Within-patient HIV mutation frequencies reveal fitness costs of CpG dinucleotides, drastic amino acid changes and G→A mutations

Marion Hartl*†, Kristof Theys *‡, Alison Feder §, Maoz Gelbart ¶, Adi Stern ¶, Pleuni S. Pennings †

June 2016

1 Abstract

HIV has a high mutation rate and exhibits remarkable genetic diversity. However, we know little about the fitness cost of HIV mutations *in vivo*. We calculated the mean frequency of mutations at 870 sites of the pol gene in 160 patients, allowing us to determine the cost of different types of mutations. We found that non-synonymous mutations that lead to drastic amino acid changes are three times more costly than those that do not, mutations that create new CpG dinucleotides are up to four times more costly than those that do not, and G→A mutations are more than twice as costly as A→G mutations. In addition, within-patient mutation frequencies are highly correlated with substitution frequencies across the global HIV pandemic. We anticipate our new frequency-based approach will provide insights into the fitness landscape and evolvability of not only HIV, but a variety of microbes.

2 Introduction

The human immunodeficiency virus (HIV) replicates with an extremely high mutation rate and exhibits significant genetic diversity within an infected host, often referred to as a "mutant cloud" or "quasispecies" [1–7]. Although mutations are crucial for all adaptive processes, they are also associated with fitness costs. Thus, to understand the evolution of HIV, it is important to know the fitness costs of mutations *in vivo*. Fitness costs influence the probability of evolution from standing genetic variation (often referred to as pre-existing mutations). Fitness costs also determine the effects of background selection (i.e., the effects of linked deleterious mutations on neutral or beneficial mutations) and thus affect optimal recombination rates. All of these processes affect drug resistance and immune escape in HIV [8–12]. Moreover, in addition to a better understanding of evolutionary processes in HIV and in general, a detailed knowledge of mutation costs could help us discover new functional elements in the HIV genome.

In infinitely large populations, mutations are present at a constant frequency equal to u/s, where u is the mutation rate from wild-type to the mutant and s is the selection coefficient that reflects the negative fitness effect, or cost, of the mutation [13,14]. In natural populations of finite size, however, the frequency of mutations is not constant; instead it fluctuates around the expected frequency of u/s, because of the stochastic nature of mutation and replication [13]. Due to these stochastic fluctuations, it is impossible to accurately infer the strength of selection acting on individual mutations (i.e., their cost) from a single observation of a single population. Thus, fitness effects are traditionally assessed either in *in vitro* systems (e.g., cell culture/competition experiments [15–18]) or in a phylogenetic framework [19–21]. Both approaches have their caveats, however. It is unclear whether *in vitro* fitness costs are similar to *in vivo* fitness costs, and because the phylogenetic framework estimates fitness costs over very long timescales, it is unclear how relevant those estimates are for current viral populations.

HIV has unique properties that allow us to study fitness effects *in vivo*: It is fast evolving [22–26] and leads to persistent infections [27–29]. This means that genetic diversity accumulates quickly and independently in every host, and samples from different patients can thus be treated as independent replicate populations [30]. With data from many replicate populations, the mean frequency of mutations will approach u/s and can be used to estimate their fitness costs, as the fluctuations in mutation frequencies represent an ergodic process [31]. Based on this logic, in this article we present a novel approach that uses observed mutation frequencies in HIV-infected patients to determine the fitness effects of mutations *in vivo*.

Theoretically, with sufficient sequencing data from enough patient samples, we could use this approach to estimate the *in vivo* fitness cost of every point mutation at every position in the HIV-1 genome. In the current study, we demonstrate its

*These authors contributed equally and are listed in alphabetical order.
†Department of Biology, San Francisco State University, San Francisco, California, USA
‡Clinical and Epidemiological Virology, Department of Microbiology and Immunology, Rega Institute for Medical Research, KU Leuven, University of Leuven, Leuven, Belgium
§Department of Biology, Stanford University, Palo Alto, California, USA
¶Department of Molecular Microbiology and Biotechnology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

utility by focusing on transition mutations (A↔G and C↔T) in 870 sites of the pol gene, which encodes HIV's protease protein and part of the reverse transcriptase (RT) protein, in 160 patients infected with HIV-1 subtype B. We selected transitions because they are much more common in HIV than transversions [24], and thus sufficient data are available to analyze; we selected the pol gene because it is highly conserved and its products experience less direct contact with the immune system than the exposed product of the much more variable env gene [27, 28]. Finally, we excluded all drug resistance-related sites, because the samples came from patients receiving several different treatments. Accordingly, most of the mutations we studied were expected to be deleterious. We report that this test of our method allowed us to quantify many known properties of mutational fitness costs, and it also revealed novel insights into the evolutionary constraints of the HIV genome.

3 Results

Mean frequency of non-synonymous mutations is lower than that of synonymous mutations

We began by examining the frequencies of the three main categories of mutations: synonymous, non-synonymous, and nonsense. Because we only considered transitions, there was exactly one possible mutation per site. The pol sequences from the Bacheler *et al* data set, which formed the basis of this study, are 984 nucleotides long, composed of 297 nucleotides that encode HIV's Protease protein and 687 nucleotides that encode its RT. However, we excluded drug resistance–related sites, determined from Johnson *et. al* [32], and protease sites that overlap with gag, which led to a total of 870 sites. For each mutation, we determined the average of 160 observed frequencies (one frequency for each patient, fewer if a site was filtered out for certain patients; see Materials and Methods).

As an example, we show the frequency spectra at codon 58 of the Protease protein, which comprises nucleotides 172 through 174 (Fig. 1A). The transition mutation at the first position (172) creates a premature stop codon. This mutation was never observed in the data and thus has a frequency of zero in all patients. A transition mutation at the second codon position (173) leads to an amino acid change (glutamine to arginine) and was found at low frequencies in some patients. A mutation at the third position of the codon (174) does not change the amino acid and was observed at higher frequencies in some patients and was even fixed in some (Fig. 1A).

We next ordered all sites according to observed mutation frequencies, which revealed that this pattern—synonymous mutations being more common than non-synonymous mutations, which were more common than nonsense mutations—was evident throughout the entire data set. The distributions of the mean frequencies for each of the three main categories of mutations were significantly different (one-sided two-sample Wilcoxon test, $p < 2.2e-16$ for nonsense vs non-synonymous mutations and for non-synonymous vs synonymous mutations; Fig. 1B). All nonsense mutations had an average frequency of zero, and so did a few non-synonymous mutations. Most non-synonymous mutations had a lower frequency than synonymous mutations (80% of non-synonymous mutations were present at a frequency lower than 0.01, whereas 82% of synonymous mutations were present at a frequency higher than 0.01). This difference in distributions probably reflects the higher cost of non-synonymous mutations, which are more likely to directly affect virus replication. This analysis therefore provides a proof of principle that our approach works: The observed frequencies reflect the relative costs we would expect for these broad categories of mutation.

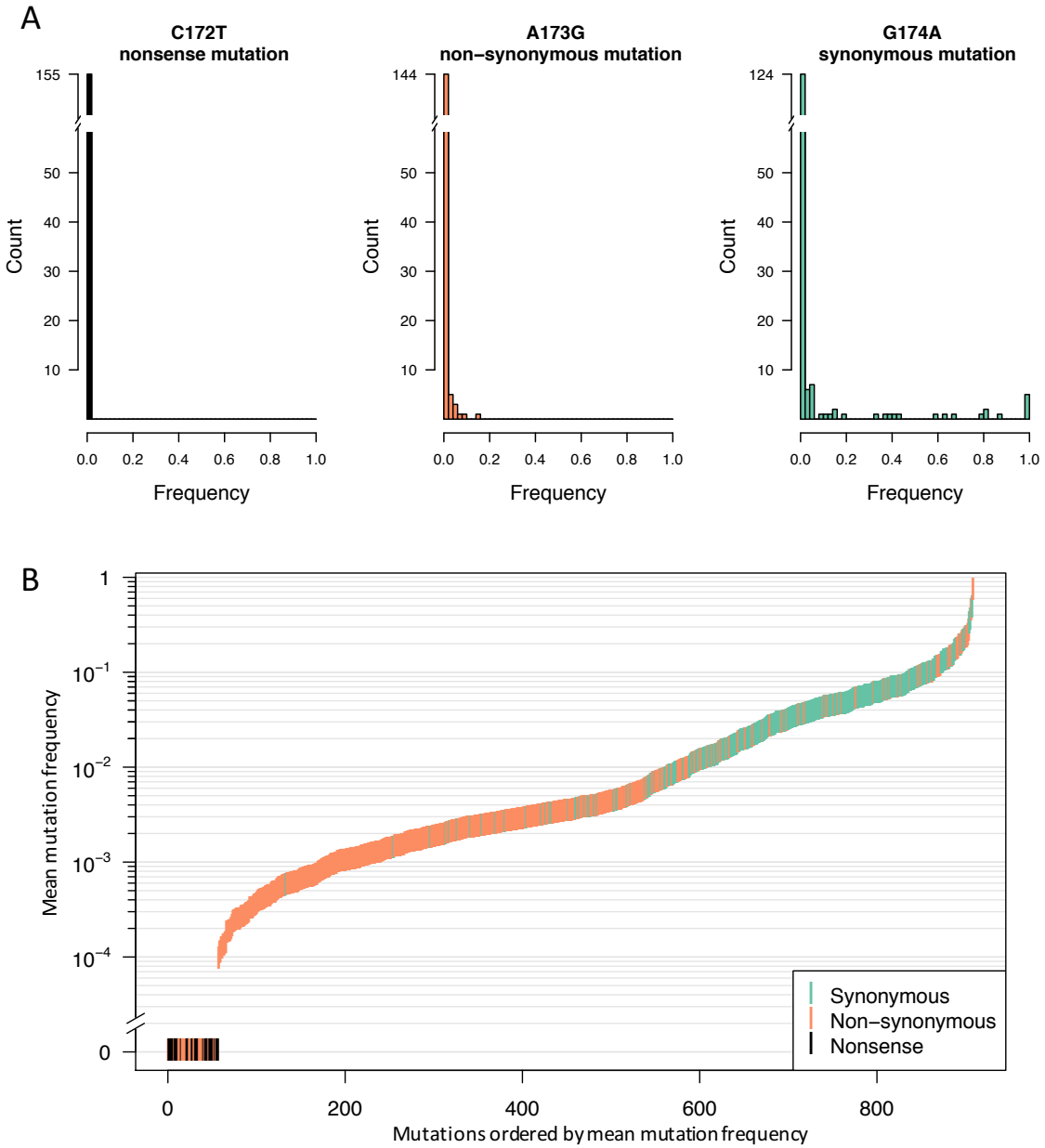

Figure 1: As expected, in the HIV pol gene, synonymous mutations occurred more frequently than non-synonymous mutations, which occurred more frequently than nonsense mutations, which were not observed at all. A) Single-site frequency spectrum for three sites in the HIV Protease protein (sites 172, 173 and 174). Note that even the synonymous mutation (at site 174) occurred at low frequency in most patients, although it reached high frequencies in some. B) Mean mutation frequencies for all sites, ordered by mutation frequency.

