Phylogenetic Evidence for Deleterious Mutation Load in RNA Viruses and Its Contribution to Viral Evolution

Oliver G. Pybus,* Andrew Rambaut,* Robert Belshaw,* Robert P. Freckleton,* Alexei J. Drummond,*† and Edward C. Holmes‡§

*Department of Zoology, University of Oxford, Oxford, United Kingdom; †Department of Computer Science, University of Auckland, Auckland, New Zealand; ‡Center for Infectious Disease Dynamics, Department of Biology, The Pennsylvania State University; and §Fogarty International Center, National Institutes of Health, Bethesda, Maryland

Populations of RNA viruses are often characterized by abundant genetic variation. However, the relative fitness of these mutations is largely unknown, although this information is central to our understanding of viral emergence, immune evasion, and drug resistance. Here we develop a phylogenetic method, based on the distribution of nonsynonymous and synonymous changes, to assess the relative fitness of polymorphisms in the structural genes of 143 RNA viruses. This reveals that a substantial proportion of the amino acid variation observed in natural populations of RNA viruses comprises transient deleterious mutations that are later purged by purifying selection, potentially limiting virus adaptability. We also demonstrate, for the first time, the existence of a relationship between amino acid variability and the phylogenetic distribution of polymorphisms. From this relationship, we propose an empirical threshold for the maximum viable deleterious mutation load in RNA viruses.

Introduction

RNA viruses represent an evolutionary system dominated by mutation (Domingo and Holland 1997). However, the fitness consequences of the majority of mutations are unknown, despite the importance of this information to understanding the intricate mechanisms of viral evolution and for developing effective methods for their control. To date, most studies of RNA virus fitness have used in vitro assays based on a small number of species; these have indicated that deleterious and slightly deleterious mutations are commonplace (Elena and Moya 1999; Sanjuan et al. 2004a; Montville et al. 2005) and that mutations can interact through positive epistasis (Bonhoeffer et al. 2004; Sanjuan et al. 2004b). Herein, we present a more comprehensive overview of the fitness of viral mutations in vivo and across genera, by analyzing patterns of genetic variation in the sequences of structural genes sampled from natural populations of 143 virus species and comprising a total data set of more than 4 million nucleotides.

Gene sequences sampled from natural populations contain 2 sources of information about the action of natural selection on mutations: 1) the ratio of nonsynonymous (d_N) to synonymous (d_S) variation per site, quantified as the ratio d_N/d_S or ω (e.g., Nei and Gojobori 1986; Yang et al. 2000), and 2) the frequency of ancestral and derived nucleotides at variable sites, quantified as a site-frequency spectrum (e.g., Bustamante et al. 2001). Recent approaches to estimating mutational fitness effects from gene sequences have used both sources of information to increase statistical power, namely variants of the Macdonald–Kreitman test (e.g., Smith and Eyre-Walker 2002; Williamson 2003) and methods that estimate Tajima's D statistic separately for synonymous and nonsynonymous polymorphisms (e.g., Hughes 2005). Unfortunately these approaches are not applicable to our data sets—which represent the species-level diversity of rapidly evolving viruses with small genomes—for several reasons: 1) the "infinite-sites" assumption (i.e., mutations

never occur at the same site) is almost always violated by such data, 2) complex nucleotide substitution models are required to accurately model RNA virus sequence evolution, and 3) the ancestral state of polymorphic sites cannot be reliably inferred.

To circumvent these problems, we employed a novel phylogenetic method that considers polymorphisms both in terms of their site frequency and their affect on amino acid sequence. Our method is based on the straightforward observation that the average age of nonsynonymous mutations increases with their selective advantage, even in the absence of recombination (Nielsen and Weinreich 1999). Thus, deleterious mutations will be young and more likely fall on the external branches of a population-level phylogeny (relative to neutral mutations), whereas advantageous mutations, if not already fixed, are likely to fall deeper in the genealogy (relative to neutral mutations). To quantify this effect, we first inferred gene genealogies for each virus species. We then estimated the ratio of nonsynonymous to synonymous substitution rates separately for internal and external tree branches (denoted ω_i and ω_e, respectively). Finally, for each data set, the phylogenetic age distribution of mutations was summarized graphically using the new "deviation from neutrality statistic" (DNS; see Materials and Methods for description). We also perform a comprehensive set of simulations to test the robustness of our approach. Note that the ω values are used here to characterize population polymorphism, not nucleotide fixation, and are interpreted accordingly throughout.

The approach taken here is related to previous studies in which ω was estimated separately for within- and among-species branches of molecular phylogenies. For example, analyses of primate mtDNA sequences have found that ω is significantly higher for within-species branches, indicating the presence of short-lived deleterious mutations (e.g., Hasegawa et al. 1998). This method was subsequently applied to the primate lentiviruses (such as the human virus HIV-1 and the chimpanzee virus SIVcpz), where it was noted that ω ratios were higher on branches nearer the tips of HIV-1/SIVcpz phylogenies (Sharp et al. 2001). Similarly, a higher ω was noted for within-host compared with among-host genetic variation in Dengue virus (Holmes 2003). Here,

we analyze 143 virus species and investigate whether these results are representative of RNA virus evolution in general. In addition, we attempt to provide a quantitative measure of the level of deleterious mutation in each species.

