Inferring mutational fitness costs in HIV from within-patient frequencies

Marion Hartl*[†] Kristof Theys *[†] Alison Feder, Maoz Gelbart, Adi Stern, Pleuni S. Pennings

May 2016

1 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 "quasi species"[1, 2, 3, 4, 5, 6, 7]. While mutations are crucial for all adaptive processes, including the evolution of drug resistance and immune escape of HIV, it is well-known that most mutations are associated with fitness costs. 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 (the effect of linked deleterious mutations on neutral or beneficial mutations) and thus affect optimal recombination rates. All of these processes were shown to affect drug resistance and immune escape in HIV [8, 9, 10, 11, 12]. In addition to a better understanding of evolutionary processes, a detailed knowledge of costs of mutations can also help us to discover new functional elements in the HIV genome.

Traditionally, fitness effects are assessed either in *in vitro* systems (cell culture / competition experiments [13, 14, 15]) or in a phylogenetic framework [16, 17, 18]. Here, we use a novel approach based on observed mutation frequencies in HIV-infected patients to determine the fitness effects of mutations. HIV has unique properties that allow us to study fitness effects *in vivo*: it is fast evolving [19, 20, 21, 22, 23] and leads to persistent infections [24, 25, 26]. 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. With data from many replicate populations, frequencies of individual mutations, averaged across populations, can be used to estimate their fitness cost, something that is not possible when only one or a few populations are available.

Costly mutations, which are also referred to as deleterious mutations, occur in populations naturally by mutation and are purged from the population by selection [27, 28]. In infinitely large populations, following the principles of mutation-selection balance, the opposing forces of mutation and selection cause mutations to be 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 the cost of the mutation). In natural populations of finite size, however, the frequency of mutations is not constant, but fluctuates around the expected frequency of u/s, because of the stochastic nature of mutation and replication [27]. Due to these stochastic fluctuations, it is impossible to accurately infer the strength of selection acting on individual mutations (i.e., the cost of mutations) from a single observation of a single population. In HIV however, it is possible to sample many independent populations, since each patient harbors an independent HIV population [29]. The mean frequency of mutations across populations will approach u/s if enough populations are sampled, essentially because the fluctuations of mutation frequencies form an ergodic process [30]. With sufficient sequencing data from a sufficient number of patient samples, we can thus estimate the *in vivo* fitness cost of every point mutation at every position in the HIV-1 genome. In the current study, we focus on transition mutations (A↔G and C↔T) in the first 984 sites of the *pol* gene, which encode for the Protease and part of the the Reverse Transcriptase proteins, because sufficient data are available for these mutations (transition mutations are much more common in HIV than transversions [21]). We exclude all drug resistance related sites from our analysis and expect that positive and balancing selection are less important at the other sites in the *pol* gene, as this is a highly conserved gene that does not come into contact with the immune system [24, 25]. Accordingly, most mutations are expected to be deleterious.

Our analysis is based on observed mutation frequencies in 160 patients who are infected with HIV-1 subtype B, with a median of 19 sequences (range: 2-69) per patient for 984 sites in the *pol* gene [31], after minimal filtering of the data (see Material and Methods). First, we find that there is a very clear separation of observed frequencies of synonymous mutations, non-synonymous mutations and nonsense mutations (i.e., mutations that create premature stop codons). Second, we find that inferred costs of mutations are strongly affected by whether the resulting amino-acid change is drastic or not, i.e., costs are lower if a mutation leads to a change to a similar amino acid. Third, we find that mutations that create new CpG sites are more costly than mutations that do not. This finding may reflect selection against CpG sites because they trigger recognition

*These authors contributed equally and are listed in reverse alphabetical order

[†]Department of Biology, San Francisco State University, San Francisco, CA 94132

by antiviral restriction enzymes [32, 33, 34]. Fourth, we find that there is a difference in fitness cost depending on which of the four nucleotides is changed, most clearly, G→A mutations tend to be more costly than the other mutations. Thus, despite the fact we have analysed only a small part of the HIV genome using a dataset with very limited sequencing depth, we succeeded in recovering and quantifying many known properties of mutational fitness costs, combined with novel findings. Our data also allow us to estimate parameters of the entire distribution of fitness effects (DFEs), which may be useful for studies of background selection. Finally, we find that within-patient frequencies and global frequencies in the subtype B clade are very similar (Pearson's product-moment correlation 0.95), we discuss the possible reasons for this in the discussion. Note that in parallel to our study, another study was done that also used HIV within-patient mutation frequencies to estimate fitness costs of mutations [23].

2 Results

Mean frequencies of non-synonymous site mutations are lower than mean frequencies of synonymous site mutations

We began by examining the frequencies of the three main categories of mutations: synonymous, non-synonymous and nonsense mutations. Since we only consider transition mutations, there is exactly one possible mutation per site. The sequences are 984 nucleotides long, but we excluded 75 drug resistance related sites and 39 sites of Protease which overlap with Gag, leaving us with 287 synonymous, 555 non-synonymous and 28 nonsense mutations and 870 sites in total. For each mutation, we have 160 observed frequencies (or fewer if a site is filtered out for certain patients, see Material and Methods). As an example, we show single-site frequency spectra at codon 58 of the Protease protein, composed of nucleotide sites 172 through 174. A transition mutation at the first position (172) creates a premature stop codon. This mutation is never observed in the data and has a frequency of zero. A transition mutation at the second codon position (173) leads to an amino acid change (Glutamine to Arginine) and is found at low frequencies. A mutation at the third position of the codon (174) does not change the amino acid and is observed at higher frequencies, on average, than the other two mutations (Fig1A). We next order all sites according to observed mutation frequencies, which reveals that the pattern is general in the entire dataset. We find that the three main categories of mutations show significantly different distributions of mean frequencies (Kolmogorov-Smirnov test, p-value $< 2.2e-16$ for nonsense vs. non-synonymous sites and for non-synonymous vs. synonymous sites), see figure 1B. Nonsense mutations, as well as a few non-synonymous mutations, are never observed in the data and have a frequency of zero. Most non-synonymous mutations have a lower frequency (80% lower than 0.01) than synonymous mutations (82% higher than 0.01). This difference in distributions likely reflects the higher cost of non-synonymous mutations, which directly affect virus replication. This analysis therefore provides a proof of principle that the observed frequencies reflect the general expectations of costs of broad categories of mutations.