High costs associated with mutations that create new CpG dinucleotides, G→A mutations and mutations that lead to drastic amino acid changes

Next, we fit a generalized linear model (GLM) to determine the ensemble characteristics that explain the observed frequencies of synonymous and non-synonymous mutations (see Table 1). The advantage of using a GLM is that we can directly analyze raw counts as opposed to averaged frequencies. That is, if we have 20 sequences for patient 1, and at a given nucleotide, we observe 18 As and 2 Gs, we can make use of this information. This approach automatically gives more weight to patients for whom we have more sequences, and it also allowed us to investigate several effects simultaneously. We then used estimated mutation rates from Abram *et al* [24] and the mutation-selection formula ($f = u/s$) to translate the observed frequencies into selection coefficients (costs). The estimated costs are shown for some groups of mutations in Figure 2. As in our first

result, synonymous mutations were found at higher frequencies than non-synonymous mutations ($p < 0.001$), which means that they are less costly.

Synonymous mutations. We then sought to determine the characteristics responsible for differences in frequencies, and thus fitness costs, among synonymous mutations. Interestingly, the strongest effect was associated with whether or not a mutation created a new CpG dinucleotide site (see Table 1). Specifically, A→G mutations and T→C mutations that created new CpG sites were found at significantly lower frequencies than A→G mutations and T→C mutations that did not ($p < 0.001$) (note that G→A and C→T mutations cannot create new CpG sites). The model predictions for the frequencies suggest that CpG-creating synonymous mutations are four times more costly (selection coefficient appr. 0.001) than non-CpG-creating synonymous mutations (selection coefficient 0.00025). This finding is consistent with the hypothesis that CpG sites are costly for RNA viruses because they trigger the host's antiviral cellular response [33–35].

We also found a surprisingly strong effect of the nucleotide in the consensus sequence (i.e., the presumed ancestral nucleotide): Synonymous G→A mutations were observed at frequencies lower than expected given their mutation rate. We could not formally test whether this effect was significant using the GLM framework, but a one-sided two-sample Wilcoxon test showed that the difference in estimated selection coefficients for G→A mutations and non-CpG-forming A→G mutations was highly significant ($p = 5.088e-09$). Indeed, the estimated selection coefficients based on model predictions suggested that synonymous G→A mutations are two-and-a-half times as costly as non-CpG-forming A→G mutations (0.0007 vs 0.00025) (Fig. 2).

Non-synonymous mutations. Subsequently, we analyzed the characteristics associated with frequency differences, and hence fitness costs, in non-synonymous mutations. Here, we distinguished between mutations that led to a drastic amino acid change and those that did not. This distinction was based on the classical grouping of amino acids into five groups (positively charged, negatively charged, uncharged, hydrophobic and special cases: cysteine, selenocysteine, glycine and proline); we defined a change in group as a drastic amino acid change. In general, mutations that led to a drastic amino acid change were found at lower frequency than mutations that did not ($p < 0.001$). For example, an A→G mutation that results in a drastic amino acid change costs roughly three times more than one that does not (0.02 vs 0.006). We observed similar fold changes for the other possible transitions.

There was also an effect of whether or not a non-synonymous mutation created a CpG site ($p < 0.001$ for both A→G and T→C mutations). The difference in frequencies suggests that, among mutations that do not lead to a drastic amino acid change, A→G mutations that create a CpG site are approximately one-and-a-half times more costly than those that do not (0.0008 vs 0.0005). Similarly, for T→C mutations that do not involve a drastic amino acid change, CpG-forming mutations are three-and-a-half times more costly (0.0017 vs 0.0005).

In addition, we found a strong effect of the nucleotide in the consensus sequence (i.e., the presumed ancestral nucleotide): C→T and G→A mutations appear to be more costly than A→G and T→C mutations (Fig. 2). We could not use the GLM framework to test whether this difference was significant, but one-sided two-sample Wilcoxon tests showed that the difference in estimated selection coefficients was highly significant ($p = 3.824e-06$ for C→T and $p = 3.181e-05$ for G→A mutations when compared with A→G mutations). Using model-predicted frequencies and known mutation rates, we estimated that, among non-synonymous mutations that do not involve a drastic amino acid change or create a CpG site, C→T mutations are six times more costly than A→G mutations (0.0031 vs 0.0005), and G→A mutations are three-and-a-half times more costly (0.0018 vs 0.0005).

Because the nature of the amino acid change and the ancestral nucleotide in the consensus sequence both had a strong effect on costs, we decided to investigate further. We found that many very costly mutations were associated with a small number of amino acid changes starting from glycine and proline. This is consistent with our knowledge of protein structure: glycine and proline are often unique and irreplaceable, as the only cyclic and smallest amino acid, respectively. Figure 3 shows the cost of non-synonymous changes ordered by ancestral and mutant amino acid. These amino acid effects may partially explain why G→A mutations are especially costly.

Reverse transcriptase vs protease mutations and effects of RNA secondary structure. According to our model predictions, mutations in the RT portion of the gene had slightly lower frequencies than those in the protease portion ($p < 0.001$), suggesting that they are somewhat more costly. Similarly, our model predicts a small but significant effect of the SHAPE parameter ($p < 0.001$), an experimentally determined measure of RNA secondary structure [36]. Specifically, sites with a lower SHAPE parameter value (i.e., those more likely to be part of an RNA structure) were associated with lower mutation frequencies (which suggests higher mutational costs), presumably because the secondary structure of the RNA molecule plays a functional role in HIV replication [36] (see Table 1).

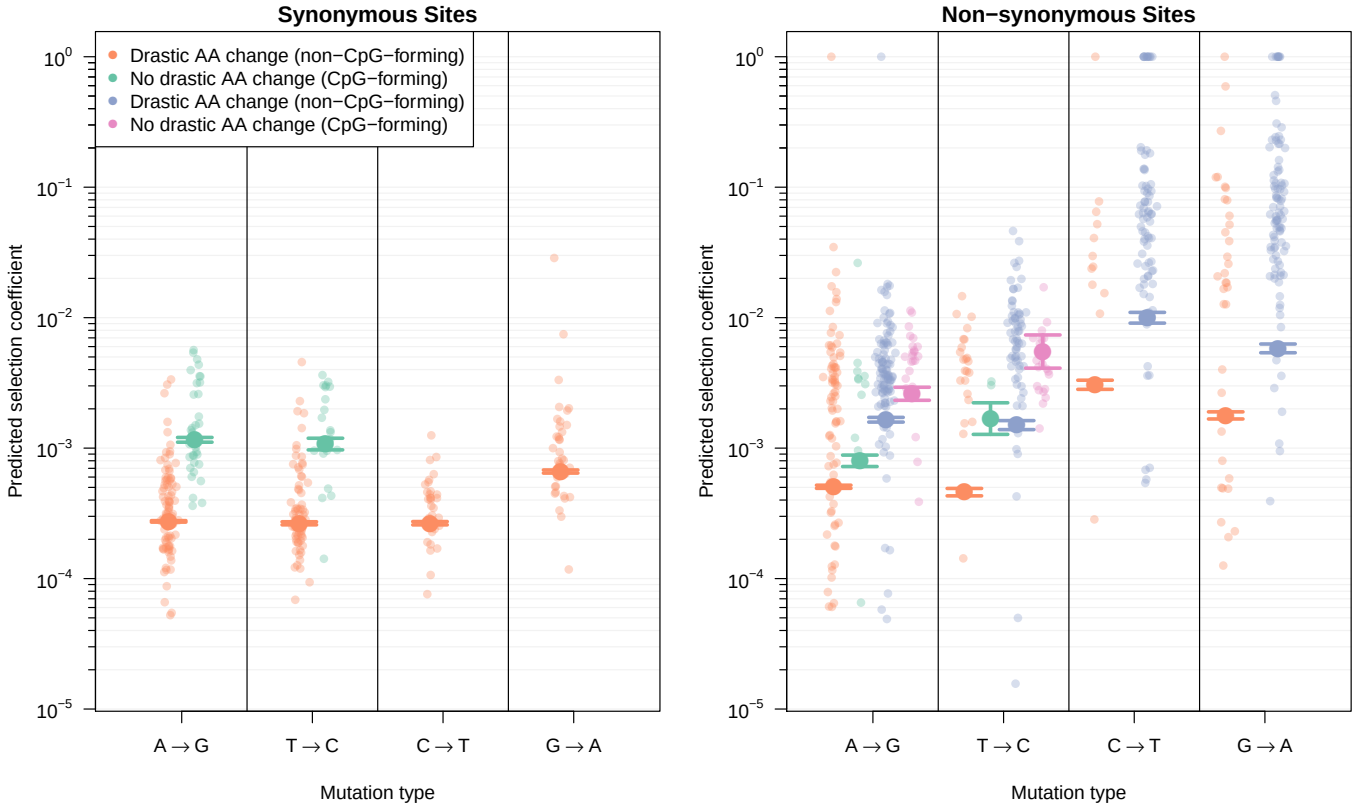

Figure 2: Selection coefficients for transitions at every nucleotide site in the pol sequence show that CpG-forming mutations are more costly than non-CpG-forming mutations and that mutations that involve a drastic amino acid change are more costly than mutations that do not. Selection coefficients were estimated using a generalized linear model and sequence data from 160 HIV-infected patients. Shown are predicted selection coefficients for synonymous (*left*) and non-synonymous (*right*) mutations that do not involve a drastic amino acid change and either create CpG sites (*green*) or do not (*orange*). For non-synonymous mutations, predictions are also shown for mutations that do involve drastic amino acid changes and either create CpG sites (*pink*) or do not (*blue*).

