In-vivo mutation rates and fitness landscape of HIV-1

Fabio Zanini[1,2], Vadim Puller[1], Johanna Brodin[3], Jan Albert[3,4], Richard A. Neher,[1*]

[1] *Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany* [2] *Department of Bioengineering, Stanford University, Stanford, USA* [3] *Department of Microbiology, Tumor and Cell Biology, Karolinska Institute, Stockholm, Sweden* [4] *Department of Clinical Microbiology, Karolinska University Hospital, Stockholm, Sweden*

(Dated: July 1, 2016)

Mutation rates and fitness costs of deleterious mutations are difficult to measure *in vivo* but essential for a quantitative understanding of evolution. Using whole genome deep sequencing data from longitudinal samples during untreated HIV-1 infection, we estimated mutation rates and fitness costs in HIV-1 from the temporal dynamics of genetic variation. At approximately neutral sites, mutations accumulate with a rate of 1.2×10^{-5} per day, in agreement with the rate measured in cell cultures. The rate from G to A is largest, followed by the other transitions C to T, T to C, and A to G, while transversions are more rare. At non-neutral sites, most mutations reduce virus replication; using a model of mutation selection balance, we estimated the fitness cost of mutations at every site in the HIV-1 genome. About half of all nonsynonymous mutations have large fitness costs (greater than 10%), while most synonymous mutations have costs below 1%. The cost of synonymous mutations is especially low in most of gag and pol, while much higher costs are observed in important RNA structures and regulatory regions. The intrapatient fitness cost estimates are consistent across multiple patients, suggesting that the deleterious part of the fitness landscape is universal and explains a large fraction of global HIV-1 group M diversity.

Introduction

HIV-1 evolves rapidly within individual hosts: mutations allow it to evade immune predation but can also impair viral replication. Genetic changes arise during reverse transcription, during forward transcription by the human RNA polymerase II, or are caused by the innate immune system (Abram *et al.*, 2010; Cuevas *et al.*, 2015; Malim, 2009; Mansky and Temin, 1995). These changes are the source of genetic diversity, from which selection amplifies beneficial variants and filters deleterious mutations. Characterization of the mutation rate matrix and the genome wide landscape of fitness effects is a prerequisite a quantitative understanding of the evolutionary dynamics of HIV and for rational design of both vaccines and resistance proof drugs.

The majority of mutations are deleterious, some mutations are neutral and have little or no effect, and a minority of mutations are beneficial. While beneficial mutations rapidly spread through the virus population within a patient, deleterious mutations stay at low frequency in a balance between mutation and selection. Beneficial mutations are often patient-specific and mediate escape from cytotoxic T-lymphocytes (CTL) and neutralizing antibodies (Bar *et al.*, 2012; Goonetilleke *et al.*, 2009; Walker and McMichael, 2012). At the same time, substitutions in response to immune selection are expected to lower intrinsic viral fitness; host-specific adaptation is a tradeoff between immune evasion and fitness costs of escape mutations.

Since HIV-1 proteins serve the same function in different hosts, the landscape of fitness costs might be expected to be similar in different hosts. However, the effect of a particular mutation can depend on other sites in the genome – a phenomenon known as epistasis (de Visser and Krug, 2014). Epistasis and interaction between mutation has been observed as compensatory evolution after CTL escape (Schneidewind *et al.*, 2009) or as covariation of amino acids (Carlson *et al.*, 2008; Dahirel *et al.*, 2011). While epistasis is clearly an important aspects of protein fitness landscapes, it is expected to be only a weak effect at short evolutionary distances: Doud *et al.* (2015) have shown that the majority of mutation effects tend to be conserved in mildly diverged influenza virus proteins. Since sequences from the same HIV-1 subtype differ at only about 10% of amino acids (Li *et al.*, 2015), the majority of residues with which a given amino acid interacts will be conserved and the fitness effects of mutations are expected to be similar across HIV strains. Consistent with such a universal fitness landscape, reversion of CTL escape mutations upon transmission to a new host is common (Friedrich *et al.*, 2004; Leslie *et al.*, 2004; Li *et al.*, 2007) and has been quantified during transmission (Carlson *et al.*, 2014) and during chronic infection (Zanini *et al.*, 2016).

Two main approaches to estimate fitness costs have been pursued. First, the cost of individual mutations can be quantified by competing mutant and wild-type viruses in cell culture (Martinez-Picado and Martinez, 2008; Parera *et al.*, 2007). Similar measurements of replication capacity are done routinely for drug resistance testing (Petropoulos *et al.*, 2000) and have been used to infer the fitness landscape of the HIV-1 protease and reverse transcriptase (Hinkley *et al.*, 2011). Recently, high-throughout methods have been developed to identify the amino acid preferences or fitness costs at every position in a protein (Acevedo *et al.*, 2014; Rihn *et al.*, 2015; Thyagarajan and Bloom, 2014). An alternative approach is to estimate the fitness landscape indirectly from large global collections of sequences (Dahirel *et al.*, 2011; Ferguson *et al.*, 2013), under the key assumption that high fitness variants are at high frequency in the global HIV-1 population. Either approach has limitations: whereas cell culture experiments are not sensitive to small costs (below 5%), models based on cross-sectional data are confounded by immune escape because they cannot differentiate between selective sweeps and absence of functional constraints.

Here, we estimate the fitness landscape and the rates and spectrum of mutations of HIV-1 using whole genome deep sequencing data from longitudinal samples (Zanini *et al.*, 2016). In contrast to previous efforts, we determine fitness costs from the *in vivo* intrapatient balance of mutation and selection against deleterious variants. Our estimates are most sensitive for small and moderate costs (between 0.1% and 10%), not affected by patterns of immune escape, and not restricted to one single protein: we estimated fitness costs at almost every position of the HIV-1 genome. This direct analysis from intrapatient diversity data can be used to quantify the relationship between sequence conservation across the HIV-1 pandemic and direct fitness costs of mutations.

Results

We previously reported whole genome deep sequencing of HIV-1 RNA from 6-12 samples from 9 untreated patients (Zanini *et al.*, 2016). RNA was reverse transcribed and amplified in six overlapping fragments and sequenced to high coverage on an Illumina MiSeq. Depending on template input, minor variation at frequencies down to 0.3% could be detected and frequencies could be reliably measured down to about 1% (see Zanini *et al.* (2016) and Methods below). For some of the analyses below, we include one additional patient (p7) described in (Brodin *et al.*, 2016), for a total of 82 plasma samples. Eight of the ten patients were infected with subtype B, one with subtype C, and one with subtype CRF01_AE.

We first discuss how we estimated the *in-vivo* mutation rate matrix of HIV-1 from the accumulation of mutations at approximately neutral sites. We then show how these rates can be used to establish a quantitative correspondence between fitness costs and global diversity at non-neutral sites, and present site specific fitness cost estimates of mutations at almost every site in the HIV-1 genome.

Neutral mutation rate matrix

Mutations at neutral sites accumulate freely over the time of infection and the average genetic distance from the founder sequence of later samples increases linearly with the time since infection. This rate of divergence at neutral sites is precisely the *in vivo* mutation rate (Kimura, 1968). (Deleterious mutations, in contrast, accumulate more slowly and we will use this saturation to estimate their fitness costs.)

To estimate the neutral mutation rate, it is crucial to identify a set of positions at which mutations are approximately neutral – otherwise the mutation rate will be underestimated. We selected a set of synonymous mutations that (i) are not part of known RNA secondary structures or overlapping reading frames, (ii) are globally unconserved (diversity > 0.3 bits), (iii) are outside gp120 which has been shown to be sensitive to synonymous mutations and recoding (Vabret *et al.*, 2014; Zanini and Neher, 2013), and (iv) align to HXB2. Fig. 1A and B show the average divergence from the approximate virus founder sequence in this neutral set, for all 12 nucleotide substitutions. We pooled data from patients p1, p2, p5, p6, p8, p9, p11 (those with early samples and without suspected dual infection); the error bars indicate standard deviations over patient bootstraps. The data confirm that divergence increases linearly, suggesting that our criteria for approximate neutrality succeed to identify a set of sites that are not strongly affected by selection. We can estimate the mutation rate matrix by linear regression – indicated by straight lines. Transition rates are about 5-fold higher than transversions, while the total mutation rate per site is about $1.2 \cdot 10^{-5}$ per site and day. The highest rate is G→A, while the lowest rates are estimated to be those between Watson-Crick binding partners. The smallest rates cannot by measured accurately because the corresponding mutations are hardly observed. The estimated rates are insensitive to the exact criteria used to select the set of neutral positions (see Fig. S2).