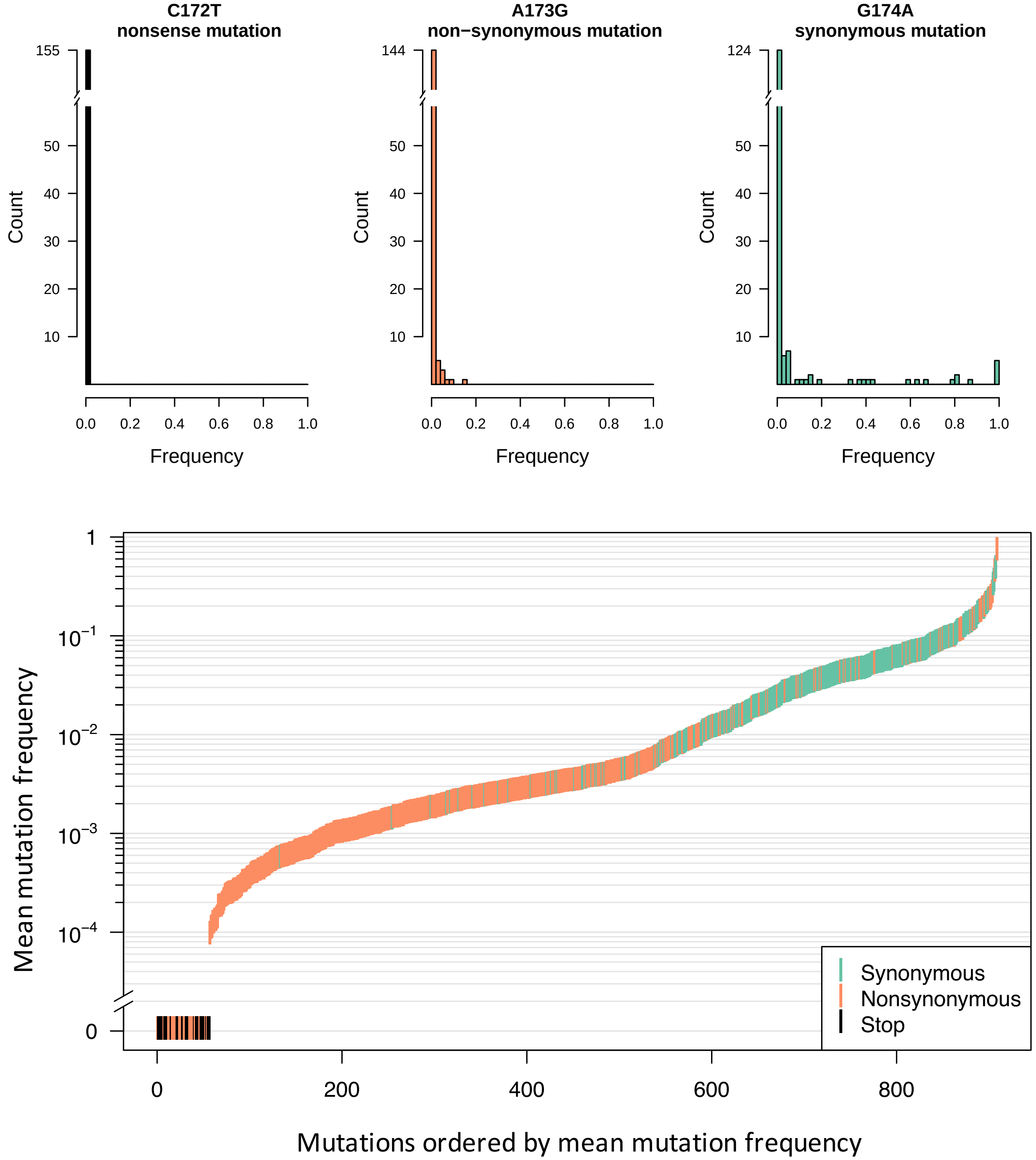

Figure 1: A) The distribution of mutation frequencies plotted for single sites in the Protease protein (site 172 = nonsense mutation, site 173 = non-synonymous mutation, site 174 = synonymous mutation). B) Mean mutation frequencies for all sites are shown ordered by mutation frequency. Nonsense mutations are shown in black, mutations that cause an amino-acid change (non-synonymous) are shown in red and mutations that do not change the amino-acid (synonymous) are shown in green.

Type of amino-acid change and whether a mutation creates a CpG site have major effect on cost of mutations

We used a Generalized Linear Model (GLM) to determine the ensemble of characteristics that can explain the observed frequencies of synonymous and non-synonymous mutations. We then use estimated mutation rates from other studies [21, 35] and the mutation-selection formula ($f = u/s$) to translate observed frequencies into selection coefficients (costs). As in our first result, we find that overall, synonymous mutations are associated with a lower cost than non-synonymous mutations (Fig 2, $p < 0.0001$).

We next sought to determine what creates differences in fitness costs among different synonymous mutations. Interestingly, the strongest effect we found for synonymous mutations was determined by whether or not a mutation creates a new CpG dinucleotide site (see Table 1). CpG sites appear to be costly for RNA viruses, potentially because they trigger the antiviral cellular response [32, 33]. Specifically, we find that A → G mutations and T → C mutations that create new CpG sites are approximately four times more costly (selection coefficient appr. 0.001) than G → A or C → T mutations that do not create new CpG sites (selection coefficient appr. 0.00025, $p < 0.0001$). In addition to a strong effect of creating CpG sites we find that synonymous G → A mutations more costly than other synonymous mutations that do not create CpG sites. Specifically, they are are about two-and-a-halve times as costly as A → G mutations (0.0007 vs 0.00025, $p < 0.0001$) (Fig 2).

For non-synonymous mutations, we made a distinction between mutations that lead to a drastic amino acid (AA) change and mutations that do not lead to a drastic amino acid change. This was based on classical grouping of AAs into five groups (positively charged, negatively charged, uncharged, hydrophobic and special cases (Cysteine, Glycine and Proline)) and on defining a change in group as a drastic change. We find, first, that mutations that lead to a drastic AA change are generally more costly than mutations that do not (three times more costly, $p < 0.0001$). Second, for A → G and T → C sites, there is an effect of whether or not the mutation leads to a CpG site (among mutations that do not lead to a drastic AA change, A → G mutations that create a CpG site are about one-and-a-halve times more costly than A → G mutations that do not create a CpG site, for similar T → C mutations the CpG forming mutations are three-and-a-halve times more costly, both $p < 0.0001$). Third, C → T and G → A mutations tend to be more costly than A → G and T → C mutations (C → T mutations that do not change AA group are six times more costly than A → G mutations that do not change the AA group, nor create a CpG site, similarly, G → A mutations are three-and-a-halve times more costly than A → G mutations, both $p < 0.0001$). Fourth, we found that overall, mutations in Reverse Transcriptase had slightly higher costs than mutations in the Protease ($p < 0.0001$). Finally, fifth, we found a small but significant effect of the SHAPE parameter ($p < 0.0001$), an experimentally determined measure of RNA secondary structure [36]. Specifically, sites with a lower SHAPE parameter (more likely to be part of an RNA structure) were associated with higher mutational costs, presumably because the secondary structure of the RNA molecule plays a functional role in HIV replication [36] (see Table 1).

Because the nature of the AA change and the ancestral nucleotide both had a significant effect on costs, we decided to look into this effect more closely. We found that many of the very costly mutations were associated with a small number of AA changes, starting from Glutamate, Glycine and Proline. This is partly in line with knowledge of protein structures, whereby Proline and Glycine are often unique and non-replaceable 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 be in part responsible for the nucleotide effect we found.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.208	0.013	-244.671	0.000
inRT	0.136	0.008	16.206	0.000
shape	0.169	0.014	12.034	0.000
t	0.034	0.014	2.422	0.015
c	0.808	0.016	50.379	0.000
g	0.717	0.014	49.936	0.000
CpG	-1.439	0.029	-49.853	0.000
t:CpG	0.041	0.047	0.875	0.381
nonsyn	-0.611	0.014	-42.644	0.000
t:nonsyn	0.061	0.024	2.551	0.011
c:nonsyn	-1.833	0.037	-49.396	0.000
g:nonsyn	-0.380	0.021	-18.065	0.000
nonsyn:CpG	0.981	0.045	21.991	0.000
t:nonsyn:CpG	-0.881	0.093	-9.447	0.000
bigAAChange	-1.183	0.014	-83.501	0.000

Table 1: Table with GLM results for observed counts of mutants. Significant predictors are listed together with their associated P-value based on a Z-test.