Table 1: Significant predictors of fitness costs for mutations in the pol sequence, estimated using a generalized linear model

	Estimate	Std. Error	z value	Pr (> | z |)
(Intercept)	-3.208	0.013	-244.671	< 0.001
In reverse transcriptase	0.136	0.008	16.206	< 0.001
SHAPE parameter	0.169	0.014	12.034	< 0.001
T→C	0.034	0.014	2.422	0.015
C→T	0.808	0.016	50.379	< 0.001
G→A	0.717	0.014	49.936	< 0.001
CpG-forming	-1.439	0.029	-49.853	< 0.001
T→C:CpG-forming	0.041	0.047	0.875	0.381
Non-syn	-0.611	0.014	-42.644	< 0.001
T→C:Non-syn	0.061	0.024	2.551	0.011
C→T:Non-syn	-1.833	0.037	-49.396	< 0.001
G→A:Non-syn	-0.380	0.021	-18.065	< 0.001
Non-syn:CpG-forming	0.981	0.045	21.991	< 0.001
T→C:Non-syn:CpG-forming	-0.881	0.093	-9.447	< 0.001
Drastic amino acid change	-1.183	0.014	-83.501	< 0.001

Parameters for gamma distribution of fitness effects

In addition to the characteristics that determine the fitness costs of mutations, we investigated the distribution of fitness effects (DFE). This distribution is of great interest to the evolutionary biology community because it affects standing genetic variation, background selection, and optimal recombination rates [20]. Moreover, the DFE affects the evolvability of a population: A DFE weighted toward neutral and adaptive mutations reflects a population with more capacity to evolve. Many viruses, however, have been found to have a DFE composed mainly of deleterious and lethal mutations. To determine

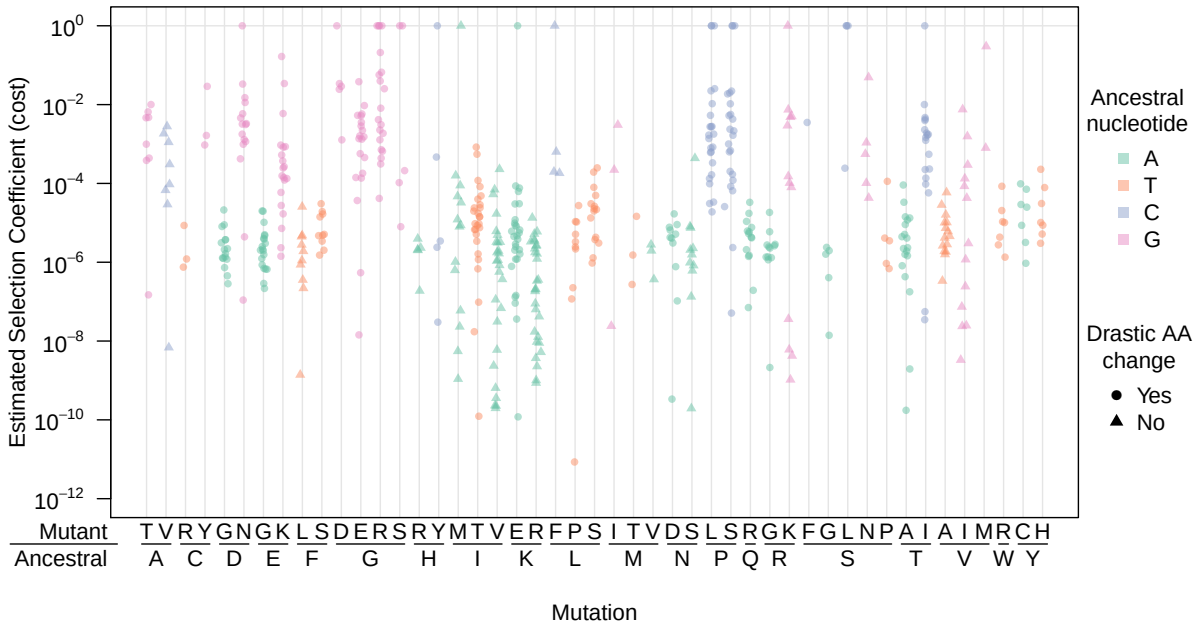

Figure 3: Estimated costs of non-synonymous mutations in the pol sequence, ordered by ancestral amino acid, show that many of the most costly mutations are concentrated at a few amino acids (e.g., proline and glycine). The selection coefficients shown are calculated directly from mean mutation frequencies and mutation rates using the mutation-selection balance formula, $f = u/s$.

the DFE of HIV, we took the average mutation frequency for each pol site and used it to directly estimate the fitness cost using the mutation-selection balance formula ($f = u/s$). In Figure 4, we show the DFEs for each of the ancestral nucleotides, stratified by synonymous and non-synonymous mutations (including nonsense mutations). Overall, there were few very deleterious and lethal mutations, except for non-synonymous C→T and G→A mutations. We also estimated parameters for the gamma distribution that best describes the entire DFE (Table 2). These parameters can be used in studies of background selection and in other studies that involve simulations of evolving populations. We extended this analysis also for the Lehman [37] and Zanini data set [38] (see Figure S5, Figure S6 and Table S1).

Table 2: Parameters for the gamma distribution of fitness effects for pol mutations in 160 HIV-infected patients, reflecting scale (κ) and shape (θ)

		Mut. rates from Abrahm 2010		Mut. rates from Zanini 2016		
	Num. sites	κ	θ	κ	θ	Lethal
Bacheler	870	0.31	0.211	0.315	0.243	0.059
		(0.222, 0.4)	(0.203, 0.223)	(0.244, 0.399)	(0.233, 0.256)	(0.044, 0.076)

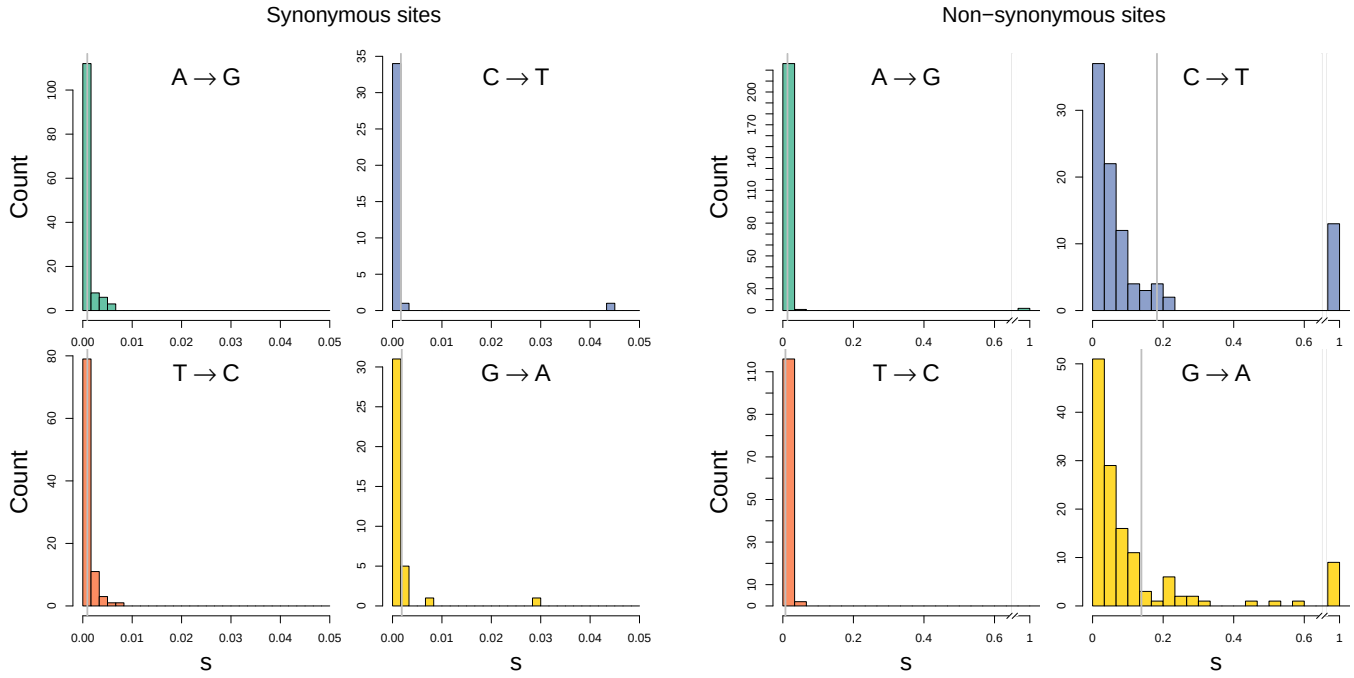

Figure 4: Distribution of fitness effects as estimated from mutation frequencies using the mutation-selection balance formula ($f = u/s$). Most synonymous mutations (*left panel*) have very low selection coefficients. For non-synonymous mutations (*right panel*), selection coefficients are higher, especially for C→T and G→A mutations. Grey vertical lines indicate median selection coefficients. Note that the scales of the x- and y-axis differ between the figures.

Relationship between mutation frequencies within patients and within the global subtype B epidemic

Next, we wanted to determine how well the observed within-patient mutation frequencies correspond with worldwide HIV mutation frequencies. All sequences in the Bacheler *et al* data set we used belonged to HIV-1 subtype B, so we assembled a comparison set of HIV-1 subtype B sequences from treatment-naive patients using the Stanford HIV Drug Resistance database (HIVdb); this set contained 23,742 protease sequences and 22,785 reverse transcriptase sequences [39]. Notably, each sequence in the comparison set represented the consensus of the "mutant cloud" evolving in a patient at a given time point. Figure 5 shows the correlation between average within-patient mutation frequencies from the 160 patients analyzed in this study and global mutation frequencies calculated from the HIVdb. A fairly high correlation coefficient was detected when comparing all 870 sites (Spearman's rank correlation coefficient $\rho = 0.76$), showing a surprising concordance between mutation frequencies within patients and in the global subtype B epidemic. Figure 5 also shows that costly mutations that occurred at low frequencies within patients were often not observed in consensus sequences from the HIVdb.

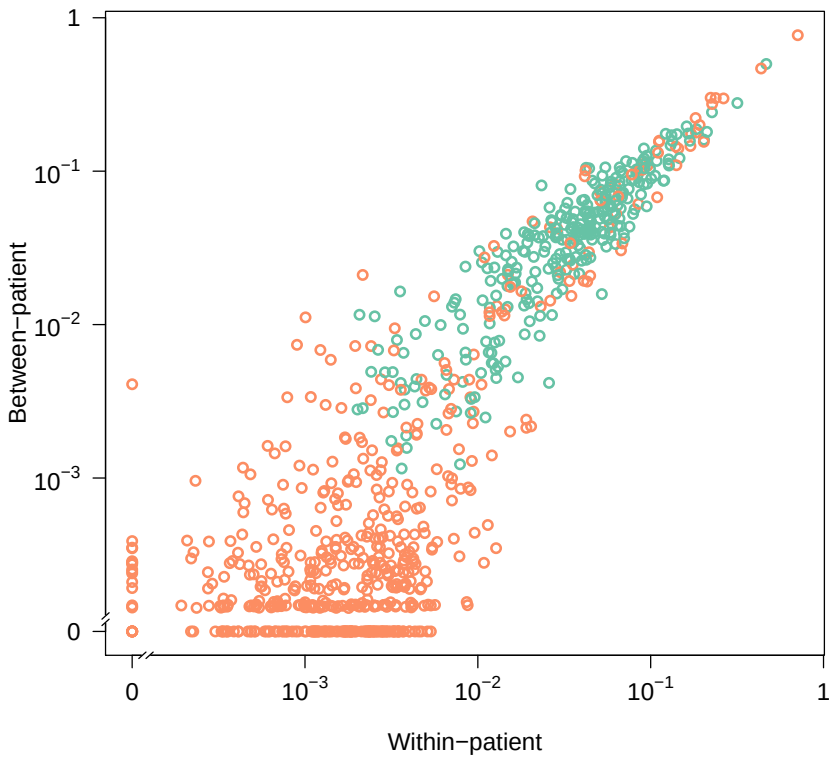

Figure 5: A strong correlation (Spearman's rank correlation coefficient $\rho = 0.76$) exists between average pol mutation frequencies at the within-patient level (in the 160 patients analyzed in this study) and mutant frequencies in the global subtype B epidemic (23,742 protease and 22,785 reverse transcriptase consensus sequences from the HIVdb [39]). Values shown on a log scale. Non-synonymous mutations are shown in red, synonymous mutations in green.