The estimated matrix (Fig. 1C) agrees well with previous estimates of HIV-1 mutation rates obtained using lacZ assays in cell culture (Abram *et al.*, 2010; Mansky and Temin, 1995), see Fig. S1. This quantitative agreement suggests that the average properties of mutations to HIV-1 depend little on the host cell. To obtain sufficient statistics, we measure the rate averaging across many sites; the mutation rates at single sites are known to depend on local sequence context (Abbotts *et al.*, 1993; Lewis *et al.*, 1999).

The high G→A rate might be partially due to the effect of human deaminases such as APOBEC3G. But the G→A rate is consistent with the rate estimated by Abram *et al.* (2010) who produced virus with an APOBEG3G negative cell line such that APOBEG3G probably only makes a minor contribution. We do not observe rates as high as estimated from integrated proviral DNA (Cuevas *et al.*, 2015). The high rate is likely due to the contribution of heavily hypermutated genomes and is discussed below.

Similarly, the functional latent reservoir of HIV-1 is unlikely to bias our estimates of mutations rates. In a recent study of proviral DNA in the same patients, we found that the latent reservoir is an accurate snapshot of the HIV-1 diversity circulating in the year prior to the sample (Brodin *et al.*, 2016). Hence we don't expect that the accumulation of diversity is delayed in substantial ways by contributions from reactivated latent virus.

Landscape of fitness costs in the HIV-1 genome

In contrast to neutral mutations, deleterious mutations reduce the replication rate of viruses carrying them. As a result, they accumulate less rapidly. The temporal dynamics of their frequency $x(t)$ is roughly described by

$$\frac{d}{dt}x(t) = \mu - sx(t) + \xi(x,t) \tag{1}$$

where μ and s are the mutation rate and fitness cost specific to the SNP in question, respectively (Haigh, 1978; Haldane, 1937). The last term $\xi(x,t)$ describes stochastic effects including genetic drift and selection on linked SNPs at other loci in the genome. Frequent recombination within HIV-1 populations (Neher and Leitner, 2010; Zanini *et al.*, 2016) reduces

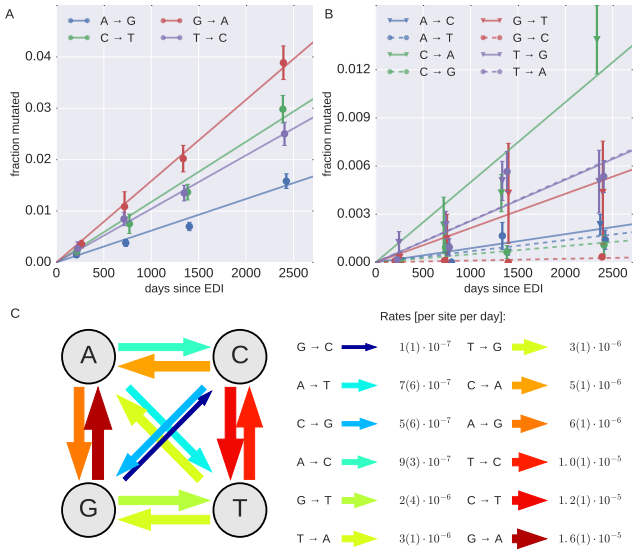

FIG. 1 **Accumulation of approximately neutral mutations over time.** Panels A&B show the accumulation of mutations at approximately neutral sites over time averaged over patients p1, p2, p5, p6, p8, p9, p11, for transitions (A) and transversions (B). EDI: estimated date of infection. (C) The slope of the individual regression lines in panel A&B provide estimates of the in vivo mutation rates. Error bars for the estimates, indicated in parenthesis as uncertainties over the last significant digit, are standard deviations over 100 patient bootstraps.

the effects of linked selection such that Eq. (1) can be a useful approximation. Depending on whether linked selection or genetic drift dominates the stochastic component, the absolute value of $\xi(x,t)$ is in average proportional to x or $\sqrt{x}$, respectively (Kimura, 1955; Neher, 2013). By definition, the average of ξ is zero.

Starting with a genetically monomorphic population, the *average* trajectory of a SNP frequency is given by

$$\langle x \rangle = \frac{\mu}{s}(1 - e^{-st}) \tag{2}$$

and saturates at $\bar{x} = \mu/s$ after a time of order s^{-1} (Haldane, 1937). If an appropriate average of the data is available, the fitness cost s can be estimated both from the approach to saturation and the level of saturation μ/s. Linear accumulation of neutral mutations is recovered in the limit $s \to 0$. This approach has been generalized to complex fitness landscapes (Seifert *et al.*, 2015).

Eq. (2) describes the *average* trajectory, but trajectories of individual SNPs are noisy. To make progress, trajectories at many sites or in many samples need to be averaged in ways that preserves important features of the fitness landscape.

Relationship of global conservation and fitness costs

In first approximation, conservation of a site across global HIV-1 diversity is expected to be a proxy for high fitness cost of mutations at that site, while mutating a site that is observed in many different states probably doesn't affect fitness much. To quantify the relationship between conservation and fitness cost s, we group sites in the HIV-1 genome by global diversity in group M (measured by Shannon entropy of an alignment column, see Methods). We chose six groups of equal size, i.e., quantiles of global diversity. Instead of estimating fitness costs for all three possible mutations at a given site, we estimated one fitness cost parameter for each site as the cost of the typical mutation away from the founder virus sequence (a more elaborate model that includes the 12 different mutation rates is described in Fig. S3). For each conservation group, we average the frequencies of non-founder nucleotides over all sites and patient samples in 7 time bins. These average divergences are indicated by dots in Fig. 2A along with a nonlinear least square fit of Eq. (2) to the data of each quantile (each color indicates a conservation group, blue to red by increasing diversity). The least conserved group accumulates divergence linearly – this is consistent with our mutation rate estimates above. With increasing conservation, divergence saturates more rapidly and at lower levels. We set $\mu = 1.2 \cdot 10^{-5}$ per site per day according to our estimate of the neutral mutation rate and fit a single parameter, the fitness cost s, for each group. The estimated average costs and their error bars from 100 bootstraps over patients are shown in Fig. 2B as a blue line ("Sat").

The fitness cost of mutations in the least conserved 1/6 of the genome is undetectably small, consistent with neutrality. More conserved sites have higher costs, up to about 1% for sites where the group M alignment entropy of 0.03 bits. For even more conserved sites, saturation is very fast and we estimated the fitness cost using a different averaging procedure (see below).

Notice that for Eq. (2) to hold, it is essential that the infection is founded by a single founder sequence. For this reason, patients p3 and p10 were excluded from this part of the analysis since there is evidence indicating that they were infected by more than one viral variant. Furthermore, it is important to exclude sites subject to immune selection and sites where the initial nucleotide differs from the global consensus. Otherwise, rapid rise of beneficial mutations driven by CTL escape or reversion increase divergence and result in underestimation of the fitness costs.

Site-specific fitness costs in the HIV-1 genome

In addition to averaging mutation trajectories across multiple sites, we also estimated site-specific fitness costs by averaging data from multiple samples during late infection. Average frequencies at sites where mutations carry large costs saturate rapidly after a time $1/s$. Frequencies of minor variants in different samples are therefore uncorrelated and can be averaged to increase the accuracy of frequency estimates, which then allows direct estimation of site specific costs s_i from the relation $\bar{x}_i = \mu/s_i$.