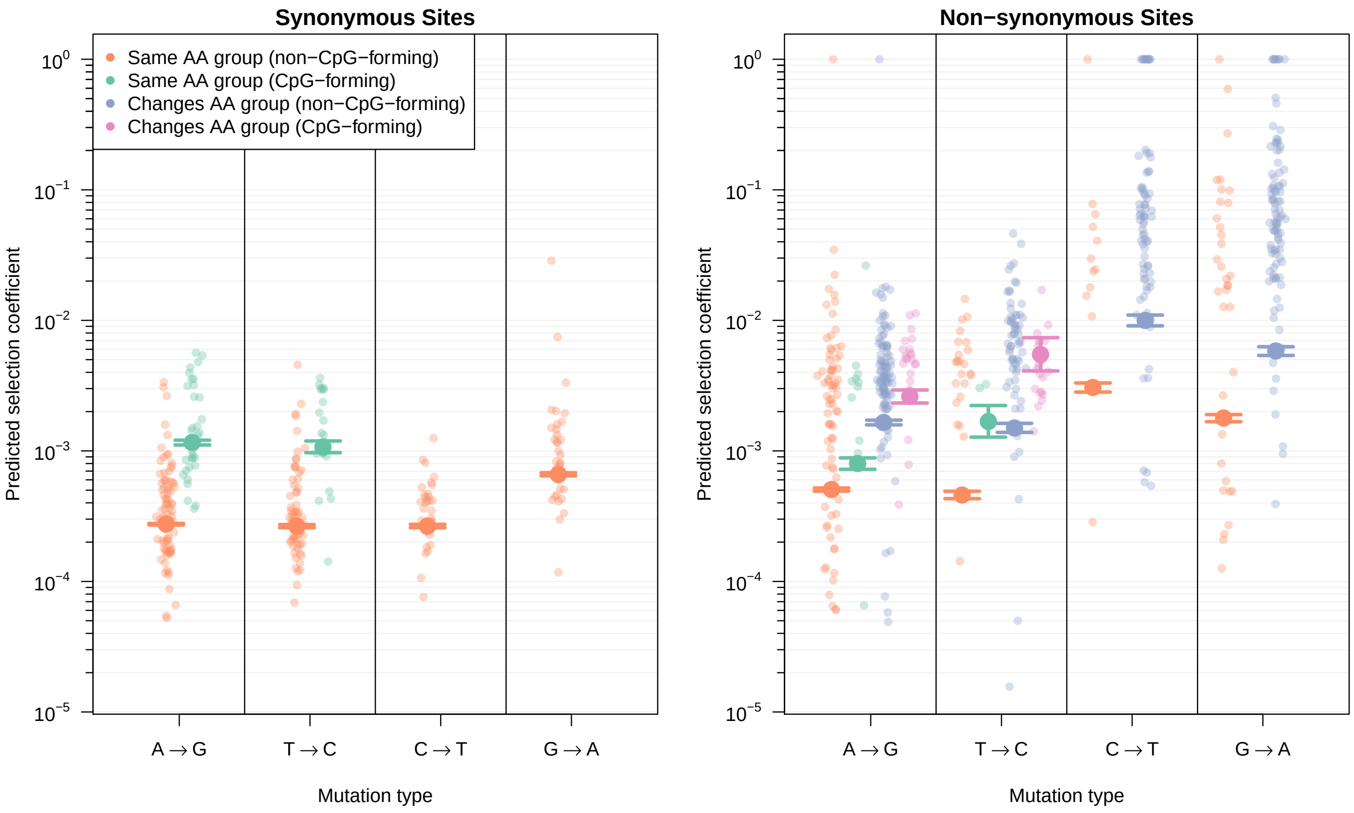

Figure 2: Generalized Linear Model (GLM) with selection coefficients estimated from the Bacheler dataset. The graph shows the model predictions for synonymous and non-synonymous mutations that either form CpG sites (green) or do not form CpG sites (orange) by preserving the same amino acid group. For non-synonymous mutations in addition, predictions are shown which change the amino acid group and form CpG sites (pink) or do not form CpG sites (blue). The prediction are shown for every single nucleotide. The cloud of dots in each panel represents the estimated selection coefficients found in the dataset.

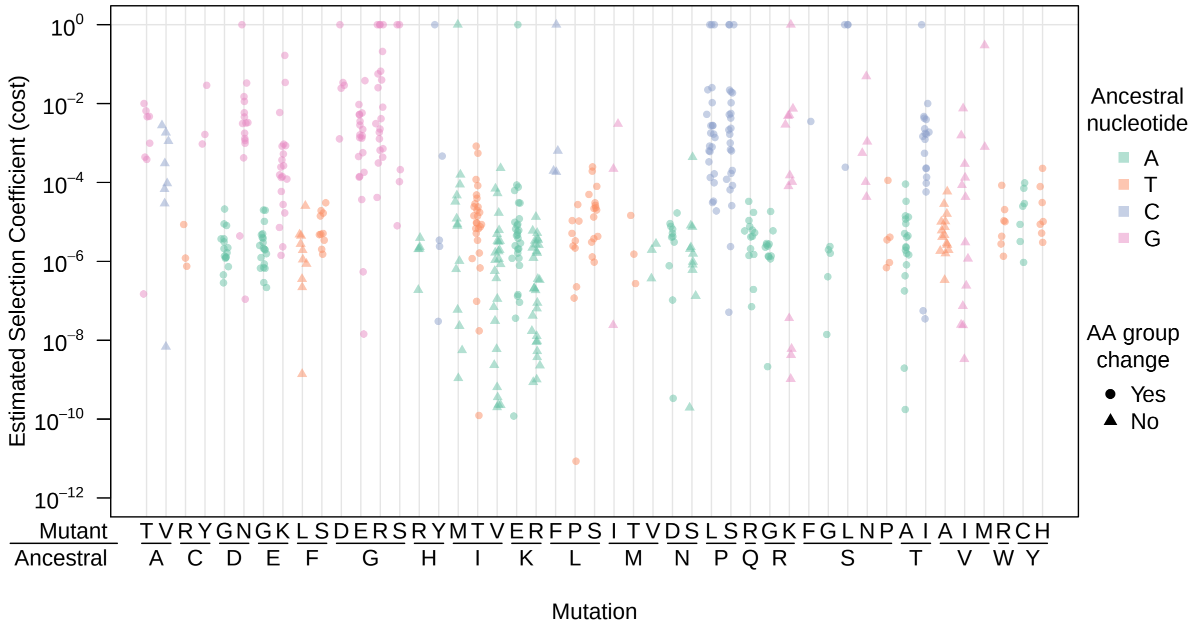

Figure 3: Estimated costs of non-synonymous mutations, ordered by ancestral and mutant amino acid. Colors indicate the different nucleotides, shapes indicate whether the AA change is drastic or not.

Parameters for Gamma distribution for distribution of fitness effects

In addition to the characteristics that determine fitness costs of mutations we also looked into 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 [17]. Moreover, it determines the evolvability of the population: A DFE that has more weight on neutral and adaptive mutations, reflects a population with more capacity to evolve. Nevertheless, despite their reputation as fast evolving entities, many viruses have a DFE which is composed mainly of deleterious and lethal mutations. To determine the DFE, we took the average mutant frequency for each site and used it to directly estimate the fitness cost using the mutation-selection formula ($f = u/s$). In Figure 4 we show the distributions for each of the ancestral nucleotides for synonymous mutations and for non-synonymous mutations (including nonsense mutations). Note that the scales of the x and y-axis differ between the figures that show synonymous and non-synonymous mutations. We also estimated parameters for a *Gamma* distribution that best describes the entire DFE (Table 2).

