Published in final edited form as:
Virology. 2009 October 10; 393(1): 127–134. doi:10.1016/j.virol.2009.07.016.

Small Effective Population Sizes and Rare Nonsynonymous Variants in Potyviruses

Austin L. Hughes
Department of Biological Sciences, University of South Carolina, Columbia SC 29205

Abstract

Analysis of nucleotide sequence polymorphism in complete genomes of 12 species of potyviruses (single-stranded, positive-sense RNA viruses, family *Potyviridae*) revealed evidence that long-term effective population sizes of these viruses are on the order of 10^4. Comparison of nucleotide diversity in non-coding regions and at synonymous and nonsynonymous sites in coding regions showed that purifying selection has acted to eliminate numerous deleterious mutations both at nonsynonymous sites and in non-coding regions. The ratio of nonsynonymous to synonymous polymorphic sites increased as a function of the number of genomes sampled, whereas mean gene diversity at nonsynonymous polymorphic sites decreased with increasing sample size at a substantially faster rate than does mean gene diversity at synonymous polymorphic sites. Very similar relationships were observed both in available genomic sequences of 12 potyvirus species and in subsets created by randomly sampling from among 98 TuMV genomes. Taken together, these observations imply that a greater proportion of nonsynonymous than of synonymous variants are relatively rare as the result of ongoing purifying selection, and thus many nonsynonymous variants are unlikely to be discovered without extensive sampling.

Keywords

effective population size; nonsynonymous substitution; potyvirus; purifying selection

RNA viruses include some of the most important pathogens of humans, domestic animals, and crop plants. Because of their high mutation rate (Drake and Holland 1999), RNA viruses have a great potential for sequence diversity; and numerous theoretical and empirical studies have been devoted to understanding the population processes affecting the fate of new mutations in these viruses (Domingo and Holland 1997; Hughes and Hughes 2007; Jenkins et al. 2001; Manrubia et al. 2005). Understanding RNA virus sequence diversity in turn has numerous practical applications, ranging from the development of vaccines against viruses infecting humans or other vertebrates (Barouch 2008) to understanding the basis of host specificity in plant viruses (Tan et al. 2005).

In the case of plant RNA viruses, García-Arenal et al. (2001) reviewed evidence suggesting that a high level of mutation has not led to a great deal of variability in viral proteins or a rapid

Austin L. Hughes, Ph.D. Department of Biological Sciences University of South Carolina Coker Life Sciences Bldg. 700 Sumter St. Columbia SC 29208 USA Tel: 1-803-777-9186 Fax: 1-803-777-4002 austin@biol.sc.edu.

rate of genomic change. In fact, plant RNA viruses show remarkable stability at the genomic level, and the extent of amino acid sequence variation in their proteins is reported to be no greater than in the proteins of their hosts (García-Arenal et al. 2001). One explanation for this observation is that plant RNA virus genomes are subject to strong and effective purifying selection (García-Arenal et al. 2001) — that is, natural selection acting to eliminate deleterious mutations — an explanation consistent with results from studies of RNA viruses in general (Domingo and Holland 1997; Hughes and Hughes 2007). In addition, it has been argued that the comparative lack of diversity of certain plant viruses can be explained by low effective population sizes, because only a few virions may be involved in initiating infection, as for example has been demonstrated in wheat streak mosaic virus (French and Stegner 2003).

For a given mutation rate, the level of selectively neutral polymorphism observed in a population is expected to be an increasing function of effective population size (Nei 1987). By contrast, deleterious variants are expected to accumulate when effective population size is low, because purifying selection becomes increasingly ineffective in removing slightly deleterious variants as effective population size decreases (Ohta 1973). The rapid accumulation of deleterious mutations in experimentally bottlenecked RNA viruses provides a dramatic confirmation of the latter prediction (Chao 1990; de la Igelsia and Elena 2007: Duarte et al. 1992). On the other hand, when the effective population size is large, purifying selection is expected to keep slightly deleterious variants at low frequencies. The frequency of recombination is an additional factor in determining the effectiveness of purifying selection in purging deleterious variants. In the absence of recombination, deleterious mutations fixed during a bottleneck are difficult to remove even after population expansion, whereas recombination allows deleterious mutations to be isolated and thus purged by selection in an expanding population (Hughes and Hughes 2007).

Statistical evidence for the action of purifying selection derives from the comparison of synonymous and nonsynonymous (amino acid-altering) nucleotide substitutions in coding regions. Such comparisons are based on the assumption that synonymous mutations are much less likely to deleterious than nonsynonymous mutations, since many of the latter will disrupt protein structure and function (Hughes 1999). The observation that, in most protein-coding genes, the number of synonymous nucleotide substitutions per synonymous site (d_S) exceeds the number of nonsynonymous substitutions per nonsynonymous site (d_N), provides strong evidence that past purifying selection has acted to eliminate nonsynonymous mutations to a much greater extent than synonymous mutations (Nei 1987). Moreover, if purifying selection against slightly deleterious nonsynonymous mutations is ongoing within a population, it will result in a pattern whereby nonsynoymous variants tend to be rare relative to synonymous variants in the same coding regions (Hughes et al. 2003; Hughes 2005). Such patterns have been observed in sequences of a wide variety of both DNA and RNA viruses (Hughes 2007, 2009; Hughes and Hughes 2007; Hughes and Piontkivska 2008; Irausquin and Hughes 2008).

Here I use statistical analyses of genomic sequence data to examine the factors responsible for the maintenance of polymorphism in *Potyviridae* (potyviruses), a large and diverse family of plant RNA viruses (Shukla et al. 1994). I estimate effective population size from nucleotide diversity; and I test the hypothesis that purifying selection is a major factor acting to reduce diversity at the amino acid sequence level in these viruses (García-Arenal et al. 2001). On the hypothesis that purifying section is ongoing, a substantial fraction of nonsynonymous variants are predicted to be rare (Hughes 2008). As a consequence, the number of polymorphic nonsynonymous sites relative to synonymous sites will be an increasing function of the number of sequences sampled, because further sampling will reveal new rare nonsynonymous variants. Conversely, the mean gene diversity ("heterozygosity") at nonsynonymous single-nucleotide polymorphic (SNP) sites is predicted to be a decreasing function of the number of sequences sampled, because the inclusion of an increasing number of rare variants will lower the average

gene diversity. I test these predictions by examining patterns of synonymous and nonsynonymous sequence polymorphism in 12 potyvirus species, all pathogens of major agricultural plant species.