Materials and Methods
Sequence Collection

All sequences were collected manually from GenBank and compiled on the following criteria: 1) they represented structural genes (1 gene per virus species), 2) they contained at least 10 sequences of ≥ 500 bp in length, 3) each sequence was taken from a different individual, and 4) alignment gene diversity (θ) was greater than 0.01. Strains subject to excessive laboratory culture were, as far as practicable, also excluded. After collection, sequences were aligned using the ClustalX program (Thompson et al. 1997), with any adjustments in the alignments then made by eye using the Se-Al program (http://evolve.zoo.ox.ac.uk). Full alignment details are provided in the Supplementary Material online.

Phylogenetic Analysis

For each virus data set, a maximum likelihood phylogeny was estimated under the General Time Reversible $+ I + \Gamma_4$ substitution model, using PAUP* (Swofford 2003). Given these trees, the program CODEML (Yang et al. 2000) was used to estimate the maximum likelihood ω value for each data set. This overall ω was calculated using all branches in the tree (the 1-ratio model). Next, a separate ω value was estimated for the external (ω_e) and internal (ω_i) branches of each phylogeny using the lineage-specific codon substitution model in CODEML (the 2-ratio model). Estimated values are provided in the Supplementary Material online.

Deviation from Neutrality Plots

We introduce a new, intuitive method to display and quantify the phylogenetic age distribution of mutations within a species, called the "deviation from neutrality plot" (see fig. 1a). For each data set, we plot $\log(\omega_i)$ and $\log(\omega_e)$ versus $\log(\omega)$. If a data set contains only neutral mutations, then both points will be superimposed at the origin ($\omega = \omega_i = \omega_e = 1$). The origin therefore represents the occurrence of strict neutrality. If a data set contains lethal mutations in addition to neutral ones, then both points will be superimposed somewhere on the dashed line. The dashed line therefore represents general neutrality; that is, all observed polymorphisms are neutral but overall ω is <1 (i.e., lethal mutations are not observed in the sequences). In this case, the distance to the origin along the dashed line represents the ratio of lethal to neutral changes (fig. 1a).

In contrast, if a data set contains variable sites that are under any form of natural selection, then the 2 plotted points will fall on either side of the dashed line and will not be superimposed. If some polymorphic sites are deleterious or slightly deleterious, then there will be an excess of recent nonsynonymous changes and $\omega_e > \omega$, hence the external branch data point will fall above the dashed line (provided that overall ω is <1). If there is an unusually strong or recurrent positive selection, then there will be an excess of high-frequency nonsynonymous changes and $\omega_i > \omega$, thus the internal branch data

point will most likely fall above the dashed line (fig. 1a). The 2 data points that represent each data set are, of course, correlated and not statistically independent.

To quantify the deviation of each data set from the null hypothesis of general neutrality, we use the DNS. The DNS value equals the perpendicular distance between each point and the dashed line and is calculated as $\sin(45)*[\log(\omega_e) - \log(\omega)]$ for external branches and $\sin(45)*[\log(\omega_i) - \log(\omega)]$ for internal branches. DNS values are positive for points above the line and negative for those below (fig. 1a). The choice of horizontal, vertical, or perpendicular distances to the dashed line in calculating DNS is not important, as all 3 distances are linearly proportional and differ only by the constant $\pm\sin(45)$. In addition to being easy to interpret, the DNS statistic is expected to be invariant to the overall level of selective constraint, under the null hypothesis. The interpretation of DNS is poorly defined if $\omega > 1$; however, this situation does not occur for any of our virus data sets.

Results
Simulations

We performed 2 comprehensive sets of simulations to investigate how well DNS plots represent the relative fitness effects of mutations. First, 143 simulated data sets were generated using a standard codon substitution model (Yang et al. 2000) under the hypothesis of strict neutrality (i.e., ω was set to 1.0). Simulations were performed using the program EVOLVER (http://abacus.gene.ucl.ac.uk). To match our empirical data as closely as possible, we used the maximum likelihood trees estimated from the real virus alignments as the guide trees upon which the simulated alignments were generated. In addition, the sequence length and sample size of each simulated data set was set equal to those of the corresponding real virus alignment. All parameters of the simulated codon substitution model (except ω) were fixed at the empirical values estimated from the corresponding real virus alignment. The simulated data sets were then analyzed in the exactly same manner as the original data (see Materials and Methods). As expected, all of the points cluster tightly around the origin (see fig. 1b).