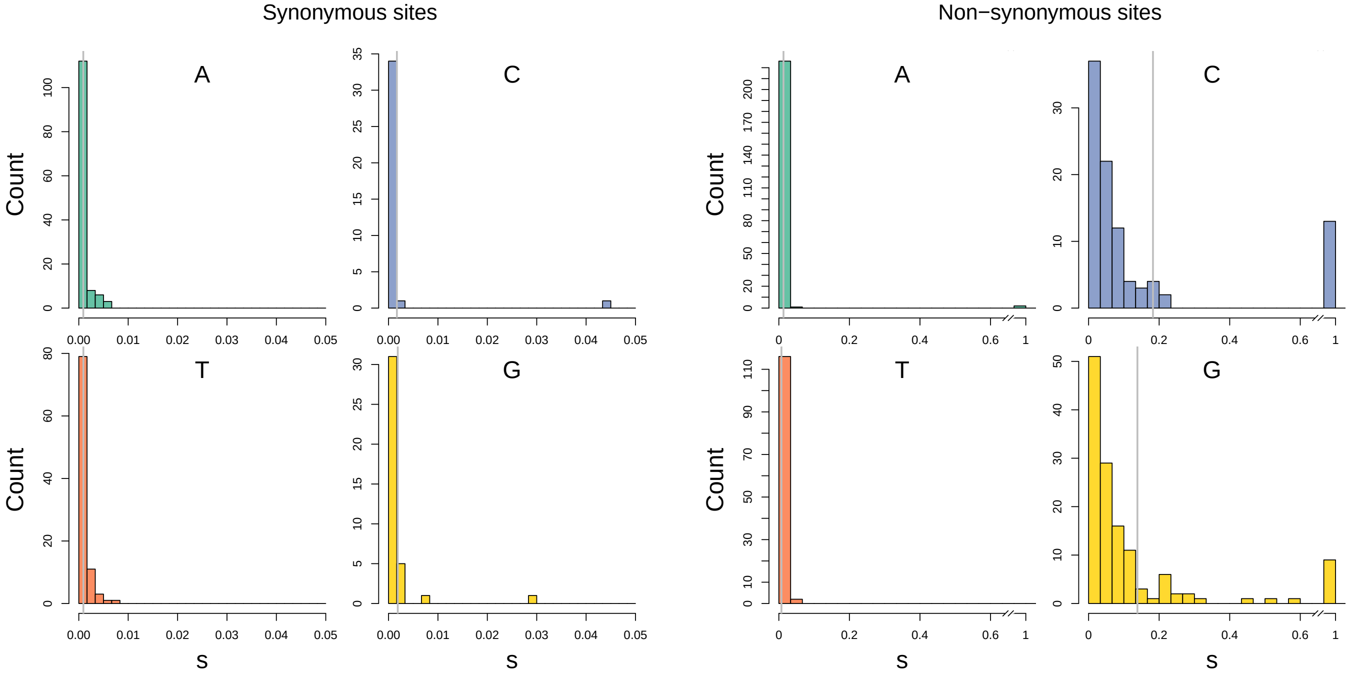

Figure 4: Distribution of fitness effects (DFE) for non-synonymous and synonymous mutations; nonsense mutations are included in the non-synonymous mutations. Note that the scales of the x and y-axis differ between the figures.

		Mut. rates from	Abrahm 2010	Mut. rates from	Zanini 2016	
	Sites	κ	θ	κ	θ	Lethal
Bacheler	870	0.317 (0.241, 0.397)	0.209 (0.202, 0.219)	0.319 (0.247, 0.395)	0.242 (0.233, 0.253)	0.065 (0.05, 0.083)

Table 2: Table with Gamma distribution parameters reflecting scale (κ) and shape (θ).

Relationship between within-patient mutation frequencies and site conservation at population level

We were interested to determine how well the observed within-patient frequencies correspond with world wide frequencies. All sequences from the Bacheler dataset belong to the HIV-1 B subtype, so as a comparison, we collected sequences from the Stanford HIV Drug Resistance database (HIVdb) that were also HIV-1 subtype B. Notably, each such sequence represents the consensus sequence present in a patient (i.e., the consensus of the "mutant cloud" evolving in a patient at a given time point). HIV subtype B sequences from treatment-naive patients were retrieved from the Stanford Drug Resistance database (HIVdb), with 23742 Protease sequences and 22785 Reverse Transcriptase sequences [37]. Figure 5 shows the correlation between average within-patient mutation frequencies calculated from the Bacheler dataset and between-patient mutation frequencies calculated from the HIVdb dataset. A high correlation coefficient was detected when comparing all 870 sites ($R^2 = 0.949$), with a higher coefficient for non-synonymous mutations ($R^2 = 0.973$) than for synonymous mutations ($R^2 = 0.884$). In the lower frequency regions, it can be seen that costly mutations occurring at low frequencies within patients are often not observed in consensus sequences of patients.

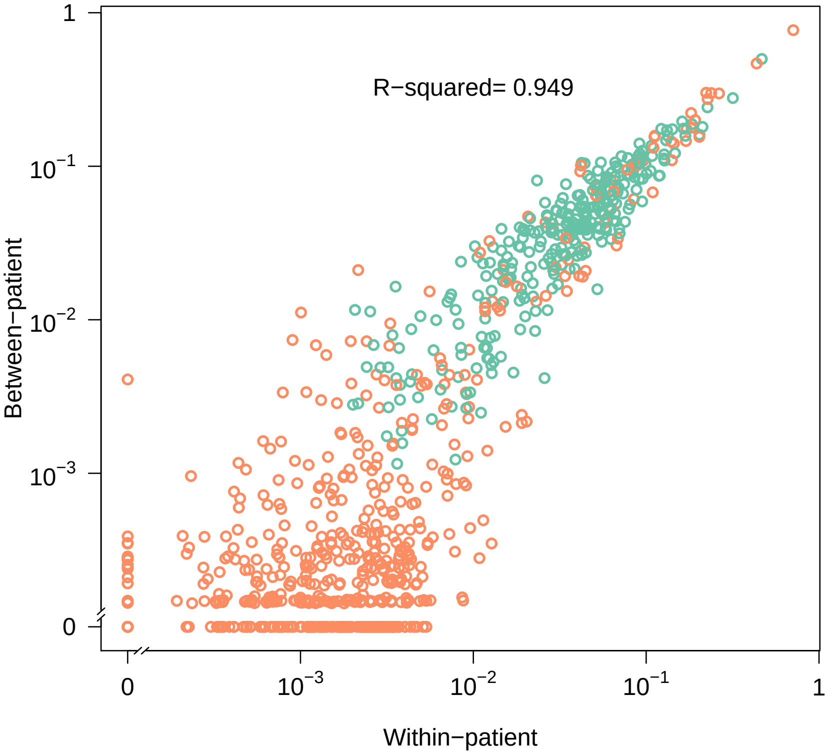

Figure 5: Correlation of average mutations frequencies at within-patient level (Bacheler data) and mutation frequencies at the between-patient level (HIVdb), with values shown on a log scale. Non-synonymous mutations are shown in red, synonymous mutations are shown in green.

3 Conclusion / Discussion

We used within-patient mutation frequencies for 870 transition mutations in the *pol* gene of HIV as observed in samples from 160 patients [31] to estimate the costs of these mutations for the virus as it evolves in a patient. Due to the high mutation rate of HIV and the large size of our dataset, 93.5% of 870 possible transition mutations were observed in at least one patient in our dataset. Only 6.5% of the mutations (57 mutations) were never observed, but half of these are nonsense mutations that create premature stop codons, so these mutations are expected to be lethal and be swiftly removed due to purifying selection. The other mutations that were never observed may also be lethal, or they may be very costly. A larger dataset would be needed to distinguish between lethal and strongly deleterious mutations. We found that the cost of a mutation depends on whether the mutation leads to an amino-acid change, whether that change is drastic or not and whether the mutation creates a new CpG site or not. We also found that G → A and C → T mutations are generally more costly than A → G and T → C mutations. Other characteristics of sites had smaller effects on the estimated fitness costs, such as the SHAPE parameter which is related to RNA secondary structure (lower SHAPE parameter is associated with a higher cost), and whether a site is part of Protease or Reverse Transcriptase. Our results show that within-patient frequencies of mutations carry a wealth of information on costs of mutations.

Comparison with other studies in viruses.

Our main results are consistent with results from the recent study by Zanini *et al* [35], in that we also find a clear difference between synonymous and non-synonymous mutations (see Supplementary figure 7). One clear difference between the data from our study (based on the Bacheler dataset) and the data from Zanini *et al* is that there are some transition mutations that were never observed in the Bacheler dataset (6.5% of the sites), whereas all transition mutations were observed in the Zanini dataset. A mutation that is not observed will have an estimated fitness cost of 1, which means that the mutation is lethal. The Bacheler dataset therefore suggests that 6.5% of the mutations are lethal, but the Zanini dataset suggests that no mutations are lethal (see Table 2). The difference between the two datasets may be due to sequencing depth and sequencing technique. Some of the mutations that were not observed in the Bacheler dataset, may have been missed because the sequencing was not very deep (median of 19 sequences per patient). If this is true, then 6.5% is an overestimate. On the other hand, in the Zanini dataset, some of the mutations that were observed may be due to sequencing errors. For example, nonsense mutations were observed at various frequencies in the Zanini dataset see Supplementary figure 7), even though they are probably lethal and we should observe them at very low frequencies (equal to the mutation rate) or not at all. All next-generation sequencing techniques have error rates that are higher than the HIV mutation rates, so the observed frequencies for such mutations are likely overestimates of the real frequencies. The zero percentage of lethal mutations in the Zanini datset is therefore likely an underestimate.