The single-stranded, positive-sense RNA genome of potyviruses encodes a single polyprotein of more than 3000 amino acids in length, which is later enzymatically cleaved into nine distinct protein products (Shukla et al. 1994). The long coding region is particularly appropriate for the type of analysis conducted here because stochastic error in estimates of population parameters is minimized. I also compare the pattern of nucleotide substitution in the non-coding regions, located 5′ and 3′ to the polyprotein gene, with that in the coding region.

Methods

Sequences and Phylogenetic Analysis

The analyses reported here involved 355 complete genome sequences of Potyviridae belonging to 62 viral species (Supplementary Table S1), which were aligned using the CLUSTAL X program (Thompson et al. 1997). The polyprotein-encoding sequences were aligned at the amino acid level and the alignment imposed on the DNA sequences. A phylogenetic tree of the 62 virus species was constructed by the neighbor-joining method (Saitou and Nei 1987) on the basis of the JTT model (Jones et al. 1992) with the assumption that rate variation among sites followed a gamma distribution. In this phylogenetic analysis, a single representative sequence was used for each species. The shape parameter of the gamma distribution was estimated by the TREE-PUZZLE program (Schmidt et al. 2004). Confidence in branching patterns in the phylogenetic tree was assessed by bootstrapping (Felsenstein 1985); 1000 bootstrap samples were used.

Analysis of Polymorphism

Sequence polymorphism was analyzed within 12 viral species, which were chosen because at least four complete genome sequences were available. Sequences derived from passaging experiments were excluded from these analyses (Tan et al. 2005; Wallis et al. 2007). For these 12 viral species, the mean transition:transversion ratio at third positions in the coding region was 6.2, indicating a strong transitional bias. The number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous (amino acid-altering) substitutions per nonsynonymous site (d_N) were estimated by Li's (1993) method. This method was used because it takes into account the effect of transitional bias, which is particularly important in the case of two-fold degenerate sites (Li 1993). The synonymous nucleotide diversity (symbolized π_S) is defined as the mean of d_S for all pairwise comparisons among a set of sequences, while the nonsynonymous nucleotide diversity (symbolized π_N) is the mean of d_N for all pairwise comparisons among a set of sequences. The maximum composite likelihood method (MCL; Tamura et al. 2007), which also takes into account transitional bias, was used to estimate the number of nucleotide substitutions per site (d) in non-coding regions; the mean of all pairwise comparisons of d is the nucleotide diversity (π). Standard errors of π_S, π_N, and π were estimated by the bootstrap method (Tamura et al. 2007); 1000 bootstrap samples were used.

In each of the 12 virus species, gene diversity ("heterozygosity") was estimated at each polymorphic site by the formula:

$$1 - \sum_{i=1}^{n} x_i^2$$

where n is the number of alleles and x_i is the frequency of the i^{th} allele in the set of sample sequences analyzed (Nei 1987, p. 177). In coding regions, single nucleotide polymorphisms (SNPs) were classified either as synonymous or nonsynonymous depending on their effect of the encoded amino acid sequence. Ambiguous sites were excluded from these analyses. The latter included sites at which both synonymous and nonsynonymous variants occurred in the set of sequences analyzed. Also excluded were certain polymorphic sites within codons with two or more polymorphic sites, when the polymorphism could be considered synonymous or nonsynoymous depending on the pathway taken by evolution. For example, consider the two codons CTA and TTT. A mutation C→T in the first position would be synonymous if there were A in the third position, but not if there were T in the third position.

The mean of gene diversities at synonymous SNP sites was designated H_S, and the mean of the gene diversities at nonsynonymous SNP sites was designated H_N. Gene diversities were not normally distributed. Therefore, in testing for differences in the mean and variance of gene diversity among categories (synonymous, nonsynonymous, and non-coding) of SNP sites, randomization tests were used. In each test, 1000 pseudo-data sets were created by sampling (with replacement) from the data; a difference between two categories was considered significant at the α level if it was greater than the absolute value of 100(1-α) % of the differences observed between the same categories in the pseudo-data sets.

The effective number of genes in the population (Ng) was estimated by the formula

$$\mathrm{Ng}=\pi_s/2u$$

where u is the mutation rate per site per generation (Lynch 2007, p. 91). Ng is equivalent to the long-term effective population size in the case of a haploid organism (Lynch 2007). The effective population size, a fundamental parameter of population genetics, corresponds to the size of an idealized population having the same properties with respect to genetic drift as a given real population (Wright 1931). In general, the effective population size is smaller than the census number of the population because of factors such as periodic bottlenecks (Nei 1987, p. 362-363). This concept is readily applicable to any population of replicating organisms, including RNA viruses (Leigh Brown 1997; Miralles et al. 2000; Pybus et al. 2001). Estimates of u were based on the estimate of the number of mutations per generation per genome (0.11) estimated for the tobacco mosaic virus, another single-stranded RNA positive-sense virus (Malpica et al. 2002).

Randomly Sampled Subsets

In order to test for the effect on population parameters of the number of genomes sampled, subsets of the data were constructed by randomly sampling (without replacement) sequences from 98 sequences of turnip mosaic virus (TuMV). Five random subsets were created for each of the following numbers of sequences: 4, 8, 16, 32, and 64. Population parameters were then estimated for each subset.

Exponential Regression

The relationship between H_S and H_N was investigated by the exponential (allometric) regression method (Sokal and Rohlf 1981). This method involves applying linear regression to the log-transformed variables, then re-expressing the resulting regression equation in exponential form. The same method was applied to examine the relationship between the ratio of the number of nonsynonymous SNPs to the number of synonymous SNPs (*N:S*) and sample size.