Specifically, we calculate a weighted average frequency from the samples from each patient. The average frequency

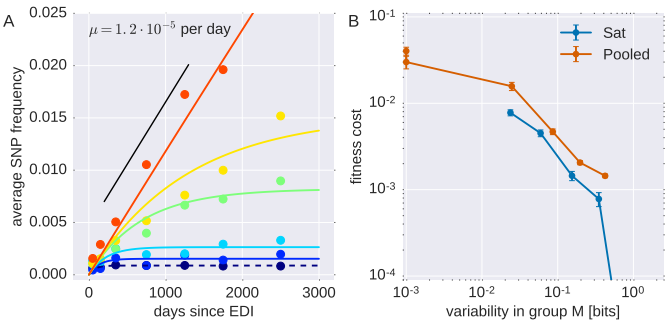

FIG. 2 **Average intrapatient fitness cost within quantiles of global HIV-1 group M diversity.** (A) Average derived SNP frequencies (1 - frequency of the ancestral state) saturate fast at positions in the conserved quantiles (blue), while intrapatient diversity keeps increasing in variable quantiles (yellow to red). The initial slope is the mutation rate $1.2 \cdot 10^{-5}$ per site per day. The solid lines show fits of Eq. (2) to the binned data, from which we estimate average selection coefficients shown in panel (B) labeled "Sat" (this method is not applicable in the most conserved third of the genome). The "Pooled" line refers to harmonic averages of site-specific cost estimates. Error bars indicate 100 bootstraps over patients: note that while error bars are small, there is substantial variation of fitness costs within each diversity quantile. Positions at which putatively adaptive mutations have swept through the population have been excluded.

of nucleotide or amino acid α at position i is then given by

$$\hat{x}_{i,\alpha} = \frac{1}{\sum_k w_k} \sum_k w_k x_{k,i,\alpha} \tag{3}$$

where $x_{k,i,\alpha}$ is the frequency in sample k (k runs over all plasma samples from the patient at least 2 years after infection). The weight w_k accounts for the variable number of HIV-1 genomes that contributed to the sequencing library as estimated by limiting dilution (see methods below and Zanini *et al.* (2016)). From individual samples, frequencies above the error rate of 0.002 are assumed to avoid inflation by sequencing and PCR errors (we never observed errors above this level in our control samples). After averaging samples within patients, we average $\hat{x}_{i,\alpha}$ over patients and sum all non-consensus nucleotides or amino acids to obtain the average non-consensus frequency $\hat{x}_i$ for each position i in the HIV-1 genome; the cost at that position is then given by $\mu/\hat{x}_i$ where μ is the mutation rate at that position.

As above, we only include data from a particular sample if the majority nucleotide agrees with the global consensus and at which no potential sweep was observed. Without this restriction, the estimated fitness costs would be biased downward by reversions and immune selection.

Notice that although the combined sequencing and PCR error can be up to 0.002 and we don't use counts below this threshold for any *single* sample, pooling many samples allows to estimate much smaller *average* frequencies: if a mutation is present at frequency 0.005 in 10% of samples, its average frequency is 0.0005. This type of averaging works precisely because frequencies of individual costly mutations are noisy and rare variants are brought to measurable frequencies occasionally by linked selection and sampling.

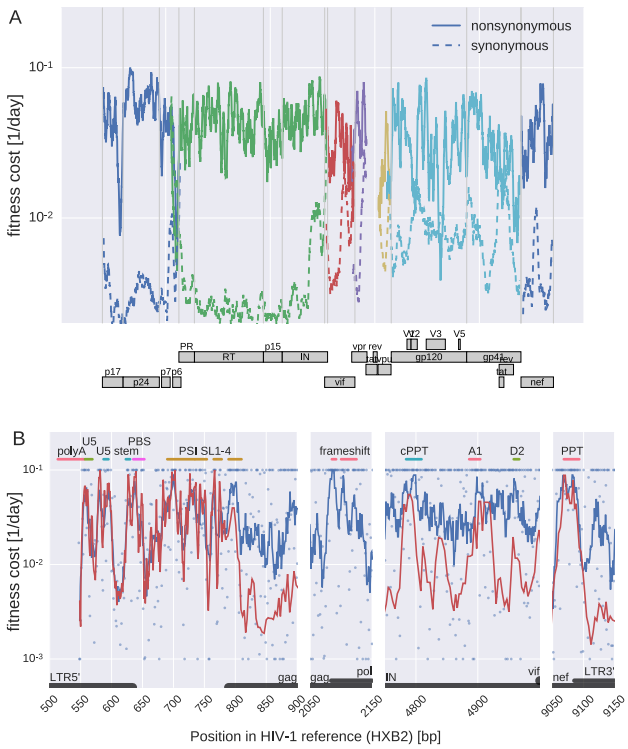

FIG. 3 **Fitness costs along the HIV-1 genome**. Panel (A) shows fitness costs of synonymous and nonsynonymous mutations in gag, pol, vif, vpu, env, and nef as a geometric sliding average with window size 30. Note that frequency estimates in gp120 are expected to be less accurate due to consistent difficulties amplifying this part of the genome. Panel (B) shows fitness costs in selected regions of the genome that contain important regulatory elements. Blue dots show estimates for individual bases, blue lines show running averages over 8 bases and red lines show running averages excluding bases where mutations cause amino acid changes. PBS: tRNA primer binding site. U5: unique 5' region. SL 1-4 PSI: stem loops of the PSI packaging signal. (c)PPT: (central) poly purine tract. A1, D2: splice sites.

Fig. 3A shows fitness costs of mutations at most positions along the HIV-1 genome (including env) separately for synonymous and nonsynonymous mutations: the numerical estimates are available for all sites in the Supplementary Materials. The costs of synonymous and nonsynonymous mutations are clearly different, and distinct peaks are observed at several locations across the genome. Before analyzing these patterns in details (see below), as a consistency check we compared in Fig. 2B the average estimates ("Pooled" line) to our previous estimates "Sat", which take into account the explicit time information of the samples. We found good agreement between the two approaches. To further assess the accuracy of our estimates, in Fig. S5 we show the variation in the fitness cost estimate after bootstrapping over patients. The variation is approximately twofold in each direction, so fitness costs above 5% are clearly separated from costs of 1% or less.

Fitness costs estimated from within patient diversity data correlate strongly with global HIV-1 group M diversity (rank

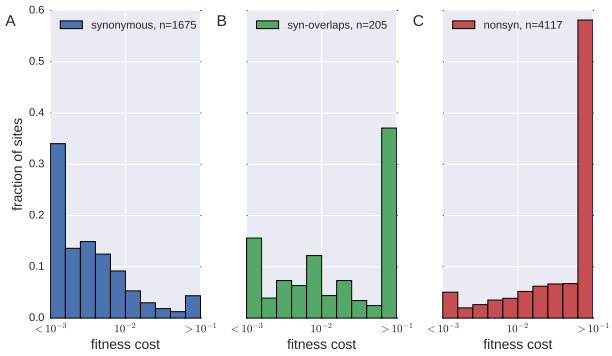

FIG. 4 **Distributions of fitness costs.** Distributions of (A) synonymous mutations, (B) mutations that are synonymous in one gene but affect another protein in a different reading frame and (C) nonsynonymous mutations (includes codons in gag, pol, vif, vpu, vpr). The extremal bins include all points larger or smaller than the axis boundary.

correlation $\rho \approx 0.7$ for per site diversity measured by entropy, see Fig. S4). Importantly, a particular site contributes to the estimate only if the founder and majority nucleotide in that sample equals the consensus variant. This condition removes any direct signal of cross-sectional diversity. The correlation increases as intrapatient variation is estimated using more patients (see Fig. S4), suggesting that fitness costs at individual sites is largely conserved between patients. Fig. S4 also shows scatter plots of global diversity vs fitness costs.

Distributions of fitness costs

We observe marked differences between the distributions of fitness costs of synonymous and non-synonymous mutations (see Fig. 4): about half of all nonsynonymous mutations have estimated fitness costs in excess of 10%, while the majority of synonymous mutations have fitness costs below 1%. The distribution of fitness costs of mutations that are synonymous in one gene, but that affect another gene in a different reading frame, resembles that of nonsynonymous mutations (see Fig. 4B). We estimate about 10% of synonymous mutations outside env to be highly deleterious; we discuss the specific costs of synonymous mutations in more detail below.

Fig. S6 shows the distribution of fitness costs for different genes. In gag and pol, the contrast between synonymous and nonsynonymous mutations is greatest. Synonymous mutations are costly in several isolated regions discussed below but have low fitness effects in much of pol and gag.

The distribution of fitness costs is consistent with those found in other viruses, where typically about 20-40% of mutations are lethal and another $\sim 40\%$ are strongly deleterious with about 30% being weakly deleterious or neutral (Sanjuán, 2010).