We also analyzed a third dataset from a recent paper by Lehman *et al* [38]. In this dataset 1.3% of the transition mutations are never observed and thus estimated to be lethal. Because of the sequencing technology and the associated error rate, we expect that this percentage is also an underestimate of the true percentage of lethal mutations.

It should be noted that even the highest number we found (6.5% of mutations lethal) is low compared to percentages from *in vitro* studies on coding sequences in viruses (see [39] for an overview). For example, Sanjuan *et al* [13] find that 40% of random mutations in the RNA virus VSV (Vesicular stomatitis virus) are lethal. Another study which used the Tobacco Etch Virus estimated that 27 mutations out of 66 were lethal (41%) [40]. A study by Rihn *et al* on the capsid of HIV found that 70% of non-synonymous mutations were lethal, which corresponds to around 47% of all mutations [41]. Two studies on bacteriophage found somewhat lower percentages. Domingo *et al.* found that 20 or 29% of mutations were lethal (depending on the species) [42] and Peris *et al.* found that 21% of mutations were lethal [43]. A recent study on Poliovirus estimated that around 26% of all mutations were lethal [14]. A small study on Phage ΦX174 by Vale *et al.* [44] found that only 1 or 3 out of 36 mutations ($3-8\%$) were lethal which is in the same range as our finding. A few small studies on non-coding regions also found lower percentages of lethal mutations. A study on fitness effects in HIV ([45]) found no lethal mutations out of 15 mutations and a study on a non-coding region in Tobacco Etch Virus found that 2 out of 20 mutations were lethal (10%) [46].

Several factors could contribute to why we find a lower percentage of lethal mutations than *in vitro* studies. First, we only looked at transition mutations, and excluded transversions from our analysis, but transversions may be more likely to be lethal since they are more often non-synonymous, more often lead to drastic amino acid changes and more often 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 dataset were only observed once, and it is possible that some of these are not true variants, we may thus underestimate 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 viral genomes. Finally, it may be that the different environments (*in vitro* vs *in vivo*) or the different genetic backgrounds (usually one genetic background in the *in vitro* studies vs many in *in vivo* studies) lead to the observed differences. We expect that 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 effect.

Some of our results are surprising. As can be seen in Figure 2, G → A mutations appear to be two-and-a-halve to three-and-a-halve times more costly than A → G or T → C mutations. C → T mutations appear to be more 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 what causes this nucleotide identity effect. First of all, we realize that the effect could be an artifact caused by spurious mutation rate estimates. However, mutation rate estimates from two very different studies Abram *et al* [21] or Zanini *et al* [23] are very similar and using one or the other does not change our main results. Secondly, the nucleotide effect we find may be related to the activity of APOBEC3 enzymes which hypermutate the HIV genome, leading to an increased proportion of G → A mutations [47, 48, 49]. Thus, it is possible that the G → A mutations we observe in the *pol* gene occur together with other G → A mutations at other regions in the genome (that we do not observe), leading to a higher fitness cost. Third, we hypothesize that the effect may 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 [21, 23], which could have led, over evolutionary times, to the well known A bias in the HIV genome [50, 51]. Specifically, sites at which it doesn't matter for viral fitness whether it carries an A or a G would become A biased over time. The result is that A sites are 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 we observed may be partly due to the specific amino acids that are involved. As can be seen in figure 3, most of the very deleterious mutations are concentrated in just a few specific amino acid changes. Studies on larger parts of the genome, or including transversions in addition to transitions, are needed to disentangle the nucleotide and amino acid effects.

Another factor we find that greatly affects the cost of a mutation is the formation of CpG sites. Depending on the type of sites, CpG forming mutations are between one-and-a-halve and four times more costly than the equivalent mutation that does not create a CpG site. In animal and plant genomes, CpG sites are usually found in promotor regions and their methylation leads to the silencing of the downstream gene [52]. It was hypothesized, that these sites occur rarely in coding regions of because methylated cytosines can randomly deaminate to thymine and cause unwanted mutations [53]. The same mechanism is unlikely to affect a retrovirus like HIV because in HIV most mutations are introduced during reverse transcription and not in the DNA stage. However, CpG sites are only found very rarely in RNA viruses [54] and strongly selected against in a wide range of viral genomes [34]. Several studies have recently suggested that CpG sites may trigger the antiviral cellular response [32, 33]. A potential reason might be that viral transcripts with CpG are more easily detected by the immune system since CpG sites in animals are usually found in promoters and not in the coding region of genes and therefore rarely transcribed [55]. It was further shown in mutational studies that introducing CpG sites in the viral genome strongly decreased the replication rate [32, 33] and hence the virus' fitness. Corroborating these previous studies, our analysis shows that CpG site formation has deleterious effects on the fitness of HIV *in vivo*. Given the ever more convincing evidence that CpG site are deleterious for viral genomes, future efforts should be geared towards the discovery of the molecular mechanism of this process which most likely differs from eukaryotic cells.

Limitations of our study.

A limitation of our study is that we only focused on a small part of the HIV genome, namely 870 sites of the *pol* gene, thereby excluding a large part of the viral genome and all drug resistance mutations, which would be of special interest. A second limitation of our analysis is that we only focused on transitions and left out transversion mutations. When deeper and more precise data become available, it would be possible to see how results change when transversions are included. For one, we expect that the distribution of fitness effects will shift towards more costly mutations when transversions are included.

Another limitation of our study is that it is unknown how long the patients in the Bacheler [31] dataset had been infected before samples were taken. This is relevant because it is known that diversity within a host increases with time [56, 57]. The reason is that most patients are infected with one or a very small number of founder viruses and therefore genetic diversity within the host is usually low right after infection. Over time, genetic diversity accumulates and plateaus after several years. At the same time, the viral population diverges from the founder virus. It is therefore reasonable to ask whether we expect our results to differ depending on when samples were taken. First of all, early samples tend to have less genetic diversity, therefore they would carry less information. Mutant frequencies are expected to be close to 0 (if the founder virus is WT at a position) or close to 1 (if the founder virus is mutant at the position). If founder viruses are a random sample from the viral population within a host, then the average mutant frequency across many samples from newly infected patients should be the same as the average mutant frequency across many sample from long term infected patients. For example, if the average frequency of a mutant in all patients in the epidemic is 1%, then we expect 1% of the founder viruses to be mutant. Out of 100 newly infected patients, we expect 1 sample to have a mutant frequency at 100% and 99 samples at 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-host frequencies are similar, then using early samples would lead to unbiased estimates of frequencies, but the variance of such estimate may be very high. It would therefore be better to work with samples from later during the infection.

Secondly, we should keep in mind that founder viruses may not be a random sample from within-host viruses. If a mutation comes with a fairly severe (but not lethal) cost, such that it is observed within patients at a frequency of 1%, but it cannot establish a new infection, then we'd expect to see a lower mean frequency in newly infected patients as compared to patients that have been infected longer. Samples from patients that were recently infected could thus be more similar to the wildtype than samples from patients that have been infected for a longer time. This scenario is likely if selection at the beginning of an infection is stronger than selection during an ongoing infection. In a future study it may be possible to compare early and late samples to determine whether such an effect exists. We should note, however, that the tight correlation between within-patient frequencies and global frequencies (see figure 5) suggests that fitness of HIV strains within hosts and at the infection stage are similar.

Future directions.

In this study, we have used observed frequencies of transition mutations to estimate fitness costs of mutations. Our results show the power of analyzing mutant frequencies from *in vivo* viral populations to study fitness effects of mutations. We also realize however, that the observed frequencies (and thus the estimated costs) between the three datasets we considered differed substantially (see Fig. 1, 6, 7) [31, 35, 38]. Three key differences exist among the the three datasets: (a) sequencing techniques and associated error rates differed, (b) the viral load among the three studies differed (mainly, in the Zanini dataset viral load was relatively low), and (c) the timing of the samples since infection differed (in the Lehman dataset most samples were at a very early stage of infection). These factors will all influence the frequency measurements, and whether a viral population is at mutation selection balance. To overcome these issues, a controlled study is necessary, at higher resolution, across more samples and across more sites, ideally using samples from untreated patients. This will allow us to get a more fine-grained and precise picture of costs of mutations at individual sites accross the entire HIV genome.