4 Discussion

Our analysis was based on observed mutation frequencies at 870 sites in the pol gene in 160 patients infected with HIV-1 subtype B, with a median of 19 sequences (range: 2–69) per patient [40]. First, as expected, we found a clear separation of observed frequencies for synonymous, non-synonymous and nonsense mutations (i.e., mutations that create premature stop codons) (Fig. 1). Second, we found that inferred costs of non-synonymous mutations were strongly affected by whether the resulting amino acid change was drastic or not; costs were lower if a mutation led to a change to a similar amino acid (Fig. 2). Third, and surprisingly, we found that mutations that created new CpG sites were up to four times more costly than mutations that did not. This finding may reflect selection against CpG sites, which trigger recognition by antiviral defense mechanisms [33–35] (Fig. 2). A fourth and also surprising finding was the substantial difference in fitness cost depending on which of the four nucleotides was altered. In particular, G→A mutations were more costly than other mutations. Thus, although we analyzed only a small part of the HIV genome using a data set with limited sequencing depth, we succeeded in recovering and quantifying many known properties of mutational fitness costs, as well as discovering novel findings. Our data also allowed us to estimate parameters of DFEs (Fig. 4, Table 2). Finally, we found that within-patient frequencies and global frequencies in the subtype B clade were very similar (Spearman's rank correlation coefficient $\rho = 0.76$).

Comparison with other studies in viruses

In general, our results are consistent with those from a recent study on HIV-1 evolution by Zanini *et al* [26], based on a data set described recently by the same authors in a previous paper [38]. Like them, we found a clear difference in the fitness costs between synonymous and non-synonymous mutations (see Supplementary Figure S2 and Figure 4 in Zanini *et.al* [26]). One clear difference between our studies is that out of the 870 pol transition mutations that we focused on, some mutations (5.9%) were never observed in our data set, whereas all transition mutations were observed in the Zanini data set [38]. A mutation that is not observed has an estimated fitness cost of 1, which means that it is lethal. Therefore, our results suggest that 5.9% of pol transition mutations are lethal, whereas the Zanini data set—if all frequencies are taken at face value—suggests that no mutations are lethal (see Table S1). This difference may simply reflect differences in sequencing depth and technique.

The Zanini data set had much deeper sequencing depth, making it more likely that even very deleterious or lethal mutations would be observed. In addition, next-generation sequencing techniques have error rates higher than HIV's mutation rate, so that even mutations that do not actually occur in a sample may be observed due to sequencing errors. For mutations at low frequencies, observed mutation frequencies are thus likely overestimates, resulting in underestimates of the percentage of lethal mutations. Indeed, when Zanini *et al* [26] filtered out mutations that occurred at a frequency less than 0.002, roughly 50% of all non-synonymous mutations no longer occurred in their data set and were therefore considered lethal. We did not perform such filtering for the Bacheler data set used in this article, but some of the mutations that were observed at low frequencies may also have been due to sequencing errors, so that the real number of lethal or very deleterious mutations may be higher than our estimate of 5.9%.

We also analyzed a third data set from a recent paper on HIV-1 by Lehman *et al* [37] (see Supplementary Figure S1). In this data set, 1.3% of transition mutations were never observed and thus estimated to be lethal. Again, because of the sequencing technology and the associated error rate, we expect that this percentage underestimates the true percentage of lethal mutations.

It should be noted that the proportion of lethal mutations estimated in our study (5.9%), though higher than in the other two HIV-1 *in vivo* studies discussed above, is low compared to percentages from *in vitro* studies on viral coding sequences (see [41] for an overview). For example, Sanjuan *et al* [15] found that 40% of random mutations in the RNA vesicular stomatitis virus were lethal, and a study of the tobacco etch virus estimated that 27 of 66 mutations were lethal (41%) [42]. Similarly, a study by Rihn *et al* [43] of the HIV capsid found that 70% of non-synonymous mutations were lethal, which corresponds to around 47% of all mutations [43]. Two studies on bacteriophage found somewhat lower percentages [44, 45], as did a study on polio virus [16].

Several factors could explain why we found a lower percentage of lethal mutations than most *in vitro* studies. First, we only looked at transition mutations, whereas transversions may be more frequently lethal, as they are more often non-synonymous, more likely to lead to drastic amino acid changes, and more likely to create premature stop codons, due to the nature of the genetic code. Second, sequencing, cloning or recording errors may obscure our results. Many low-frequency variants in our data set were only observed once, and it is possible that some of these were not true variants; we may thus have underestimated the percentage of lethal mutations. Third, we looked only at one gene, and this gene may have a different fitness landscape than other parts of the viral genome. Finally, different environments (*in vitro* vs *in vivo*) or different genetic backgrounds (usually one genetic background in the *in vitro* studies vs many in *in vivo* studies) may explain the observed differences. Future studies with more sequences and more sites will have better power to determine the true proportion of lethal mutations in HIV *in vivo*.

Factors associated with high fitness costs: nucleotide, amino acid and CpG effects

Some of our results were surprising. As can be seen in Figure 2, G→A mutations appear to be two-and-a-half to three-and-a-half times more costly than A→G or T→C mutations. In addition, C→T mutations appear to be particularly costly if they are non-synonymous (six times more costly than A→G mutations), but not when they are synonymous. It is difficult to determine the cause of this ancestral nucleotide effect. It could be an artifact caused by spurious mutation rate estimates. However, mutation rate estimates from two very different studies, Abram *et al* [24] and Zanini *et al* [26], are very similar, and using one or the other did not change our finding. Alternatively, this effect may be related to the activity of APOBEC3 enzymes, which hypermutate the HIV genome, leading to an increased proportion of G→A mutations [46–48]. Perhaps the G→A mutations we observed in the pol gene occurred alongside G→A mutations at other regions in the genome (that we did not observe), leading to a higher overall fitness cost. The effect might also be related to a strong mutation bias in the HIV genome. G→A mutations are three to five times more common than A→G mutations [24, 26], which may have led, over long evolutionary timescales, to the well known A bias in the HIV genome [49, 50]. Specifically, sites at which having an A or G does not affect viral fitness would become A-biased over time. Thus, A sites would be enriched for (nearly) neutral sites, and G sites would be depleted of neutral sites, which could lead to G→A mutations being more costly, on average, than A→G mutations. Finally, the effect may be partially due to the specific amino acids involved. As can be seen in Figure 3, most very deleterious mutations are concentrated in just a few specific amino acid changes (especially mutations away from glycine and proline). Studies on larger parts of the genome, or on transversions in addition to transitions, are needed to disentangle the nucleotide and amino acid effects.

Another factor that greatly affected the cost of a mutation in this study was the formation of CpG sites. Depending on the type of site, we found that CpG-creating mutations are between one-and-a-half and four times more costly than equivalent mutations that do not create CpG sites. CpG sites are found very rarely in RNA viruses [51] and are strongly selected against in a wide range of viral genomes [34]. The results of several recent studies suggest that CpG sites trigger the host's antiviral cellular response [33–35]. Viral transcripts with CpGs may be more easily detected by the host immune system, as CpG sites in animals are usually found in gene promoters and are therefore rarely transcribed [52]. Indeed, introducing additional CpG sites in the viral genome strongly decreases a virus's replication rate and hence its fitness, as has been shown experimentally in two different mammalian viruses [33, 35]. Given the increasing evidence that CpG site are deleterious for viral genomes, future efforts should be geared toward discovering the molecular mechanism responsible for anti-CpG selection, which most likely differs from the mechanism present in eukaryotic cells.

Study limitations

One limitation of our study is that we only focused on a small part of the HIV genome, namely the 870 sites of the pol gene for which the best data were available [40]. Because the patients in the Bacheler *et al* study were treated with a variety of antiviral treatments, we had to exclude drug resistance mutations, as they would have been under positive selection in at least some of the patients. To study the costs of resistance mutations, it would be necessary to analyze samples from untreated patients. A second limitation is that we focused only on transitions and excluded transversions. Transversions are observed less often because their mutation rates are lower [24, 26]. When deeper and more precise data on transversions in HIV become available, we expect the DFE to shift toward more costly mutations. A third limitation is that we assumed one mutation rate for all A→G mutations, and one rate for all C→T mutations, etc. However, some evidence exists that mutation rates are highly variable, which would mean that selection coefficient estimates for individual mutations may be unreliable [53].

Another limitation of our study is that it is unknown how long the patients in our data set were infected before samples were collected. If samples were taken soon after infection, genetic diversity in the viral population may have been low. Most patients are infected with one or a small number of founder viruses, and therefore genetic diversity within the host is initially low. Over time, genetic diversity accumulates [54, 55] before plateauing after several years. Early samples therefore carry less information than later samples, because mutant frequencies in early samples are expected to be close to 0 (if the founder virus is wild type at a position) or close to 1 (if the founder virus is mutant at the position). However, if founder viruses are a random sample from the viral population within the individual who transmits the infection, then the *average* mutant frequency across many samples from newly infected patients should actually be the same as the average mutant frequency across many samples from patients with longer-term infections. For example, if the average frequency of a mutant in all patients in the epidemic is 1%, then we would expect 1% of founder viruses to be mutant. Of 100 newly infected patients, we would expect 1 to have a mutant frequency of 100% and 99 to have a frequency of 0%, leading to a 1% average frequency among newly infected patients. The variance of such an average frequency, however, is expected to be very high, because each sample has a frequency of either 0% or 100%. Thus, if frequencies in founder viruses and within-patient frequencies are similar, then using early samples should lead to unbiased estimates of frequencies, but the variance of such estimates may be very high. It would therefore be better to work with samples from later in infection. Conversely, we know that over time, within-patient viral populations diverge from their founder virus [38,54], and it may be that certain mutants cannot establish an infection, but do occur later on in infections and reach fairly high frequencies. If this is true, then founder viruses may not represent a random sample of later within-host viruses. Samples from recently infected patients might thus possess genetic diversity distinct from that of samples from patients with long-term infections. Thus, it is unclear at present whether the timing of sample collection might affect our results. In a future study, it may be possible to compare early and late samples to determine whether such an effect exists.

