



Detecting natural selection in RNA virus populations using sequence summary statistics

Samir Bhatt^a, Aris Katzourakis^{a,b}, Oliver G. Pybus^{a,*}

^a Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, United Kingdom

^b Institute for Emergent Infections, The James Martin 21st Century School, University of Oxford, Oxford, United Kingdom

ARTICLE INFO

Article history:

Received 26 February 2009

Received in revised form 26 May 2009

Accepted 2 June 2009

Available online 11 June 2009

Keywords:

RNA virus

McDonald–Kreitman test

Natural selection

Tajima's *D*

ABSTRACT

At present, most analyses that aim to detect the action of natural selection upon viral gene sequences use phylogenetic estimates of the ratio of silent to replacement mutations. Such methods, however, are impractical to compute on large data sets comprising hundreds of complete viral genomes, which are becoming increasingly common due to advances in genome sequencing technology. Here we investigate the statistical performance of computationally efficient tests that are based on sequence summary statistics, and explore their applicability to RNA virus data sets in two ways. Firstly, we perform extensive simulations in order to measure the type I error of two well-known summary statistic methods – Tajima's *D* and the McDonald–Kreitman test – under a range of virus-like mutational and demographic scenarios. Secondly, we apply these methods to a compilation of ~100 RNA virus alignments that represent natural RNA virus populations. In addition, we develop and introduce a new implementation of the McDonald–Kreitman test and show that it greatly improves the test's statistical reliability on typical viral data sets. Our results suggest that variants of the McDonald–Kreitman test could prove useful in the analysis of very large sets of highly diverse viral genetic data.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

One of the main goals of viral evolutionary genetics is to understand to what extent natural selection – as opposed to mutation and random genetic drift – determines the genetic variability and evolution of viruses. Various methods of gene sequence analysis have been developed to detect and measure natural selection, the most popular of which can be categorised as either *dn/ds*-based methods (e.g. Nei and Gojobori, 1986) or methods based on site-frequency summary statistics (e.g. Tajima, 1989; McDonald and Kreitman, 1991a). The former calculate the ratio of non-synonymous to synonymous genetic changes, which is typically denoted *dn/ds* or ω . A ratio greater than one indicates the action of positive selection, while a ratio of less than one can indicate purifying selection. In contrast, summary statistic methods depend on the frequency at which polymorphisms are found in a sample of sequences. These statistics may be computed from within-species polymorphisms (Tajima, 1989) or from both polymorphisms and among-species fixations (McDonald and Kreitman, 1991a).

Currently, most studies of viral genetic data use phylogenetic *dn/ds* methods as a means to detect selection (e.g. Yang, 2007; Pond and Frost, 2005), which are based on statistical models of

codon evolution (Goldman and Yang, 1994; Yang et al., 2000). Examples of this approach are too numerous to list here, but one of the most influential was Nielsen and Yang's (1998) investigation of positive selection in the HIV-1 *env* gene. Phylogenetic *dn/ds* methods do not require users to make specific assumptions about the sampled population and can therefore provide robust evidence for the directionality of selection. In addition, simulations show *dn/ds* methods to have good statistical power under models of both positive and negative selection (Zhai et al., 2009), although in practice such methods are likely more powerful in detecting recurrent or reciprocal selection than single, historical selective sweeps (Pybus and Shapiro, 2009). However, the interpretation of *dn/ds* can be potentially misleading when recombination has been operating (Wilson and McVean, 2006) and the application of phylogenetic *dn/ds* methods to within-population data sets has recently been criticised (Kryazhimskiy and Plotkin, 2008). Crucially, phylogenetic *dn/ds* methods can be time consuming or impractical to compute on large data sets. Recent developments in sequencing technology (Margulies et al., 2005) will make commonplace the publication of data sets containing hundreds or thousands of complete viral genomes, and therefore it is sensible to investigate the potential utility of alternative methods.

Site-frequency summary statistics, such as Tajima's *D* (Tajima, 1989) have occasionally been used to analyse viral data sets. For example, Edwards et al. (2006), and Shriner et al. (2004a,b) applied versions of Tajima's *D* to HIV-1 and Tsompana et al. (2005) employed

* Corresponding author.

E-mail address: oliver.pybus@zoo.ox.ac.uk (O.G. Pybus).

the test on the Tomato spotted wilt virus. In addition, tests that consider patterns of both polymorphism and divergence, notably the McDonald–Kreitman (MK) test, have been applied to the Bovine immunodeficiency virus (Cooper et al., 1999), beak and feather disease virus (Ritchie et al., 2003) and North American Powassan virus (Ebel et al., 2001). Most pertinent to virus evolution, Williamson (2003) demonstrated that the MK test can be applied to “serially-sampled” sequences that are obtained from the same population at different time points, thereby estimating the rate of viral adaptation through time. Summary statistic methods are computationally very efficient, can potentially be applied to very large whole genome data sets, and perhaps are more robust to the effects of recombination than phylogenetic dn/ds methods. However, summary statistic methods typically assume that multiple mutations do not occur at the same nucleotide site, which may explain why they are rarely employed on rapidly evolving viral data sets, but commonly applied to species with relatively low evolutionary rates, such as *Drosophila* (McDonald and Kreitman, 1991a; Smith and Eyre-Walker, 2002; Andolfatto, 2005).

In this paper we investigate the utility and performance of two common summary statistic methods, Tajima’s D statistic (Tajima, 1989) and the MK test (McDonald and Kreitman, 1991a), when applied to RNA virus sequences. We perform extensive simulations of virus-like alignments in order to measure the type I error of these tests (i.e. the chance of falsely rejecting the hypothesis of neutral evolution). Second, we apply the two tests to a collection of almost 100 RNA virus alignments that represent natural viral populations. Third, we develop and implement a new algorithm for computing the MK test that improves the performance of the test on data sets containing much genetic variation.

2. Background

2.1. Tajima’s D statistic

The Tajima’s D test is based on two different estimates of θ , the genetic diversity of a sequence alignment: (i) the mean number of pairwise differences ($\hat{\theta}_k$) and (ii) the scaled number of segregating sites ($\hat{\theta}_s$), otherwise known as the Watterson estimate (Watterson, 1975). The units of θ are substitutions per site. The premise of Tajima’s D test is that under neutral evolution these two measures should be equal, hence the difference between them should be zero. For a neutrally evolving haploid population, θ is expected to equal $2N_e\mu$, where N_e is effective population size and μ is the rate of nucleotide substitution. Tajima’s D statistic is defined as:

$$D = \frac{\hat{\theta}_k - \hat{\theta}_s}{\sqrt{\alpha s + \beta s^2}}, \quad (1)$$

where $\hat{\theta}_s = s\gamma$, s is the number of segregating sites and α , β and γ are constants that depend on the number and length of the sequences. The denominator is a normalizing term equal to the standard error of the numerator. Under neutrality, the mean and variance of the D statistic should be approximately zero and one, respectively. Tajima’s D critically depends on the shape of the genealogy that relates the sampled sequences. For a star-like tree (long terminal branches and short internal branches), $\hat{\theta}_k < \hat{\theta}_s$, hence D is negative. This may occur during population growth or as a result of a selective sweep, which both generate more low-frequency polymorphisms than expected under neutrality. If the tree has long internal and short terminal branches (which may occur if, for example, there is strong population subdivision) then $\hat{\theta}_k > \hat{\theta}_s$ and D is positive, signifying an excess of mid-frequency polymorphisms. Tajima’s D does not require an outgroup sequence, that is, the ancestral or derived state of each polymorphism is not relevant.

2.2. The McDonald–Kreitman test

The McDonald–Kreitman (MK) test compares the pattern of polymorphism within a group (population or species) to that between two closely related groups. Under neutrality, the ratio of the number of replacement polymorphisms (r_p) to silent polymorphisms (s_p) within a group should equal the ratio of the number of replacement differences (r_d) to silent differences (s_d) between groups, such that

$$\frac{r_p}{s_p} = \frac{r_d}{s_d}. \quad (2)$$

If an excess of replacement differences between groups is observed then adaptive fixation and positive selection is inferred (McDonald and Kreitman, 1991a). The MK test is expected to be less affected by the shape of the underlying genealogy and should therefore be more robust to changes in demography (Nielsen, 2001).