The second set of simulations was performed in the same manner as described above, except that the ω parameter of the codon substitution model was fixed to the empirical ω value estimated from the corresponding real virus alignment (as opposed to being fixed to 1). These simulations therefore represent data sets generated under the general neutrality hypothesis, with varying ratios of lethal to neutral mutations. Again the simulations behaved as predicted; as shown in figure 2a, all the points fall close to the dashed line. In addition, we calculated DNS values from these points. As expected in the absence of deleterious or advantageous mutation, DNS values for the simulated data sets are centered on zero and are almost identical for external and internal branches (see fig. 2b). The simulations therefore indicate that our empirical results are not adversely affected by estimation error or bias.

Empirical Data Sets

Figure 3a shows the DNS plot for our empirical alignments, representing 143 RNA virus species. The observed
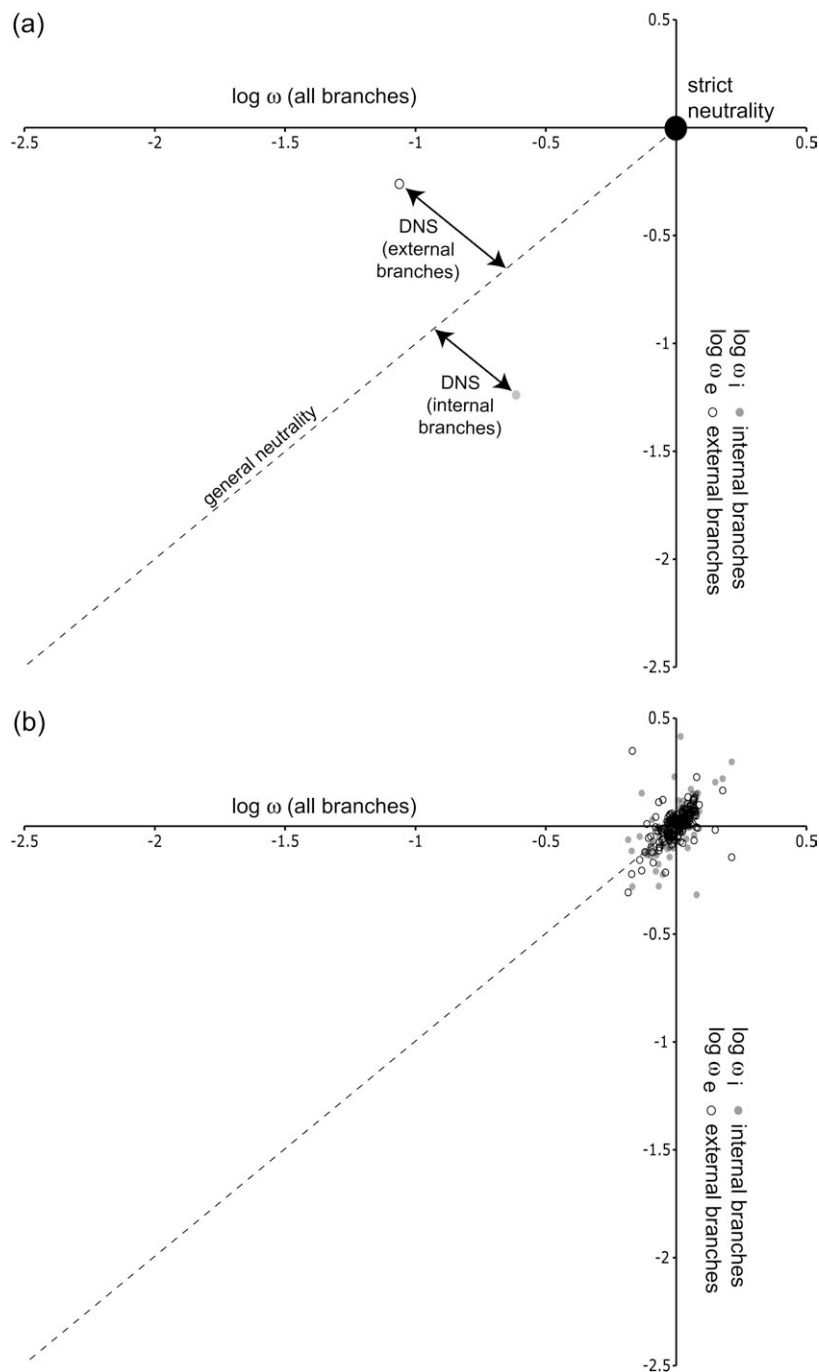
FIG. 1.—(a) The structure of a DNS plot. ω_i (filled circles) and ω_e (open circles) values are plotted against ω (horizontal axis). The dashed line shows the expected relationship under general neutrality (only neutral and lethal mutations occur). The origin corresponds to strict neutrality (neutral mutations only). DNS values are calculated as perpendicular distances to the dashed line. (b) The DNS plot for data sets simulated under strict neutrality.