Strengths of our study and future directions

We used a new approach to study costs of individual mutations based on average within-patient mutant frequencies. HIV is especially well suited for this approach because the genetic diversity within each patient accumulates quickly and independently from HIV populations in other patients. Due to HIV's high mutation rate and the large number of patients in this study (160), 94.1% of all possible 870 transition mutations (818) were observed in at least one patient.

Only 5.9% of possible mutations (51) were never observed; more than half of these (28) were nonsense mutations expected to be lethal and thus swiftly removed by purifying selection.

The other mutations that were never observed (23) were non-synonymous, and most of them (20) led to a drastic amino acid change, making it quite likely that they are very costly and that we did not observe them because they only reach very low frequencies. Thus, not only were we able to estimate fitness costs *in vivo*, we were also able to estimate costs for many more mutations than can typically be done in site-directed mutagenesis studies [41]. Because we were able to study a large number of mutations, it was possible to determine how characteristics of mutations affected their costs in much more detail then has previously been possible. This, in turn, allowed us to quantify the effects of drastic amino acid changes, the creation of new CpG dinucleotides and the ancestral nucleotide at a site.

The current study should be seen as a proof of concept. Our results demonstrate the power of analyzing mutant frequencies from *in vivo* viral populations to study the fitness effects of mutations. We expect that this method will soon be applied to the entire HIV genome and the genomes of other fast-evolving microbes. For HIV specifically, we expect that patient samples with high viral loads will soon be sequenced much more deeply than in any of the studies analyzed in this article. Transversion mutations can then be analyzed in addition to transition mutations. Such a data set will allow us to get a more fine-grained and precise picture of the costs of mutations at individual sites across the entire HIV genome, including for mutations in other genes and non-coding regions of the virus and for drug resistance mutations in pol and elsewhere. Because our method makes it possible to estimate *in vivo* costs, the results will contribute to our understanding of drug resistance evolution and immune escape and may also contribute to vaccine design.

5 Methods

Description of the data/filtering

We used sequences from a data set collected by Bacheler *et al.* [40]. This study focused on patients in three clinical trials of different treatments, all based on efavirenz (a non-nucleoside RT inhibitor) in combination with NRTIs (nucleoside RT inhibitors) and/or protease inhibitors. The treatments in this study were not very effective, in part because some patients were initially prescribed monotherapy, which almost always leads to drug resistance, and in part because patients had previously been treated with some of the drugs, so their viruses were already resistant to some components of the treatment. The result was that viral loads in these patients were typically not suppressed, which made it possible to sequence samples even during therapy. We have previously used part of this data set to study soft and hard selective sweeps [30].

The Bacheler *et al.* [40] samples were cloned and Sanger-sequenced, leading to sequences with a negligible error rate. For each patient, all available pol sequences were treated as one sample, even when they came from different time points. Patients with only a single sequence were excluded from the analysis, leaving us with a median of 19 sequences per patient (3,572 sequences in total). Sequences were 984 nucleotides long and were composed of the 297 nucleotides that encode the HIV protease protein and the 687 that encode RT. We excluded 75 drug resistance–related sites [32] and 39 protease sites that overlap with gag, leaving 287 synonymous, 555 non-synonymous and 28 nonsense mutations, for a total of 870 sites.

Sequences were retrieved from Genbank under accession numbers AY000001 to AY003708. The Lehman data set [37] was downloaded from the NCBI website using accession number SRP049715 (www.ncbi.nlm.nih.gov/sra/?term=SRP049715). The Zanini data [38] was accessed through the website of the Neher laboratory (http://hiv.tuebingen.mpg.de/data/).

Calculation of mutation frequencies

To identify mutations, we compared the sequences to the consensus HIV-1 subtype B reference sequence, available at www.hiv.lanl.gov/content/sequence/HIV/CONSENSUS/Consensus. Occasionally, mutations had a frequency of (near) 100% (see the example of a synonymous mutation in Fig. 1A). We still included these observations in our analysis, assuming that the high frequency was due to a strong bottleneck (e.g., at infection) or a selective sweep at a nearby resistance mutation, and not due to positive selection on the mutation itself. Bottlenecks and selective sweeps increase the variance of mutation frequencies, but not the expected value of the frequencies [13].

We only considered transition mutations (A↔G and C↔T), excluding transversion mutations, which are less common. For example, for a site with an A in the ancestral state (according to the subtype B consensus sequence), the frequency of a transition mutation was calculated for each patient as the number of sequences with a G divided by the number of sequences with a G or an A. These frequencies were then averaged over all 160 patients. Sequences with a C or a T were thus not considered at all if the ancestral state was an A. In addition, if in a given sequence there was more than one mutation in a given triplet, this triplet was removed for that specific sequence, so that all mutations could be clearly classified non-synonymous or synonymous. Occasionally this meant that a sample from a patient had to be excluded for a given site, so for some mutations we had fewer than 160 frequencies to analyze.

Selection coefficients were estimated for each mutation by dividing the nucleotide-specific mutation rate by the observed average frequency (based on the mutation-selection balance formula $f = u/s$). We used mutation rates as estimated by Abram *et al.* [24], and in a few cases we compared the outcomes using Abram *et al.* [24] with those obtained using the mutation rates from Zanini *et al.* [26].

Generalized linear model analysis

Using a GLM, we predicted mutant frequencies for certain categories of mutations (e.g., synonymous, non-CpG-forming, A↔G) and then used the mutation-selection formula ($f = u/s$) to predict the costs of these groups of mutations (see Fig. 2). Specifically, we fit a binomial GLM to model the state (ancestral or mutant) of each position based on the nucleotide in the consensus sequence, its SHAPE value, whether or not the position was in the RT protein and the types of changes resulting from a transition at that position. These changes included whether a transition was non-synonymous, changed the amino acid group (i.e., between the positive-charged, negative-charged, uncharged and hydrophobic groups, or to or from the special amino acids: cysteine, selenocysteine, glycine and proline) or formed a new CpG site. We also fit interactions between the ancestral nucleotides, whether a transition was non-synonymous, and whether the transition formed a CpG site. For the GLM, actual counts were considered as opposed to frequencies (i.e., if we had 20 sequences for patient 1, and at a given nucleotide, we observed 18 As and 2 Gs, we used those counts). The count approach automatically gives more weight to the patients for whom we have more sequences. Each position in each sequence from each patient was treated as an independent observation.

To explicitly test whether two categories of mutations with different mutation rates had different selection coefficients, we used a one-sided two-sample Wilcoxon test (also known as a Mann-Whitney test). This was necessary because a GLM can only test whether a mutant of a certain category is more likely to be present than a mutant of another category. We were interested, however, in whether a mutant of a certain category is more costly than a mutant of another category. For example, synonymous C↔T mutations are more common than synonymous, non-CpG forming A↔G mutations (see Figure

S4), but this difference can be explained entirely by the higher mutation rate of C↔T mutations, so the estimated selection coefficients are not different (see Figure 2). Note that C↔T and G↔A mutations cannot create new CpG sites. We did not include amino acid position in the statistical analysis because there does not seem to be any effect of position, as can be seen in Supplementary Figure S3.

Estimating a gamma distribution to fit the distribution of fitness effects

We fit a gamma distribution to the DFE (based directly on averaged frequencies and the mutation-selection balance formula $f = u/s$). Transitions that were never observed (frequency of 0) and mutations at resistance-related sites were not considered when fitting the gamma distribution. The most likely shape and scale parameters for the data were found using the subplex algorithm implemented in the R package nloptr [56]. Bootstrapped confidence intervals were created by resampling the data with replacement and re-estimating the gamma distribution parameters. Selection coefficients were estimated using the mutations rates given in Abram *et al.* [24] and Zanini *et al.* [26].

Comparison with the global epidemic

A large HIV-1 sequence data set was retrieved from the HIVdb (http://hivdb.stanford.edu/pages/geno-rx-datasets.html) [39]. This data set contains a single sequence per patient. Protease and RT sequences were downloaded in separate files. Sequences that met the following criteria were included in the analysis: treatment-naive host status and classification as HIV-1 subtype B. In total, 23,742 protease and 22,785 RT sequences were collected. Average mutation frequencies for each site were calculated as explained above (e.g., including only transitions, excluding triplets with more than one mutation). Spearman's rank correlation coefficient (ρ) was used to quantify the correlation between within-patient and global mutation frequencies.

Additional data sets

In order to test, how transferable our method is, we repeated parts of our analysis with the Lehman *et al.* data set [37] and the Zanini *et al.* data set [38].

The Lehman samples were 454-sequenced, but the resulting sequences exhibited a very high error rate. The samples were collected at seroconversion and one month later, but we only included the time point one month after seroconversion in our analysis, as we expected that the samples from the earliest time point would contain almost no genetic diversity. The sequences span approximately 600 sites in the RT protein. Since the Lehman data [37] also contained HIV subtypes C and A, we only considered sites that were conserved between subtypes A, B and C (621 sites).

The Zanini [38] samples came from nine patients. There were multiple samples per patient (72 samples in total), typically collected at least a few months apart. Thus we followed Zanini *et al* in treating those samples as if they were completely independent. The sequencing method used was Illumina. We downloaded mutation frequencies for each sample (http://hiv.tuebingen.mpg.de/data/) and averaged frequencies across all 72 samples. The Zanini data cover the whole HIV genome, but we only considered the regions that overlap with the Bacheler data [40]. In addition, the Zanini data [38] contain sequences for different HIV subtypes (B, C and CRF01-AE); we only considered sites that were conserved between subtypes B, C and CRF01-AE (758 sites).

6 Acknowledgments

The authors wish to thank Dmitri Petrov, Arbel Harpak, David Enard, Alan Bergland and Ryan Taylor for helpful discussions; Richard Neher and Fabio Zanini for comments on an earlier version of the manuscript; and Scott Roy for help aligning the Lehman *et al* sequences.

References

[1] E. Batschelet, E. Domingo, and C. Weissmann, "The proportion of revertant and mutant phage in a growing population, as a function of mutation and growth rate," *Gene*, vol. 1, no. 1, pp. 27–32, 1976.

[2] E. Domingo, D. Sabo, T. Taniguchi, and C. Weissmann, "Nucleotide sequence heterogeneity of an RNA phage population," *Cell*, vol. 13, no. 4, pp. 735–744, 1978.

[3] M. Eigen, "Viral quasispecies," *Scient Am*, vol. 269, pp. 32–32, 1993.