Fitness costs peak at functional RNA elements

The HIV-1 genome contains a number of well characterized RNA elements that regulate different stages of the replication cycle. Many of these elements are embedded in protein-coding sequence and because selection reduces genetic diversity (Mayrose *et al.*, 2013; Ngandu *et al.*, 2008) we expect to estimate higher fitness costs in these regions. Indeed, in Fig. 3B important regulatory elements are clearly visible as well defined peaks in the running averages of fitness costs along the genome. In the 5' LTR the largest fitness costs overlap with the hairpin containing the poly-A signal, the U5 sequence (Lu *et al.*, 2011), the base of the following hairpin, the primer binding site (PBS) and the 1-4 for the PSI element (LANL HIV sequence data base, 2016). The frameshift region (slippery sequences plus hairpin), the splice acceptor site A1, and the polypurine tracts (PPT) in integrase and at the 3' LTR show similarly high fitness costs (the TAR element is only partially covered by the sequencing data set and hence not shown here). Mutations within the fourth stem loop of PSI are almost never observed, while synonymous sites are almost free to vary beyond the end of the stem. Synonymous mutations in the RRE are costly, but not as deleterious as those in PPT, the splice acceptor site A1, or the PSI element, indicating a higher evolutionary plasticity. Among the more striking patterns is also the drop in synonymous cost at the beginning of gag. Beyond these known elements, the correlation of fitness costs at synonymous mutations with cross-sectional diversity suggests that there are a number of additional regions with important function on the RNA level, for example a double peak in p24 and three more peaks in pol. While well characterized RNA elements correspond to clear patterns in the estimated fitness costs, RNA secondary structure prediction correlate poorly with fitness costs (see Fig. S11 and discussion below).

Fitness costs and immune selection

Among sites that are globally variable (Shannon entropy above 0.1 bits), nonsynonymous mutations are enriched despite having a high fitness cost (cost > 0.03 per day, odds ratio 15). This enrichment is most pronounced in pol, gag and nef with little enrichment in env. This observation is consistent with host-specific selection pressures (CTL selection) at sites with a large fitness cost; the resulting adaptations revert quickly when transmitted to a new host (Friedrich *et al.*, 2004; Leslie *et al.*, 2004; Li *et al.*, 2007; Zanini *et al.*, 2016).

Such patient-specific selection has the potential to blur the relationship between fitness cost and diversity, as shown in Fig. 5A for nef (see Fig. S4 for other genes). The majority of sites with high fitness costs and high cross-sectional diversity (upper right corner) have been reported to be associated with HLA type ((Carlson *et al.*, 2012), shown in red) or with low viral load ((Bartha *et al.*, 2013), annotated dots). HLA-associated sites that fall into the top right corner of Fig. 5A are of particular interest since they are expected to result in virus control (Pereyra *et al.*, 2014).

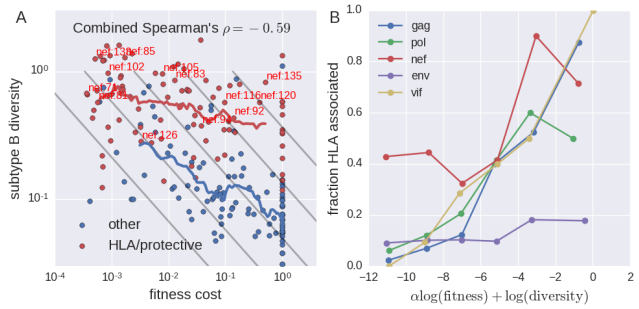

FIG. 5 **CTL selection blurs the relationship between fitness costs and diversity.** (A) The majority of sites in nef with high diversity despite high fitness costs are associated with HLA types (red) (Carlson *et al.*, 2012) or with low viral load (annotated dots) (Bartha *et al.*, 2013). (B) Quantification of the fraction of HLA associated sites in bins of increasing diversity and fitness costs (indicated by straight grey lines in (A) with $\alpha = 2$). This figure uses data from subtype B patients only.

To quantify the overrepresentation of HLA associated sites among diverse positions where mutations incur large fitness costs, we plotted the fraction of HLA associated sites in bins indicated by diagonal straight lines in Fig. 5A for the genes gag, pol, vif, env, and nef. Bin boundaries are defined by $\alpha\log(\text{fitness}) + \log(\text{diversity}) = \text{const.}$ with $\alpha = 2$. For all genes test other than env, the fraction of HLA associated sites increases strongly in bins corresponding to high diversity and fitness cost indicating that CTL selection pressure is responsible for global diversity that is deleterious to virus replication.

HLA associations can only be detected for sites with some global variation. Hence there is a strong ascertainment bias and almost all HLA associated are found in the top half of Fig. 5A. Without independent characterization of this bias, a statistical assessment of the relation between CTL selection pressure, fitness cost, and global diversity remains challenging.

Fitness costs are weakly correlated with protein disorder and solvent accessibility

Perturbations to protein structure are expected to reduce virus fitness. Hence mutations that decrease stability, occur in tightly packed regions, or are deeply buried in the protein are expected to incur the greatest fitness costs. Disorder scores and solvent accessibility have been compared to cross-sectional diversity by Li *et al.* (2015). We correlated these *in-silico* derived scores with intrapatient diversity, finding rank correlation coefficients of about 0.2-0.4 for disorder scores and solvent accessibility. While highly statistically significant, the fraction of variation in diversity explained by these scores is low; this is consistent with previous observations by Meyer and Wilke (2015). By far the best correlate of fitness cost is cross-sectional conservation, see Table I.

The distribution of fitness costs depends strongly on the consensus amino acid. Mutations of cysteins (C), histidines (H), prolines (P), tryptophans (W), and tyrosines (Y) tend to be most costly, while mutations of glutamic acid (E), lysine (K), aspartic acid (D) and arginine (R) are less often very deleterious. These patterns are consistent in gag, pol, and env, see Fig. S7.

gene	group M	subtype B	disorder	accessibility	RNA
gag	-0.51	-0.59	-0.23	-0.26	0.13
pol	-0.56	-0.59	-0.13	-0.31	0.09
nef	-0.54	-0.59	-0.30	-0.19	0.11
env	-0.47	-0.46	0.00	0.07	0.09
vif	-0.57	-0.69	-0.08	-0.16	0.06

TABLE I Spearman's correlation coefficients of fitness estimates with cross-sectional diversity (measured as entropy in group M and subtype B alignments), disorder scores and solvent accessibility values obtained from (Li *et al.*, 2015). The column "RNA" contains rank correlation coefficients of fitness at synonymous mutations with the pairing probability predicted by (Siegfried *et al.*, 2014). Fig. S4 shows how intrapatient/global diversity correlations improve when basing intrapatient estimates on larger numbers of patients.

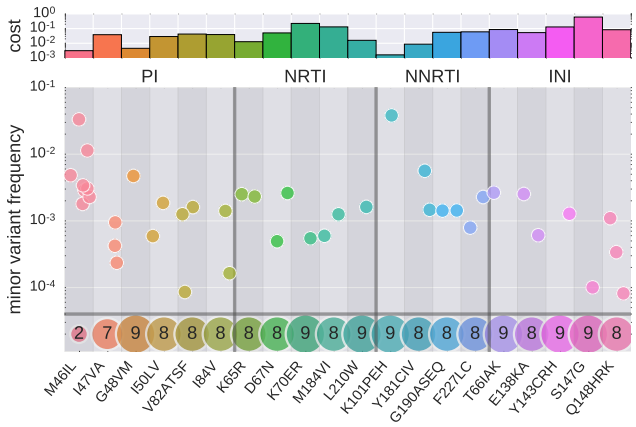

FIG. 6 **Pre-existing drug resistance mutations.** Each point shows the time averaged frequency of minor amino acids in individual patients. The bottom row indicates in how many out of 10 patients each mutation is not observed. Most mutations are observed only in a minority of patients suggesting high fitness costs. The top panel shows the estimated fitness costs associated with the mutations. The following mutations were never found at frequencies above 0.1% in any patient, indicating a large fitness cost: PI: L24I, V32I, I54VTAM, L76V, N88S, L90M; NRTI: M41L, K70ER, L74VI, Y115F, T215YF, K219QE; NNRTI: L100I, K103N, V106AM, E138K, V179DEF, Y188LCH, M230L; INI: E92Q, N155H.

Most drug resistance mutations have a large fitness cost

Of particular interest are the fitness costs of mutations that confer resistance against anti-retroviral drugs. The most commonly administered drugs are nucleoside analog reverse transcriptase inhibitors (NRTI), non-nucleoside analog reverse transcriptase inhibitors (NNRTI), protease inhibitors (PI), and integrase inhibitors (INI). Resistance mutations against these drugs are well known (Johnson *et al.*, 2011).

Pre-existing low frequency drug resistance mutations have been associated with failing therapy (Johnson *et al.*, 2008; Li *et al.*, 2011). Some deep-sequencing studies have characterized such pre-existing variation in treatment-naive patients and found that drug-resistance mutations are usually below the detection limit, suggesting relatively high fitness costs (Gianella *et al.*, 2011; Hedskog *et al.*, 2010; Li *et al.*, 2011). Fig. 6 shows estimated frequencies of several drug resistance mutations in the different patients. The majority of mutations are not seen at all, while most of the remainder is observed in only one or two patients (pooled across all time points of each patient). Only the protease mutation M46I is observed consistently across several patients.