ω, ω_i, and ω_e values are all strongly correlated ($P < 0.01$), reflecting the natural variation in selective constraint among genes and species. The corresponding DNS values (fig. 3b) are considerably different from those obtained under general neutrality (fig. 2b). Most data sets (88%) have DNS > 0 for external branches. Correspondingly, few data sets (12%) have DNS > 0 for internal branches. The mean DNS values for external and internal branches were 0.14 and -0.16, respectively (fig. 3b). The symmetry between the DNS results for internal and external branches is expected, as the pairs of values for each data set are not independent.

These results reveal a significant excess of low-frequency nonsynonymous polymorphisms on external phylogeny branches; deleterious or slightly deleterious non-synonymous mutations are therefore a major component of the segregating genetic variation in RNA viruses, although the magnitude of this component varies among species. Some synonymous polymorphisms may be deleterious,
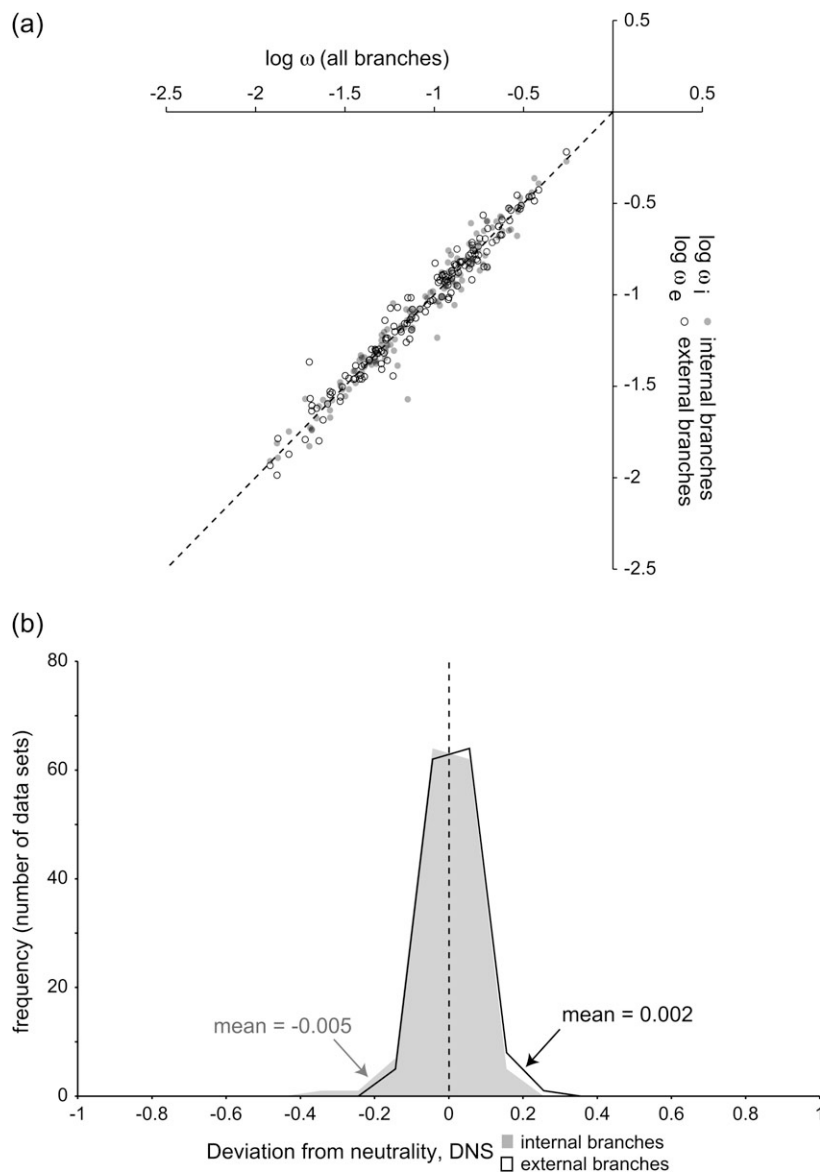
Fig. 2.—(a) The DNS plot for data sets simulated under the general neutrality model. ω values for internal branches (ω_i; filled circles) and for external branches (ω_e; open circles) are plotted against ω (horizontal axis). See text for further details. (b) The frequency distribution of DNS values for the simulated data sets shown in part (a). The dashed vertical line in (b) denotes DNS=0. The shaded distribution represents internal branch DNS values and the black line outlines the distribution of external branch DNS values.

due to such factors as RNA secondary structure (Simmonds et al. 2004). However, the possible action of selection on synonymous sites does not invalidate our approach as ω ratios are best interpreted as describing the difference in selective pressure between synonymous and nonsynonymous sites (see, e.g., Yang 2006). It is also clear that the rate of recombination or reassortment, which varies greatly among RNA viruses, does not have a major bearing on DNS, as largely clonal RNA viruses (such as negative-sense RNA viruses) and those with high rates of recombination (such as retroviruses) commonly both have DNS > 0 for external branches. We note that even though DNS values are mostly above zero, the absolute ω values for external branches are typically low; $\omega_e < 0.25$ in 78% data sets examined, indicating that purifying selection is in general the most powerful evolutionary force at any phylogenetic scale.