4 Methods

Most of our analysis is done on the Bacheler *et al.* dataset [31]. In addition, we look at the Lehman *et al.* data [38] and the Zanini *et al.* data [35].

Description of the data / filtering

The Bacheler *et al.* data [31] is cloned and sequenced, so it is of high quality. For each patient, we lump together all sequences, even though they come from different time points. Patients with only a single sequence were excluded from the analysis, leaving us with median 19 sequences per patient for 160 patients. We don't do much filtering, except that we remove sites from specific sequences from specific patients if there is a mutation in one or both of the other two sites that make up a triplet. This last filtering is done because if there is a mutation elsewhere in the triplet, it is no longer clear whether a given mutation should be counted as synonymous or non-synonymous.

The Lehman data is 454 sequences, much deeper, but higher error rate, there are multiple time points for most patients, but we only use one time point per patient (one month after seroconversion). The sequences span around 600 sites in RT. Because the Lehman data [38] contain different HIV subtypes (mostly C and A), we only consider sites that are conserved between A, B and C.

The Zanini data [35] is Illumina sequenced, multiple time points per patient. Time points are usually a few months or more apart and are treated as independent samples. Though the Zanini data are genome wide, we only consider the sites for which we also have data from the Bacheler data [31]. Because the Zanini data [35] contain different HIV subtypes (B and C), we only consider sites that are conserved between B and C.

For all three datasets, we only consider transition mutations because they are more common in HIV than transversion mutations. For example, for a site that is A in the ancestral state, the frequency of a transition mutation is calculated for each patient (and each time point in the case of the Zanini data [35]) as the number of sequences that carry a G divided by the number of sequences that carry a G or an A. These frequencies are then averaged over all patients (and time points in the case of the Zanini data [35]). Sequences with a C or a T are thus not considered if the ancestral state is A. Selection coefficients are estimated by dividing the nucleotide-specific mutation rate by the average frequency, bootstrapping is done over sites (not patients). For the Generalized Linear Model, actual counts are considered as opposed to frequencies.

Generalized Linear Model

We fit a binomial generalized linear model to model the state (derived or ancestral) of each position based on its ancestral nucleotide, its SHAPE value, whether or not the position is in Reverse Transcriptase 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 if the transition 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. Each position in each sequence from each patient was treated as an independent observation. The results are reported in Table 1.

Estimating the gamma distribution

Gamma distributions were estimated separately for each of the three datasets. Transitions that did not appear and resistance mutation 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 nloptr [58]. Bootstrapped confidence intervals were created by resampling the data with replacement and re-estimating the gamma distribution parameters. Selection coefficients estimated using the mutations rates given both in Abram *et al.* [21] and Zanini *et al.* [23].

Comparison with the global epidemic

A large HIV-1 sequence dataset was retrieved from the Stanford HIV Drug Resistance Database (HIVdb) [37]. Protease and reverse transcriptase sequences were downloaded in separate files. Inclusion criteria for the analysis were the treatment-naive status of the host, classification as HIV-1 subtype B and a single sequence per host (selected at random), with a total of 23742 protease and 22785 reverse transcriptase sequences used in the analysis. Average mutation frequencies for each site were calculated as explained previously. To determine to what extent frequency values between hosts correlate with frequency values obtained from within host, the Pearson product-moment correlation coefficient (R^2) was used to quantify the strength of the relationship.

5 Acknowledgements

Dmitri Petrov, Arbel Harpak, David Enard, Ryan Taylor, Dara Lehman.

6 References

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," *SCIENTIFIC AMERICAN-AMERICAN EDITION-*, 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," *Microbiology and molecular biology reviews*, vol. 65, no. 1, pp. 151–185, 2001.

[5] C. O. Wilke, "Quasispecies theory in the context of population genetics," *BMC evolutionary biology*, 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," *The Journal of infectious diseases*, 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," *Proceedings of the National Academy of Sciences*, vol. 108, no. 14, pp. 5661–5666, 2011.

[13] R. Sanjuán, A. Moya, and S. F. Elena, "The distribution of fitness effects caused by single-nucleotide substitutions in an rna virus," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 22, pp. 8396–8401, 2004.

[14] 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.

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

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

[17] A. Eyre-Walker and P. D. Keightley, "The distribution of fitness effects of new mutations," *Nature Reviews Genetics*, vol. 8, no. 8, pp. 610–618, 2007.

[18] 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 evolutionary biology*, vol. 13, no. 1, p. 1, 2013.

[19] 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.

[20] 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.," *Journal of virology*, vol. 69, no. 8, pp. 5087–5094, 1995.

[21] 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," *Journal of virology*, vol. 84, no. 19, pp. 9864–9878, 2010.

[22] 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.

[23] 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.

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

[25] 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.

[26] D. C. Douek, L. J. Picker, and R. A. Koup, "T cell dynamics in hiv-1 infection*," *Annual review of immunology*, vol. 21, no. 1, pp. 265–304, 2003.

[27] D. L. Hartl, A. G. Clark, and A. G. Clark, *Principles of population genetics*, vol. 116. Sinauer associates Sunderland, 1997.

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

[29] 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.

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

[31] 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," *Antimicrobial agents and chemotherapy*, vol. 44, no. 9, pp. 2475–2484, 2000.

[32] 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," *Journal of virology*, vol. 83, no. 19, pp. 9957–9969, 2009.

[33] 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 research*, vol. 42, no. 7, pp. 4527–4545, 2014.

[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] 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.

[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] 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 research*, vol. 31, no. 1, pp. 298–303, 2003.

[38] 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," *Journal of Infectious Diseases*, p. jiu677, 2015.

[39] R. Sanjuán, "Mutational fitness effects in rna and single-stranded dna viruses: common patterns revealed by site-directed mutagenesis studies," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 365, no. 1548, pp. 1975–1982, 2010.

[40] 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," *Journal of virology*, vol. 81, no. 23, pp. 12979–12984, 2007.

[41] 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.

[42] 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.

[43] 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.

[44] P. F. Vale, M. Choisy, R. Froissart, R. Sanjuán, and S. Gandon, "The distribution of mutational fitness effects of phage φx174 on different hosts," *Evolution*, vol. 66, no. 11, pp. 3495–3507, 2012.

[45] T. van Opijnen, M. C. Boerlijst, and B. Berkhout, "Effects of random mutations in the human immunodeficiency virus type 1 transcriptional promoter on viral fitness in different host cell environments," *Journal of virology*, vol. 80, no. 13, pp. 6678–6685, 2006.

[46] G. P. Bernet and S. F. Elena, "Distribution of mutational fitness effects and of epistasis in the 5'untranslated region of a plant rna virus," *BMC evolutionary biology*, vol. 15, no. 1, p. 274, 2015.

[47] 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.

[48] 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.

[49] 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.

[50] 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 biology*, vol. 10, no. 2, pp. 211–215, 2013.

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

[52] J. A. Law and S. E. Jacobsen, "Establishing, maintaining and modifying dna methylation patterns in plants and animals," *Nature Reviews Genetics*, vol. 11, no. 3, pp. 204–220, 2010.

[53] E. Scarano, M. Iaccarino, P. Grippo, and E. Parisi, "The heterogeneity of thymine methyl group origin in dna pyrimidine isostichs of developing sea urchin embryos," *Proceedings of the National Academy of Sciences*, vol. 57, no. 5, pp. 1394–1400, 1967.

[54] B. K. Rima and N. V. McFerran, "Dinucleotide and stop codon frequencies in single-stranded rna viruses.," *Journal of general virology*, vol. 78, no. 11, pp. 2859–2870, 1997.

[55] 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.

[56] 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," *Journal of virology*, vol. 73, no. 12, pp. 10489–10502, 1999.

[57] R. D. Kouyos and H. F. Günthard, "The irreversibility of hiv drug resistance," *Clinical Infectious Diseases*, p. civ400, 2015.