The frequency of drug resistance mutations is expected to be inversely proportional to their fitness cost in absence of treatment and of some these costs have been measured in cell cultures (see e.g. Chow *et al.* (1993); Cong *et al.* (2007); Martinez-Picado and Martinez (2008)). Many resistance mutations quickly revert upon treatment interruption suggesting high fitness costs (Deeks, 2003; Hedskog *et al.*, 2010; Joos *et al.*, 2008). Indeed, for most drug resistance mutations, we estimate fitness costs in excess of 5% (sites where minor variation is not or only sporadically observed), see top panel in Fig. 6. Note that the costs of very deleterious mutations tend to be underestimated if the mutations are only observed in a small number of patients. For instance, G48VM in the protease and K101PEH in the reverse transcriptase are attributed a low cost but are only observed in one patient, so their actual cost might be larger.

Discussion

Sequence evolution of HIV-1 is the determined by the rate and spectrum of mutations as well as their phenotypic effects. Many studies have focused on beneficial mutations that sweep across the intrapatient HIV-1 population (Asquith *et al.*, 2006; Ganusov *et al.*, 2011; Kessinger *et al.*, 2013; Neher and Leitner, 2010), and we observe similar patterns in our study subjects (see Fig. S8 and Fig. S9). The majority of mutations, however, are deleterious and stay at low frequencies within hosts; selection is constantly pruning deleterious variation from the population to maintain a functional genome. Deleterious mutations contribute substantially to sequence evolution due to their large number: if 5000 sites accumulate deleterious variation at frequencies of 1%, the typical HIV-1 genome will contain 50 such mutations. Here, we used longitudinal whole genome deep sequencing data from (Zanini *et al.*, 2016) to quantify the in vivo mutation rates of HIV-1 and the fitness costs of deleterious mutations.

The accumulation of mutations at approximately neutral sites is consistent with the mutation rates of HIV-1 measured in cell culture using lacZ assays (Abram *et al.*, 2010; Mansky and Temin, 1995). This agreement suggests that the mutation rate of HIV-1, which is the joint rate of the HIV-1 RT, mutagenesis by the innate immune system, and the human DNA-dependent RNA polymerase II, is largely independent of cell type. Because the cell culture studies used an exogenous template while we monitor mutations on the HIV-1 genome itself, it appears also that the mutation rate does not depend, in average, on the nature of the template. The mutation rate at specific genomic sites, however, is likely to depend on the sequence context, similar to other polymerases and as indicated by previous studies (Abbotts *et al.*, 1993; Lewis *et al.*, 1999). The highest rate is $G \rightarrow A$ and transitions are about 5-fold faster than transversions; the lowest rates are between base pairing partners, e.g. $G \leftrightarrow C$, see Fig. 1. If the human RNA pol II has similar error rates as its *C. elegans* homologue (error rate 4×10^{-6} per site (Gout *et al.*, 2013)) roughly a fifth of all mutations observed in HIV are due to the RNA polymerase (assuming an HIV generation time of 1-2 days).

While consistent with cell culture estimates, the rates we estimate are incompatible with those reported by Cuevas *et al.* (2015). Whereas we measure mutations in the population of RNA virions, Cuevas *et al.* (2015) counted nonsense mutations in proviral DNA integrated into host cell genomes and estimated a rate of 4×10^{-3} per site and replication – more than 100 times higher than our estimate. Unlike in circulating viral RNA, a large fraction of proviral HIV DNA is heavily hypermutated by enzymes of the APOBEC family (Malim, 2009). Hypermutation is approximately an all-or-nothing phenomenon in which either a sequence contains dozens of stop codons or none (Armitage *et al.*, 2012; Cuevas *et al.*, 2015; Delviks-Frankenberry *et al.*, 2016).

Because of this bimodal nature, hypermutation and reverse transcriptase mutation can not be meaningfully described by one mutation rate matrix. In the former case, a sequence with dozens of stops integrates into the host genome as a defective virus, the in the latter rare independent mutations (about 0.2 per genome) can lead to gradual evolution and adaptation. Sporadic deamination by APOBEG enzymes might still contribute to the G→A mutation rate and is included in our estimate, but heavily hypermutated sequences are likely "dead on arrival" and make a minor contribution to genetic diversity, as also argued by others (Armitage *et al.*, 2012; Delviks-Frankenberry *et al.*, 2016).

Furthermore, proviral HIV DNA is enriched for hypermutated sequences. While functional proviruses rapidly lead to death of the infected cell, hypermutated proviruses tend to accumulate latently in HIV-1 target cells over the many months a T-cell can live. This accumulation likely results in a multifold overrepresentation of hypermutated sequences compared to the probability at which hypermutation happens in a single reverse transcription. Because we measure mutations from RNA data from plasma, our estimates are not affected by this accumulation bias.

With the time calibrated mutation rate estimates, we estimated absolute fitness costs from mutation selection balance and quantify the relationship between group M diversity and fitness cost. Overall, fitness costs explains about half of the diversity in global alignments of HIV-1 sequences, while the remainder might be linked to patient-specifc processes such as immune escape. The relationship between logarithmic group M diversity (measured as entropy) and logarithmic fitness costs is approximately linear.

Our site-specific fitness landscape highlights a number of

known functional elements across the HIV-1 genome, including regulatory elements at the RNA level. Constraints on synonymous mutations appear to be stronger and more prevalent in env than in gag or pol, consistent with earlier results that many synonymous mutations in gp120 tend to be weakly deleterious (Zanini and Neher, 2013) and that env recoding results in non-infectious virus (Vabret *et al.*, 2014). However, comparison of our fitness cost estimates with genome wide RNA structure predictions by Siegfried *et al.* (2014) and Sükösd *et al.* (2015) show little correlation. While mutations in validated RNA structure elements are associated with high fitness costs, genome wide predictions of RNA structure explain little variation in fitness costs of synonymous mutations (see Fig. S11 and Tab. I). This lack of strong correlation is consistent with the observation that (predicted) pairing patterns evolve rapidly in most of the genome (Pollom *et al.*, 2013) or might reflect inaccuracies in RNA structure prediction: only a minority of pairings agree between the predictions by Siegfried *et al.* (2014) and Sükösd *et al.* (2015).

Several groups have estimated fitness costs within HIV-1 proteins using experimental approaches (Martinez-Picado and Martinez, 2008; Rihn *et al.*, 2015; Thyagarajan and Bloom, 2014). Our estimates presented here are complementary to those studies in two ways. First, because of the short but dense temporal sampling, cell culture experiments are sensitive to large fitness costs, typically above $> 5\%$, while estimates from natural variation are most accurate for effects below a few percent. Second, ex vivo estimates are not affected by the specific conditions of cell culture systems.

Computational methods to estimate fitness landscapes from cross-sectional data have also been proposed (Dahirel *et al.*, 2011; Ferguson *et al.*, 2013), including a recent effort to include intrapatient diversity via shallow sequencing (Hartl *et al.*, 2016). The relationship between fitness cost and diversity, however, might be blurred since sites that are costly to mutate might still be globally diverse due to frequent escape from CTL pressure. Indeed, we have shown in Fig. 5 that globally polymorphic sites that we estimate to have high fitness costs are over-represented among sites known to be HLA associated (Carlson *et al.*, 2012). Barton *et al.* (2016) have shown that the rate of CTL escape depends on fitness costs. More generally the cross-sectional inferences and our intra-patient inferences reinforce the notion that HIV-1 evolution is governed by a fitness landscape that consists of a universal component determining the replicative capacity of the virus, and a host specific component responsible of escape mutations (Shekhar *et al.*, 2013). Our approach based on longitudinal deep intrapatient data allows to explicitly disentangle these two contributions, since we can condition on the founder sequence and the absence of host-specific selective sweeps. Purely cross-sectional inferences of the fitness landscape likely underestimate the fitness cost of mutations at HLA associated positions.

In the future, as whole genome deep sequencing becomes more common, estimates of mutation rates and the fitness landscape could be extended to a higher number of samples. A much larger sample pool might allow site-specific inference of the mutation rates. Furthermore, by providing more accurate minor SNP frequencies, estimates of their associated fitness costs will improve, leading to a deeper understanding of the selective forces that shape viral evolution.

Materials and Methods

Code and data availability

The sequences from the longitudinal samples were taken from Zanini *et al.* (2016) and analyzed using the library `hivevo_access` (https://github.com/neherlab/HIVEVO_access) and custom scripts.

