likelihood phylogenetic method. This analysis revealed a significant relationship between genetic divergence and isolation time for an extensive array of RNA viruses, although more rate variation was usually present among lineages than would be expected under the constraints of a molecular clock. Despite the lack of a molecular clock, the range of statistically significant variation in overall substitution rates was surprisingly narrow for those viruses where a significant relationship between genetic divergence and time was found, as was the case when synonymous sites were considered alone, where the molecular clock was rejected less frequently. An analysis of the ecological and genetic factors that might explain this rate variation revealed some evidence of significantly lower substitution rates in vector-borne viruses, as well as a weak correlation between rate and genome length. Finally, a simulation study revealed that our maximum likelihood estimates of substitution rates are valid, even if the molecular clock is rejected, provided that sufficiently large data sets are analyzed.

Key words: RNA viruses — Substitution rate — Molecular clock — Maximum likelihood — Codon bias

Correspondence to: E.C. Holmes; *email:* Edward.Holmes@zoo.ox.ac.uk

Introduction

Despite the rapid accumulation of gene sequence data from RNA viruses, relatively little is known about the determinants of rates of nucleotide substitution in these infectious agents. Yet such information is fundamental if we are to understand the processes governing viral evolution and to predict their response to treatment with vaccines and drugs.

When studying RNA viruses, it is often assumed that because of their error prone replication they will mutate quickly and hence evolve quickly. However, this interpretation may be overly simplistic because the rate of nucleotide substitution is a function of both the rate of mutation and the rate of replication. As a case in point, although the mutation rate per genome replication of human immunodeficiency virus (HIV-1) is less than that of lytic RNA viruses (Mansky and Temin 1995; Drake et al. 1998), this retrovirus still has one of the highest measured overall rates of nucleotide substitution as a result of undergoing many more genome replications per unit time (Perelson 1996; Albert and Leitner 1999). Furthermore, if viral RNA is not evolving neutrally—which may be true of both synonymous and nonsynonymous substitutions given the largely undetermined effects of codon usage and RNA secondary structure—then fluctuations in adaptive environments and population size will also affect substitution rates and their constancy over time. Conversely, if mutation and replication rates are constant and most substitutions neutral, then molecular evolution would be expected to follow a molecular clock. Analyzing rates of nucleotide substitution can therefore provide valuable information regarding the relative importance

of genetic drift versus natural selection in driving viral evolution, as well as the accuracy with which divergence dates may be estimated directly from gene sequence data.

Thus far, substitution rates have been estimated for only a limited number of RNA viruses. Values typically lie close to 1×10^{-3} substitution/site/year, although considerable variation exists. For example, HIV-1, influenza virus A and foot-and-mouth disease virus have been reported to have substitution rates in excess of 1×10^{-3} (Gojobori et al. 1990; Gorman et al. 1990; Martinez et al. 1992; Albert and Leitner 1999), while estimates for measles virus, influenza virus C, GBV-C virus, and many vector-transmitted viruses are lower, ranging from 1×10^{-6} to 1×10^{-3} (Weaver et al. 1992; Muraki et al. 1996; Rima et al. 1997; Holland and Domingo 1998; McGuire et al. 1998; Suzuki et al. 1999). Furthermore, the evolution of some viruses appears to fit a molecular clock, notable cases being influenza A virus and HIV-1 (Gojobori et al. 1990; Albert and Leitner 1999), although for other viruses such as vesicular stomatitis virus, no relationship between time and extent of sequence divergence has been observed (Nichol et al. 1993; Rodriguez et al. 1996).

Unfortunately, the true extent of variation in rates of nucleotide substitution is difficult to determine since these rates are frequently estimated using methods that are not statistically rigorous. For example, one commonly used method involves fitting a linear regression to a scatterplot of genetic distances from a reference sequence and times of isolation for a sample of temporally distinct virus isolates (Gojobori et al. 1990; Gorman et al. 1990; Martinez et al. 1992; Albert and Leitner 1999). However, the distance data points cannot be considered independent because isolates would be expected to share some phylogenetic history. An alternative approach involves comparing phylogenetically independent pairs of sequences sampled at different times with an outgroup sequence (Li et al. 1988, Krushkal and Li 1995; McGuire et al. 1998; Suzuki et al. 1999). A rate can be determined for each pair of sequences by dividing the divergence that has occurred between the sampling of the earlier and that of the later sequence by the difference in isolation times. However, the number of pairs of sequences that are suitable for estimating a substitution rate in this manner is usually very limited. Furthermore, the standard error of the mean rate for an entire data set cannot easily be used to calculate confidence intervals because the distribution of rates would be expected to have a nonnormal distribution determined by the difference of two Poisson processes, since substitution rates are calculated from the difference between two branch lengths. Finally, if dates of divergence events are known, substitution rates can also be calculated from the genetic divergences of contemporary sequences (Smith et al. 1997; Salemi et al. 1999), although such studies are rare.

The aim of our study was to use the wealth of data available in gene sequence databases (such as GenBank) together with a recently developed maximum likelihood method to estimate substitution rates with correct confidence intervals for a wide range of RNA viruses. We focused mainly on those viruses that replicate via an RNA-dependent RNA polymerase, although HIV, which utilizes reverse transcriptase, was also included for comparison. Our analytical method requires as input data viruses from which samples of sequences have been isolated at different times (Rambaut 2000). A similar approach has also been used recently to date the ancestor of HIV-1 (Korber et al. 2000). The maximum likelihood approach is a powerful framework for analyzing substitution rates because it allows assumptions to be tested directly, in the case that substitution rates are constant among lineages. Furthermore, the method uses all the sequences within the input data set, takes into account phylogenetic dependencies among samples to estimate the substitution rate, and calculates confidence intervals to indicate the level of error due to the inherently stochastic nature of nucleotide substitution and any irregularity of the molecular clock. Finally, we sought to identify genetic or ecological factors that might be responsible for any observed variation in substitution rates among RNA viruses.

Materials and Methods

Sequence Data

Data sets of 50 RNA viruses were created using sequences obtained from GenBank. The contents of each data set are summarized in Table 1, and all alignments and accession numbers are available at http://evolve.zoo.ox.ac.uk/data. The region of the genome used in each case was chosen to maximize a combination of the following criteria—number of samples, sequence length, sampling time period, evenness of sampling, and relatedness among samples—so that the sampling time period would be as great a component of the total divergence time period as possible. On average, each data set contained 24 sequences of length 1.2 kb isolated over a time period of 38 years. Noncoding regions or regions containing overlapping reading frames were excluded. Also, sequences of viruses known to have been extensively passaged between isolation and sequencing were excluded, as were sequences of viruses that have been artificially manipulated (e.g., vaccine strains and laboratory-generated mutants).