[4] I. M. Rouzine, A. Rodrigo, and J. Coffin, "Transition between stochastic evolution and deterministic evolution in the presence of selection: general theory and application to virology," *Microbiol Mol Biol Rev*, vol. 65, no. 1, pp. 151–185, 2001.

[5] C. O. Wilke, "Quasispecies theory in the context of population genetics," *BMC Evol Biol*, vol. 5, no. 1, p. 44, 2005.

[6] C. K. Biebricher and M. Eigen, "What is a quasispecies?," in *Quasispecies: Concept and Implications for Virology*, pp. 1–31, Springer, 2006.

[7] A. S. Lauring and R. Andino, "Quasispecies theory and the behavior of RNA viruses," *PLoS Pathog*, vol. 6, no. 7, p. e1001005, 2010.

[8] P. S. Pennings, "Standing genetic variation and the evolution of drug resistance in HIV," *PLoS Comput Biol*, vol. 8, no. 6, p. e1002527, 2012.

[9] R. Paredes, C. M. Lalama, H. J. Ribaudo, B. R. Schackman, C. Shikuma, F. Giguel, W. A. Meyer, V. A. Johnson, S. A. Fiscus, R. T. D'Aquila, R. M. Gulick, and D. R. Kuritzkes, "Pre-existing minority drug-resistant HIV-1 variants, adherence, and risk of antiretroviral treatment failure," *J Infect Dis*, vol. 201, pp. 662–671, 03 2010.

[10] J. Z. Li, R. Paredes, H. J. Ribaudo, E. S. Svarovskaia, K. J. Metzner, M. J. Kozal, K. H. Hullsiek, M. Balduin, M. R. Jakobsen, A. M. Geretti, *et al.*, "Low-frequency HIV-1 drug resistance mutations and risk of NNRTI-based antiretroviral treatment failure: a systematic review and pooled analysis," *JAMA*, vol. 305, no. 13, pp. 1327–1335, 2011.

[11] R. A. Neher and T. Leitner, "Recombination rate and selection strength in HIV intra-patient evolution," *PLoS Comput Biol*, vol. 6, no. 1, p. e1000660, 2010.

[12] R. Batorsky, M. F. Kearney, S. E. Palmer, F. Maldarelli, I. M. Rouzine, and J. M. Coffin, "Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection," *PNAS*, vol. 108, no. 14, pp. 5661–5666, 2011.

[13] D. L. Hartl, A. G. Clark, and A. G. Clark, *Principles of Population Genetics*, vol. 116. Sinauer Associates, Sunderland, MA, 1997.

[14] M. V. Trotter, "Mutation–selection balance," *eLS*, 2014.

[15] R. Sanjuán, A. Moya, and S. F. Elena, "The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus," *PNAS*, vol. 101, no. 22, pp. 8396–8401, 2004.

[16] A. Acevedo, L. Brodsky, and R. Andino, "Mutational and fitness landscapes of an RNA virus revealed through population sequencing," *Nature*, vol. 505, no. 7485, pp. 686–690, 2014.

[17] B. Thyagarajan and J. D. Bloom, "The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin," *Elife*, p. e03300, 2014.

[18] T. Hinkley, J. Martins, C. Chappey, M. Haddad, E. Stawiski, J. M. Whitcomb, C. J. Petropoulos, and S. Bonhoeffer, "A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase," *Nature Genet*, vol. 43, no. 5, pp. 487–489, 2011.

[19] D. S. Lawrie and D. A. Petrov, "Comparative population genomics: power and principles for the inference of functionality," *Trends Genet*, vol. 30, no. 4, pp. 133–139, 2014.

[20] A. Eyre-Walker and P. D. Keightley, "The distribution of fitness effects of new mutations," *Nat Rev Genet*, vol. 8, no. 8, pp. 610–618, 2007.

[21] I. Mayrose, A. Stern, E. O. Burdelova, Y. Sabo, N. Laham-Karam, R. Zamostiano, E. Bacharach, and T. Pupko, "Synonymous site conservation in the HIV-1 genome," *BMC Evol Biol*, vol. 13, no. 1, p. 1, 2013.

[22] J. D. Roberts, K. Bebenek, and T. A. Kunkel, "The accuracy of reverse transcriptase from HIV-1," *Science*, vol. 242, no. 4882, pp. 1171–1173, 1988.

[23] L. M. Mansky and H. M. Temin, "Lower *in vivo* mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase.," *J Virol*, vol. 69, no. 8, pp. 5087–5094, 1995.

[24] M. E. Abram, A. L. Ferris, W. Shao, W. G. Alvord, and S. H. Hughes, "Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication," *J Virol*, vol. 84, no. 19, pp. 9864–9878, 2010.

[25] J. M. Cuevas, R. Geller, R. Garijo, J. López-Aldeguer, and R. Sanjuán, "Extremely high mutation rate of HIV-1 *in vivo*," *PLoS Biol*, vol. 13, no. 9, p. e1002251, 2015.

[26] F. Zanini, V. Puller, J. Brodin, J. Albert, and R. Neher, "*In-vivo* mutation rates and fitness landscape of HIV-1," *arXiv preprint arXiv:1603.06634*, 2016.

[27] J. M. Coffin, "HIV population dynamics *in vivo*: implications for genetic variation, pathogenesis, and therapy," *Science*, vol. 267, no. 5197, pp. 483–489, 1995.

[28] J. M. Coffin, S. H. Hughes, H. E. Varmus, J. Boeke, and J. Stoye, *Retrotransposons, endogenous retroviruses, and the evolution of retroelements*. Cold Spring Harbor Laboratory Press, 1997.

[29] D. C. Douek, L. J. Picker, and R. A. Koup, "T cell dynamics in HIV-1 infection*," *Annu Rev Immunol*, vol. 21, no. 1, pp. 265–304, 2003.

[30] P. S. Pennings, S. Kryazhimskiy, and J. Wakeley, "Loss and recovery of genetic diversity in adapting populations of HIV," *PLoS Genet*, vol. 10, no. 1, p. e1004000, 2014.

[31] S. Karlin, "A first course in stochastic processes," *Elsevier*, 2014.

[32] V. A. Johnson, V. Calvez, H. F. Günthard, R. Paredes, D. Pillay, R. Shafer, A. M. Wensing, and D. D. Richman, "2011 update of the drug resistance mutations in hiv-1," *HIV Med*, vol. 18, pp. 156–163, 2010.

[33] C. C. Burns, R. Campagnoli, J. Shaw, A. Vincent, J. Jorba, and O. Kew, "Genetic inactivation of poliovirus infectivity by increasing the frequencies of CpG and UpA dinucleotides within and across synonymous capsid region codons," *J Virol*, vol. 83, no. 19, pp. 9957–9969, 2009.

[34] X. Cheng, N. Virk, W. Chen, S. Ji, S. Ji, Y. Sun, and X. Wu, "CpG usage in RNA viruses: data and hypotheses," *PloS One*, vol. 8, no. 9, p. e74109, 2013.

[35] N. J. Atkinson, J. Witteveldt, D. J. Evans, and P. Simmonds, "The influence of CpG and UpA dinucleotide frequencies on RNA virus replication and characterization of the innate cellular pathways underlying virus attenuation and enhanced replication," *Nucleic Acids Res*, vol. 42, no. 7, pp. 4527–4545, 2014.

[36] J. M. Watts, K. K. Dang, R. J. Gorelick, C. W. Leonard, J. W. Bess Jr, R. Swanstrom, C. L. Burch, and K. M. Weeks, "Architecture and secondary structure of an entire HIV-1 RNA genome," *Nature*, vol. 460, no. 7256, pp. 711–716, 2009.

[37] D. A. Lehman, J. M. Baeten, C. O. McCoy, J. F. Weis, D. Peterson, G. Mbara, D. Donnell, K. K. Thomas, C. W. Hendrix, M. A. Marzinke, *et al.*, "Risk of drug resistance among persons acquiring hiv within a randomized clinical trial of single-or dual-agent preexposure prophylaxis," *J Infect Dis*, p. jiu677, 2015.

[38] F. Zanini, J. Brodin, L. Thebo, C. Lanz, G. Bratt, J. Albert, and R. A. Neher, "Population genomics of intrapatient HIV-1 evolution," *eLife*, vol. 4, p. e11282, 2016.

[39] S.-Y. Rhee, M. J. Gonzales, R. Kantor, B. J. Betts, J. Ravela, and R. W. Shafer, "Human immunodeficiency virus reverse transcriptase and protease sequence database," *Nucleic Acids Res*, vol. 31, no. 1, pp. 298–303, 2003.

[40] L. T. Bacheler, E. D. Anton, P. Kudish, D. Baker, J. Bunville, K. Krakowski, L. Bolling, M. Aujay, X. V. Wang, D. Ellis, *et al.*, "Human immunodeficiency virus type 1 mutations selected in patients failing efavirenz combination therapy," *Antimicrob Agents Chemother*, vol. 44, no. 9, pp. 2475–2484, 2000.

[41] R. Sanjuán, "Mutational fitness effects in RNA and single-stranded DNA viruses: common patterns revealed by site-directed mutagenesis studies," *Philos Trans R Soc Lond B Biol Sci*, vol. 365, no. 1548, pp. 1975–1982, 2010.

[42] P. Carrasco, F. de la Iglesia, and S. F. Elena, "Distribution of fitness and virulence effects caused by single-nucleotide substitutions in tobacco etch virus," *J Virol*, vol. 81, no. 23, pp. 12979–12984, 2007.

[43] S. J. Rihn, S. J. Wilson, N. J. Loman, M. Alim, S. E. Bakker, D. Bhella, R. J. Gifford, F. J. Rixon, and P. D. Bieniasz, "Extreme genetic fragility of the HIV-1 capsid," *PLoS Pathog*, vol. 9, no. 6, p. e1003461, 2013.

[44] P. Domingo-Calap, J. M. Cuevas, and R. Sanjuán, "The fitness effects of random mutations in single-stranded DNA and RNA bacteriophages," *PLoS Genet*, vol. 5, no. 11, p. e1000742, 2009.

[45] J. B. Peris, P. Davis, J. M. Cuevas, M. R. Nebot, and R. Sanjuán, "Distribution of fitness effects caused by single-nucleotide substitutions in bacteriophage f1," *Genetics*, vol. 185, no. 2, pp. 603–609, 2010.

[46] A. M. Sheehy, N. C. Gaddis, J. D. Choi, and M. H. Malim, "Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein," *Nature*, vol. 418, no. 6898, pp. 646–650, 2002.

[47] K.-M. Chen, E. Harjes, P. J. Gross, A. Fahmy, Y. Lu, K. Shindo, R. S. Harris, and H. Matsuo, "Structure of the DNA deaminase domain of the HIV-1 restriction factor APOBEC3G," *Nature*, vol. 452, no. 7183, pp. 116–119, 2008.