The nucleotide and amino acid cross-sectional alignments of HIV-1 group M were downloaded from the Los Alamos National Laboratory HIV database and filtered for short or otherwise problematic sequences and are available as supplementary material.

Disorder and solvent accessibility scores amino acids for different HIV proteins were provided by the authors of (Li *et al.*, 2015) (available at `www.virusface.com`). These scores were mapped to homologous positions in the virus populations via alignments to the reference sequence NL4-3. Positions without scores were discarded.

Our analysis scripts, as well as the resulting data for the mutation rate and fitness cost estimates, are available online at https://github.com/iosonofabio/HIV_fitness_landscape.

Mutation rate estimation

For each patient, a set of nucleotide sites is identified, for which (i) the entropy in a group M alignment is higher than 0.1 bits and (ii) the consensus nucleotide of the earliest sample corresponds with the HIV-1 group M consensus. Derived alleles at those sites are considered if (i) they are translated in a single reading frame, (ii) they are synonymous changes, (iii) they are outside of known RNA structures or overlapping reading frames. The frequencies of these variable synonymous changes are grouped by mutation (e.g. $A \to G$) and averaged across the genome and different samples with the following time bins: [0, 500, 1000, 1750, 3000]. Variations of the parameters have been tested and yielded similar results. The time-binned average frequencies are modeled by a linear fit with zero intercept, so the inferred rate $\hat{\mu}$ is:

$$\hat{\mu} = \frac{\sum_i t_i \cdot x_i}{\sum_i t_i^2},$$

where (t_i, x_i) are the time and frequency of each point (see Fig. 1A&B). Different mutations are estimated (independently) to obtain the entire mutation rate matrix. The whole procedure is repeated for 100 bootstraps over patients to estimate the uncertainty of the rates, shown as $\pm$ errors in Fig. 1C. An error of ± 0.0 means an uncertainty smaller than ± 0.1. See the supplementary script `mutation_rate.py` for the estimate implementation.

Estimation of selection coefficients

The selection coefficients were estimated using two different approaches, called "Sat" and "Pooled" in Fig. 2B.

Nonlinear least squares on saturation curves

To estimate the fitness costs as in the "Sat" curve of Fig. 2B, we considered all sites in genomes from viral populations of all patient at which (i) the majority nucleotide at the earliest time point equals the global HIV-1 group M consensus and (ii) the majority nucleotide does not change during the infection. The latter criterion is necessary to ensure we exclude sites under positive selection. At each site, instead of modeling the whole set of 4 possible nucleotides, we used a simplified 2-state model: the subtype M consensus state and the sum of the derived mutations. We collected the frequencies of the derived states from all sites and patients and averaged into two-dimensional bins, by entropy category and time since Estimated Date of Infection (EDI). The averages in each entropy group are shown in Fig. 2A as dots: each color indicates a different entropy group (from blue to red, low to high). We fitted those points via nonlinear least squares to equation (2) with a single fit parameter, s. The resulting fits are shown in Fig. 2A and the fitness costs s in Fig. 2B.

Pooled SNP frequencies from late samples

To obtain site specific estimates, we averaged SNP frequencies at individual sites according to Eq. (3). The average is weighted to ensure that samples contribute approximately proportionally to the number of template molecules present in the sample. This weight is calculated from the estimated template input T_k as $w_k = (0.002 + 1/T_k)^{-1}$, where 0.002 is the combined error rate of RT-PCR and sequencing. Samples contribute proportionally to the number of RNA templates is small if T_k is small, while for large T_k the sequencing error rate is limiting and the per sample contribution is capped at 500. The weighted average is performed within each patient. To average SNP frequencies further over patients, we use the alignment of each patient to the NL4-3 reference sequence to identify homologous positions to average. As before, we exclude sites that don't agree with the global HIV-1 consensus and sites that sweep (i.e. where the majority state changes during infection). These exclusions are particularly important, since sites from different patients are combined and minor frequencies are only meaningful when measured relative to the same reference nucleotide or amino acid. To determine uncertainties, bootstrap distributions are constructed by resampling the patients contributing the average. Estimates of fitness costs for nucleotide and amino acid mutations where done in very similar ways.

Selection coefficients are estimates via $\mu/\hat{x}$, where μ is the sum of mutation rates away from the consensus nucleotide or amino acid estimated above. Amino acid mutation rates are calculated specifically for each patient on the bases of the codon coding for the amino acid in the founder sequence of that patient (amino acid changes requiring two nucleotide changes were ignored).

To determine the uncertainty of fitness cost estimates, we picked sites within small slices of the distribution of selection coefficients and constructed bootstrap distributions for the estimates at each of the positions. Fig. 4D shows the combined distributions for each of the positions contained in these initial slices.

Acknowledgements

We thank Lina Thebo and Crista Lanz for excellent technical assistance and Pleuni Pennings and Nate Cira for helpful comments on the manuscript. This work was supported by the European Research Council through grant Stg. 260686 and partly by grant NSF PHY11-25915 to KITP and the Swedish Research Council through grant K2014-57X-09935-23-5.

References

Abbotts, J., K. Bebenek, T. A. Kunkel, and S. H. Wilson, 1993, Journal of Biological Chemistry **268**(14), 10312, ISSN 0021-9258, 1083-351X.

Abram, M. E., A. L. Ferris, W. Shao, W. G. Alvord, and S. H. Hughes, 2010, Journal of virology **84**(19), 9864.

Acevedo, A., L. Brodsky, and R. Andino, 2014, Nature **505**(7485), 686, ISSN 0028-0836.

Armitage, A. E., K. Deforche, C.-h. Chang, E. Wee, B. Kramer, J. J. Welch, J. Gerstoft, L. Fugger, A. McMichael, A. Rambaut, and A. K. N. Iversen, 2012, PLoS Genet **8**(3), e1002550.

Asquith, B., C. T. T. Edwards, M. Lipsitch, and A. R. McLean, 2006, PLoS Biol **4**(4), e90.

Bar, K. J., C.-y. Tsao, S. S. Iyer, J. M. Decker, Y. Yang, M. Bonsignori, X. Chen, K.-K. Hwang, D. C. Montefiori, H.-X. Liao, P. Hraber, W. Fischer, *et al.*, 2012, PLoS Pathog **8**(5), e1002721, URL `http://dx.doi.org/10.1371/journal.ppat.1002721`.

Bartha, I., J. M. Carlson, C. J. Brumme, P. J. McLaren, Z. L. Brumme, M. John, D. W. Haas, J. Martinez-Picado, J. Dalmau, C. Lpez-Galndez, C. Casado, A. Rauch, *et al.*, 2013, eLife Sciences **2**, e01123, ISSN 2050-084X.

Barton, J. P., N. Goonetilleke, T. C. Butler, B. D. Walker, A. J. McMichael, and A. K. Chakraborty, 2016, Nat Commun **7**, 11660.

Brodin, J., F. Zanini, L. Thebo, C. Lanz, G. Bratt, R. Neher, and J. Albert, 2016, bioRxiv , 053983.

Carlson, J. M., C. J. Brumme, E. Martin, J. Listgarten, M. A. Brockman, A. Q. Le, C. Chui, L. A. Cotton, D. J. H. F. Knapp, S. A. Riddler, R. Haubrich, G. Nelson, *et al.*, 2012, J. Virol. , JVI.01998ISSN 0022-538X, 1098-5514.

Carlson, J. M., Z. L. Brumme, C. M. Rousseau, C. J. Brumme, P. Matthews, C. Kadie, J. I. Mullins, B. D. Walker, P. R. Harrigan, P. J. R. Goulder, and D. Heckerman, 2008, PLOS Comput Biol **4**(11), e1000225, ISSN 1553-7358.

Carlson, J. M., M. Schaefer, D. C. Monaco, R. Batorsky, D. T. Claiborne, J. Prince, M. J. Deymier, Z. S. Ende, N. R. Klatt, C. E. DeZiel, T.-H. Lin, J. Peng, *et al.*, 2014, Science **345**(6193), 1254031.

Chow, Y.-K., M. S. Hirsch, D. P. Merrill, L. J. Bechtel, J. J. Eron, J. C. Kaplan, and R. T. D'Aquila, 1993, Nature **361**(6413), 650.

Cong, M.-e., W. Heneine, and J. G. Garca-Lerma, 2007, Journal of Virology **81**(6), 3037, ISSN 0022-538X, 1098-5514.