Data Analysis

Sequences were aligned manually and the maximum likelihood tree topology was constructed under an HKY85 codon position substitution model (Hasegawa et al. 1985) using PAUP* version 4 (Swofford 1998). The parameters of this model are the base frequencies, ratio of transitions to transversions, and relative rates of substitution at each codon position. This model was also employed in the subsequent maximum likelihood analysis of substitution rates. It was chosen since it contains an intermediate number of parameters, ensuring that our analysis would not be seriously under- or overparameterized for any given data set. We also verified that using more complex models of nucleotide substitution did not significantly alter the substitution rate estimate or outcome of the molecular clock test for a sample of viruses (results available upon request).

Maximum likelihood rates of substitution were calculated using the program TipDate (Rambaut 2000), available at http://evolve.zoo.ox.

Table 1. Overall and synonymous substitution rates for a range of 50 RNA viruses

Virus	Abbreviation	Gene[a]	Substitutions/site/year[b] $\times 10^{-3}$	
			All sites	Synonymous sites
Picornaviridae				
Foot and mouth disease A	FMDV-A	VP1	1.4 (0.88, 1.9)	2.5 (0.36, 4.6)
Foot and mouth disease C	FMDV-C	P1	0.73 (0.40, 1.2)	3.2 (1.8, 4.7)
Foot and mouth disease O	FMDV-O	VP1	1.3 (0.66, 1.9)	0 (0, 2.4)
Human enterovirus 71	EV71	VP1	3.4 (2.7, 4.0)	12 (7.9, 18)
Swine vesicular disease	SVDV	3BC	3.4 (2.8, 4.0)	7.1 (3.6, 15)
Calciviridae				
E. brown hare syndrome	EBHSV	VP60 (p)	0 {0, 0.73)[c]	0 (0, 3.3)[c]
Rabbit hemorrhagic disease	RHDV	VP60 (p)	1.3 (0.59, 2.1)	0.19 (0, 3.6)[c]
Flaviviridae				
Dengue 1	DENV-1	E	0.49 (0.19, 0.79)	1.7 (0.16, 3.0)
Dengue 2	DENV-2	E	0.62 (0.49, 0.74)	1.5 (0.94, 2.1)
Dengue 3	DENV-3	E	0.72 (0.53, 0.91)	2.1 (1.3, 3.0)
Dengue 4	DENV-4	E	0.77 (0.59, 1.0)	2.2 (1.5, 3.3)
Japanese encephalitis	JPEV	E	0.35 (0.089, 0.53)	0.80 (0.24, 1.3)
Louping ill	LIV	E	0.098 (0, 0.26)[c]	0 (0, 0.50)[c]
St. Louis encephalitis	SLEV	E (p)	0.16 (0, 0.38)[c]	0.37 (0, 1.5)[c]
Tick-borne encephalitis	TBEV	E	0 (0, 0.12)[c]	0.29 (0, 0.41)[c]
Yellow fever	YFV	E	0.28 (0.10, 0.44)	1.26 (0, 1.31)[c]
Classical swine fever	CSFV	NS5B (p)	2.0 (1.5, 2.7)	4.9 (2.5, 7.7)
Hepatitis C	HCV	E1 (p)	0.79 (0.61, 1.0)	0.49 (0.082, 1.5)
Togaviridae				
Barmah Forest	BFV	E2 (p)	0.010 (0, 0.16)[c]	0.096 (0, 1.0)[c]
Eastern equine encephalitis	EEEV	26S	0.20 (0.16, 0.26)	0.27 (0.21, 0.68)
Highlands J	HJV	E1	0.14 (0.086, 0.26)	0.36 (0.091, 0.90)
Ross River	RRV	E2	0.24 (0, 0.51)[c]	0.78 (0, 1.6)[c]
Sinbis	SINV	E2 (p)	0.19 (0, 0.50)[c]	0 (0, 0.83)[c]
Venezuelan equine encephalitis	VEEV	E3-E2 (p)	0.13 (0, 0.32)[c]	0 (0, 1.0)[c]
Western equine encephalitis	WEEV	E1 (p)	0.055 (0.015, 0.16)	0.010 (0, 0.53)[c]
Rubella	RUBV	E1	0.61 (0.45, 0.76)	0.51 (0, 1.2)[c]
Coronaviridae				
Equine arteritis	EAV	GL (p)	0.67 (0.39, 0.96)	0 (0, 1.1)[c]
Rhabdoviridae				
Rabies	RABV	G (p)	0 (0, 1.8)[c]	0 (0, 0.76)[c]
Vesicular stomatitis	VSV	P (p)	0.015 (0, 0.29)[c]	0 (0, 0.75)[c]
Paramyxoviridae				
Bovine respiratory syncytial	BRSV	G	0.79 (0.43, 1.6)	2.8 (0.56, 5.4)
Human respiratory syncytial A	HRSV-A	G	1.6 (1.3, 2.2)	3.3 (1.8, 5.2)
Human respiratory syncytial B	HRSV-B	G	0.27 (0, 0.88)[c]	4.8 (3.0, 7.2)
Measles	MV	HE	0.40 (0.31, 0.49)	1.0 (0.71, 1.4)
Mumps	MUV	F-SH-HN	0.25 (0, 0.43)[c]	0.65 (0, 1.4)[c]
Newcastle disease	NDV	M-F (p)	0 (0, 0.075)[c]	0 (0, 0.32)[c]
Human parainfluenza 1	HPIV-1	HN	0.22 (0.037, 0.41)	0.70 (0.35, 1.2)
Filoviridae				
Ebola	EBOV	G (p)	0 (0, 0.20)[c]	0.36 (0, 1.6)
Orthomyxoviridae				
Avian influenza A	AFLUV-A	NP	1.1 (0.94, 1.3)	3.3 (2.4, 4.2)
Classical swine influenza A	CSFLUV-A	NP	1.5 (1.3, 1.7)	3.6 (3.0, 5.5)
Equine influenza A	EQFLUV-A	NP	1.1 (0.46, 1.8)	4.2 (1.3, 7.7)
European swine influenza A	ESFLUV-A	NP	0 (0, 0.72)[c]	0.38 (0, 6.3)[c]
Human influenza A	HFLUV-A	NP	1.8 (1.6, 2.1)	3.4 (2.3, 4.5)
Influenza B	FLUV-B	HA (p)	1.6 (1.3, 1.9)	2.3 (1.7, 3.7)
Human influenza C	HFLUV-C	HE	0.24 (0.13, 0.34)	0.80 (0.42, 1.3)
Bunyaviridae				
Rift valley fever	RVFV	NSs (p)	0 (0, 0.12)[c]	0.79 (0.010, 1.7)
Reoviridae				
Bluetongue	BTV	S3	0.36 (0.28, 0.48)	0.60 (0.36, 1.1)
Human rotavirus	HROTAV	VP4	0.58 (0.23, 0.95)	1.4 (0, 3.4)[c]
Reovirus	REOV	S3	0.30 (0, 0.64)[c]	1.0 (0, 2.4)[c]
Birnaviridae				
Infectious bursal disease	IBDV	VP2 (p)	0.44 (0, 0.90)[c]	0.71 (0, 2.6)[c]
Retroviridae				
Human immunodeficiency 1	HIV-1	Gag-Env (p)	2.5 (1.1, 4.0)	0 (0, 0.48)[c]