The MK test requires that sites in a sequence alignment are assigned to one of the four categories defined above. Therefore an additional ‘outgroup’ sequence (or sequences) is needed to determine which sites are fixed differences (Figs. 1 and 2). Typically, this outgroup represents a closely related population or sister species (McDonald and Kreitman, 1991a; Fig. 1b) but for rapidly evolving viruses sampled at different times, the outgroup can represent the same population at an earlier time point (Williamson, 2003; Fig. 1a). The four totals (r_p , s_p , r_d and s_d) are

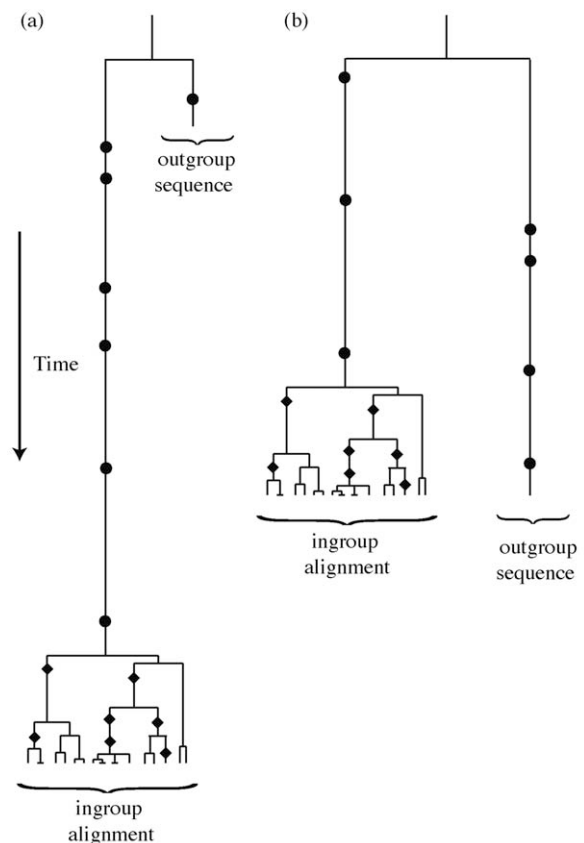


Fig. 1. An illustration of the rationale of the McDonald–Kreitman test. Sequences are sampled from the study population (ingroup alignment). In order to identify the direction of evolutionary change, and outgroup sequence is also obtained. (a) Outgroup is sampled from the study population at an earlier time point, *sensu* Williamson (2003). (b) Outgroup is obtained from a contemporaneous sister population or sister species, *sensu* McDonald and Kreitman (1991a). The circles represent fixed differences between the ingroup and outgroup. The diamonds represent ingroup polymorphisms.

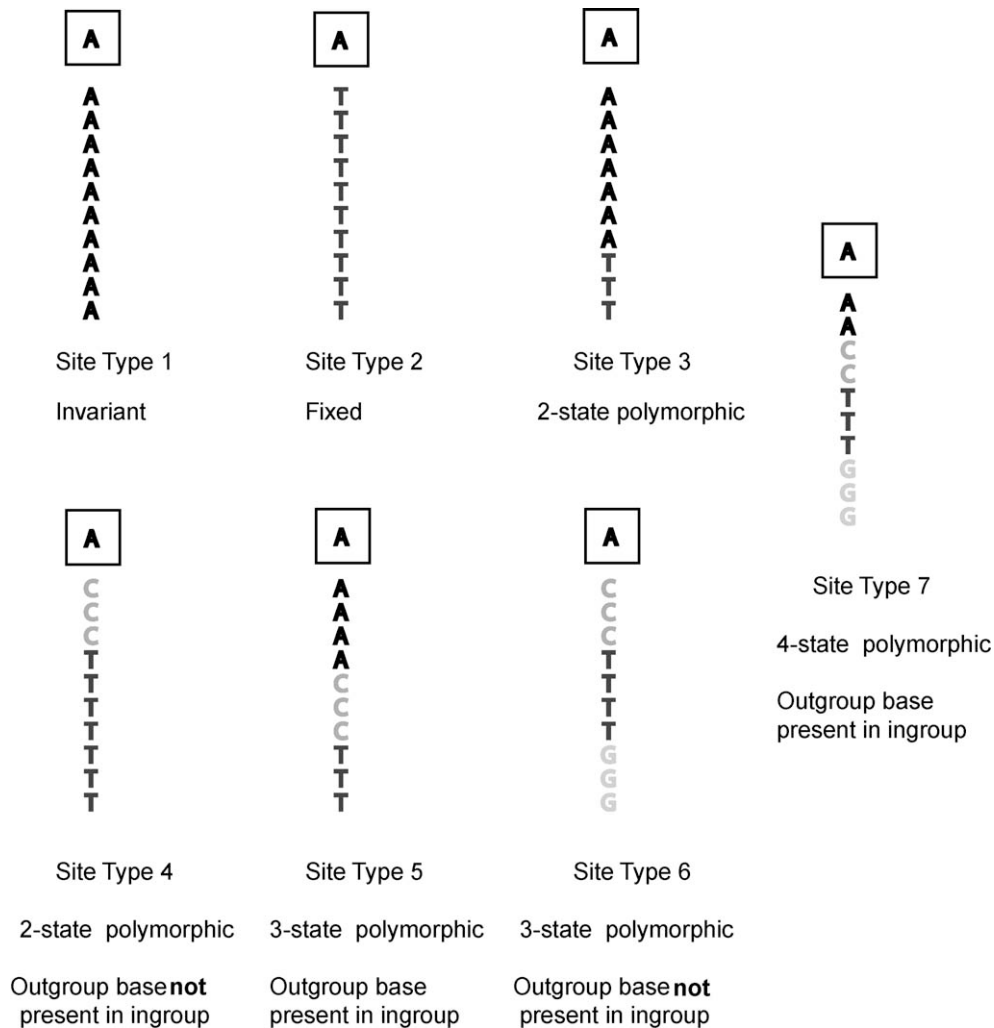


Fig. 2. An illustration of the seven “site types” defined in the main text. In each case, the box contains the nucleotide observed in the outgroup sequence. Below this is shown the site pattern observed in the ingroup alignment.

summarized in a contingency table and a non-parametric test of independence, such as the χ^2 -test, can then be used to test for a statistically significant deviation from neutrality.

3. Methods

3.1. Investigating the performance of Tajima's *D*

To explore the reliability and type I error rate of Tajima's *D* statistic, we simulated alignments of neutrally evolving sequences under various scenarios. Simulation was a two-step process. First, for each scenario, 500 neutral coalescent trees with 50 taxa were simulated. Second, one alignment of sequences, 6000 nt in length, was simulated along each tree.

Neutral coalescent trees were simulated using standard approaches (e.g. Hudson, 1990) which were implemented in the Java Evolutionary Biology Library (JEBL; available from <http://sourceforge.net/projects/jeb1>). Coalescent trees were simulated under two scenarios, constant population size and exponential growth. The latter scenario was chosen because many viral populations of interest undergo a sustained increase in population size, either during an epidemic or, at a smaller scale, immediately following transmission to a new host.

For the constant population size scenario, trees were simulated under 28 logarithmically spaced values of θ , ranging

from 0.00001 to 70. For the exponential growth scenario, trees were simulated under the same θ values plus a scaled growth rate $\rho = 200$. [Note that $\rho = r/\mu$, where r is the exponential growth rate of the population, hence $\theta\rho = N_e r$. If $\theta\rho \gg 1$ then very star-like trees are generated; see Pybus et al., 1999]. These parameter ranges were chosen to include the range of values typical for RNA virus data sets.