Cuevas, J. M., R. Geller, R. Garijo, J. Lpez-Aldeguer, and R. Sanjun, 2015, PLoS Biol **13**(9), e1002251.

Dahirel, V., K. Shekhar, F. Pereyra, T. Miura, M. Artyomov, S. Talsania, T. M. Allen, M. Altfeld, M. Carrington, D. J. Irvine, B. D. Walker, and A. K. Chakraborty, 2011, PNAS **108**(28), 11530, ISSN 0027-8424, 1091-6490, URL http://www.pnas.org/content/108/28/11530.

Deeks, S. G., 2003, Lancet **362**(9400), 2002, ISSN 1474-547X.

Delviks-Frankenberry, K. A., O. A. Nikolaitchik, R. C. Burdick, R. J. Gorelick, B. F. Keele, W.-S. Hu, and V. K. Pathak, 2016, PLOS Pathog **12**(5), ISSN 1553-7374.

Doud, M. B., O. Ashenberg, and J. D. Bloom, 2015, Mol Biol Evol **32**(11), 2944, ISSN 0737-4038, 1537-1719.

Ferguson, A., J. Mann, S. Omarjee, T. Ndungu, B. Walker, and A. Chakraborty, 2013, Immunity **38**(3), 606, ISSN 1074-7613, URL http://www.sciencedirect.com/science/article/pii/S1074761313001076.

Friedrich, T. C., E. J. Dodds, L. J. Yant, L. Vojnov, R. Rudersdorf, C. Cullen, D. T. Evans, R. C. Desrosiers, B. R. Moth, J. Sidney, A. Sette, K. Kunstman, *et al.*, 2004, Nat Med **10**(3), 275.

Ganusov, V. V., N. Goonetilleke, M. K. P. Liu, G. Ferrari, G. M. Shaw, A. J. McMichael, P. Borrow, B. T. Korber, and A. S. Perelson, 2011, J. Virol **85**(20), 10518.

Gianella, S., W. Delport, M. E. Pacold, J. A. Young, J. Y. Choi, S. J. Little, D. D. Richman, S. L. K. Pond, and D. M. Smith, 2011, J. Virol. **85**(16), 8359, ISSN 0022-538X, 1098-5514.

Goonetilleke, N., M. K. P. Liu, J. F. Salazar-Gonzalez, G. Ferrari, E. Giorgi, V. V. Ganusov, B. F. Keele, G. H. Learn, E. L. Turnbull, M. G. Salazar, K. J. Weinhold, S. Moore, *et al.*, 2009, J Exp Med **206**(6), 1253, ISSN 0022-1007, 1540-9538.

Gout, J.-F., W. K. Thomas, Z. Smith, K. Okamoto, and M. Lynch, 2013, PNAS **110**(46), 18584, ISSN 0027-8424, 1091-6490.

Haigh, J., 1978, Theoretical Population Biology **14**(2), 251, ISSN 0040-5809.

Haldane, J. B. S., 1937, The American Naturalist **71**(735), 337, ISSN 0003-0147, 1537-5323.

Hartl, M., K. Theys, A. Feder, M. Gelbart, A. Stern, and P. S. Pennings, 2016, bioRxiv , 057026.

Hedskog, C., M. Mild, J. Jernberg, E. Sherwood, G. Bratt, T. Leitner, J. Lundeberg, B. Andersson, and J. Albert, 2010, PLoS ONE **5**(7), e11345, URL http://dx.doi.org/10.1371/journal.pone.0011345.

Hinkley, T., J. Martins, C. Chappey, M. Haddad, E. Stawiski, J. M. Whitcomb, C. J. Petropoulos, and S. Bonhoeffer, 2011, Nat Genet **43**(5), 487, ISSN 1061-4036.

Johnson, J. A., J.-F. Li, X. Wei, J. Lipscomb, D. Irlbeck, C. Craig, A. Smith, D. E. Bennett, M. Monsour, P. Sandstrom, E. R. Lanier, and W. Heneine, 2008, PLoS Med **5**(7), e158.

Johnson, V., V. Calvez, H. Gnthard, R. Paredes, D. Pillay, R. Shafer, A. Wensing, and D. Richman, 2011, Top Antivir Med **19**(4), 156, ISSN 2161-5861, URL http://europepmc.org/abstract/med/22156218.

Joos, B., M. Fischer, H. Kuster, S. K. Pillai, J. K. Wong, J. Bni, B. Hirschel, R. Weber, A. Trkola, H. F. Gnthard, and T. S. H. C. Study2, 2008, PNAS **105**(43), 16725, ISSN 0027-8424, 1091-6490.

Kessinger, T. A., A. S. Perelson, and R. A. Neher, 2013, Front. Immunol. **4**, 252.

Kimura, M., 1955, Cold Spring Harb Symp Quant Biol **20**, 33.

Kimura, M., 1968, Nature **217**(5129), 624.

LANL HIV sequence data base, 2016, HXB2 genome annotation, URL http://www.hiv.lanl.gov/content/sequence/HIV/MAP/annotation.html.

Leslie, A. J., K. J. Pfafferott, P. Chetty, R. Draenert, M. M. Addo, M. Feeney, Y. Tang, E. C. Holmes, T. Allen, J. G. Prado, M. Altfeld, C. Brander, *et al.*, 2004, Nat. Med. **10**(3), 282.

Lewis, D. A., K. Bebenek, W. A. Beard, S. H. Wilson, and T. A. Kunkel, 1999, Journal of Biological Chemistry **274**(46), 32924, ISSN 0021-9258, 1083-351X.

Li, B., A. D. Gladden, M. Altfeld, J. M. Kaldor, D. A. Cooper, A. D. Kelleher, and T. M. Allen, 2007, J. Virol. **81**(1), 193.

Li, G., S. Piampongsant, N. R. Faria, A. Voet, A. a.-C. Pineda-Pea, R. Khouri, P. Lemey, A.-M. Vandamme, and K. Theys, 2015, Retrovirology **12**(1), 18, ISSN 1742-4690.

Li, J. Z., 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, R. Thiebaut, L. Ostergaard, *et al.*, 2011, JAMA **305**(13), 1327.

Lu, K., X. Heng, L. Garyu, S. Monti, E. L. Garcia, S. Kharytonchyk, B. Dorjsuren, G. Kulandaivel, S. Jones, A. Hiremath, S. S. Divakaruni, C. LaCotti, *et al.*, 2011, Science **334**(6053), 242, ISSN 1095-9203.

Malim, M. H., 2009, Philosophical Transactions of the Royal Society of London B: Biological Sciences **364**(1517), 675, ISSN 0962-8436, 1471-2970.

Mansky, L. M., and H. M. Temin, 1995, J. Virol. **69**(8), 5087, ISSN 0022-538X, 1098-5514.

Martinez-Picado, J., and M. A. Martinez, 2008, Virus Research **134**(12), 104, ISSN 0168-1702.

Mayrose, I., A. Stern, E. O. Burdelova, Y. Sabo, N. Laham-Karam, R. Zamostiano, E. Bacharach, and T. Pupko, 2013, BMC Evolutionary Biology **13**, 164, ISSN 1471-2148.

Meyer, A. G., and C. O. Wilke, 2015, Journal of The Royal Society Interface **12**(111), 20150579, ISSN 1742-5689, 1742-5662.

Neher, R. A., 2013, Annual Review of Ecology, Evolution, and Systematics **44**(1), null.

Neher, R. A., and T. Leitner, 2010, PLoS Comput Biol **6**(1), e1000660.

Ngandu, N. K., K. Scheffler, P. Moore, Z. Woodman, D. Martin, and C. Seoighe, 2008, Virology Journal **5**, 160, ISSN 1743-422X.

Parera, M., G. Fernndez, B. Clotet, and M. A. Martnez, 2007, Mol Biol Evol **24**(2), 382, ISSN 0737-4038, 1537-1719.

Pereyra, F., D. Heckerman, J. M. Carlson, C. Kadie, D. Z. Soghoian, D. Karel, A. Goldenthal, O. B. Davis, C. E. DeZiel, T. Lin, J. Peng, A. Piechocka, *et al.*, 2014, J. Virol. **88**(22), 12937, ISSN 0022-538X, 1098-5514.

Petropoulos, C. J., N. T. Parkin, K. L. Limoli, Y. S. Lie, T. Wrin, W. Huang, H. Tian, D. Smith, G. A. Winslow, D. J. Capon, and J. M. Whitcomb, 2000, Antimicrob. Agents Chemother. **44**(4), 920, ISSN 0066-4804, 1098-6596.