[48] L. G. Holden, C. Prochnow, Y. P. Chang, R. Bransteitter, L. Chelico, U. Sen, R. C. Stevens, M. F. Goodman, and X. S. Chen, "Crystal structure of the anti-viral APOBEC3G catalytic domain and functional implications," *Nature*, vol. 456, no. 7218, pp. 121–124, 2008.

[49] F. J. van Hemert, A. C. van der Kuyl, and B. Berkhout, "The A-nucleotide preference of HIV-1 in the context of its structured RNA genome," *RNA Biol*, vol. 10, no. 2, pp. 211–215, 2013.

[50] F. van Hemert, A. C. van der Kuyl, and B. Berkhout, "On the nucleotide composition and structure of retroviral RNA genomes," *Virus Res*, vol. 193, pp. 16–23, 2014.

[51] B. K. Rima and N. V. McFerran, "Dinucleotide and stop codon frequencies in single-stranded RNA viruses.," *J Gen Virol*, vol. 78, no. 11, pp. 2859–2870, 1997.

[52] A. C. van der Kuyl and B. Berkhout, "The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus," *Retrovirology*, vol. 9, no. 1, pp. 1–14, 2012.

[53] R. Geller, Ú. Estada, J. B. Peris, I. Andreu, J.-V. Bou, R. Garijo, J. M. Cuevas, R. Sabariegos, A. Mas, and R. Sanjuán, "Highly heterogeneous mutation rates in the hepatitis C virus genome," *Nat Microbiol*, p. 16045, 2016.

[54] R. Shankarappa, J. B. Margolick, S. J. Gange, A. G. Rodrigo, D. Upchurch, H. Farzadegan, P. Gupta, C. R. Rinaldo, G. H. Learn, X. He, *et al.*, "Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection," *J Virol*, vol. 73, no. 12, pp. 10489–10502, 1999.

[55] R. D. Kouyos and H. F. Günthard, "The irreversibility of HIV drug resistance," *Clin Infect Dis*, p. civ400, 2015.

[56] S. G. Johnson, "The NLopt nonlinear-optimization package," *(R package)*, 2008.

7 Supplementary Figures

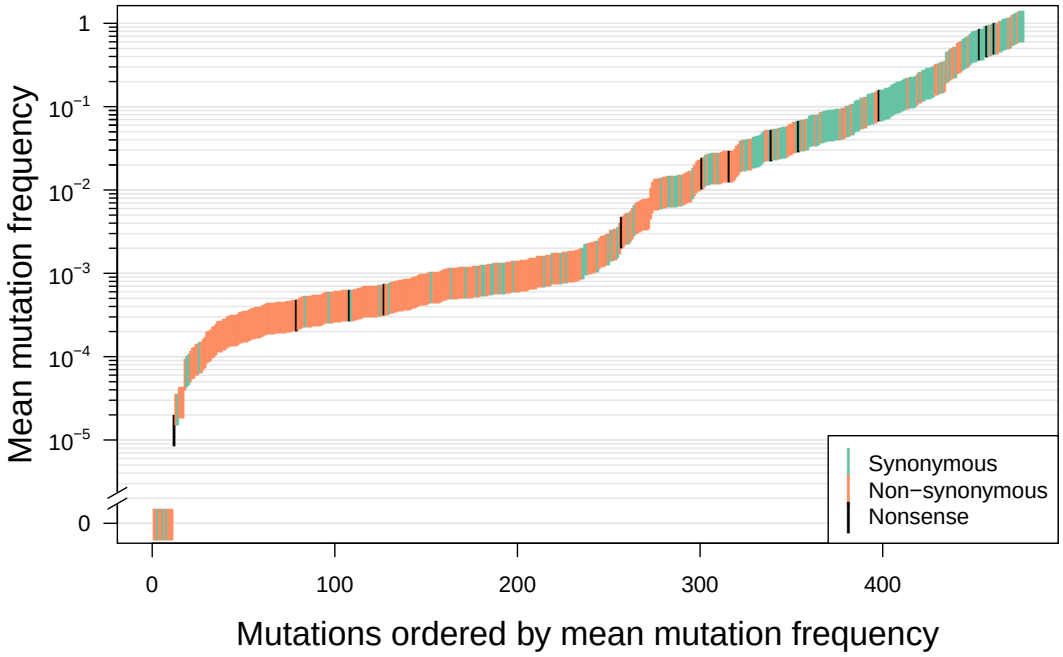

Figure S1: Mutation frequency for 621 reverse transcriptase sites from the Lehman data set [37], ordered by mutation frequency.

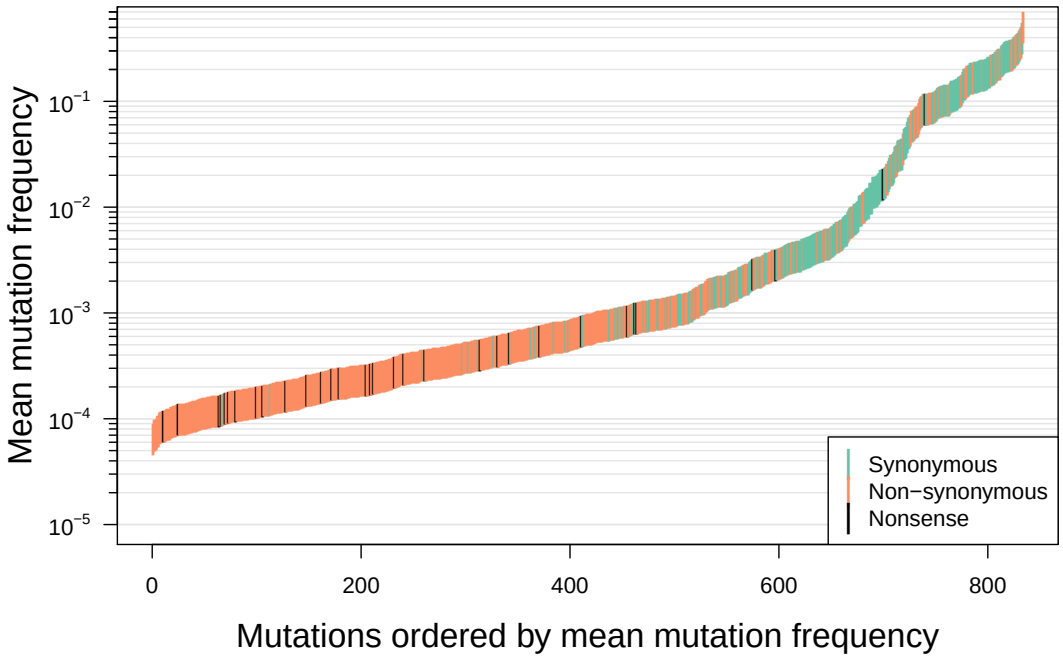

Figure S2: Mutation frequency for 758 pol sites from the Zanini data set [38], ordered by mutation frequency.

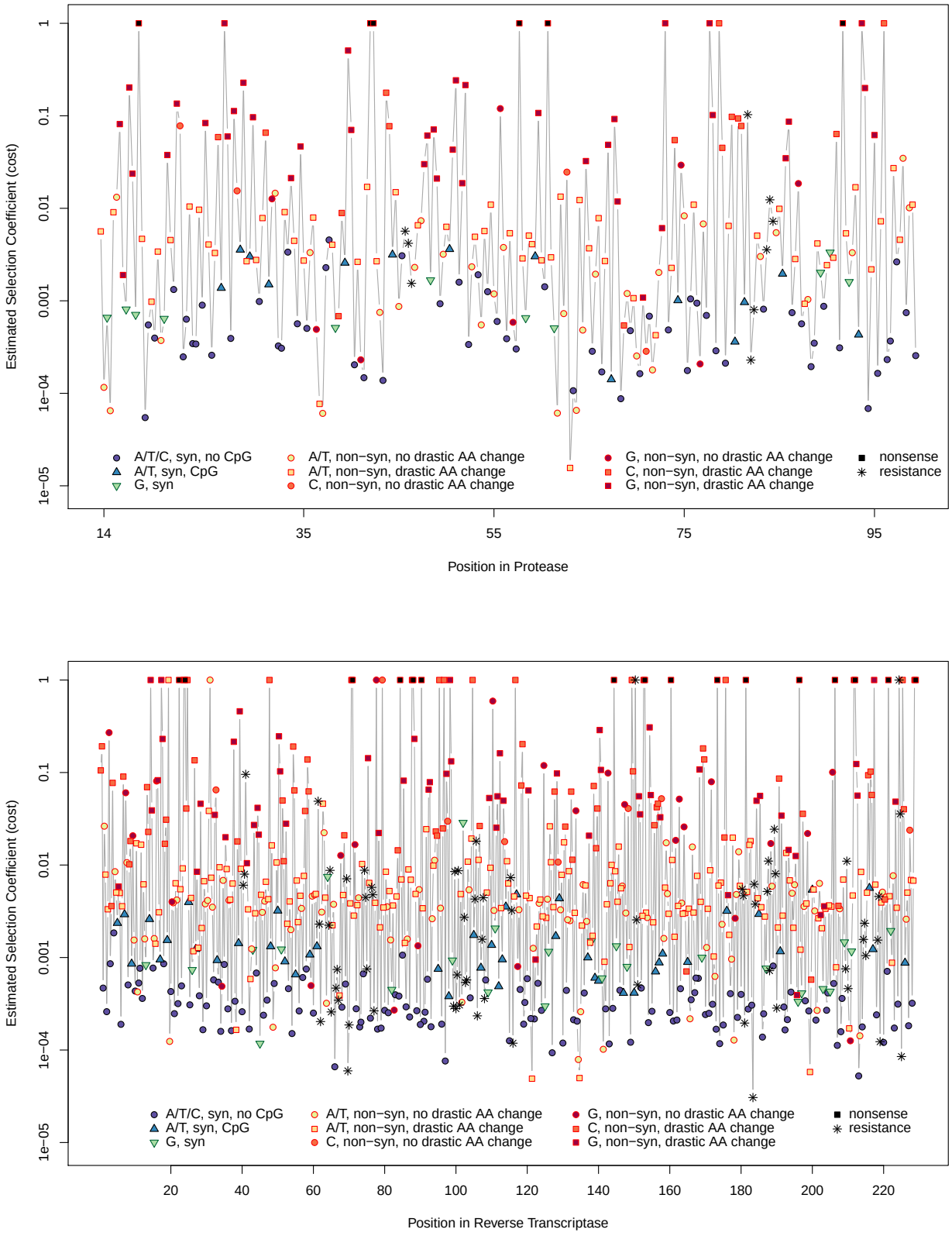

Figure S3: Estimated selection coefficients for each transition mutation along the protease (*top panel*) and reverse transcriptase (*bottom panel*) proteins.