A codon-based Markov substitution model (Goldman and Yang, 1994) was used to simulate neutrally evolving sequences along the coalescent trees, as implemented in PAML (Yang, 2007). The sequences were generated under $dn/ds = 1$ and with equal rates of transitions and transversions. One sequence alignment was generated for every simulated tree, meaning that for each value of θ , 500 alignments of 50 sequences were generated.

Tajima's *D* statistic was calculated for each simulated data set (Tajima, 1989; computer program available on request). Although Tajima (1989) used the beta distribution to calculate critical values for the test, Simonsen et al. (1995) argue that this approach leads to conservative values and a reduction in statistical power. Therefore we used parametric bootstrapping to obtain a null distribution and 95% critical values for *D*, as follows: (i) $\hat{\theta}_s$ was calculated from the target data set, (ii) given this $\hat{\theta}_s$ value, 1000 constant population size coalescent trees were simulated using the methods above, (iii) for each tree generated in step (ii) a sequence alignment was generated under the infinite sites assumption, following the

method described in Simonsen et al. (1995), (iv) Tajima's D was calculated for each alignment generated in step (iii), resulting in a null distribution of the statistic, (v) the null hypothesis was rejected if the D value of the target data set fell outside the 95% critical values obtained in step (vi). The type I error of the test was then calculated as proportion of the 500 target data sets that rejected the null hypothesis.

3.2. Investigating the performance of the MK test

As explained above, the MK test needs to discriminate between polymorphisms and fixed differences and therefore requires an outgroup sequence, taken from either a closely related species (Fig. 1b) or an earlier time point (Fig. 1a). We chose to simulate the latter situation, which can be easily represented using the serial-sample coalescent model (Rodrigo and Felsenstein, 1999) and also corresponds to situation investigated by Williamson (2003). Crucially, the results we obtain are applicable to both situations, because the MK test depends on the genetic distance between the outgroup and ingroup, not on their relative positions in time (see Fig. 1).

As before, simulations were undertaken on both constant population size and exponential growth scenarios. For the former, two parameters were required to simulate the serial-sample coalescent trees, θ and $\tau = t\mu$, where t is the time elapsed between the earlier time point and the ingroup (Fig. 1). A range of 13 logarithmically spaced θ values were chosen, ranging from 0.00001 to 1. For each θ value, 500 trees were simulated under 12 different τ values, ranging from 0.1 to 5. Each tree comprised 50 ingroup sequences plus one outgroup sequence sampled τ time units into the past. These serial-sample coalescent trees were simulated using JEBL (see above). For the exponential growth scenario, phylogenies were simulated under the range of θ and τ values described immediately above, with the addition of a scaled exponential growth rate of $\rho = 200$.

As before, a codon-based Markov substitution model (Goldman and Yang, 1994) was used to simulate neutrally evolving sequences along the coalescent trees. One sequence alignment (6000nt long) was generated for every simulated serial-sample tree, meaning that for each value of θ or τ , 500 alignments of 51 sequences were generated. As before, sequences were generated under $dn/ds = 1$ and with equal rates of transitions and transversions.

For each simulated alignment, the total number of sites in each category (r_p , s_p , r_d and s_d) were computed. We developed a new approach to this computation, explained below, and analysis of simulated alignments was performed using both the standard method and our new approach. A χ^2 test of independence was applied to the site totals for each of the 500 replicate alignments. Hence, for each specific combination of θ and τ , the type I error equals the proportion of the 500 χ^2 tests that were significant at the $p = 0.05$ level.

3.3. New proportional counting algorithm for the McDonald–Kreitman test

The MK test requires that the number of sites belonging to different categories (r_p , s_p , r_d , s_d) are computed accurately. When sequence diversity (θ) is low this is straightforward, as the majority of sites will be either fixed or 1-state polymorphic (Fig. 2), that is, each mutation occurs at a different site. Furthermore, variable sites are unlikely to fall within the same codon hence each mutation can be easily categorised as silent or replacement. Therefore a simple count will suffice when the test is applied to animal genomes (e.g. McDonald and Kreitman, 1991b; Eyre-Walker, 2006). However, a more sophisticated approach is needed for viral alignments, which can be highly diverse. As θ increases, there is a greater chance of observing sites which are 2, 3 or 4-state polymorphic (Fig. 2), hence

multiple mutations may occur at the same site or within the same codon. The categorisation of sites therefore becomes more ambiguous—not accounting for this ambiguity could potentially introduce biases into the MK test. To avoid such biases we have developed a “proportional” counting approach, described below, that incorporates the ambiguity in site categorisation. A different, but related, approach was employed by Egea et al. (2008).

For a given ingroup alignment plus outgroup sequence (Fig. 1), we define seven ‘site types’ that describe all the possible nucleotide patterns that could occur, illustrated in Fig. 2. Rather than unambiguously assigning sites as fixed or polymorphic, we give each site i a “fixation score” F_i and a “polymorphism score” $P_i = (1 - F_i)$. If the site is definitely fixed then $F_i = 1$ and $P_i = 0$. Uncertainty in the status of a site is represented by assigning values between zero and one, as follows.

- **SITE TYPE 1:** All ingroup bases identical to the outgroup (invariant sites). $F_i = 0$ and $P_i = 0$.
- **SITE TYPE 2:** All ingroup bases identical but different from the outgroup (fixed sites). $F_i = 1$ and $P_i = 0$.
- **SITE TYPE 3:** Ingroup contains two bases, one of which is identical to the outgroup. $F_i = 0$ and $P_i = 1$.
- **SITE TYPE 4:** Ingroup contains two bases, neither of which is identical to the outgroup. McDonald and Kreitman (1991a) would classify this site as polymorphic (i.e. $F_i = 0$, $P_i = 1$). However, as no ancestral base is observed, the most plausible explanation is that an earlier fixation event has been followed by another mutation at the same site. Classifying such sites as polymorphic would under-estimate the number of fixations. Therefore $F_i = 0.5$ and $P_i = 0.5$.
- **SITE TYPE 5:** Ingroup contains three bases, one of which is identical to the outgroup. Observing an outgroup base increases the likelihood that neither of the two polymorphic bases has yet fixed. Therefore $F_i = 0$ and $P_i = 1$.
- **SITE TYPE 6:** Ingroup contains three bases, none is identical to the outgroup. As with site type 4, no ancestral bases are observed hence the most likely scenario is an earlier fixation followed by further mutations at the same site. Therefore $F_i = 1/3$ and $P_i = 2/3$.
- **SITE TYPE 7:** Ingroup contains all four bases. No reliable conclusion can be drawn, so we conservatively assign the site as $F_i = 0$ and $P_i = 1$. If such sites are common then the MK test should not be applied.

We also developed a proportional approach to evaluating whether a variable site is silent or replacement. Rather than unambiguously assigning sites as fixed or polymorphic, we give each site i a “silent score” S_i and a “replacement score” $R_i = (1 - S_i)$. For each site, the silent score is simply the proportion of ingroup bases that, if hypothetically inserted into the outgroup sequence, would not change the amino acid coded by the corresponding codon.

For a given alignment, the number of sites in different categories (r_p , s_p , r_d , s_d) are straightforwardly computed from the proportional site scores as follows:

$$\begin{aligned} r_p &= \sum_{i=1}^n P_i R_i, & s_p &= \sum_{i=1}^n P_i S_i \\ r_d &= \sum_{i=1}^n F_i R_i, & s_d &= \sum_{i=1}^n F_i S_i \end{aligned} \quad (3)$$

This algorithm was implemented in a Java computer program (available on request).