Results

Nucleotide Sequence Diversity

Patterns of nucleotide substitution were analyzed in 12 species of potyvirus for which four or more genome sequences were available. These 12 species were scattered throughout a phylogenetic tree of potyviruses, based on complete polyprotein sequences (Figure 1). Overall there was not a close relationship between the phylogeny of potyviruses and that of their host plants. For example, bean yellow mosaic virus and bean common mosaic virus clustered very far apart in the tree (Figure 1). The phylogeny thus supported a history involving numerous host transfer events.

In the 12 virus species for which polymorphism was analyzed, π_S was significantly greater than π_N in every case (Table 1). The same pattern was seen when π_S and π_N were estimated separately for each of the nine proteins making up the polyprotein (data not shown). Thus, there was evidence that potyvirus polyproteins have been subjected to strong past purifying selection. Likewise, in all of the viruses, π_S was significantly greater than π in non-coding regions (Table 1). On the other hand, in 11 of the 12 viruses (all except LMV), π in non-coding regions was significantly greater than π_N in the coding region (Table 1). Thus, there was evidence that non-coding regions of potyvirus genomes have been subject to purifying selection, but not as strongly as nonsynonymous sites in the polyprotein gene. On the basis of π_S, the long-term effective population size (Ng) was estimated to be on the order of 10^4 in the case of each of the 12 viruses (Table 1).

The mean gene diversity at synonymous polymorphic nucleotide (SNP) sites was significantly greater than at nonsynonymous SNP sites in 11 of the 12 viruses. This pattern is indicative of ongoing purifying selection, acting to reduce the population frequency of certain nonsynonymous variants. Similarly, in 4 of the viruses, the mean gene diversity at noncoding SNP sites was significantly lower than that at synonymous SNP sites, indicative of ongoing purifying selection on non-coding sites in these viruses.

Effects of Sample Size

For the 12 viruses, there was not a significant correlation between the number of genomes sampled and either π_S or π_N (Figure 2A). The values of π_S showed considerable spread in the case of viruses represented by small numbers of sequences (less than about 20), which seems likely to reflect stochastic error due to small sample size. The fact that this stochastic error is greater in the case of π_S than in the case of π_N apparently reflects the greater error in the former, due to the smaller number of synonymous sites. This interpretation was strongly supported by random sampling of TuMV sequences, which also showed higher variation of smaller samples, particularly in the case of π_S (Figure 2B). Even with small sample sizes, the values of π_S were clustered around the true mean value for all 98 TuMV sequences (0.5463; Table 1).

I fitted an exponential relationships between mean gene diversity at SNP sites and the number of sequences. In the case of synonymous SNP sites (H_S), the relationship was not statistically significant; the equation relating gene diversity at synonymous SNP sites (H_S) to the number of sequences (n) was the following: $H_S = 0.428\ n^{-0.119}$ ($R^2 = 0.236$; n.s.). The mean diversity at nonsynonymous SNP sites (H_N) was related to n by the following equation: $H_N = 0.624\ n^{-0.355}$ ($R^2 = 0.871$; $P < 0.001$). The exponent was negative, indicating a decreasing relationship, and significantly different from −1 ($P < 0.001$), indicating a decrease at a less than linear rate. In the case of nonsynonymous SNPs, 87.1% of the variance among the 12 viruses with respect to H_N was explainable by differences in sample size.

The importance of sample size in explaining the relationship between mean gene diversity at SNP sites and sample size was confirmed by random sampling of TuMV sequences (Figure

3B). In the case of the randomly sampled subsets of the TuMV sequences, there was a strong relationship between H_S and sample size: $H_S = 0.474\ n^{-0.109}$ ($R^2 = 0.828$; $P < 0.001$). The relationship between H_N and sample size was even stronger: $H_N = 0.676\ n^{-0.350}$ ($R^2 = 0.979$; $P < 0.001$). In the randomly sampled subsets, the exponents were negative for both H_S and H_N; and in each case the exponent was significantly different from −1.0 ($P < 0.001$ in each case), indicating a decrease at a less than linear rate. The similarity between the regression equations for the 12 viruses and those for the randomly sampled subsets was striking, particularly the similarity of the exponents (about −0.35) in the two equations for H_N.

I also examined the relationship between sample size and the ratio of the numbers of nonsynonymous to synonymous SNP sites (*N:S*). For the 12 viruses, the following exponential regression equation was obtained: $N:S = 0.164\ n^{0.237}$ ($R^2 = 0.632$; $P = 0.001$; Figure 4A). The exponent was significantly less than 1.0 ($P < 0.001$), indicating that *N:S* increases with *n* at a less than linear rate. A similar relationship was observed in the case of the randomly sampled subsets of TuMV: $N:S = 0.134\ n^{0.267}$ ($R^2 = 0.887$; $P < 0.001$; Figure 4B). Taken together with the results on the relationship between H_N and sample size, these results suggest that, as sample size increases, the number of nonsynonymous SNP sites increases relative to the number of synonymous SNP sites, but that this increase is due mainly to the inclusion of more rare nonsynonymous SNPs in larger samples, thus in turn lowering H_N.

The existence of numerous rare nonsynonymous SNP sites in turn is expected to reflect the action of purifying selection. In order to provide an additional test of this hypothesis, I added to the analysis the *N:S* value for a sample of 16 TuMV sequences that were serially passaged in experimental adaptation to *Raphanus sativus* (Tan et al. 2005). Since these viruses were bottlenecked during passage, the ability of purifying selection to remove slightly deleterious variants was expected to have been reduced in this lineage. Consistent with this prediction, *N:S* for the passaged TuMV was much higher than expected for a sample of this size (Figure 4A). In fact, the 16 passaged TuMV genomes showed a higher *N:S* than any of the 12 samples from natural potyvirus samples, higher even than that of the 98 naturally samples TuMV (Figure 4A). When the passaged TuMV sample was included in the exponential regression, the studentized residual for this point was 3.39 ($P = 0.006$), indicating that removing this point from the data set provides a highly significant improvement in the regression. No other point in the dataset had a significant studentized residual. These results support the hypothesis that the observed relationship between *N:S* and sample size reflects the presence of rare nonsynonymous variants in potyvirus populations and thus the action of purifying selection.