[a] (p) indicates partial gene sequences.
[b] Numbers in parentheses are 0.95 confidence limits.
[c] In this case the SRDT model was not significantly better than the SR model.

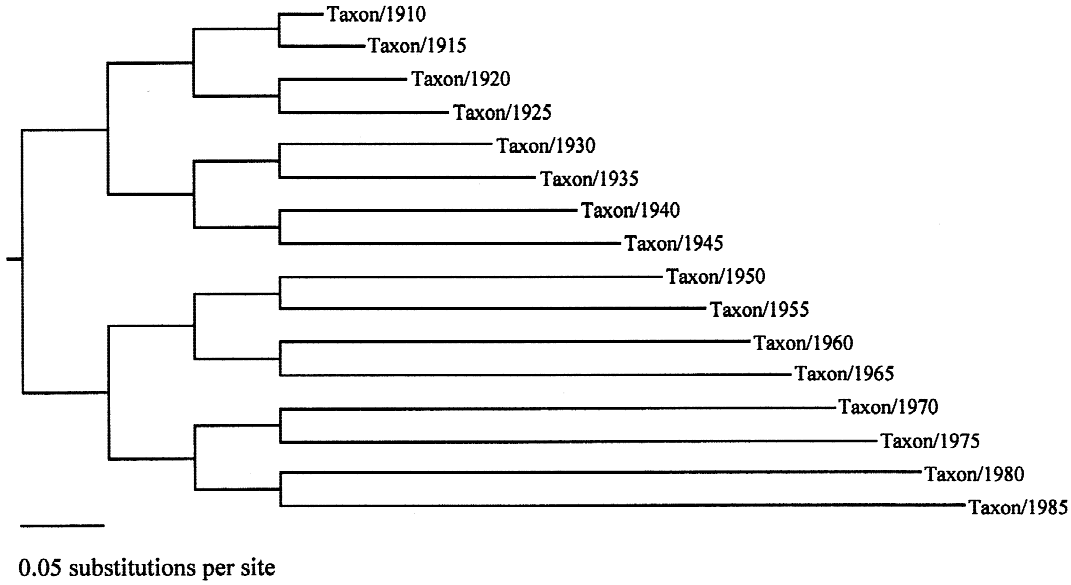


Fig. 1. Phylogeny used to simulate initial data sets.

ac.uk/software/TipDate. Briefly, the method estimates substitution rates by optimizing the branch lengths of a rooted phylogenetic tree under the constraints of a single rate of substitution applied to every branch so that the relative positions of the tips are consistent with isolation dates. This model is called the single rate dated tip (SRDT) model. Trees were rooted at the position that maximized the likelihood of the single rate model, and 0.95 confidence intervals for the rate were obtained by finding the values on either side of the maximum likelihood value that give a likelihood score 1.92 less than the maximum value. The molecular clock hypothesis was tested using a likelihood ratio test in which the test statistic is the difference in likelihood between the SRDT model and the different rate (DR) model in which branch lengths are no longer constrained, with the difference in free parameters equal to $N - 3$, where N represents the number of tips. An exploratory rate analysis was also performed for each data set using a linear regression of genetic divergence against time to verify that no outliers were included that could dramatically alter the substitution rates or the outcome of the molecular clock test. Synonymous substitution rates in this analysis were computed using the third codon position nucleotides that were fourfold degenerate at the amino acid level.

For human immunodeficiency virus (HIV), the data set of Leitner et al. (1996) was used. This contains 13 sequences of two separate gene fragments, p17*gag* and *env*V3, isolated from nine Swedish patients where the true phylogenetic history is known. The gene fragments were analyzed together since the rates of evolution of each fragment were found not to differ significantly in our analysis. The known phylogeny was also utilized. For hepatitis C virus (HCV), the data set of Smith et al. (1997) was used. This comprises a sample of contemporary sequences isolated from patients infected with a common source of contaminated blood products 17 years earlier. To estimate substitution rates in this case, a modified version of TipDate was employed which uses a star phylogeny where all branches are constrained to have equal lengths. This can be compared to the likelihood of the unconstrained model to test the molecular clock using $N - 1$ as the difference in free parameters (N again equals the number of tips). The rate of substitution for HCV will therefore represent that which occurs within hosts rather than between them.

The relative contribution of synonymous (d_S) and nonsynonymous (d_N) substitutions to the overall rate was calculated by obtaining d_N/d_S as well as the number of synonymous and nonsynonymous sites for each data set using the CODEML program in the PAML package (Yang 1997).

Codon usage, which may influence substitution rates, was measured using the effective codon index, N_C (Wright 1990). This index is a measure of overall codon bias analogous to the effective number of alleles measure used in population genetics and was calculated using the codon W program (available at http://www.molbiol.ox.ac.uk/cu/codonW.html). The reported value of N_C is always between 20 (when only one codon is effectively used for each amino acid) and 61 (when all codons are used randomly).