3.4. Comparative analysis of RNA virus data sets

To investigate the performance of Tajima's D and the MK test on viral sequences, we utilized a previously published and curated

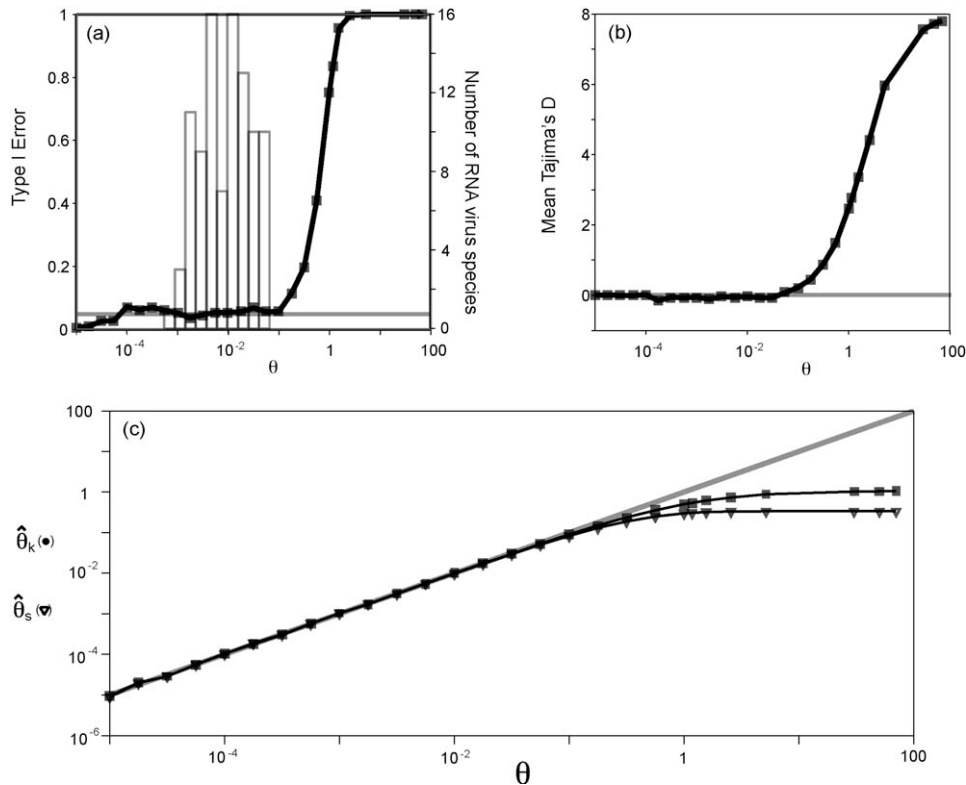


Fig. 3. The behaviour of Tajima's D under a constant population size coalescent model. Simulations were conducted under a range of θ values (horizontal axes). (a) The type I error of Tajima's D test for different values of θ . Each point represents the mean error of 500 simulations, and the grey line marks the expected 5% error rate. The superimposed histogram represents the distribution of θ values from 96 empirical RNA virus data sets (see text and Table 1). (b) The mean values of Tajima's D statistic for each values of θ . Each point represents the average of 500 simulations. The grey line marks the expected value of D , zero. (c) Mean values $\hat{\theta}_k$ and $\hat{\theta}_s$ for each simulated value of θ . Each point represents the average of 500 simulations. The gray line represents the expected relationship $\hat{\theta}_k = \hat{\theta}_s = \text{true}$.

collection of alignments from ~ 100 different RNA virus species (see Shapiro et al., 2006; Pybus et al., 2007 for details). The alignments represent partial or complete structural gene sequences and should well represent the behaviour and diversity of RNA virus data sets. For each alignment, gene diversity (θ) was calculated using the Watterson estimator ($\hat{\theta}_s$) and Tajima's D statistic was calculated as described above. In order to perform the MK test, it was first necessary to identify an outgroup sequence for each data set. We chose to use sister-species for outgroups, as serial-sample outgroups were less common. Sister-species outgroups were identified as follows: (i) representative sequences from each species were used as queries in a nBLAST search against the non-redundant database, resulting in a set of candidate outgroups; (ii) distance-based phylogenies and the viral taxonomic literature were used to choose the most closely related candidate outgroup; (iii) the chosen outgroup was profile-aligned to the curated alignment using ClustalW2 (Larkin et al., 2007) and subsequently inspected and edited by hand, paying particular attention to codon structure. Using this approach, 96 RNA virus data sets were given a reliable sister-species outgroup and were subjected to the MK test, as described above. In addition, the mean pairwise genetic distance between the outgroup and ingroup sequences was calculated for each data set, using the Jukes–Cantor method (Jukes and Cantor, 1969).

4. Results

4.1. Investigating the performance of Tajima's D

Fig. 3 shows the performance of the Tajima's D test on neutral sequences simulated under different θ values and sampled from a constant-sized population. Fig. 3a shows the type I error of the test,

Fig. 3b shows the average D value and Fig. 3c shows the mean values of $\hat{\theta}_k$ and $\hat{\theta}_s$ for each simulated value of θ . The statistical performance of the test depends greatly on θ . For explanatory convenience, we divide the range of θ into three regions.

- **REGION ONE** ($\theta < 10^{-4}$): Alignments generated under these low θ values have very few polymorphic sites (0–4 per alignment). Although the error rate of the test appears low in this region (Fig. 3a), alignments with such small amounts of variation are not suitable for analysis. Furthermore, simulations in this region are conditionally distributed, not random, because we discard alignments with zero polymorphisms. Therefore the simulations from this region are ignored.
- **REGION TWO** ($10^{-4} < \theta < 0.1$): Tajima's D test performs very well in this region, with type I error rates close to 5% (Fig. 3a) and a mean D value close to zero (Fig. 3b). A small amount of measurement error is noticeable, as only 500 simulations were performed for each point.
- **REGION THREE** ($\theta > 0.1$): The error rate in this region rises rapidly as θ increases. If $\theta > 5$, the error rate is 100% (Fig. 3a) and $\hat{\theta}_k$ and $\hat{\theta}_s$ reach maximal values because all sites are polymorphic (Fig. 3c). In this region multiple changes at the same site are observed, violating the 'infinite sites' assumption of the test and generating error. Both $\hat{\theta}_k$ and $\hat{\theta}_s$ under-estimate true θ . However the under-estimation is greater for $\hat{\theta}_s$, hence mean $D > 0$ (Fig. 3b).

Fig. 4 shows the performance of the Tajima's D test on neutral sequences sampled from exponentially growing populations. These results differ from those simulated under constant population size (Fig. 3) in several ways. Firstly, at high θ values, we no longer observe multiple mutations at the same site. This is because, on average, the underlying phylogeny becomes shorter as $\theta\rho$

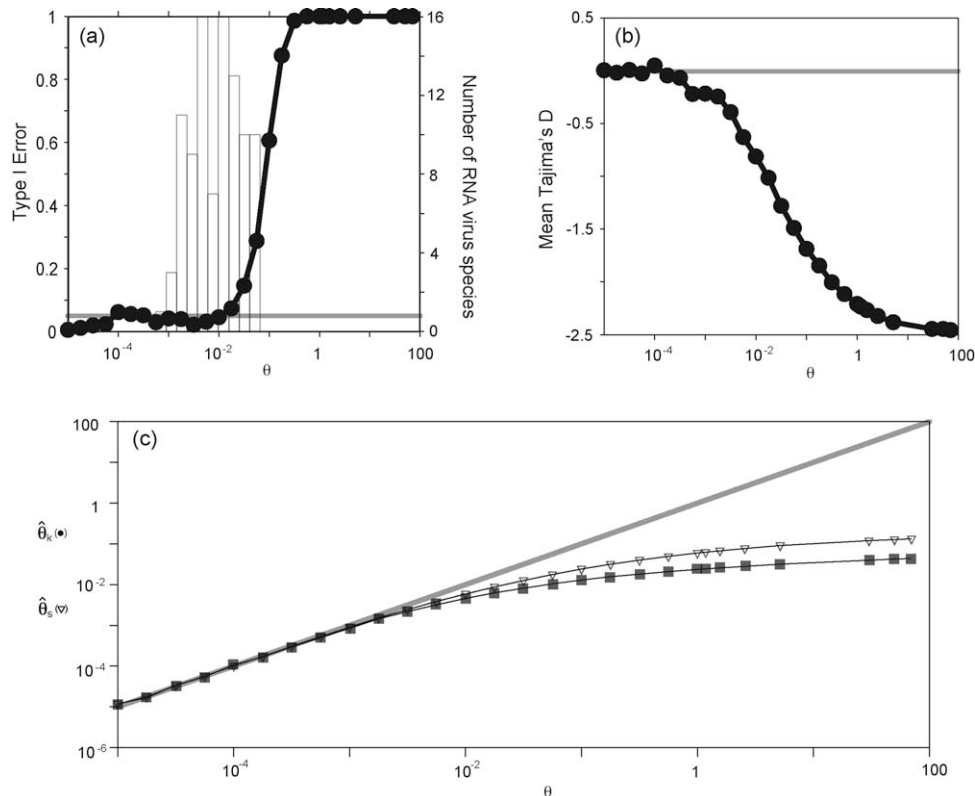


Fig. 4. The behaviour of Tajima's D under an exponential growth coalescent model. Simulations were identical to those shown in Fig. 2, except that a scaled growth rate of $\rho = 200$ was used. See Fig. 2 legend for further details.