It is also notable that some virus data sets have a DNS value close to zero, which might reflect neutral evolution or, more likely, a combination of advantageous and deleterious mutation. The opposing effects of advantageous and deleterious mutations on DNS therefore "cancel out," reducing the statistical power of the DNS to detect deleterious polymorphisms. Our results are therefore conservative, and disadvantageous mutations are likely to be even more abundant in RNA viruses than the DNS values suggest.

The large size of our data set enables us to test the hypothesis that there is an inverse relationship between the ratio ω_e/ω_i and the level of nonsynonymous variation, which was first proposed by Hasegawa et al. (1998) on the basis of 11 data sets of primate mtDNA. We therefore plotted the ratio ω_e/ω_i against L_N for each virus, where L_N is the total nonsynonymous branch length of each virus
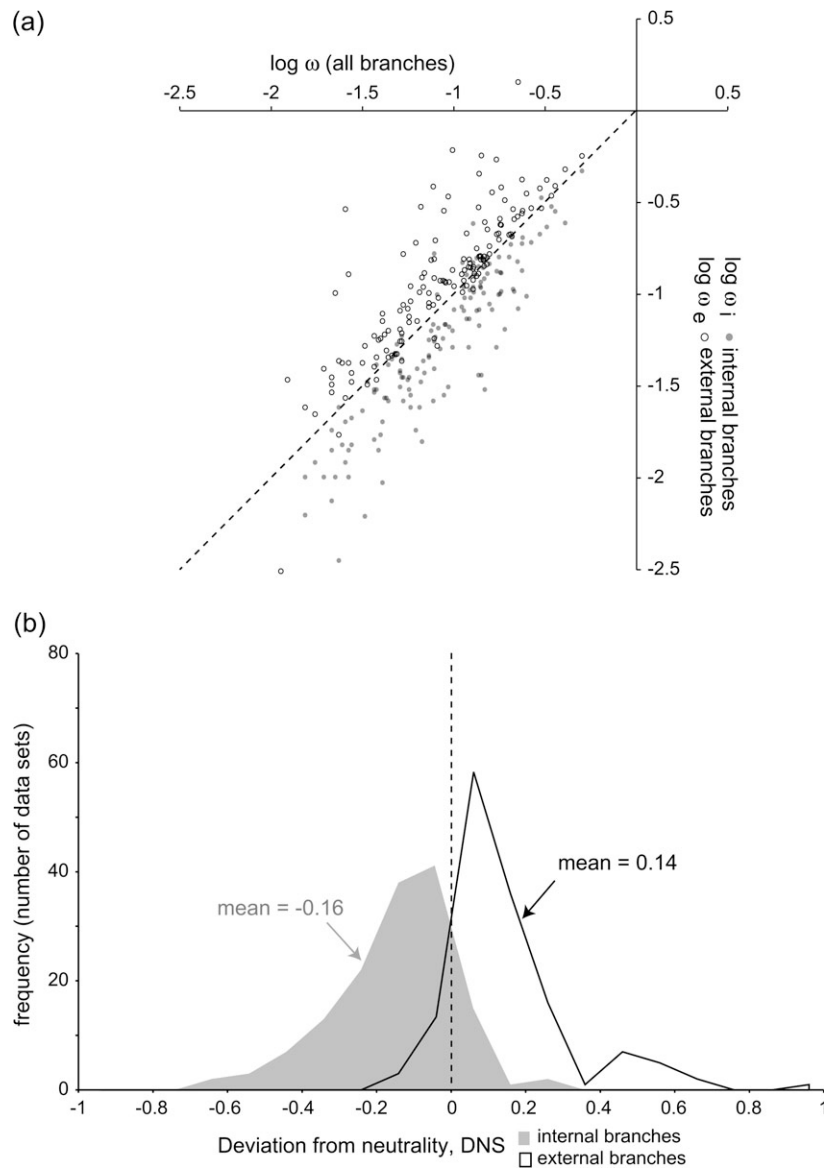
FIG. 3.—(a) The DNS plot for the empirical data sets, representing 143 RNA virus species. ω values for internal branches (ω_i; filled circles) and for external branches (ω_e; open circles) are plotted against ω (horizontal axis). See text for further details. (b) The frequency distribution of DNS values for the empirical data sets shown in part (a). The dashed vertical line in (b) denotes DNS=0. The shaded distribution represents internal branch DNS values and the black line outlines the distribution of external branch DNS values.

phylogeny estimated under the 2-ratio model (fig. 4a; this plot is exactly equivalent to Figure 2 of Hasegawa et al. [1998]). Our larger data set also demonstrates a nonlinear decline in ω_e/ω_i with increasing L_N (the nonlinear model fitted better than a linear model by 6.6 Akaike Information Criterion units; Burnham and Anderson 2002). Very similar results were obtained if DNS is used instead of ω_e/ω_i (which is unsurprising because DNS and $\log(\omega_e/\omega_i)$ are strongly correlated; $P < 0.001$). Figure 4b shows exactly the same plot for our simulated data sets (simulations performed under the general neutrality model; see above). As expected, the ratio ω_e/ω_i is approximately 1.0, and the variation in L_N among data sets matches that of the empirical sequences (fig. 4b). However, there is no inverse relationship, indicating that the pattern shown in figure 4a is not the result of estimation bias or phylogenetic tree shape.