[58] S. G. Johnson, "The nlopt nonlinear-optimization package."

7 Supplementary figures

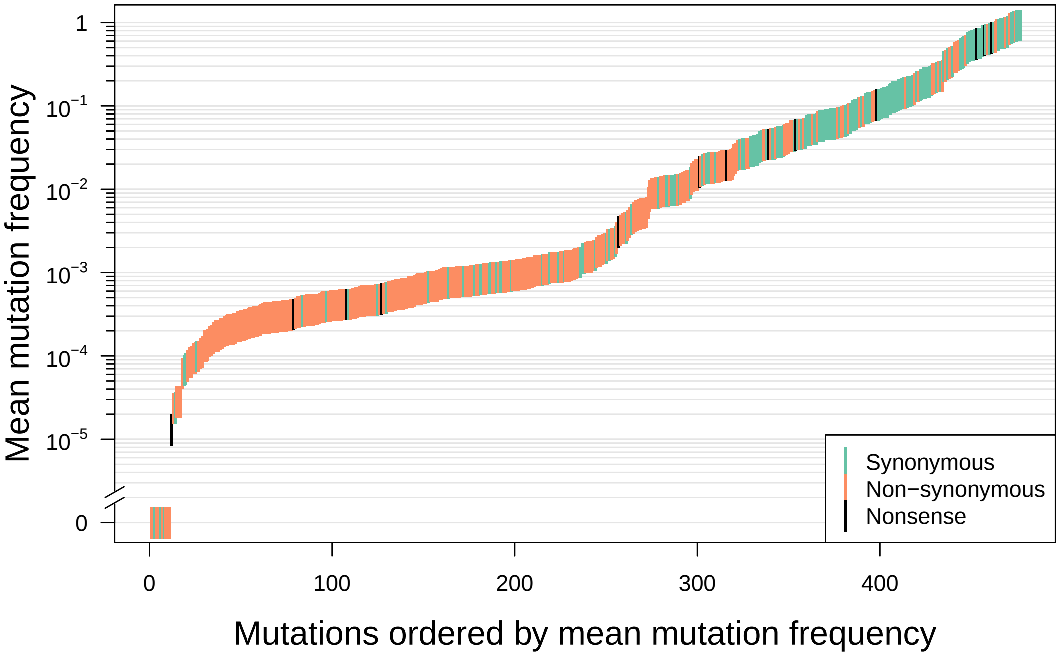

Figure 6: Mutation frequency for 687 sites from the Lehman data, ordered by mutation frequency. Mutations that cause a premature stop codon are shown in black, mutations that cause an amino acid change (non-synonymous) are shown in red and mutations that do not change the amino acid (synonymous) are shown in green.

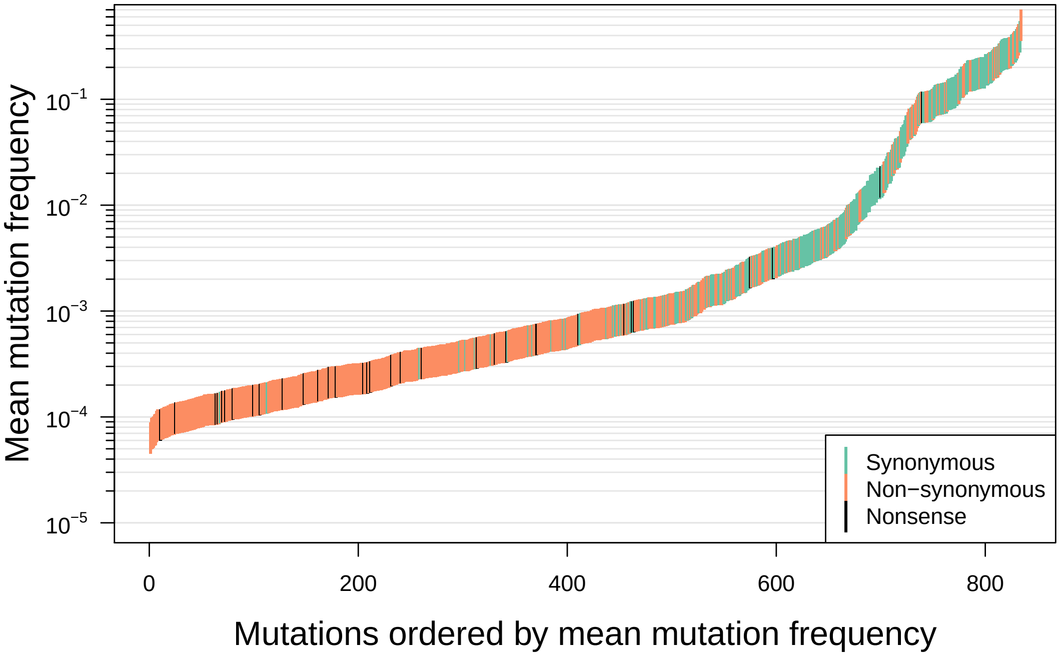

Figure 7: Mutation frequency for 984 pol sites from the Zanini data, ordered by mutation frequency. Nonsense mutations are shown in black, mutations that cause an amino acid change (non-synonymous) are shown in red and mutations that do not change the amino acid (synonymous) are shown in green.

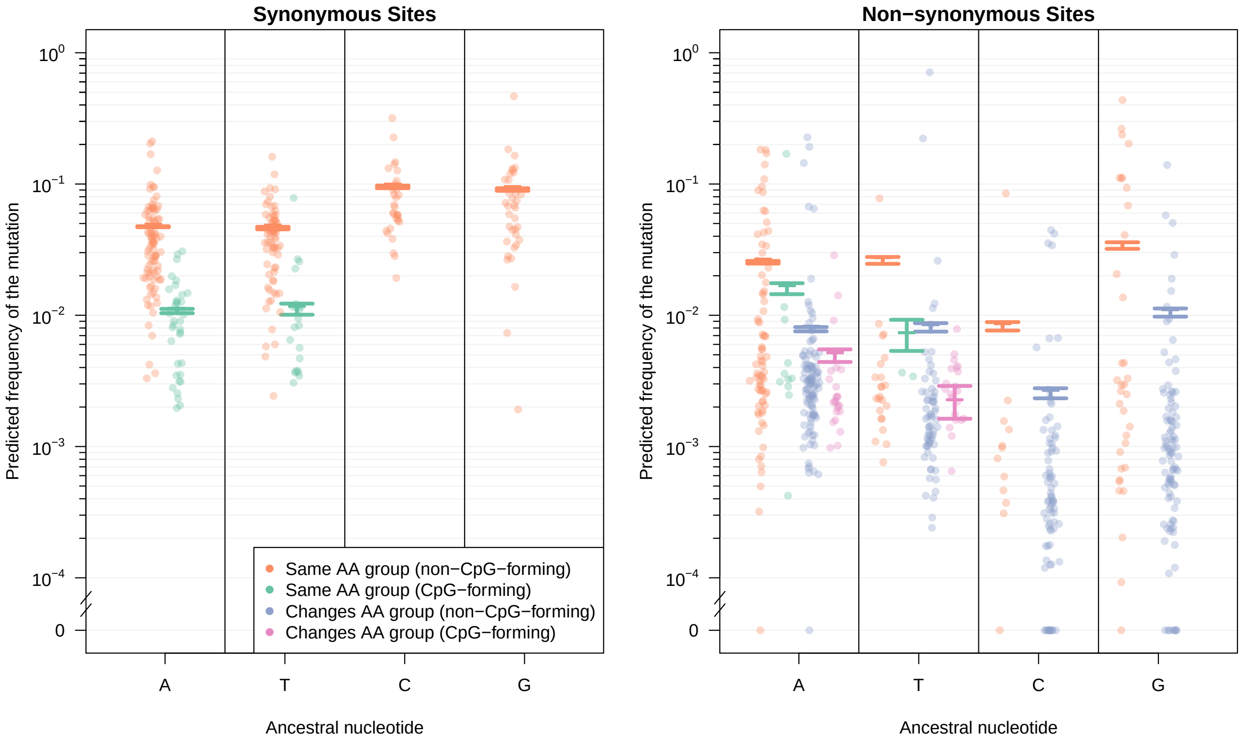

Figure 8: Generalized Linear Model (GLM) with frequencies estimated from the Bacheler dataset. The graph shows the model predictions for synonymous and non-synonymous mutations that either form CpG sites (green) or do not form CpG sites (orange) by preserving the same amino acid group. For non-synonymous mutations in addition, predictions are shown which change the amino acid group and form CpG sites (pink) or do not form CpG sites (blue). The predictions are shown for the type of nucleotide. The cloud of dots in each panel represents the calculated frequencies found in the Bacheler dataset.