increases and therefore fewer polymorphisms are seen in the sample (Slatkin and Hudson, 1991). Secondly, as θ rises, average D becomes increasingly negative (Fig. 4b) because exponential growth causes the phylogeny to become more star-like (Slatkin and Hudson, 1991). Therefore, as $\theta\rho$ increases, the level of diversity stabilizes (Fig. 4c) and polymorphisms are more commonly seen at low frequencies, which results in comparatively lower values for $\hat{\theta}_k$ than for $\hat{\theta}_s$ (Fig. 4c). Previous studies have shown that the transition from structured 'constant-size' phylogenies to star-like 'exponential growth' phylogenies occurs around $\theta\rho = 1$ (Slatkin and Hudson,

1991; Pybus et al., 1999). In our simulations we used $\rho = 200$, hence in Fig. 4 this transition occurs around $\theta = 0.005$. Above this value, the error rate rises rapidly (Fig. 4a) and mean D values become significantly negative (Fig. 4b), as expected by theory.

To test whether real RNA virus data sets are suitable for analysis using Tajima's D , we calculated θ and Tajima's D statistic for each of 96 RNA virus data sets (Table 1). We find that 17 out of 96 (17.7%) of empirical data sets rejected the null hypothesis of neutral evolution. In addition, we superimposed the frequency distribution of these empirical θ values onto the type I error plots (Figs. 3a

Table 1

Summary statistics and McDonald–Kreitman test p -values for RNA virus data sets.

RNA virus (gene)	Sequence diversity ^a (θ_s)	Tajima's D statistic	Mean pairwise ingroup/outgroup genetic distance ^b	p -Value of MK test
Australian bat lyssavirus (G)	0.0011	-1.6204	0.3398	0.0004
Acute bee paralysis virus (C)	0.0038	0.4417	0.3273	<0.0001
Akabane virus (NP)	0.0125	-0.2751	0.3808	<0.0001
Avian influenza A, serotype H5N1 (NP)	0.0252	-0.5852	0.0439	0.7663
Avian influenza A, serotype H7N1 (HA)	0.0486	-0.1258	0.0979	0.5382
Avian pneumovirus (N)	0.0070	0.8327	0.2779	0.0008
Barley yellow mosaic virus (CP)	0.0015	-1.2145	0.3923	0.0189
Bean yellow mosaic virus (CP)	0.0043	1.1846	0.3666	0.0335
Bluetongue virus (VP7)	0.0463	2.7726	0.5042	<0.0001
Bovine rotavirus (VP7)	0.0660	1.8774	0.2520	0.3057
Crimean-Congo haemorrhagic fever virus (NP)	0.0473	-0.2565	0.4981	<0.0001
Canine distemper virus (H)	0.0336	-1.1216	0.3627	0.0079
Chikungunga virus (E1)	0.0030	-0.3933	0.2935	0.7942
Classical swine fever virus (E2)	0.0046	0.7598	0.4454	<0.0001
Clover yellow vein virus (CP)	0.0428	0.1443	0.3433	<0.0001
Coxsackievirus B4 (VP1)	0.0387	2.0103	0.4862	<0.0001
Curcubit yellow stunting disease virus (CP)	0.0015	0.6448	0.4175	<0.0001
Dengue virus, serotype 1 (E)	0.0337	-0.4532	0.3672	<0.0001
Dengue virus, serotype 1 (CM)	0.0023	0.0696	0.3928	<0.0001
Dengue virus, serotype 2 (E)	0.0057	-0.7046	0.4410	<0.0001
Dengue virus, serotype 3 (E)	0.0282	-0.2544	0.3700	<0.0001
Dengue virus, serotype 4 (E)	0.0035	-0.5530	0.4298	<0.0001

Table 1 (Continued)