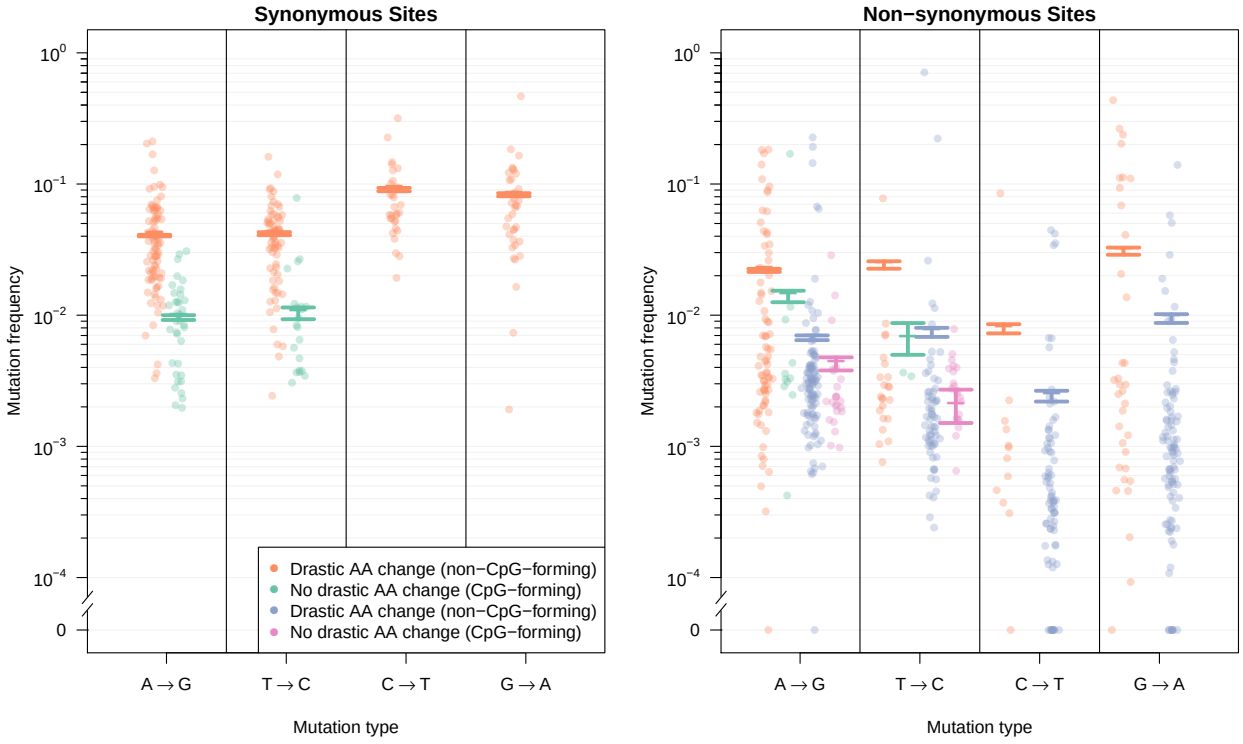

Figure S4: Mutation frequencies, estimated using a generalized linear model and the Bacheler data set [40], which formed the basis of this paper. The graph shows the model predictions for synonymous and non-synonymous mutations that do not involve a drastic amino acid change and either form CpG sites (*green*) or do not (*orange*). In addition, for non-synonymous mutations, predictions are shown for mutations that do involve a drastic amino acid change and either form CpG sites (*pink*) or do not (*blue*).

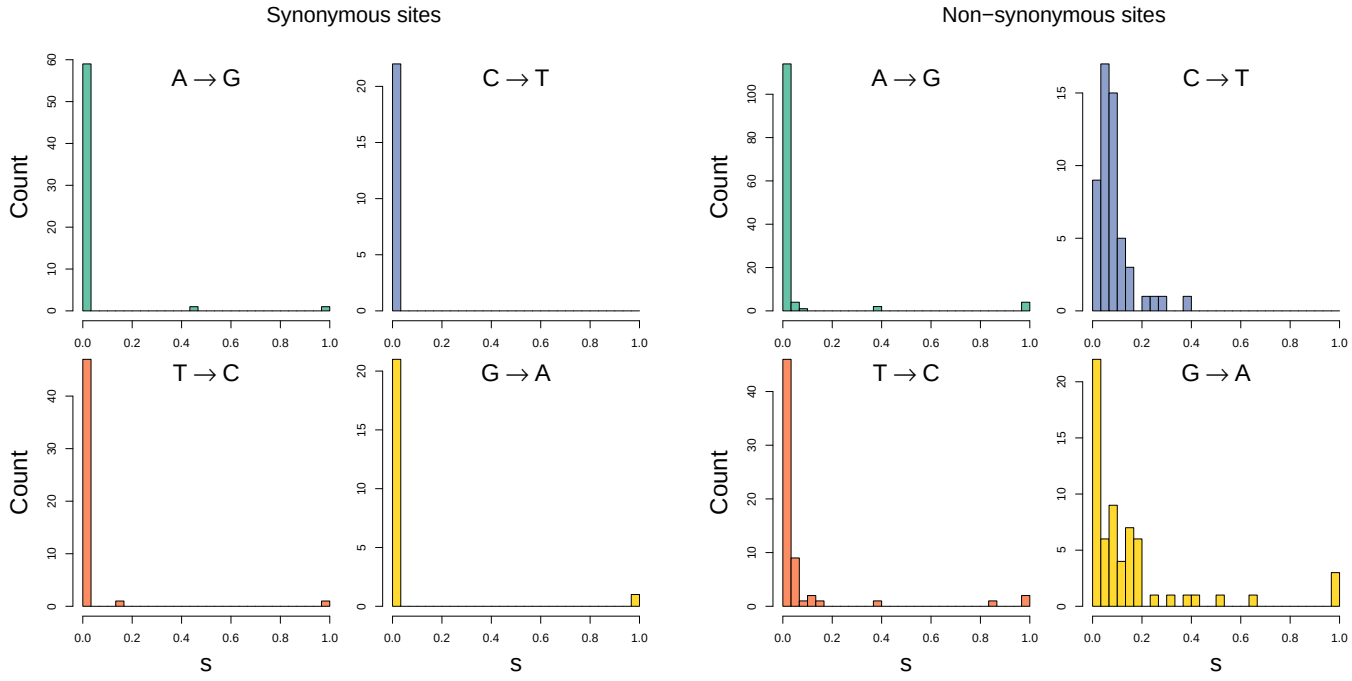

Figure S5: Distribution of fitness effects for non-synonymous and synonymous reverse transcriptase mutations from the Lehman data set [37]; nonsense mutations are included in the non-synonymous mutation category. Note that the scales of the x- and y-axis differ between the graphs.

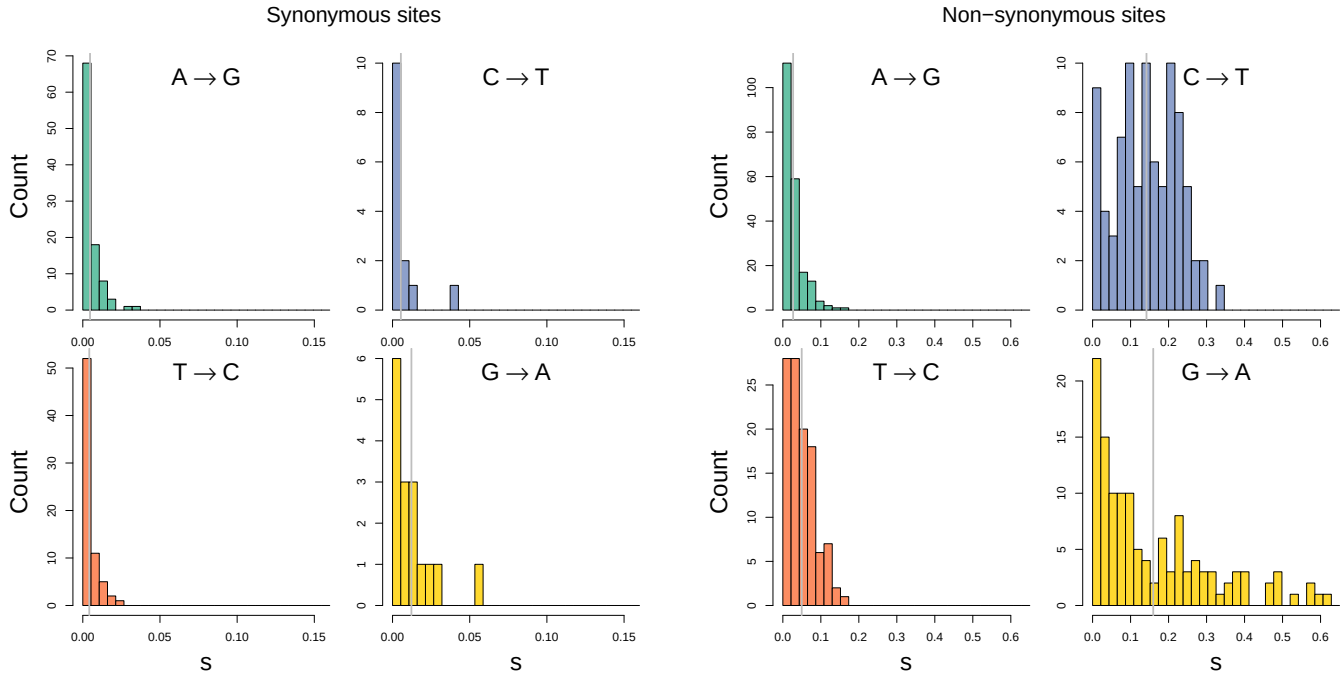

Figure S6: Distribution of fitness effects for non-synonymous and synonymous mutations for the Zanini data set [38]; nonsense mutations are included in the non-synonymous mutations. Note that the scales of the x- and y-axis differ between the graphs.

Table S1: Table with gamma distribution parameters reflecting scale (κ) and shape (θ) for Bacheler [40], Zanini [38] and Lehman [37] data sets.

	Num. sites	Mut. rates from Abrahm 2010 κ	θ	Mut. rates from Zanini 2016 κ	θ	Proportion lethal
Bacheler	870	0.31	0.211	0.315	0.243	0.059
		(0.222, 0.4)	(0.203, 0.223)	(0.244, 0.399)	(0.233, 0.256)	(0.044, 0.076)
Zanini	758	0.054	0.486	0.144	0.445	0
		(0.048, 0.059)	(0.453, 0.524)	(0.127, 0.162)	(0.417, 0.48)	(0, 0)
Lehman	621	0.173	0.237	0.249	0.258	0.019
		(0.108, 0.251)	(0.217, 0.262)	(0.187, 0.331)	(0.24, 0.281)	(0, 0)