Hasegawa et al. (1998) suggested that the inverse trend between ω_e/ω_i and the level of nonsynonymous polymorphism is caused by variation in selective constraint; mutations in constrained genes (which have little nonsynonymous variation) are more likely to be deleterious and therefore younger, leading to an increased ω_e/ω_i value, whereas mutations in less-constrained genes are more likely to be neutral, resulting in a greater level of nonsynonymous diversity and a lower ω_e/ω_i value. In our data sets, adaptive evolution will also contribute to this inverse relationship, such that a proportion of the nonsynonymous variation on internal branches will be positively selected, rather than neutral.

The inverse nature of the relationship between ω_e/ω_i and L_N suggests a simple evolutionary threshold for RNA viruses: natural virus populations are not able to exist or persist if $L_N(\omega_e/\omega_i) > x$, where x is approximately 2.5
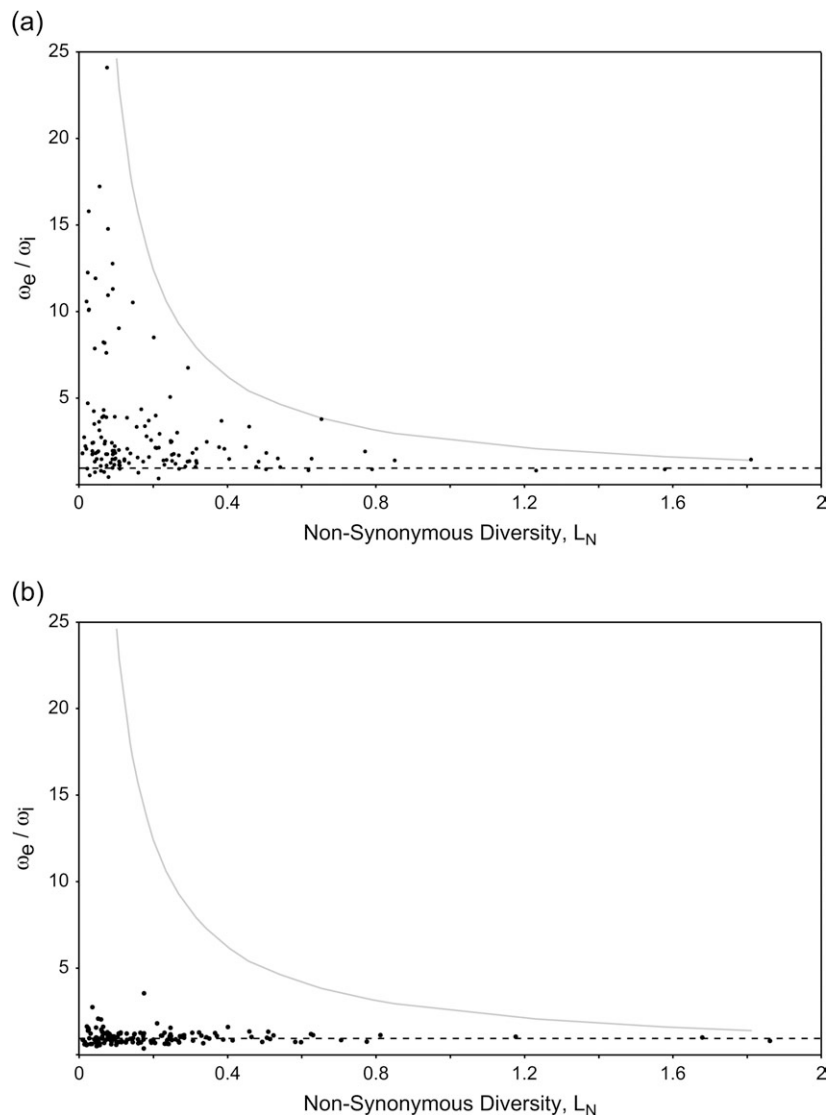
(a)

(b)

FIG. 4.—Plots of the ratio ω_e/ω_i against non-synonymous diversity, L_N, for each virus. L_N is the total non-synonymous branch length of each virus phylogeny. Part (*a*) shows our empirical data sets, representing 143 RNA virus species. Part (*b*) shows the corresponding simulated data sets, generated under the general neutrality model. The dashed line represents $\omega_e/\omega_i = 1$. The gray curve denotes $L_N(\omega_e/\omega_i) = 2.5$. The value 2.5 was chosen as the smallest value x for which all empirical data set satisfy the threshold condition $L_N(\omega_e/\omega_i) < x$.