RNA virus (gene)	Sequence diversity ^a (θ_s)	Tajima's <i>D</i> statistic	Mean pairwise ingroup/outgroup genetic distance ^b	<i>p</i> -Value of MK test
Dobrava virus (N)	0.0120	0.5137	0.1438	0.7854
Eastern equine encephalitis virus (C)	0.0091	-1.5820	0.2968	0.0044
Eastern equine encephalitis virus (E1)	0.0059	-1.2072	0.5139	< 0.0001
Enterovirus 71 (VP1)	0.0047	1.6338	0.5572	< 0.0001
Equine influenza, serotype H3N8 (HA)	0.0140	-1.1273	0.2701	< 0.0001
Feline immunodeficiency virus (Gag)	0.0026	1.2008	0.4259	< 0.0001
Human influenza A virus, serotype H3N2 (HA)	0.0192	-0.7710	0.1378	0.3342
Human influenza A virus, serotype H3N2 (NP)	0.0184	-0.5076	0.0541	0.6635
Garlic latent virus (CP)	0.0059	2.2698	0.6762	< 0.0001
Hepatitis C virus 1b (C)	0.0022	-1.3132	0.1630	0.6081
Hepatitis C virus 1b (E1E2)	0.0101	-0.0676	0.4250	0.3703
HIV type 1, subtype B (Env)	0.0133	-1.1718	0.0737	0.9962
HIV type 1, subtype B (Gag)	0.0138	-1.5351	0.2527	0.1574
Human polio virus type 2 (VP)	0.0068	-0.3438	0.3390	< 0.0001
Human respiratory syncytial virus A (G)	0.0057	-1.0333	0.4381	0.2953
Human respiratory syncytial virus A (N)	0.0400	2.0114	0.1711	0.3360
Human respiratory syncytial virus B (G)	0.0032	-0.6702	0.4301	0.1642
Hantaan virus (G1)	0.0218	0.4124	0.3797	< 0.0001
Hantaan virus (N)	0.0284	1.5128	0.3060	< 0.0001
Viral hemorrhagic septicaemia virus (GP)	0.0335	-0.9365	0.6637	0.0211
Viral hemorrhagic septicaemia virus (N)	0.0019	-1.2589	0.6738	< 0.0001
Highlands J virus (E1)	0.0028	-1.8739	0.2687	0.5030
Human astrovirus (C)	0.0005	2.2044	0.1921	1.0000
Human parainfluenza virus type 1 (HN)	0.0115	-0.7474	0.4116	0.0363
Human parainfluenza virus type 3 (HN)	0.0020	-0.2687	0.6420	< 0.0001
Infectious pancreatic necrosis virus (VP2)	0.0019	2.0689	0.2008	0.2963
Japanese encephalitis virus (CP)	0.0232	-0.2415	0.3979	< 0.0001
Japanese encephalitis virus (E)	0.0023	-0.2765	0.3743	< 0.0001
Junin virus (NP)	0.0057	-1.0405	0.2784	0.2412
Leek yellow stripe virus (CP)	0.0235	1.4963	0.5699	< 0.0001
Lettuce mosaic virus (CP)	0.0057	-0.9932	0.4807	0.0652
Maize dwarf mosaic virus (CP)	0.0023	0.3379	0.2797	< 0.0001
Measles virus (HA)	0.0038	-0.2067	0.4955	< 0.0001
Measles virus (N)	0.0179	-1.3717	0.3815	0.0081
Mumps virus (NP)	0.0148	-0.0931	0.5786	< 0.0001
Onion yellow dwarf virus (CP)	0.0011	2.1820	0.2895	0.8045
Oropouche virus (NP)	0.0043	0.1407	0.3822	< 0.0001
Pea seed-borne mosaic virus (CP)	0.0155	0.0979	0.5284	< 0.0001
Peanut stripe virus (CP)	0.0183	-0.2940	0.0391	0.9102
Polio virus, serotype 1 (VP1)	0.0096	1.7809	0.3881	< 0.0001
Porcine rotavirus (VP7)	0.0484	2.4807	0.2811	0.6181
Potato virus A (CP)	0.0533	-1.1543	0.0350	1.0000
Potato virus S (CP)	0.0136	-0.8098	0.4966	< 0.0001
Potato virus X (CP)	0.0456	-1.1908	0.8398	< 0.0001
Prunus necrotic ringspot virus (CP)	0.0029	-1.1711	0.6136	0.4305
Puumala virus (G2)	0.0254	1.4645	0.2651	0.0236
Puumala virus (N)	0.0100	2.1947	0.2207	0.1226
Rabies virus (G)	0.0309	0.3381	0.3610	0.0033
Rabies virus (N)	0.0073	0.7230	0.2571	0.7655
Rice black streaked dwarf virus (CP)	0.0442	-0.2376	0.1347	0.8355
Ross River Virus (E2)	0.0190	-0.2745	0.3393	0.0120
Rotavirus A (VP7)	0.0031	-1.7368	0.1587	0.0005
Rotavirus C (VP7)	0.0350	-1.3955	0.1797	0.1248
St. Louis encephalitis virus (E)	0.0178	-0.3246	0.4187	< 0.0001
Sendai virus (NP)	0.0048	1.5436	0.3313	0.5678
Simian foamy virus (Env)	0.0022	-1.5075	0.4515	0.3615
Soybean mosaic virus (CP)	0.0014	0.0102	0.1811	< 0.0001
Sugarcane mosaic virus (CP)	0.0023	0.4636	0.2232	0.2193
Sweet potato feathery mottle virus (CP)	0.0118	0.1388	0.3707	0.0183
Swine influenza virus, serotype H3N2 (HA)	0.0244	0.5441	0.1832	< 0.0001
Tick-borne encephalitis virus (E)	0.0245	0.7927	0.1540	1.0000
Tomato spotted wilt virus (N)	0.0609	-0.8911	0.2742	0.4981
Tula virus (NP)	0.0148	1.5002	0.1838	0.3145
Turnip mosaic virus (CP)	0.0159	-1.6141	0.3746	0.0873
Venezuelan equine encephalitis virus (C)	0.0027	1.8437	0.6133	< 0.0001
Venezuelan equine encephalitis virus (E)	0.0048	0.6638	0.5943	< 0.0001
Western equine encephalitis virus (E1)	0.0063	-1.6399	0.2774	0.6705
West Nile virus (E)	0.0108	-0.9561	0.3944	0.0026
Wheat streak mosaic virus (CP)	0.0100	-2.0964	0.3162	0.0003
Wheat yellow mosaic virus (CP)	0.0039	-1.5503	0.3900	0.4392
Yellow fever virus (E)	0.0129	2.0628	0.5578	< 0.0001
Yam mosaic virus (CP)	0.0068	0.4330	0.5368	< 0.0001
Zucchini yellow mosaic virus (CP)	0.0045	-0.0098	0.4334	< 0.0001

Significant MK test *p*-values (<0.05) are shown in bold.^a Calculated using the Watterson (1975) method.^b Calculated using the Jukes and Cantor (1969) method.

and 44a). If the sequences are assumed to come from a constant-sized population then all of the empirical data sets have θ values that lie within the working range of Tajima's D test (Fig. 3a). If the sequences are assumed to come from an exponentially growing population then the suitability of the test drops dramatically (Fig. 4a). Hence the primary problem arising when Tajima's D is applied as a neutrality test to RNA viruses is not the invalidation of the infinite sites assumption caused by high mutation rates, but rather the sensitivity of the test to changing population size or population structure (Simonsen et al., 1995). Indeed, the sensitivity of Tajima's D to population size change means that it can be used to detect population growth (e.g. Ramos-Onsins and Rozas, 2002), although this necessitates the user to assume that the sequences in question have evolved neutrally.

4.2. Investigating the performance of the MK test

Fig. 5 shows the performance of the MK test on simulated neutral sequences, performed using both the standard counting method (McDonald and Kreitman, 1991a; Eyre-Walker, 2006) and our new "proportional" counting approach (see Section 2). For both counting methods, the test was applied to neutral sequences from both constant-sized (Fig. 5a and b) and exponentially growing (Fig. 5c and d) populations. In order to directly compare the simulation results with our empirical RNA virus data sets, we plot θ against the mean pairwise genetic distance between the ingroup and outgroup sequences (rather than against τ , which is

unknown for our real data). The performance of the MK test under each set of parameters (θ and τ) is represented as a circle whose diameter is proportional to type I error. Error rates between 5% and 10% are coloured orange and error rates greater than 10% are coloured red.

We begin by considering the results obtained for constant-sized populations (Fig. 5a and b). For explanatory convenience, we again divide parameter space into different regions:

- **REGION 1** ($\theta < 10^{-2}$): Alignments generated under these θ values exhibit few differences between the ingroup and outgroup, and of this variation, almost all is described by site types 2 and 3 (see Fig. 2). As discussed in Section 2, the interpretation of these is unambiguous, resulting in low type I error rates (below or around 5%) for both implementation methods (Fig. 5a and b). Many non-viral data sets that have been analysed using the MK test have θ values that lie in this region (e.g. Andolfatto, 2005; Smith and Eyre-Walker, 2002; McDonald and Kreitman, 1991b).
- **REGION 2** ($10^{-2} < \theta < 3$): In this range of θ values, the number of polymorphisms and fixed differences increases as $\theta\tau$ increases, as does the frequency of potentially ambiguous sites (site types 4, 5 and 6). Therefore type I error rises towards the upper-right-hand corner of this region. When $\theta\tau$ is comparatively low, the standard counting method produces satisfactory error rates, but as $\theta\tau$ increases this method becomes unusable (Fig. 5a). In contrast, our new proportional counting method performs very well on all but the very highest values of $\theta\tau$ (Fig. 5b).