Simulations

To investigate the effects of rate heterogeneity and the power of the input data on the inferred maximum likelihood substitution rate, the program Seq-Gen (Rambaut and Grassly 1997) was employed to simulate 100 data sets of 16 1.0-kb sequences down a balanced phylogeny at a constant rate of 0.005 substitution/site/year using the Jukes–Cantor model of sequence evolution. This phylogeny had internal branch lengths of 0.05 substitution/site and terminal branch lengths of 0.025, 0.05,. . ., 0.4, corresponding to isolation dates of 1910, 1915, . . ., 1985, respectively (Fig. 1). Substitution rates were then estimated using TipDate and compared with those obtained from similar data sets simulated under different amounts of rate heterogeneity. This was achieved by varying the length of one of the middle terminal branches of the phylogeny without changing the corresponding isolation date. Substitution rates were then compared with those obtained from simulated data sets of variable alignment length, number of taxa, isolation time period, internal branch length, tree topology, and evenness of sampling.

Results

Maximum likelihood rates of nucleotide substitution were calculated and the molecular clock tested for 50 RNA viruses using the single rate dated tip (SRDT) model of sequence evolution (Table 1). For 32 of these viruses, the substitution rate was significantly greater than zero. Since a substitution rate of zero is equivalent to a conventional single substitution rate (SR) model in which sequences are contemporary and the tips of the tree equidistant from the root, it follows that incorporat-

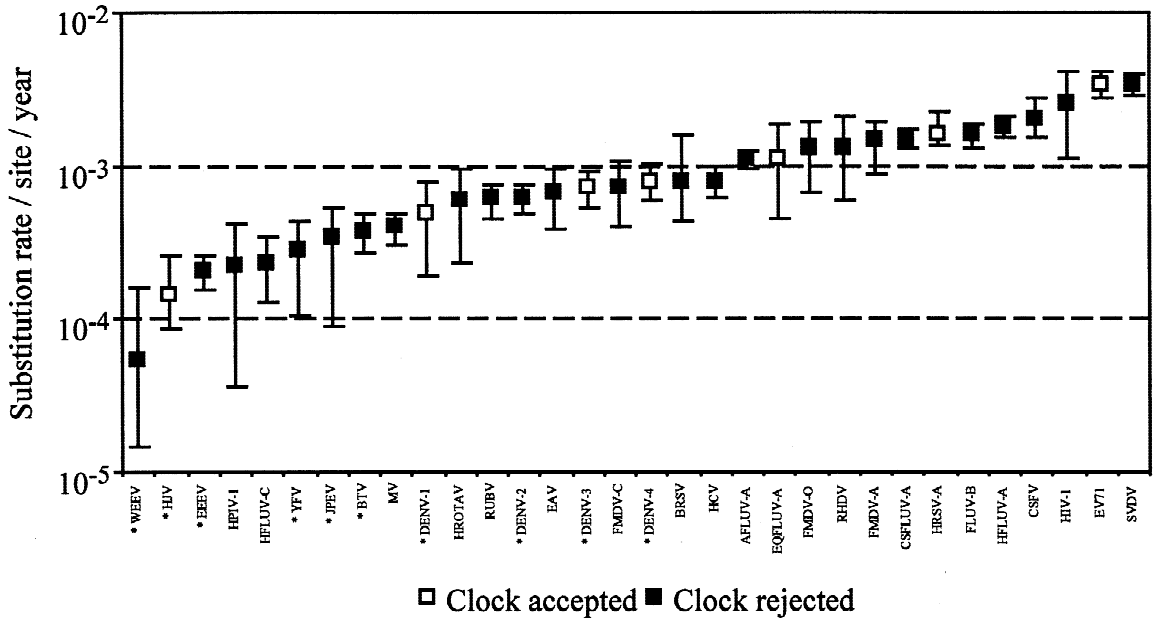


Fig. 2. Nucleotide substitution rates and results of the molecular clock test for the 32 viruses where correcting for differences in isolation times significantly improved the likelihood of a single rate model of evolution. Vector-transmitted viruses are indicated by *asterisks*, and error bars represent 0.95 confidence intervals.

ing isolation dates into a single rate model significantly improves the likelihood in these cases. For the remaining 18 viruses, where the lower confidence limit was zero, the SRDT model is not significantly better than the SR model. The sequences in these data sets cannot therefore be considered temporally distinct so that using differences in isolation times to estimate substitution rates and test the molecular clock is unjustified. Consequently, these 18 data sets were excluded from our subsequent comparative analysis of substitution rates. This invariably meant the removal of data sets that were expected to be less informative, e.g., Ebola virus, where the sequences were extremely divergent; European swine influenza A virus, where only five sequences were available; and European brown hare syndrome virus, where the sequences were very short (327 bp).

The results of the substitution rate and molecular clock analysis are summarized in Fig. 2. Many substitution rates are in agreement with previous estimates using similar sequence data, e.g., influenza viruses (Yashimata 1988; Gorman et al. 1990; Muraki et al. 1996), measles virus (Rima et al. 1997), and hepatitis C virus (Smith et al. 1997). However, the SRDT model was significantly worse than one which allows different rates along each branch (the DR model) in a majority of cases (25 of 32), rejecting a constant substitution rate for RNA viruses in general. Interestingly, the DR model was even significantly better for human influenza A virus, which supposedly represents an archetypal example of the molecular clock (Gojobori et al. 1990; Gorman et al. 1990).

Considering the confidence intervals for the substitution rates, the range over which they differ significantly is very narrow. Swine vesicular disease virus had the highest lower confidence limit (2.8×10^{-3} substitution/site/year), and Western equine encephalitis virus had the lowest upper confidence limit (0.16×10^{-3}). This corresponds to a difference of only 1.24 orders of magnitude, which is surprising given that these rates were calculated from a variety of coding regions belonging to genetically and ecologically diverse viruses.