(curves in fig. 4). This threshold states that populations can have either a large amount of nonsynonymous variation or an excess of nonsynonymous variation on external branches, but not both. If the pattern in figure 4*a* is the result of segregating deleterious mutations, then the threshold could be interpreted as the largest viable deleterious mutation load for an RNA virus population. An interesting question for future research would be to investigate if the parameter *x* varies among different types of organism; sufficient data may be available to estimate *x* for animal mtDNA (Hasegawa et al. 1998).

Discussion

An abundance of deleterious polymorphisms has important consequences for viral evolution. A high mutational load reduces the rate of adaptive evolution because deleterious changes prevent the fixation of linked sporadic advantageous mutations (Peck 1994). Thus, the high mutation rates experienced by RNA viruses do not automatically translate into enhanced adaptability (Elena and Sanjuán 2005). This is particularly important for viral emergence as the adaptation of viruses to new host species, such as avian influenza virus in humans, may be slowed or prevented by their frequent production of deleterious mutations. In this context, it is striking that despite being common agents of disease, most emergent RNA viruses in humans generate dead-end infections with no onward transmission, revealing that there are major constraints on their ability to colonize new hosts despite their high mutation rates (Webby et al. 2004).

As discussed earlier, the presence of advantageous mutations most likely causes our methods to underestimate the abundance of deleterious changes, increasing the number

of virus species with DNS values close to zero. HIV-1 is one such data set ($\omega_e = 0.29$, $\omega_i = 0.33$, DNS = -0.007 for external branches). Although a high ω_i is to be expected for HIV-1, given its rapid rate of adaptation (Williamson 2003), this observation appears to conflict with a prior analysis that reported higher ω values for branches closer to the tips of the phylogeny (Sharp et al. 2001). However, the internal branches in the 2 analyses are not comparable as the phylogeny of Sharp et al. included both HIV-1 and SIVcpz, whereas our data set only includes HIV-1. We also note that adaptation to cell-culturing conditions has previously been shown to increase ω_e in the hemagglutinin gene of influenza A virus (Bush et al. 1999). However, selection for culture adaptation will be most pronounced on external glycoproteins that interact directly with host cell receptors, whereas our analysis deliberately concentrated as far as possible on the structural (and often internal) genes of RNA viruses. More generally, low-quality sequences may remain in large-scale comparative analyses such as ours despite efforts to identify and remove them. Comparative data sets are typically robust to sources of random error due to their large size; however, their conclusions should still be considered with care to ensure they are robust. Here, we note 2 further factors that make it unlikely that our conclusion (the existence of widespread deleterious polymorphism in RNA virus populations) is affected by cell culture adaptation: 1) our comparative analysis agrees with previous experimental and quantitative studies that have examined particular virus species in detail (e.g., Sanjuan et al. 2004a; Edwards et al. 2006) and 2) the inverse relationship shown in figure 4*a* between overall nonsynonymous diversity and ω_e/ω_i can be explained in terms of deleterious mutation load, but would not be expected if cell culture adaptations were a significant presence in our data.

Further, we have identified a clearly defined viability threshold for RNA viruses based on deleterious mutation load; viruses closest to this threshold would theoretically make the best candidates for antiviral lethal mutagenesis therapy.

Future work may help to uncover the factors that determine variation in DNS among RNA viruses. The segregating deleterious mutation load of a species is expected to be primarily determined by the genomic mutation rate and the probability that mutations are deleterious (i.e., the distribution of selection coefficients). The load may be reduced by synergistic epistasis in recombining populations or by back mutation. The per-base per-generation mutation rate is exceptionally high for RNA viruses (Drake and Holland 1999) and is likely to be the primary explanation for the abundance of deleterious polymorphisms we observed. Likewise, the selection coefficients of viral mutations will depend on a multitude of factors, including the nature of the host immune response and the genome structure and life history strategy of each virus. For example, vector-borne RNA viruses, which must replicate in 2 host species, are less likely to experience adaptive evolution than viruses transmitted by other mechanisms (Woelk and Holmes 2002). Recent theoretical work has considered the tolerance or "robustness" of genomes to deleterious mutation (e.g., Krakauer and Plotkin 2002; Wilke and Adami 2003; Monteville et al. 2005). The theoretical basis of robustness

is essentially the same as that of selective constraint; both concepts relate to the topology of the adaptive landscape of a species. However, the small and highly compressed character of RNA virus genomes means that there is little opportunity for (or perhaps little selection for) genomic redundancy, so robustness is expected to be generally low. An alternative explanation is that if RNA virus population sizes are high then the strength of selection for redundancy is reduced, because the genetic load at high population sizes is less dependent on the magnitude of selection coefficients (Krakauer and Plotkin 2002). It is also possible that positive epistasis among mutations, which has been shown to be widespread among RNA viruses (Bonhoeffer et al. 2004; Burch and Chao 2004; Shapiro et al. 2006), helps to reduce the fitness consequences of a high deleterious mutation load (Sanjuan et al. 2004b). Furthermore, within RNA virus infected hosts, infected cells often produce a vast number of offspring virions, so "soft selection" may lessen the effect of deleterious mutation on parental fitness. However, deleterious mutations are likely to become fixed in the within-host virus population during the bottleneck caused by transmission. When viruses from different hosts are compared, these fixations will be apparent as segregating deleterious changes. Therefore, the joint effects of population size and population structure on mutation load in viruses (and other parasite species) are expected to be complex and are important areas for future research.