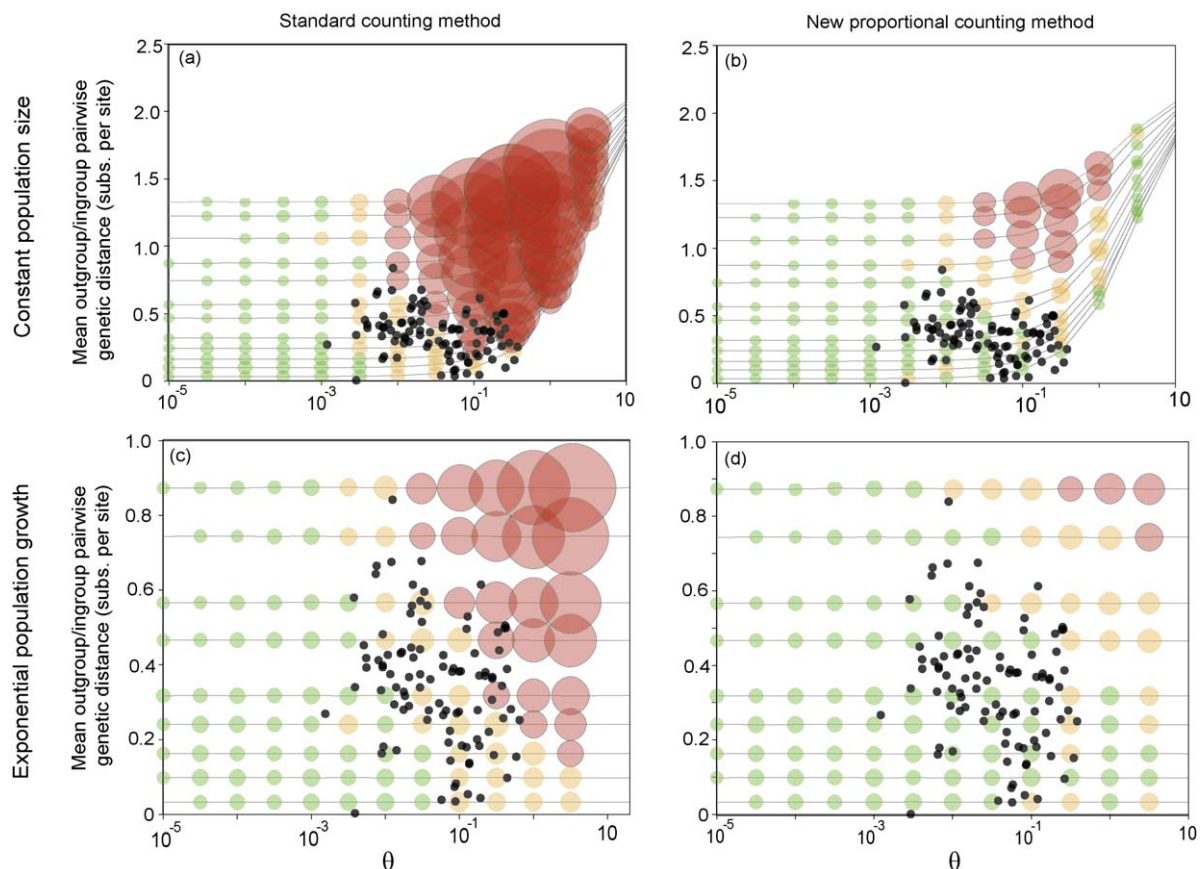


Fig. 5. The behaviour on the MK test on neutral sequences. Each plot shows the type I error of the MK test under different parameter values. For a given value of θ and τ , the diameter of each circle is proportional to the type I error of the test, averaged across 500 simulations. Green circles represent type I error rate $<5\%$, orange circles between 5% and 10%, and red circles $>10\%$. Thin grey lines connect simulations performed under the same value of τ but different values of θ . The superimposed black points represent the distribution in parameter space of the 96 empirical RNA virus data sets (see text and Table 1). (a) Standard counting method applied to sequences simulated under constant population size. (b) New proportional counting method applied to sequences simulated under constant population size. (c) Standard counting method applied to sequences simulated under exponential population growth. (d) New proportional counting method applied to sequences simulated under exponential population growth (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.).

- **REGION 3** ($\theta > 3$): This region is characterized by exceptionally high θ values, such that almost every ingroup site is polymorphic and contains multiple mutations (site types 6 and 7). Under these conditions it is impossible to accurately estimate the number of fixed differences (r_d , s_d). MK test is unlikely to generate meaningful and robust results in this region, even if the type I error sometimes appears low.

The performance of the MK test on neutral sequences from exponentially growing populations is shown in Fig. 5c and d. Under exponential growth the underlying phylogeny becomes star-like and its total length is reduced. As a result of the latter change, fewer mutations accrue and potentially ambiguous sites (site types 4, 5, 6 and 7; see Fig. 2) are rarer, hence type I errors are lower than those obtained for constant populations (Fig. 5). Under exponential growth, only 40% of the sites were variable at the highest values of $\theta\tau$, whereas under constant population size, 95–100% of sites were variable when $\theta\tau$ was very high. As before, our proportional counting method exhibits lower type I error than the corresponding values obtained using the standard method.

We also calculated $\hat{\theta}_s$ and the mean outgroup–ingroup pairwise genetic distance for 96 empirical RNA virus data sets (Table 1). These empirical values are superimposed as black dots on each plot in Fig. 5. It is clear that most RNA viral populations have comparatively high θ values ($0.001 < \theta < 0.5$ substitutions per site). Using the standard counting method, many data sets correspond to regions of parameter space with high type I error (Fig. 5a and c, red and orange circles). However, our proportional counting method reduces type I error in this region so that almost all empirical data sets correspond to parameter regions with low error (Fig. 5b and d, green circles). Table 1 also shows that the MK test rejected the null hypothesis of neutrality for 58 out of 96 (60.4%) empirical RNA virus data sets.

5. Discussion

It is widely acknowledged that Tajima's D is sensitive to changes in population size or the existence of population structure (e.g. Simonsen et al., 1995; Nielsen, 2005). A further concern with using Tajima's D test on viral populations is that their high evolutionary rates would invalidate the test's key assumption that each mutation occurs at a different site (the 'infinite sites' assumption). In our study we simulated sequences under an exhaustive range of θ values to assess the type I error of Tajima's D and also analysed a compilation of 96 alignments in order to determine the empirical range of θ values for RNA viruses (Table 1). We conclude that for constant-size populations, violation of the infinite sites assumption is not the primary problem—all our RNA viral data sets have θ values that lie in the working range of the Tajima's D test. In contrast, we find that exponential population growth causes a reduction in the working range of Tajima's D test (Figs. 3 and 4), which is expected given that critical values for the test are obtained under the assumption of constant population size. It is usually impossible to know *a priori* if a sampled population meets this and other assumptions of Tajima's D test. For example, it is likely that many viral populations of interest will have experienced a complex form of population growth (during an epidemic or directly following transmission to a new host) or been subject to population structure. In our study, it was not feasible to perform simulations under all possible population histories. Therefore we chose two representative scenarios (constant-size and exponential growth) that give rise to very different phylogenies and which likely span the range of empirically observed tree shapes. A more general solution to this problem was developed by Edwards et al. (2006), who, for each data set, simulated a null distribution of D upon phylogenies whose shapes are highly

supported by the data. This approach will reduce the error rate of Tajima's D test, but at the cost of reduced computational efficiency.

We examined the error rate of the MK test using both the standard site counting method (McDonald and Kreitman, 1991a) and our new proportional counting approach (see Section 2). Using the former method, we observed raised type I errors when θ was larger than ~ 0.01 (Fig. 5a). In this region of parameter space most variable sites are 2, 3 or 4-state polymorphic (site types 4–7; Fig. 2) and the standard method classifies such sites as polymorphic. McDonald and Kreitman (1991b), in response to Whittam and Nei (1991) and Graur (1991), justified their implementation by arguing that any algorithm for typing substitutions as fixed or polymorphic will affect the numbers of replacement and silent substitutions equally and therefore not affect their test, which depends on ratios. This argument appears correct when the infinite sites assumption is valid. However, if $\theta\tau$ is large then nucleotide saturation occurs and the alignment contains more sites of types 4 and 6 (Fig. 2). Because these sites are, on average, more likely to be replacement sites than silent, the standard counting method (which always classes such sites as polymorphic) will tend to overestimate the number of replacement polymorphisms relative to silent polymorphisms. Simulation results show that our new proportional counting method is more robust, allowing the MK test to be applied even when $\theta\tau$ is very high (Fig. 5b) and reducing the average type I error (over all parameter space) from 20.4% to 5.3%.

The MK test has a lower error rate on sequences sampled from a growing population than on corresponding sequences sampled from a constant population (Fig. 5). This is because population growth results in shorter trees that accrue fewer mutations, leading to proportionally fewer 2, 3 and 4-state polymorphic sites (Fig. 2). Lastly, we note that the empirical RNA virus data sets are placed in a region of parameter space associated with high type I error if the standard counting method is used (Fig. 5a). However, the MK test has good statistical properties in this region if the new proportional counting method is used, indicating that this approach can be reliably applied to RNA virus genomes (Fig. 5b).