Given such a narrow range of rate variation among these viruses and the often large confidence intervals for these rates, it is unlikely that our analysis would support any striking associations between substitution rates and the different properties of the viruses and genes used. This was verified by grouping substitution rates according to viral genome polarity and segmentation, genome length, presence of an envelope, method of transmission, viral persistence within individual hosts, principal host species, and whether the proteins encoded were structural or nonstructural. The only clear difference was between vector-transmitted viruses and other viruses, where the average upper confidence interval for vector-transmitted viruses was 0.21 order of magnitude lower than the average lower confidence interval for all other viruses. While this difference is not large, it is statistically significant according to a *t* test ($p = 0.042$). However, $p > 0.05$ using a Wilcoxon rank test (a nonparametric test), and all but one of the vector-transmitted pathogens belongs to either the Togaviridae or the Flaviviridae, so it is possible that this reduction in rate reflects shared ancestry, rather than many independent reductions. Also, a weak relationship exists between substitution rate and genome length (Fig. 3). This was significant in a linear regression analysis ($p = 0.028$) but was not significant according to a Spearman's rank cor-

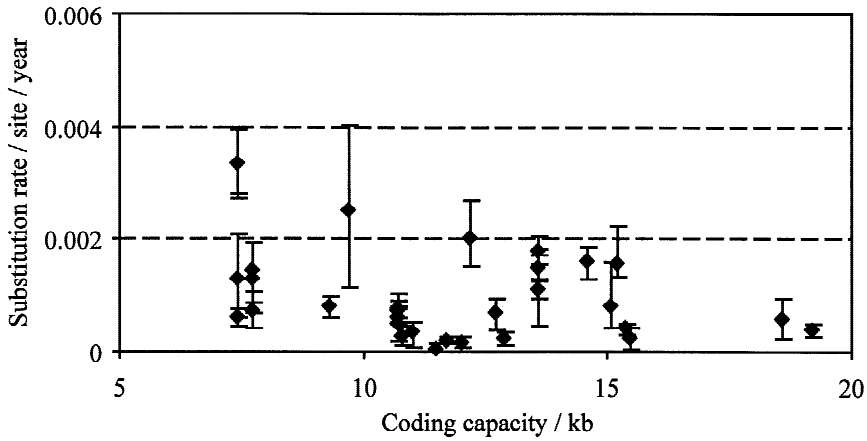


Fig. 3. Relationship between nucleotide substitution rate and viral genome size. Error bars represent 0.95 confidence intervals.

relation ($p > 0.05$). Finally, the three viruses with the lowest substitution rates have passerine birds as their principal vertebrate hosts, although the small number of avian viruses available rules out any reliable test of whether they evolve at significantly lower rates.

We then asked whether synonymous or nonsynonymous substitutions were the major contributor to the evolution of these viruses? On average for the 50 data sets, synonymous substitutions accounted for the bulk of substitutions (73%), and in only four cases were nonsynonymous substitutions more frequent. These were HIV-1 (26%), bovine respiratory syncitial virus (46%), and human respiratory syncitial viruses A (47%) and B (45%), all of which are viruses where positive selection has been proposed to play a role in evasion of the host immune response (Sullender et al. 1991; Yang 1997; Sullender et al. 1998).

Given that synonymous substitutions are the major contributor to the evolution of these viruses and that these are more likely to reflect the underlying mutation and replication rates, we tested the molecular clock using those third codon position nucleotides that are fourfold degenerate at the amino acid level and therefore potentially represent truly neutral sites. These results are summarized in Fig. 4. The SRDT model was significantly better than the SR model in 26 cases. This is a lower proportion than when all sites were used, which is probably due to fewer sites being compared in the current analysis. However, the SRDT model was rejected in favor of the DR model in a lower proportion of cases (13 of 26). This supports a slightly more clock-like pattern of evolution for synonymous sites, although this could also be a consequence of using fewer sites. Furthermore, that the molecular clock is still rejected in half the data sets is interesting and indicates that a significant number of synonymous sites in these viruses may not be evolving in a neutral manner.

The range over which the synonymous substitution rates differed significantly was 1.07 orders of magnitude, similar to the range obtained for overall substitution rates. Enterovirus 71 had the highest lower confidence limit (7.9×10^{-3}), and Eastern equine encephalitis virus the lowest upper confidence limit (0.68×10^{-3}). Again, we were unable to observe any striking associations between the substitution rates and the different properties of the viruses or type of gene used, and while the synonymous substitution rates of vector-transmitted viruses was on average lower, the difference was not statistically significant (*t* test: $p = 0.46$). Furthermore, no correlation between codon usage and synonymous substitution rates could be found. In fact, none of 50 data sets exhibited extreme biases in codon usage: the effective number of codons was on average equal to 52.1 and ranged from 39.5 for human rotavirus (medium bias) to 61.0 for Sindbis virus (no bias).

Finally, we undertook a simulation study to determine how rate heterogeneity and the power of the input data set affects the estimation of the maximum likelihood rate of substitution. Figure 5 shows the consequence of changing one of the middle terminal branch lengths of a 16-taxon tree where the genetic distance from the root was exactly proportional to the isolation time, i.e., disrupting a perfect molecular clock. It indicates that introducing fluctuations to clock-like evolution reduces the probability that the confidence limits will include the correct underlying substitution rate. However, they do not systematically bias the rate in one particular direction, since reducing the branch length in question to zero had an equal and opposite effect to increasing its length to saturation levels: the average rate for 100 simulations down each tree was 0.00492, which is not significantly different from 0.005, the underlying substitution rate (*t* test: $p > 0.05$). Also, it was found that changes to the inferred rate caused by varying one branch length could be opposed by making compensatory changes to other branches. Next, we verified that reducing the power of the input data decreases the probability that the SRDT model will be significantly better than the SR model and increases the size of the confidence intervals (Fig. 6). This was done initially by reducing the length of the input alignment and then by reducing the number of taxa and sampling time period and increasing the length of the