Discussion

Plant RNA viruses show remarkable sequence conservation, which has been attributed both to low effective population sizes and to strong purifying selection (García-Arenal et al. 2001). An analysis of nucleotide sequence polymorphism in 12 potyviruses provided evidence for both factors. Estimated long-term effective population sizes in these viruses were estimated to be on the order of 10^4. This value is close to the typical long-term effective population sizes of land plants (Lynch 2007), suggesting that the effective population sizes of potyviruses may be similar to those of their hosts. This in turn supports the conclusion that bottlenecks in host-to-host transmission may be very severe (French and Stegner 2003). Potyviruses are transmitted by aphids in a non-persistent manner (Shukla et al. 1994); i.e., the virus is retained by the vector for a short period, usually a few hours or less (Pirone and Harris 1977). It seems plausible that this process might easily give rise to severe bottlenecks in transmission of the virus from one host plant to another.

In addition, many infections of individual plants may not lead to any further infection, thus leading to extinction of certain viral lineages, a factor that will decrease the overall long-term

effective population size (Maruyama and Kimura 1980). A further factor in reducing long-term effective population sizes of potyviruses may be bottlenecks in the process of transfer to new host species. Consistent with previous phylogenetic analyses (e.g., Gibbs et al. 2008), a phylogenetic analysis of potyviruses did not show clustering on the basis of host taxa (Figure 1). These results support the hypothesis that transfers to new hosts have been a recurring feature of the history of these viruses. If new potyvirus species have typically arisen through a fortuitous transfer to a new host of a small number of virus particles, the long-term effective population size of the species might remain low even if the population size eventually increases greatly in the new host (Nei et al. 1975).

The estimates of effective population size presented here depend on the estimation of the genomic mutation rate. The value used here (0.11 per genome per replication), derived from tobacco mosaic virus (Malpica et al. 2002) is at the low end of published estimates for RNA viruses. If one were to use the mean value of the genomic mutation rate for RNA viruses reported by Drake and Holland (1999), about 0.67 per genome per replication), the estimated effective population sizes for potyviruses would be closer to 10^3. Thus effective population sizes based on the present sequence datasets would be larger than 10^4 only if the mutation rate in these viruses is substantially lower than any yet reported for an RNA virus.

These estimates depend on the assumption that the values of π_S (synonymous nucleotide diversity) obtained from the available genomic sequences are representative of each viral species as a whole. The wide scatter of π_S values for small sample sizes (Figure 2A) suggests that there is substantial stochastic error when the number of genomes available was small (less than about 20). Thus, it seems likely that estimates of effective population size based on larger samples (generally about 2×10^4; Table 2) are more likely to be accurate for most potyviruses.

The quasispecies model, based on the work of Eigen (1971), has been widely cited as applicable to RNA viruses (Domingo and Holland 1997). This model assumes very large population sizes and that, as a consequence, genetic drift is not an important factor, assumptions that have been questioned in the case of RNA viruses (Jenkins et al. 2001). The present estimates of effective population sizes in potyviruses are much lower than those assumed by the quasispecies model and support its inapplicability to understanding these viruses at the population level. However, it is possible that this model still may provide valuable insights regarding the within-host evolutionary dynamics of these viruses. There is evidence that plum pox virus, a potyvirus infecting a perennial host, can build up substantial within-host genetic diversity (Jridi et al. 2006). Thus, the short-term within-host effective population size may be much greater than the long-term between-host effective population size, because bottlenecks in transmission lead to loss of most of the diversity accumulated within the host. Indeed, in the case of RNA viruses infecting vertebrates, there is evidence that within-host and between-host patterns of sequence polymorphism can differ substantially (Irausquin and Hughes 2008).

The present results also provided evidence of both past and ongoing purifying selection on potyvirus genomes, leading to the elimination or reduction in frequency of deleterious variants both at nonsynonymous sites in coding regions and at sites in non-coding regions. Given the relatively small effective population sizes of potyviruses, patterns of nucleotide diversity revealed that purifying selection is surprisingly effective. An important factor in maintaining the effectiveness of purifying selection is recombination, which prevents the build-up of deleterious mutations due to “Muller’s ratchet” (Hughes and Hughes 2007). Sequence analyses have supported the hypothesis that recombination is frequent in potyviruses (Tan et al. 2004; Ohshima et al. 2007), consistent with the present evidence of efficient purifying selection.

The frequent occurrence in a population of relatively rare nonsynonymous SNPs is evidence of ongoing purifying selection acting to remove slightly deleterious nonsynonymous variants

(Hughes et al. 2003). The present analyses showed that, in potyviruses, the ratio of nonsynonymous to synonymous polymorphic sites (*N:S*) increases as a function of the number of genomes sampled. Moreover, mean gene diversity at nonsynonymous polymorphic sites (H_N) decreases with increasing sample size at a substantially faster rate than does mean gene diversity at synonymous polymorphic sites (H_S). Both of these relationships were observed both in the case of available genomic sequences of 12 potyvirus species and in data subsets created by randomly sampling from among 98 TuMV genomes. Taken together, these observations imply that a greater proportion of nonsynonymous than of synonymous variants are relatively rare as the result of ongoing purifying selection; and thus many nonsynonymous variants are unlikely to be discovered without extensive sampling.

Although it might be tempting to attribute the occurrence of rare variants to sequencing errors, several lines of evidence argue against sequencing errors as a general explanation for their occurrence. First, the substantially different patterns observed in synonymous and nonsynonymous variants (Figures 3-4) would not be expected in the case of sequencing errors, which would be expected to occur at random with respect to the reading frame. Second, the fact that the same pattern was seen in available virus genomes and in randomly sampled subsets of TuMV is not easily explained on the hypothesis that sequencing errors account for rare variants in the former, since the sampling process accounts for the rarity or abundance of a given variant in the latter.