Our analyses focussed on the type I error of Tajima's D and the MK test and did not directly measure the probability of failing to reject the null hypothesis when it is false (type II error, or statistical power). Measuring the statistical power of neutrality tests is a formidable task, owing to the computational difficulties of simulating sequences under selection. Furthermore, the multiplicity of possible scenarios under which selection could occur (see Zhai et al., 2009) make it difficult to obtain results of general applicability. However, we applied Tajima's D and the MK test to a compilation of 96 RNA virus data sets (Table 1) and found that the MK test rejected the null hypothesis of neutrality more than three times as often as Tajima's D (61.4% and 17.7% of data sets, respectively). Since our simulation results demonstrate that the MK test (when used with the proportional counting method) has correct type I error, we conclude that the MK test has significantly greater statistical power than the Tajima's D test. We observed a positive correlation between the p -value of the MK test and the mean ingroup/outgroup pairwise genetic distance (results not shown). The failure of the MK test to reject neutrality for $\sim 40\%$ of the RNA virus data sets may be a consequence of the presence of low frequency, slightly deleterious mutations that have yet to be purged by purifying selection. Such mutations appear to be common in RNA virus populations (Pybus et al., 2007; Hughes and Hughes, 2007). Charlesworth and Eyre-Walker (2008) show that deleterious mutations can reduce the power of the MK test to detect adaptive evolution and propose the removal of low-frequency polymorphisms as a solution to this problem. The effect of more complex phenomena – such as epistasis and complementation within co-infected cells – on the power on the MK test remains to be investigated.

In summary, our results indicate that the MK test with proportional site counting is suitable for analysis on RNA virus data sets. This test is quick to compute and makes a minimal number of assumptions, making it potentially more useful for the analysis of very large-scale genomic data sets than dn/ds methods. However, unlike dn/ds methods, the MK test cannot be used to pinpoint specific sites under selection, although it can estimate the rate of adaptive substitution of a gene (Smith and Eyre-Walker, 2002). The MK test can also be applied to intra-species data sets that have been sampled serially through time (Williamson, 2003). When applied to intra-species data, the MK test will likely have low type I error (as $\theta\tau$ will be small) but could be statistically weak if the ingroup/outgroup genetic distance is low. In future work we aim to study the application of the MK test to serially sampled intra-species sequences.

Acknowledgements

We thank Eddie Holmes for generating the original collection of RNA virus data sets. SB is funded by NERC UK. OGP is supported by the Royal Society.

References

- Andolfatto, P., 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437, 1149–1152.
- Charlesworth, J., Eyre-Walker, A., 2008. The McDonald–Kreitman test and slightly deleterious mutations. *Molecular Biology and Evolution* 25, 1007–1015.
- Cooper, C.R., Hanson, L.A., Diehl, W.J., Pharr, G.T., Coats, K.S., 1999. Natural selection of the pol Gene of bovine immunodeficiency virus. *Virology* 255 (2), 294–301.
- Ebel, G.D., Spielman, A., Telford, S.R., 2001. Phylogeny of North American powassan virus. *Journal of General Virology* 82, 1657–1665.
- Edwards, C.T.T., Holmes, E.C., Pybus, O.G., Wilson, D.J., Viscidi, R.P., Abrams, E.J., Phillips, R.E., Drummond, A.J., 2006. Evolution of the human immunodeficiency virus envelope gene is dominated by purifying selection. *Genetics* 174, 1441–1453.
- Egea, R., Casillas, S., Barbadilla, A., 2008. Standard and generalized McDonald–Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Research* 36, 157–162.
- Eyre-Walker, A., 2006. The genomic rate of adaptive evolution. *Trends in Ecology & Evolution* 21, 569–575.
- Goldman, N., Yang, Z., 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* 11, 725–736.
- Graur, D., 1991. Neutral mutation hypothesis test. *Nature* 354, 114–115.
- Hudson, R.R., 1990. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* 7, 1–44.
- Hughes, A.L., Hughes, M.A.K., 2007. More effective purifying selection on RNA viruses than in DNA viruses. *Gene* 404, 117–125.
- Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. *Mammalian Protein Metabolism* 3, 21–132.
- Kryazhimskiy, S., Plotkin, J.B., 2008. The population genetics of dN/dS . *PLoS Genetics* 4, e1000304.
- Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., 2007. ClustalW2 and ClustalX version 2. *Bioinformatics* 23, 2947–2948.
- Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380.
- McDonald, J.H., Kreitman, M., 1991a. Adaptive protein evolution at the *adh* locus in *Drosophila*. *Nature* 351, 652–654.
- McDonald, J.H., Kreitman, M., 1991b. Neutral mutation hypothesis test. *Nature* 354, 116.
- Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* 3, 418–426.
- Nielsen, R., 2001. Statistical tests of selective neutrality in the age of genomics. *Heredity* 86, 641–647.
- Nielsen, R., 2005. Molecular signatures of natural selection. *Annual Review of Genetics* 39, 197–218.
- Nielsen, R., Yang, Z., 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148, 929–936.
- Pybus, O.G., Holmes, E.C., Harvey, P.H., 1999. The mid-depth method and HIV-1: a practical approach for testing hypotheses of viral epidemic history. *Molecular Biology and Evolution* 16, 953–959.
- Pybus, O.G., Rambaut, A., Belshaw, R., Freckleton, R.P., Drummond, A.J., Holmes, E.C., 2007. Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. *Molecular Biology and Evolution* 24, 845–852.
- Pybus, O.G., Shapiro, B., 2009. Natural selection and adaptation of molecular sequences, Chapter 6. In: Lemey, P., Salemi, M., Vandamme, A.M. (Eds.), *The Phylogenetics Handbook*, second ed. Cambridge University Press.
- Ramos-Onsins, S.E., Rozas, J., 2002. Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution* 19, 2092–2100.
- Ritchie, P.A., Anderson, I.L., Lambert, D.M., 2003. Evidence for specificity of psittacine beak and feather disease viruses among avian hosts. *Virology* 306, 109–115.
- Rodrigo, A.G., Felsenstein, J., 1999. Coalescent approaches to HIV population genetics. In: Crandall, K.A. (Ed.), *The Evolution of HIV*. pp. 233–272.
- Pond, S.L.K., Frost, W.S.D., 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* 21, 2531–2533.
- Shapiro, B., Rambaut, A., Pybus, O.G., Holmes, E.C., 2006. A phylogenetic method for detecting positive epistasis in gene sequences and its application to RNA virus evolution. *Molecular Biology and Evolution* 23, 1724–1730.
- Shriner, D., Rodrigo, A.G., Nickle, D.C., Mullins, J.L., 2004a. Pervasive genomic recombination of HIV-1 in vivo. *Genetics* 167, 1573–1583.
- Shriner, D., Shankarappa, R., Jensen, M.A., Nickle, D.C., Mittler, J.E., Margolick, J.B., Mullins, J.L., 2004b. Influence of random genetic drift on human immunodeficiency virus type 1 env evolution during chronic infection. *Genetics* 166, 1155–1164.
- Simonsen, K.L., Churchill, G.A., Aquadro, C.F., 1995. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* 141, 413–429.
- Slatkin, M., Hudson, R.R., 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129, 555–562.
- Smith, N.G., Eyre-Walker, A., 2002. Adaptive protein evolution in *Drosophila*. *Nature* 415, 1022–1024.
- Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595.
- Tsompana, M., Abad, J., Purugganan, M., Moyer, J.W., 2005. The molecular population genetics of the Tomato spotted wilt virus (TSWV) genome. *Molecular Ecology* 14, 53–66.
- Watterson, G.A., 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* 7, 256–276.
- Williamson, S., 2003. Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. *Molecular Biology and Evolution* 20, 1318–1325.
- Wilson, D., McVean, J.G., 2006. Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics* 72, 1411–1425.
- Whittam, T.S., Nei, M., 1991. Neutral mutation hypothesis test. *Nature* 354, 115–116.
- Yang, Z., Nielsen, R., Goldman, N., Pedersen, A.-M.K., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155, 431–449.
- Yang, Z., 2007. Paml 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24, 1586–1591.
- Zhai, W., Nielsen, R., Slatkin, M., 2009. An investigation of the statistical power of neutrality tests based on comparative and population genetic data. *Molecular Biology and Evolution* 26, 273–283.