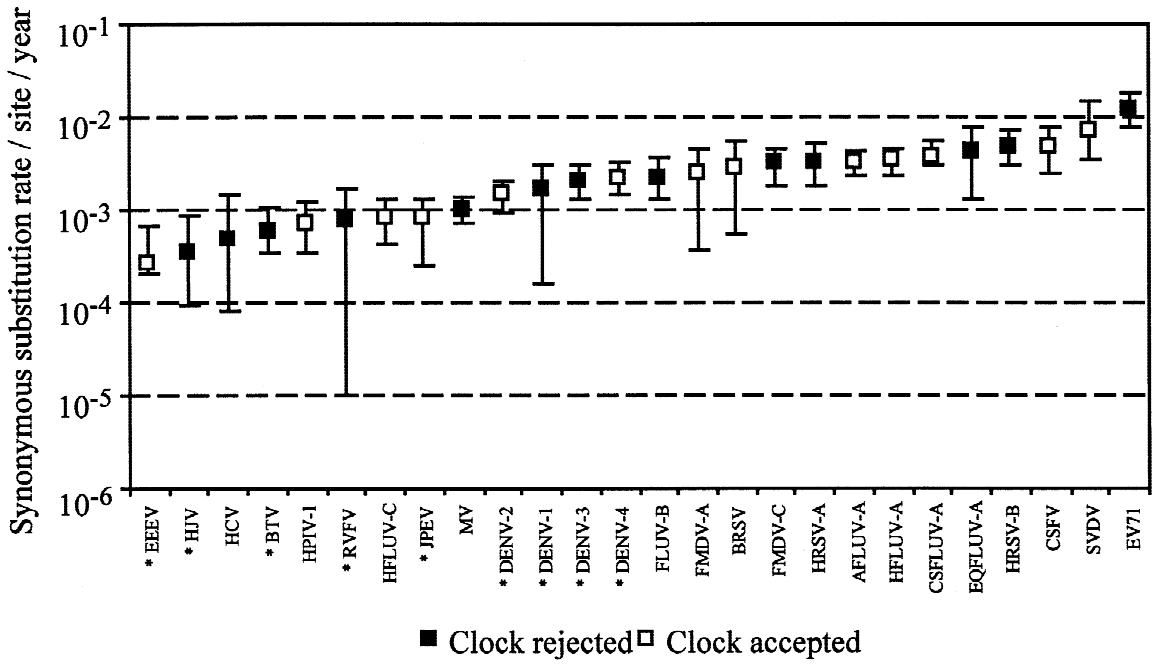


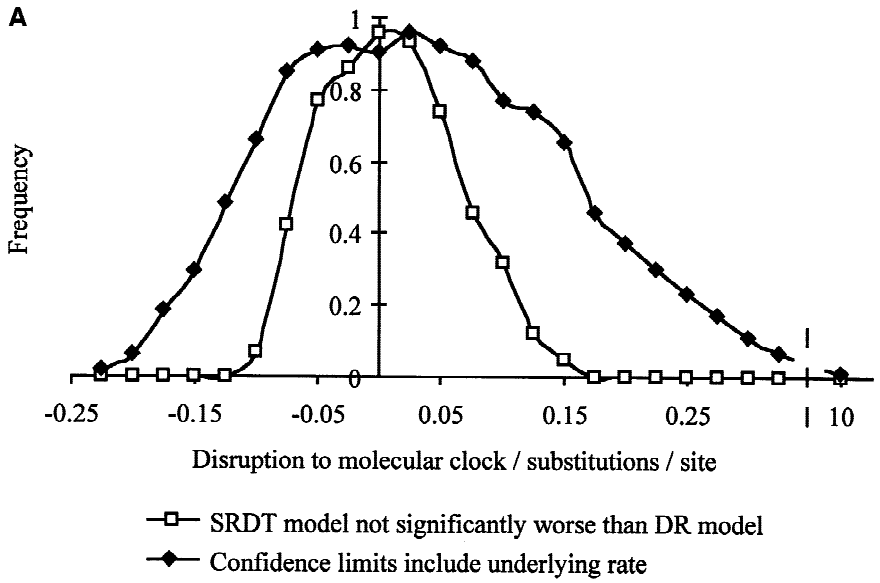
Fig. 4. Synonymous substitution rates and results of the molecular clock test for viruses whose synonymous substitution rates were significantly greater than zero. Vector-transmitted viruses are indicated by *asterisks*, and error bars represent 0.95 confidence intervals.

internal branches to saturation levels (results not shown). However, changing the topology of the phylogeny used to simulate the sequences and estimate the substitution rate influenced only slightly the size of the confidence intervals for the substitution rate, as did changing the evenness of the isolation dates. Finally, we confirmed that reducing the information contained within the input data, in this case by using shorter alignments, of 5 and 100 bp, also decreases the probability that the SRDT model will be rejected in favor of the DR model when different levels of rate heterogeneity are present.

Discussion

This study of rates of molecular evolution in RNA viruses is the most comprehensive of any such analysis undertaken to date. Our results reveal a significant relationship between genetic divergence and time of isolation for a wide range of RNA viruses, although, in most cases, more rate variation exists among lineages than would be expected under the constraints of a molecular clock. We may therefore conclude that either mutation rates, replication rates, or undefined selective constraints vary to some extent among lineages, which might be expected if RNA viruses infect a variety of host species, or that positive selection has played a role in determining sequence evolution. With respect to the latter, any selective pressure is likely to be limited to a small proportion of sites because d_S is usually greater than d_N. The general lack of a molecular clock clearly reduces the power of gene sequences to estimate divergence times for RNA viruses accurately and highlights the need for caution in such exercises. In contrast, the presence of only slight biases in codon usage for most data sets is consistent with the view that synonymous substitutions are more likely to be selectively neutral, although the general effects of this constraint, as well as that of RNA secondary structure (Simmonds and Smith 1999), remain undefined for most RNA viruses.

While our analysis provided evidence of rate variation within most RNA viruses, there was a surprisingly narrow range of statistically significant rate variation among them despite their obvious genetic and ecological differences, although it must be noted that for a minority of viruses there was no significant relationship between divergence and isolation time. Since all RNA viruses apart from retroviruses share the same basic RNA-dependent RNA polymerase and need continual productive replication for their survival, it could be that most RNA viruses evolve rapidly due simply to rapid and erroneous replication. Alternatively, there may be slower-evolving viruses among those for which the SRDT model was not significantly better than the SR model. For example, Suzuki et al. (1999) estimated the rate of nucleotide substitution in GBV-C virus to be less than 9×10^{-6} substitution/site/year, although this involved a relatively small number of sequence data (two pairs of temporally isolated sequence from single patients), which we verified was insufficient to reject the SR model in favor of the SRDT model. A slow rate of nucleotide substitution has also been proposed for human T-cell lymphotropic virus subtypes I and II (Salemi et al. 1999; Van Dooren et al. 2001), although, as these are retroviruses, the slow