Additional support for the role of purifying selection in reducing the frequency of nonsynonymous variants was the fact that *N:S* in a sample of 16 experimentally passaged TuMV was much higher than expected on the basis of sample size. Since this TuMV population was passaged in order to adapt it to a new host (*Raphanus sativus*), some of the nonsynonymous variants appearing in the passaged population might have been positively selected because of an effect on viability of the virus to growth in *R. sativus* (Tan et al. 2005). However, certain other nonsynonymous variants probably represent slightly deleterious mutations that could not be eliminated because of “Muller’s ratchet” during the passaging process (Duarte et al. 1992). This type of passaging may thus resembled that involved in the generation of live attenuated virus vaccines, which results in the accumulation of slightly deleterious nonsynonymous variants (Hughes 2009).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

This research was supported by grant GM43940 from the National Institutes of Health.

References

Barouch DH. Challenges in the development of HIV-1 vaccine. Nature 2008;455:613–639. [PubMed: 18833271]

Chao L. Fitness of RNA virus decreased by Muller’s ratchet. Nature 1990;348:454–455. [PubMed: 2247152]

Domingo E, Holland JJ. RNA virus mutations for fitness and survival. Annu. Rev. Microbiol 1997;51:151–178. [PubMed: 9343347]

Drake JW, Holland JJ. Mutation rates among RNA viruses. Proc. Natl. acad. Sci. USA 1999;96:13910–13913. [PubMed: 10570172]

Duarte E, Clarke D, Moya A, Domingo E, Holland J. Rapid fitness losses in mammalian RNA virus clones due to Muller’s ratchet. Proc. Natl. Acad. Sci. USA 1992;89:6015–6019. [PubMed: 1321432]

Eigen M. Self-organization of matter and the evolution of biological macromolecules. Naturwissenschaften 1971;58:465–523. [PubMed: 4942363]

Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 1985;39:783–791.

French R, Stenger DC. Evolution of wheat streak mosaic virus: dynamics of population growth within plants may explain limited variation. Annu. Rev. Phytopathol 2003;41:199–214. [PubMed: 12730393]

García-Arenal F, Fraile A, Malpica JM. Variability and genetic structure of plant virus populations. Annu. Rev. Phytopathol 2001;39:157–186. [PubMed: 11701863]

Gibbs AJ, Mackenzie AM, Wei K-J. The potyviruses of Australia. Arch. Virol 2008;153:1411–1420. [PubMed: 18566735]

Hughes, AL. Adaptive evolution of genes and genomes. Oxford University Press; New York: 1999.

Hughes AL. Evidence for abundant slightly deleterious polymorphisms in bacterial populations. Genetics 2005;169:553–558.

Hughes AL. Micro-scale signature of purifying selection in Marburg virus genomes. Gene 2007;392:266–272. [PubMed: 17306473]

Hughes AL. Near neutrality: leading edge of the neutral theory of molecular evolution. Ann. N.Y. Acad. Sci 2008;1133:162–179. [PubMed: 18559820]

Hughes AL. Relaxation of purifying selection on live attenuated vaccine strains of the family Paramyxoviridae. Vaccine 2009;27:1685–1690. [PubMed: 19195493]

Hughes AL, Hughes MA. More effective purifying selection in RNA viruses than in DNA viruses. Gene 2007;404:117–125. [PubMed: 17928171]

Hughes AL, Packer B, Welsch R, Bergen AW, Chanock SJ, Yeager M. Widespread purifying selection at polymorphic sites in human protein-coding loci. Proc. Natl. Acad. Sci. USA 2003;100:15754–15757. [PubMed: 14660790]

Hughes AL, Piontkivska H. Nucleotide sequence polymorphism in circoviruses. Infect. Genet. Evol 2008;8:130–138. [PubMed: 18093882]

Irausquin SJ, Hughes AL. Distinctive pattern of sequence polymorphism in the NS3 protein of hepatitis C virus type 1b reflects conflicting evolutionary pressures. J. Gen. Virol 2008;89:1921–9. [PubMed: 18632963]

Jenkins GM, Worobey M, Woelk CH, Holmes EC. Evidence for the non-quasispecies evolution of RNA viruses. Mol. Biol. Evol 2001;18:987–994. [PubMed: 11371587]

Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 1992;8:275–282. [PubMed: 1633570]

Jridi C, Martin J-F, Marie-Jeanne V, Labonne G, Blanc S. Distinct viral populations differentiate and evolve independently in a single perennial plant. J. Virol 2006;80:2349–2357. [PubMed: 16474141]

Brown, A.J. Leigh Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population. Proc. Natl. Acad. Sci. USA 1997;94:1862–1865. [PubMed: 9050870]

Li W-H. Unbiased estimates of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol 1993;36:96–99. [PubMed: 8433381]

Lynch, M. The origins of genome architecture. Sinauer; Sunderland MA: 2007.

Malpica JM, Fraile A, Moreno I, Obies CI, Drake JW, García-Arenal F. The rate and character of spontaneous mutation in an RNA virus. Genetics 2002;162:1505–1511. [PubMed: 12524327]

Manrubia SC, Escarmís C, Domingo E, Lázaro E. High mutation rates, bottlenecks, and robustness of RNA viral quasispecies. Gene 2005;347:273–282. [PubMed: 15777632]

Maruyama T, Kimura M. Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent. Proc. Natl. Acad. Sci. USA 1980;77:6710–6714. [PubMed: 16592920]

Miralles R, Moya A, Elena SF. Diminishing returns of population size in the rate of RNA virus adaptation. J. Virol 2000;74:3566–3571. [PubMed: 10729131]

Nei, M. Molecular evolutionary genetics. Columbia University Press; New York: 1987.

Nei M, Maruyama T, Chakraborty R. The bottleneck effect and genetic variability in populations. Evolution 1975;29:1–10.