Supplementary Material

Full alignment details and parameter values are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/). The sequence alignments used in this paper can be obtained from http://evolve.zoo.ox.ac.uk/data.

Acknowledgments

Literature Cited

Bonhoeffer S, Chappey C, Parkin NT, Whitcomb JM, Petropoulos CJ. 2004. Evidence for positive epistasis in HIV-1. Science. 306:1547–1550.

Burch CL, Chao L. 2004. Epistasis and its relationship to canalization in the RNA virus phi 6. Genetics. 167:559–567.

Burnham KP, Anderson DR. 2002. Model selection and multimodel inference. New York: Springer.

Bush RM, Fitch WM, Bender CA, Cox NJ. 1999. Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol Biol Evol. 16:1457–1465.

Bustamante CD, Wakeley J, Sawyer S, Hartl DL. 2001. Directional selection and the site-frequency spectrum. Genetics. 159:1779–1788.

Domingo E, Holland JJ. 1997. RNA virus mutations for fitness and survival. Annu Rev Microbiol. 51:151–178.

Drake JW, Holland JJ. 1999. Mutation rates among RNA viruses. Proc Natl Acad Sci USA. 96:13910–13913.

Edwards CT, Holmes EC, Pybus OG, Wilson DJ, Viscidi RP, Abrams EJ, Phillips RE, Drummond AJ. 2006. Evolution of

the human immunodeficiency virus envelope gene is dominated by purifying selection. Genetics. 174:1441–1453.

Elena SF, Moya A. 1999. Rate of deleterious mutation and the distribution of its effects on fitness in vesicular stomatitis virus. J Evol Biol. 12:1078–1088.

Elena SF, Sanjuán R. 2005. Adaptive value of high mutation rates of RNA viruses: separating causes from consequences. J Virol. 79:11555–11558.

Hasegawa M, Cao Y, Yang Z. 1998. Preponderance of slightly deleterious polymorphism in mitochondrial DNA: replacement/synonymous rate ratio is much higher within species than between species. Mol Biol Evol. 15:1499–1505.

Holmes EC. 2003. Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. J Virol. 77:11296–11298.

Hughes AL. 2005. Evidence for abundant slightly deleterious polymorphisms in bacterial populations. Genetics. 169:533–538.

Krakauer DC, Plotkin JB. 2002. Redundancy, antiredundancy, and the robustness of genomes. Proc Natl Acad Sci USA. 99:1405–1409.

Montville R, Froissart R, Remold SK, Tenaillon O, Turner PE. 2005. Evolution of mutational robustness in an RNA virus. PLoS Biol. 3:e381.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.

Nielsen R, Weinreich DM. 1999. The age of nonsynonymous and synonymous mutations in animal mtDNA and implications for the mildly deleterious theory. Genetics. 153:497–506.

Peck JR. 1994. A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex. Genetics. 137:597–606.

Sanjuan R, Moya A, Elena SF. 2004a. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc Natl Acad Sci USA. 101:8396–8401.

Sanjuan R, Moya A, Elena SF. 2004b. The contribution of epistasis to the architecture of fitness in an RNA virus. Proc Natl Acad Sci USA. 101:15736–15739.

Shapiro B, Rambaut A, Pybus OG, Holmes EC. 2006. A phylogenetic method for detecting positive epistasis in gene sequences and its application to RNA virus evolution. Mol Biol Evol. 23:1724–1730.

Sharp PM, Bailes E, Chaudhuri RR, Rodenburg CM, Santiago MO, Hahn BH. 2001. The origins of acquired immune deficiency syndrome viruses: where and when? Philos Trans R Soc Lond B. 356:867–876.

Simmonds P, Tuplin A, Evans DJ. 2004. Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: implications for virus evolution and host persistence. RNA. 10:1337–1351.

Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. Nature. 415:1022–1024.

Swofford DL. 2003. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25:4882.

Webby R, Hoffman E, Webster R. 2004. Molecular constraints to interspecies transmission of viral pathogens. Nat Med. 10:S77–S81.

Wilke CO, Adami C. 2003. Evolution of mutational robustness. Mut Res. 522:3–11.

Williamson S. 2003. Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. Mol Biol Evol. 20:1318–1325.

Woelk CH, Holmes EC. 2002. Reduced positive selection in vector-borne RNA viruses. Mol Biol Evol. 19:2333–2336.

Yang Z. 2006. Computational molecular evolution. Oxford: Oxford University Press.

Yang Z, Nielsen R, Goldman N, Pedersen AMK. 2000. Codon substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 155:431.

William Martin, Associate Editor