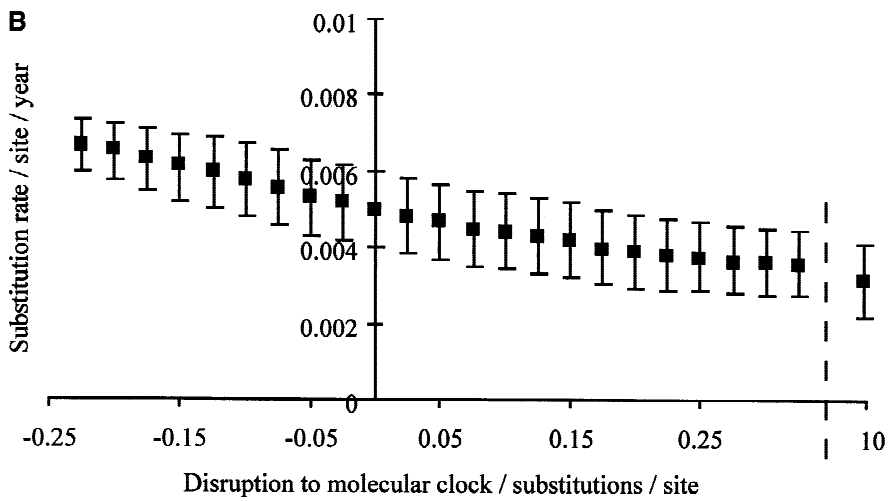


Fig. 5. **A** Effect of deviations from rate constancy on the maximum likelihood substitution rate. The *x*-axis indicates the change to the middle terminal branch length on a phylogeny where the genetic distance is exactly proportional to the isolation time. The *y*-axis indicates the frequency at which the molecular clock is accepted and at which the confidence intervals for the substitution rate include the correct underlying rate based on 100 simulations. **B** Average substitution rate and 0.95 confidence intervals for different levels of rate heterogeneity (again indicated as the change in length to one of the middle branches on a tree where the genetic distance is proportional to the isolation time), where the true underlying rate equals 0.005 substitution/site/year. Each data point represents the average of 100 simulations.

rate of nucleotide substitution could be due to long periods of latency within the host genome, and these results also rely partly on correlating virus divergence times with anthropologically documented human migrations.

Clearly, more sequence data would be necessary to demonstrate the existence of slower-evolving viruses from temporally sampled sequences. Indeed, given that stored samples are available for no more than ~80 years and that RNA virus genomes are typically no longer than 15 kb, current sequence data limit us to observing nucleotide substitutions for viruses evolving at a rate no slower than 0.8×10^{-6} substitution/site/year, assuming 1 substitution over the sampling time period. Longer sampling periods will also be required to test whether the rates estimated in this study correspond to long-term evolutionary rates or to artificially high short-term rates where purifying selection has not had enough time to eliminate deleterious mutations. However, if unusually slowly evolving RNA viruses do exist, it should be possible to show complete genomes remaining unchanged over long periods of regular sampling. Evidence that viruses have cospeciated with their hosts over very long time periods, as has been proposed for GBV-C virus and some plant viruses, would likewise provide support for low rates of nucleotide substitution (Charrel et al. 1999; Gibbs et al. 1999), although it should be noted that a cophylogeny does not necessarily prove host–virus codivergence.

Regarding our analysis of correlations between viral substitution rates and genetic and ecological factors, the possibility that vector-borne viruses evolve at a slightly lower rate is consistent with previous observations

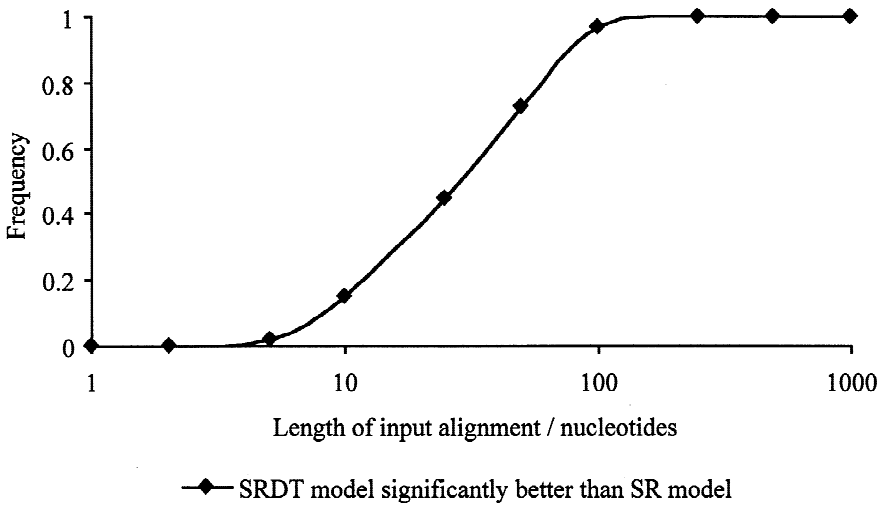


Fig. 6. Relationship between rejection of the SR model in favor of the SRDT model and input alignment length. Here, no rate heterogeneity is present, and each data point represents the average frequency for 100 simulations.

(Zanotto et al. 1996) and needs to be explored further. However, that this effect was most pronounced when all sites were included in the analysis, rather than in synonymous sites taken alone, suggests that it might be due in part to increased selective constraints imposed on the amino acid sequences of viruses that need to replicate in two very different sets of host species. The possibility of a weak relationship between substitution rates and genome length is also interesting given the proposal that an inverse relationship exists between viral mutation rates and genome length (Eigen 1993).

Our simulation study also addresses an important issue. If the molecular clock is rejected, as it was in most cases, is the maximum likelihood substitution rate a meaningful measure? As we have shown, fluctuations in clock-like behavior increase the error in rate estimates. However, since the error is random, multiple random deviations from rate constancy would be expected to have no net effect on the overall rate estimate. Therefore, substitution rates estimated from large data sets should still be reliable indicators of the average speed of evolution, even if rate heterogeneity is present. Conversely, rates estimated from less informative data sets, particularly those where incorporating isolation dates into a single substitution rate model does not increase the likelihood, may be less trustworthy. In such cases, the molecular clock test will also be unreliable, since the ability to reject a molecular clock likewise depends on the information contained within the input data.

Future advances in our understanding of the determinants of substitution rates in RNA viruses will also rely on the accurate measurement of replication rates in vivo, a barely explored area. Similarly, for viruses that cause chronic infections, it is still unclear whether within-host substitution rates differ substantially from between-host rates, particularly as selection pressures might be expected to differ within and among hosts (Bonhoeffer and Nowak 1994). Finally, the frequency and effects of localized positive selection pressure, which would cause a highly specific increase in d_N over d_S, need to be explored further, especially whether they can be correlated with wider ecological differences as perhaps we see in the vector-borne RNA viruses.