Ohshima K, Tomitaka Y, Wood JT, Minematsu Y, Kajiyama H, Tomimura K, Gibbs AJ. Patterns of recombination in turnip mosaic virus genomic sequences indicate hotspots of recombination. J. Gen. Virol 2007;88:298–315. [PubMed: 17170463]

Ohta T. Slightly deleterious mutant substitutions in evolution. Nature 1973;246:96–98. [PubMed: 4585855]

Pirone TP, Harris KF. Nonpersistent transmission of plant viruses by aphids. Annu. Rev. Phytopathol 1977;15:55–73.

Pybus OG, Charleston MA, Gupta S, Rambaut A, Holmes EC, Harvey PH. The epidemic behavior of the hepatitis C virus. Science 2001;292:2323–2325. [PubMed: 11423661]

Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol 1987;4:406–425. [PubMed: 3447015]

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 2002;18:502–504. [PubMed: 11934758]

Shukla, DD.; Ward, CW.; Brunt, AA. The potyviridae. CAB International; Wallingford UK: 1994.

Sokal, RR.; Rohlf, FJ. Biometry. Vol. 2 nd ed. W.H. Freeman; San Francisco: 1971.

Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol 2007;24:1596–1599. [PubMed: 17488738]

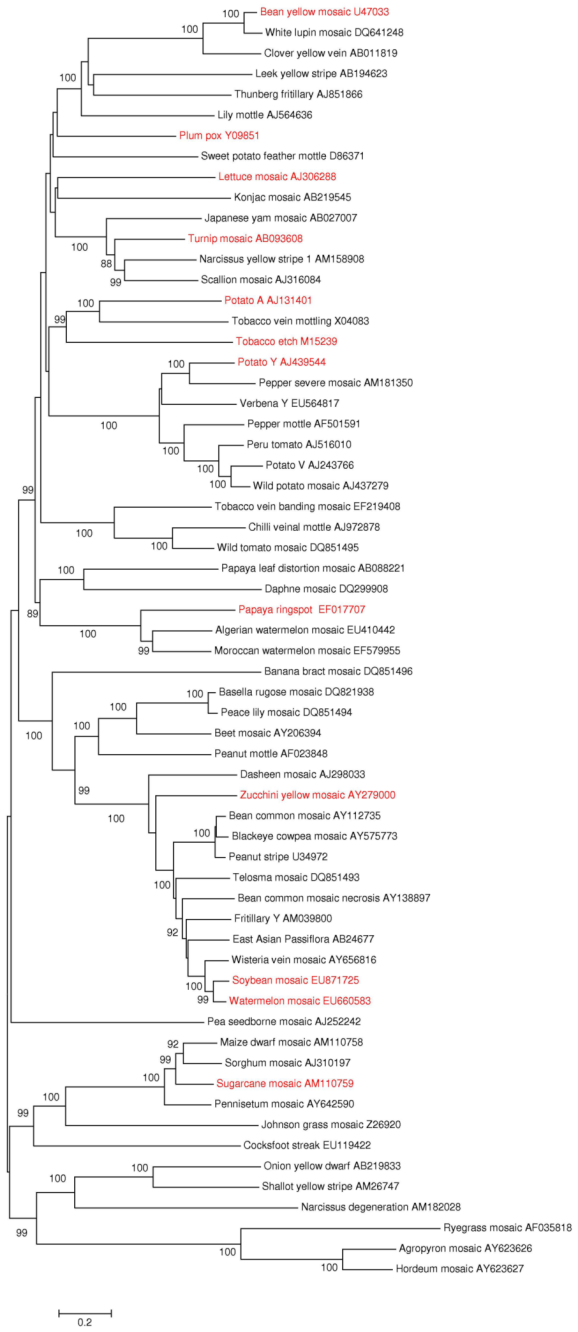
Tan Z, Wada Y, Chen J, Ohshima K. Inter- and intralineage recombinants are common in natural populations of Turnip mosaic virus. J. Gen. Virol 2004;85:2683–2696. [PubMed: 15302962]

Tan Z, Gibbs AJ, Tomitaka Y, Sánchez F, Ponz F, Ohshima K. Mutations in *Turnip mosaic virus* genomes that have adapted to *Raphanus sativus*. J. Gen. Virol 2005;86:501–510. [PubMed: 15659771]

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Diggins DG. The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997;25:4876–4882. [PubMed: 9396791]

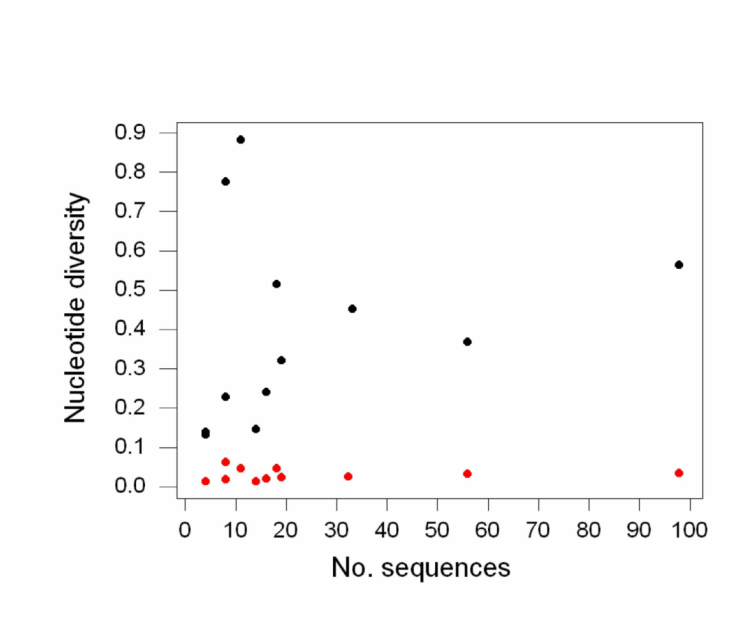
Wallis CM, Stone AL, Sherman DJ, Damsteegt VD, Gildow FE, Scheider WL. Adaptation of plum pox virus to a herbaceous host (Pisum sativum) following serial passages. J. Gen. Virol 2007;88:2839–2845. [PubMed: 17872538]