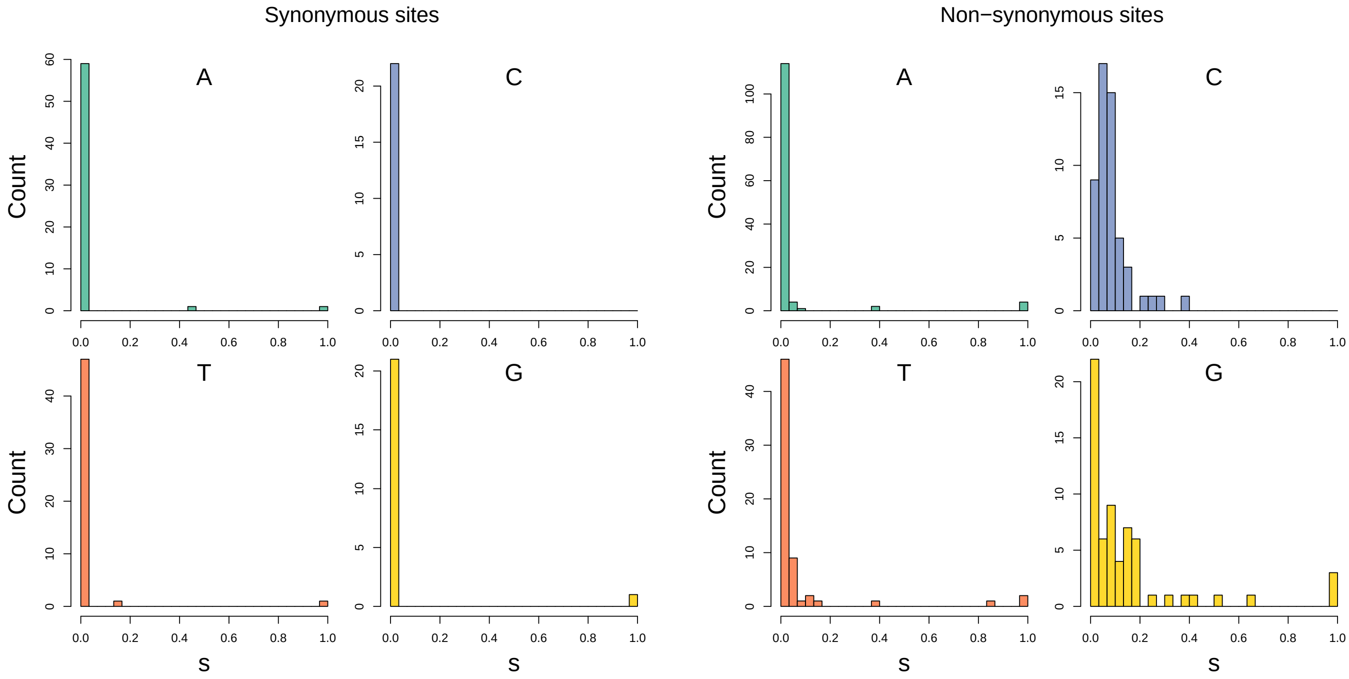

Figure 9: Distribution of fitness effects (DFE) for non-synonymous and synonymous mutations for the Lehman dataset; nonsense mutations are included in the non-synonymous mutations. Note that the scales of the x and y-axis differ between the figures.

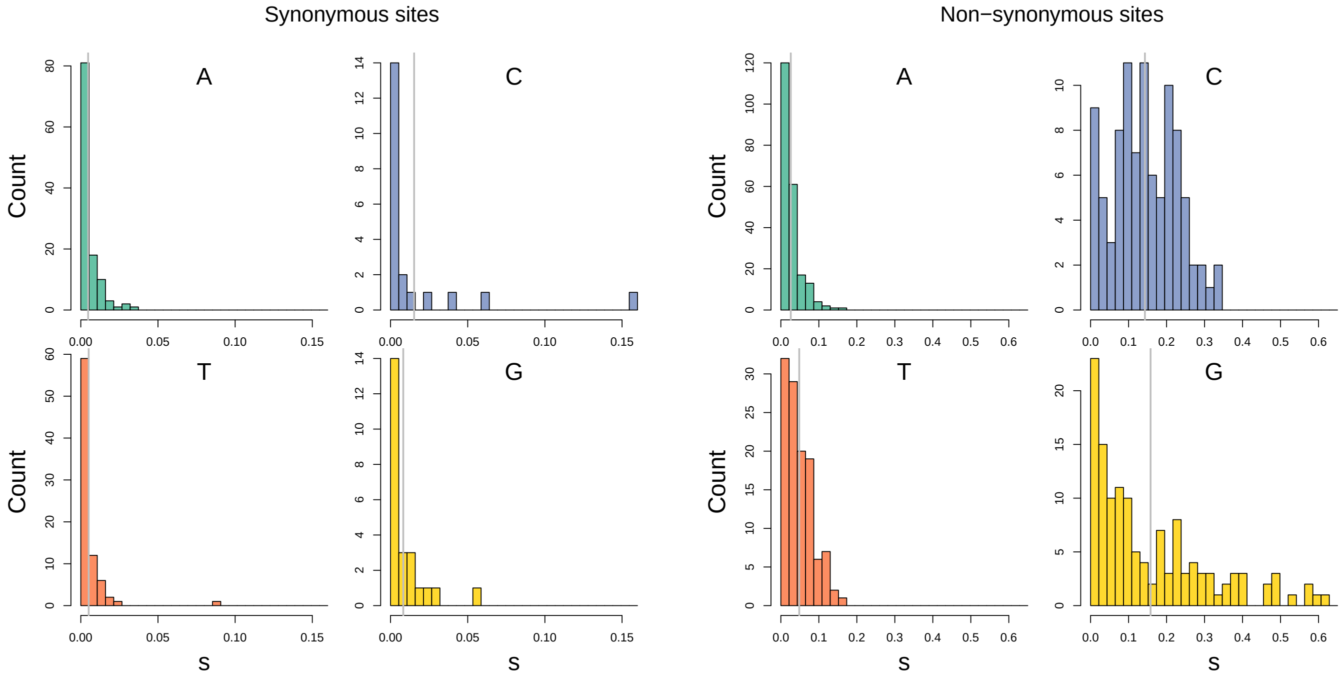

Figure 10: Distribution of fitness effects (DFE) for non-synonymous and synonymous mutations for the Zanini dataset; nonsense mutations are included in the non-synonymous mutations. Note that the scales of the x and y-axis differ between the figures.

		Mut. rates from	Abrahm 2010	Mut. rates from	Zanini 2016	
	Sites	κ	θ	κ	θ	Lethal
Bacheler	870	0.317	0.209	0.319	0.242	0.065
		(0.241, 0.397)	(0.202, 0.219)	(0.247, 0.395)	(0.233, 0.253)	(0.05, 0.083)
Zanini	903	0.056	0.447	0.146	0.414	0
		(0.05, 0.061)	(0.421, 0.481)	(0.129, 0.163)	(0.388, 0.443)	(0, 0)
Lehman	540	0.155	0.238	0.228	0.258	0.013
		(0.101, 0.221)	(0.22, 0.261)	(0.169, 0.297)	(0.242, 0.277)	(0.007, 0.021)

Table 3: Table with Gamma distribution parameters reflecting scale (κ) and shape (θ) for Bacheler, Zanini and Lehman datasets.