Acknowledgments. This work was supported by research grants from the Royal Society and the Wellcome Trust.

References

Albert J, Leitner T (1999) The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proc Natl Acad Sci USA 96:10752–10757

Bonhoeffer S, Nowak MA (1994) Intra-host versus inter-host selection—Viral strategies of immune function impairment. Proc Natl Acad Sci USA 91:8062–8066

Charrel RN, De Micco P, Lamballerie X (1999) Phylogenetic analysis of GB viruses A and C: Evidence for cospeciation between virus isolates and their primate hosts. J Gen Virol 80:2329–2335

Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. Genetics 148:1667–1686

Eigen M (1993) The origin of genetic information: viruses as models. Gene 135:37–47

Gibbs AJ, Keese PL, Gibbs MJ, Garcia-Arenal F (1999) Plant virus evolution: Past, present and future. In: E Domingo, R Webster, J Holland (eds) Origin and evolution of viruses. Academic Press, London, pp 263–285

Gojobori T, Moriyama EN, Kimura M (1990) Molecular clock of viral evolution, and the neutral theory. Proc Natl Acad Sci USA 87: 10015–10018

Gorman OT, Bean WJ, Kawaoka Y, Webster RG (1990) Evolution of the nucleoprotein gene of influenza A virus. J Virol 64:1487–1497

Hasegawa M, Kinshino H, Yano TA (1985) Dating of the human-ape splitting by a molecular clock of a mitochondrial DNA. J Mol Evol 22:160–174

Holland J, Domingo E (1998) Origin and evolution of viruses. Virus Genes 16:13–21

Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S, Bhattachrya T (2000) Timing the ancestor of the HIV-1 pandemic strains. Science 288:1789–1796

Krushkal J, Li WH (1995) Substitution rates in hepatitis delta virus. J Mol Evol 41:721–726

Leitner T, Escanilla D, Franzen C, Uhlen M, Albert J (1996) Accurate

reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proc Natl Acad Sci USA 93:10864–10869

Li WH, Tanimura M, Sharp PM (1988) Rates and dates of divergence between AIDS virus nucleotide sequences. Mol Biol Evol 5:313–330

Mansky LM, Temin HM (1995) Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J Virol 69:5087–5094

Martinez MA, Dopazo J, Hernandez J, Mateu MG, Sobrino F, Domingo E, Knowles NJ (1992) Evolution of the capsid protein genes of foot-and-mouth disease virus: Antigenic variation without accumulation of amino acid substitutions over six decades. J Virol 66: 3557–3565

McGuire K, Holmes EC, Gao GF, Reid HW, Gould EA (1998) Tracing the origins of louping ill virus by molecular phylogenetic analysis. J Gen Virol 79:981–988

Muraki Y, Hongo S, Sugawara K, Kitame F, Nakamura K (1996) Evolution of the haemagglutinin-esterase gene of influenza C virus. J Gen Virol 77:673–679

Nichol ST, Rowe JE, Fitch WM (1993) Punctuated equilibrium and positive Darwinian evolution in vesicular stomatitis virus. Proc Natl Acad Sci USA 90:10424–10428

Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD (1996) HIV-1 dynamics in vivo: Virion clearance rate, infected cell lifespan, and viral generation time. Science 271:1582–1586

Rambaut A (2000) Estimating the rate of molecular evolution: Incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics 16:395–399

Rima BK, Earle JAP, Baczko K, Ter Meulen V, Liebert UG, Carstens C, Carabana J, Caballero M, Celma ML, Fernandez-Munoz R (1997) Sequence divergence of measles virus haemagglutinin during natural evolution and adaptation to cell culture. J Gen Virol 78:97–106

Rodriguez LL, Fitch WM, Nichol ST (1996) Ecological factors rather than temporal factors dominate the evolution of vesicular stomatitis virus. Proc Natl Acad Sci USA 93:13030–13035

Salemi M, Lewis M, Egan JF, Hall WW, Desmyter J, Vandamme AM (1999) Different population dynamics of human T cell lymphotropic virus type II in intravenous drug users compared with endemically infected tribes. Proc Natl Acad Sci USA 96:13253–13258

Simmonds S, Smith DB (1999) Structural constraints on RNA virus evolution. J Virol 73:5787–5794

Smith DB, Pathirana S, Davidson F, Lawlor E, Power J, Yap PL, Simmonds P (1997) The origin of hepatitis C genotypes. J Gen Virol 78:321–328

Sullender WM, Mufson MA, Anderson LJ, Wertz GW (1991) Genetic diversity of the attachment protein of subgroup B respiratory syncytial viruses. J Virol 65:5425–5434

Sullender WM, Mufson MA, Prince GA, Anderson LJ, Wertz GW (1998) Antigenic and genetic diversity among the attatchment proteins of group A respiratory syncytial viruses that have caused repeat infections in children. J Infect Dis 178:925–932

Suzuki Y, Katayama K, Fukushi S, Kageyama T, Oya A, Okamura H, Tanaka Y, Mizokami Y, Gojobori T (1999) Slow evolutionary rate of GB virus C/hepatitis G virus. J Mol Evol 48:383–389

Swofford DL (1998) PAUP*. Phylogenetic analysis using parsimony (* and other methods), version 4. Sinauer Associates, Sunderland, MA

Van Dooren S, Salemi M, Vandamme AM (2001) Dating the origin of the African human T-cell lymphotropic virus type-1 (HTLV-I) subtypes. Mol Biol Evol 18:661–672

Weaver SC, Rico-Hesse R, Scott TW (1992) Genetic diversity and slow rates of evolution in new-world alphaviruses. Curr Top Microbiol Immunol 176:99–117

Wright F (1990) The effective number of codons used in a gene. Gene 87:23–29

Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. CABIOS 13:555–556

Yashimata M, Krystal M, Fitch WM, Palese P (1988) Influenza B virus evolution: Co-circulating lineages and comparison of evolutionary pattern with those of influenza A and C viruses. Virology 163:112–122

Zanotto PM de A, Gould EA, Gao GF, Harvey PH, Holmes EC (1996) Population dynamics of flaviviruses revealed by molecular phylogenies. Proc Natl Acad Sci USA 93:548–553