Wright S. Evolution in Mendelian populations. Genetics 1931;16:97–159. [PubMed: 17246615]

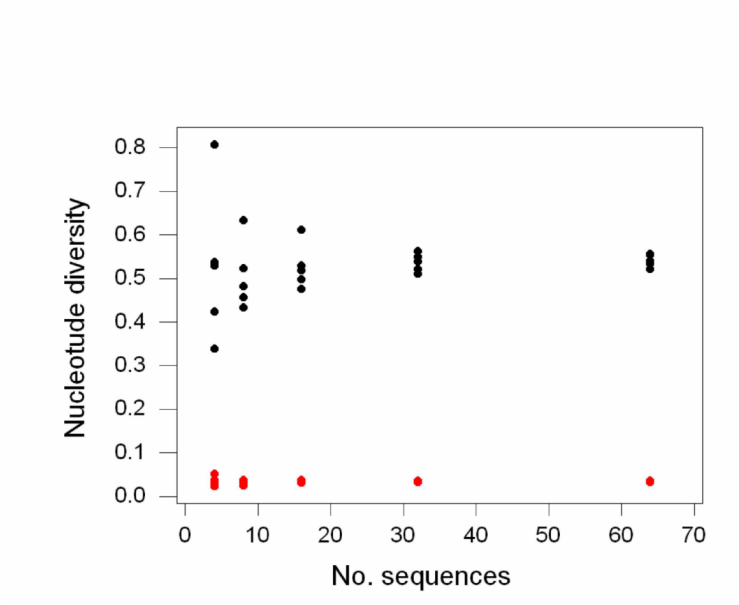


1. .

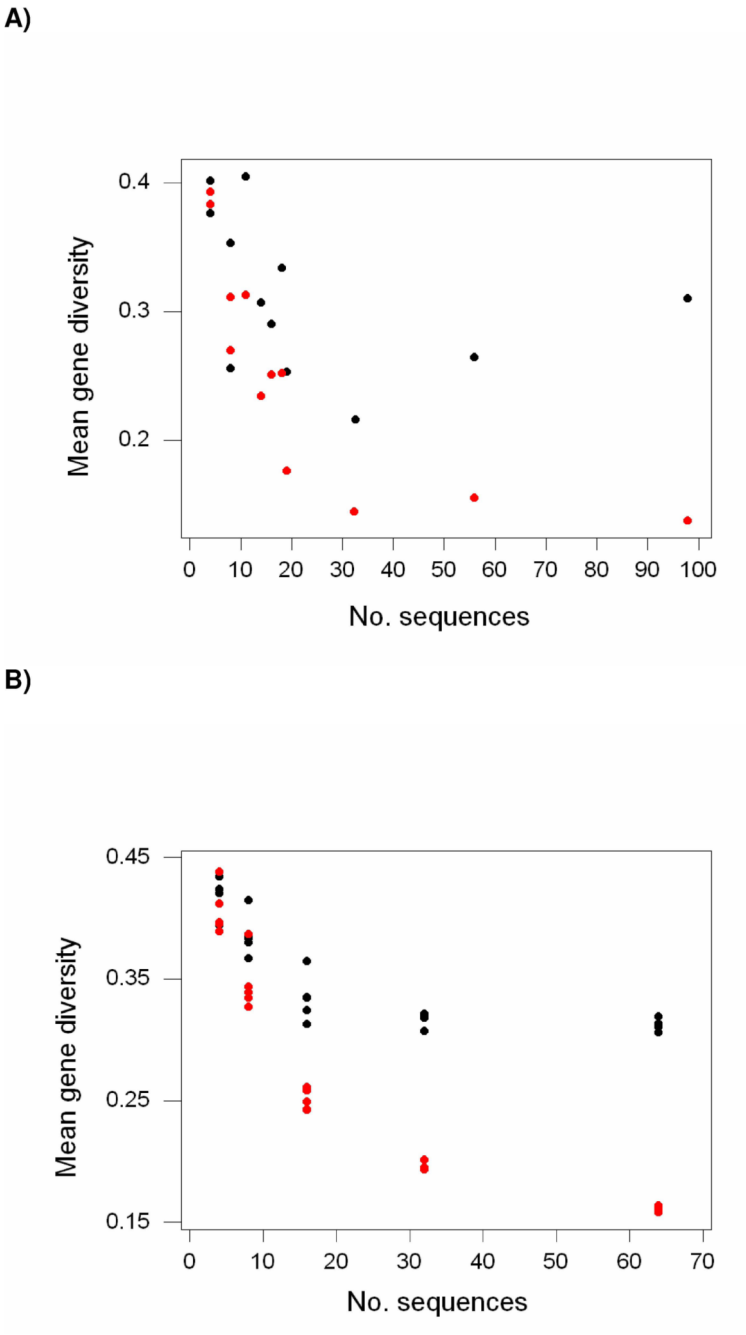
A)



B)

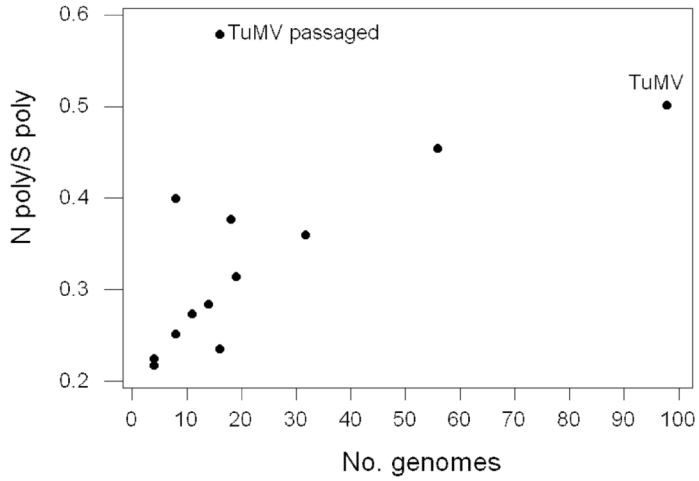


2. .

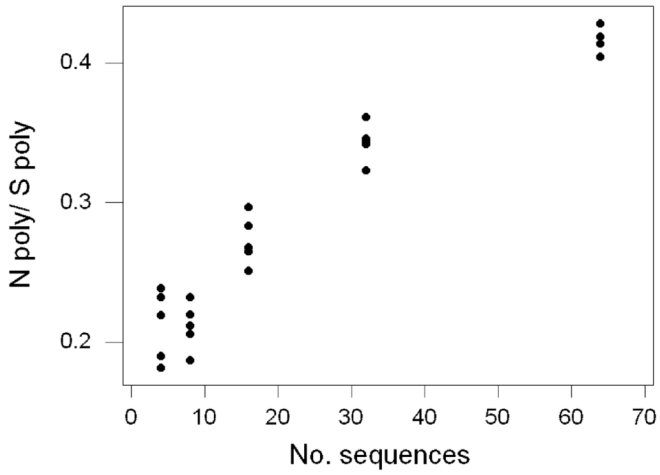


3. .

A)



B)



4. .

Table 1

Nucleotide diversity at synonymous (πS), nonsynonymous (πN), and noncoding (π) sites of 12 potyviruses, with estimates of Ng.

Virus (abbreviation)	No. genomes	Coding		Noncoding	Ng
		π_S ± S.E.	π_N ± S.E.	π ± S.E.	
Turnip mosaic virus (TuMV)	98	0.5463 ± 0.0098	0.0333 ± 0.0014[a]	0.0679 ± 0.0088[a, b]	2.5×10^4
Potato virus Y (PVY)	56	0.3675 ± 0.0087	0.0312 ± 0.0015[a]	0.1863 ± 0.0140[a, b]	1.6×10^4
Plum pox virus (PPV)	33	0.4522 ± 0.0114	0.0254 ± 0.0010[a]	0.0994 ± 0.0097[a,b]	2.0×10^4
Zucchini yellow mosaic virus (ZYMV)	19	0.3209 ± 0.0070	0.0233 ± 0.0012[a]	0.1014 ± 0.0099[a, b]	1.4×10^4
Papaya ringspot virus (PRSV)	18	0.5154 ± 0.0108	0.0450 ± 0.0019[a]	0.1452 ± 0.0163[a, b]	2.4×10^4
Watermelon mosaic virus (WMV)	16	0.2407 ± 0.0061	0.0194 ± 0.0011[a]	0.0625 ± 0.0082[a, b]	1.1×10^4
Soybean mosaic virus (SMV)	14	0.1457 ± 0.0046	0.0128 ± 0.0009[a]	0.0991 ± 0.0097[a, b]	6.3×10^3
Sugarcane mosaic virus (SCMV)	11	0.8839 ± 0.0227	0.0450 ± 0.0026[a]	0.1169 ± 0.0124[a, b]	3.8×10^4
Potato virus A (PVA)	8	0.2272 ± 0.0081	0.0172 ± 0.0010[a]	0.0422 ± 0.0060[a, b]	9.9×10^3
Bean yellow mosaic virus (BYMV)	8	0.7767 ± 0.0205	0.0061 ± 0.0023[a]	0.1890 ± 0.0171[a, b]	3.4×10^4
Lettuce mosaic virus (LMV)	4	0.1382 ± 0.0053	0.0129 ± 0.0009[a]	0.0161 ± 0.0053[a]	6.3×10^3
Tobacco etch virus (TEV)	4	0.1324 ± 0.0060	0.0112 ± 0.0010[a]	0.0491 ± 0.0093[a, b]	5.7×10^3

Z-tests of the hypothesis that π_N or π equals π_S:

Z-tests of the hypothesis that π equals π_N:

[a] $P < 0.001$.

[b] $P < 0.001$.

Table 2

Mean gene diversity at polymorphic synonymous, nonsynonymous, and noncoding sites in potyvirus genomes.

Virus	No. genomes	Synonymous		Nonsynonymous		Noncoding	
		No. polymorphic sites	Mean gene diversity ± S.E.	No. polymorphic sites	Mean gene diversity ± S.E.	No. polymorphic sites	Mean gene diversity ± S.E.
TuMV	98	2419	0.3096 ± 0.0035	1212	0.1365 ± 0.0046 [b]	116	0.1533 ± 0.0163 [b]
PVY	56	2266	0.2639 ± 0.0039	1029	0.1553 ± 0.0053 [b]	295	0.2300 ± 0.0032
PPV	53	2297	0.2160 ± 0.0022	827	0.1442 ± 0.0035 [b]	129	0.1772 ± 0.0098 [b]
ZYMV	19	2112	0.2528 ± 0.0032	663	0.1758 ± 0.0045 [b]	116	0.2602 ± 0.0148
PRSV	18	2437	0.3344 ± 0.0033	919	0.2515 ± 0.0054 [b]	94	0.3117 ± 0.0029
WMV	16	1789	0.2895 ± 0.0036	421	0.2512 ± 0.0073 [b]	74	0.2816 ± 0.0198
SMV	14	1046	0.3070 ± 0.0044	297	0.2344 ± 0.0080 [b]	122	0.2550 ± 0.0139 [b]
SCMV	11	2226	0.4050 ± 0.0031	609	0.3127 ± 0.0062 [b]	96	0.3724 ± 0.0154 [a]
PVA	8	1316	0.2558 ± 0.0025	331	0.2699 ± 0.0055 [a]	53	0.2587 ± 0.0023
BYMV	8	2094	0.3526 ± 0.0026	837	0.3112 ± 0.0041 [b]	128	0.3614 ± 0.0125
LMV	4	712	0.4017 ± 0.0020	160	0.3930 ± 0.0035	9	0.4167 ± 0.0295
TEV	4	622	0.3762 ± 0.0005	135	0.3833 ± 0.0027 [b]	30	0.3792 ± 0.0042

Randomization tests of the hypothesis that mean gene diversity in nonsynonymous or noncoding sites equals the corresponding value in synonymous sites:

[a] $P < 0.05$

[b] $P < 0.001$.