Pollom, E., K. K. Dang, E. L. Potter, R. J. Gorelick, C. L. Burch, K. M. Weeks, and R. Swanstrom, 2013, PLOS Pathog **9**(4), e1003294, ISSN 1553-7374.

Rihn, S. J., J. Hughes, S. J. Wilson, and P. D. Bieniasz, 2015, Journal of Virology **89**(1), 552, ISSN 0022-538X, 1098-5514.

Sanjuán, R., 2010, Philosophical Transactions of the Royal Society of London B: Biological Sciences **365**(1548), 1975.

Schneidewind, A., Z. L. Brumme, C. J. Brumme, K. A. Power, L. L. Reyor, K. O'Sullivan, A. Gladden, U. Hempel, T. Kuntzen, Y. E. Wang, C. Oniangue-Ndza, H. Jessen, *et al.*, 2009, J. Virol. **83**(8), 3993, ISSN 0022-538X, 1098-5514, URL http://jvi.asm.org/content/83/8/3993.

Seifert, D., F. Di Giallonardo, K. J. Metzner, H. F. Gnthard, and N. Beerenwinkel, 2015, Genetics **199**(1), 191, ISSN 1943-2631.

Shekhar, K., C. F. Ruberman, A. L. Ferguson, J. P. Barton, M. Kardar, and A. K. Chakraborty, 2013, Phys. Rev. E **88**(6), 062705.

Siegfried, N. A., S. Busan, G. M. Rice, J. A. E. Nelson, and K. M. Weeks, 2014, Nat Meth **11**, 959, ISSN 1548-7091.

Sükösd, Z., E. S. Andersen, S. E. Seemann, M. K. Jensen, M. Hansen, J. Gorodkin, and J. Kjems, 2015, Nucl. Acids Res. , gkv1039.

Thyagarajan, B., and J. D. Bloom, 2014, eLife Sciences **3**, e03300, ISSN 2050-084X.

Vabret, N., M. Bailly-Bechet, A. Lepelley, V. Najburg, O. Schwartz, B. Verrier, and F. Tangy, 2014, J. Virol. **88**(8), 4161, ISSN 1098-5514.

de Visser, J. A. G. M., and J. Krug, 2014, Nat Rev Genet **15**(7), 480, ISSN 1471-0056.

Walker, B., and A. McMichael, 2012, Cold Spring Harb Perspect Med **2**(11).

Zanini, F., J. Brodin, L. Thebo, C. Lanz, G. Bratt, J. Albert, and R. A. Neher, 2016, eLife Sciences **4**, e11282, ISSN 2050-084X, URL `http://elifesciences.org/content/4/e11282`.

Zanini, F., and R. A. Neher, 2013, J. Virol. **87**(21), 11843, ISSN 0022-538X, 1098-5514, URL `http://jvi.asm.org/content/87/21/11843`.

Appendix: Supplementary material

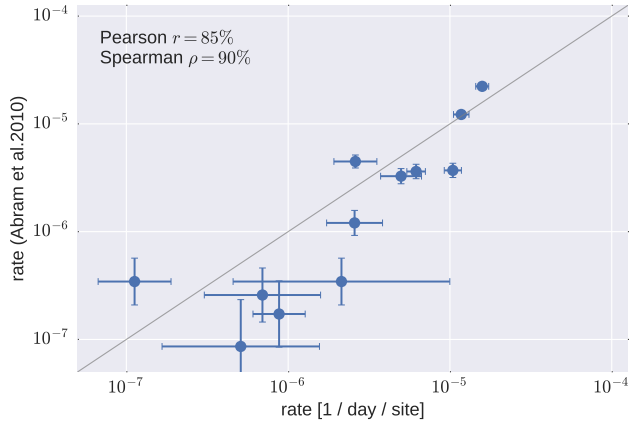

FIG. S1 Comparison of our estimates for the neutral mutation rates to *in vitro* estimates by Abram *et al.* (2010). Error bars for the estimates are standard deviations over 100 patient bootstraps. Error bars for the values from Abram et al. (2010) are standard deviations of binomial sampling noise (low-frequency mutations were observed 1-2 times only in that study).

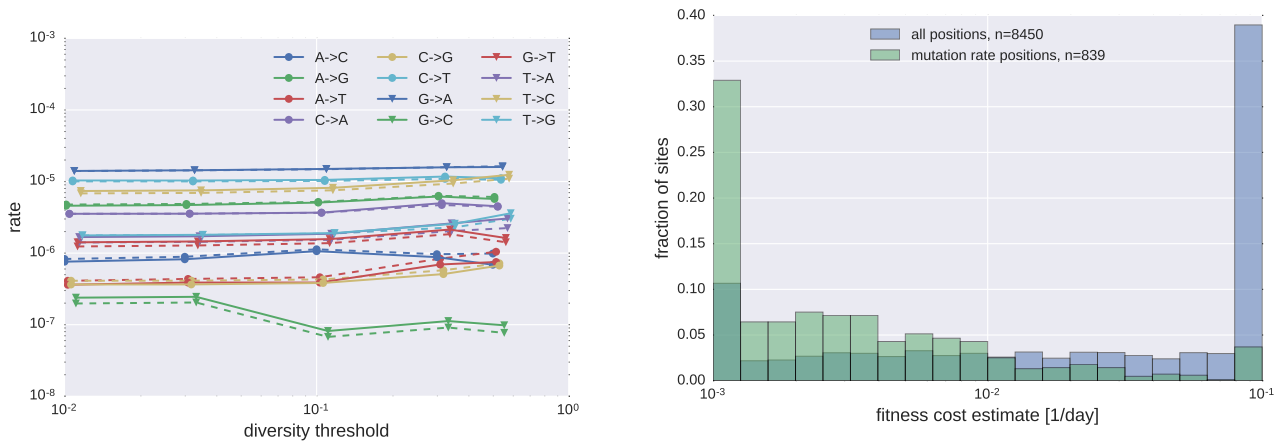

FIG. S2 Sensitivity of mutation rate estimates on the criteria used to define the set of approximately neutral positions. (A) Mutation rate estimates depend only weakly on the threshold used to define the approximately neutral set of positions or whether gp120 is included or not. (B) The positions chosen to estimate the neutral mutation rate are among the most neutral positions as estimated by the saturation of intrapatient frequencies. Note that frequencies of neutral mutations don't saturate and can be less diverse than expected due to linked selection and drift; this is not a problem for our estimates as we do not infer site-specific mutation rates.

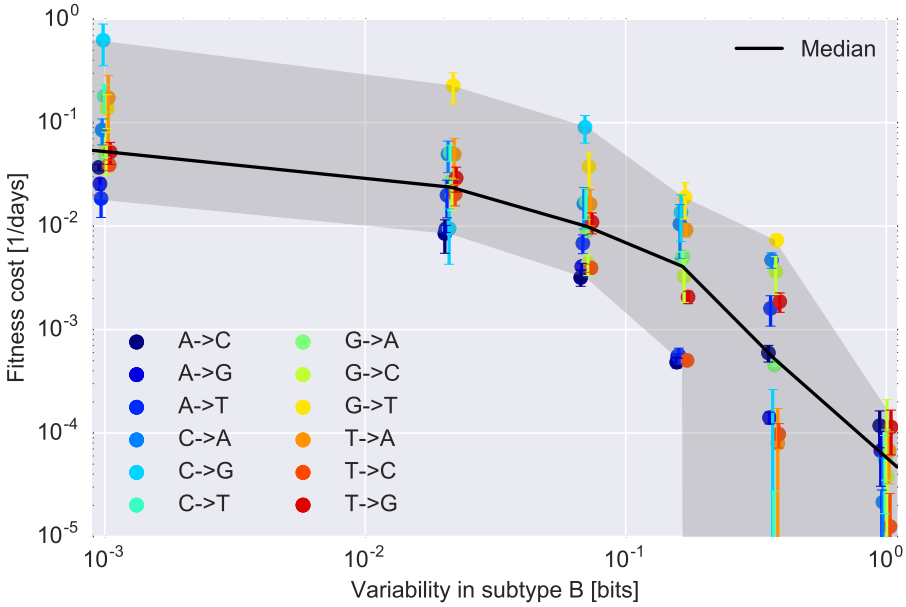

FIG. S3 Fitness cost estimates as a function of subtype conservation, from saturation curves similar to Fig. 2 ”Sat” but separate for each of the 12 mutations. The general picture is the same like shown in Fig. 2, but some mutations appear to be slightly more or less suppressed than the average.

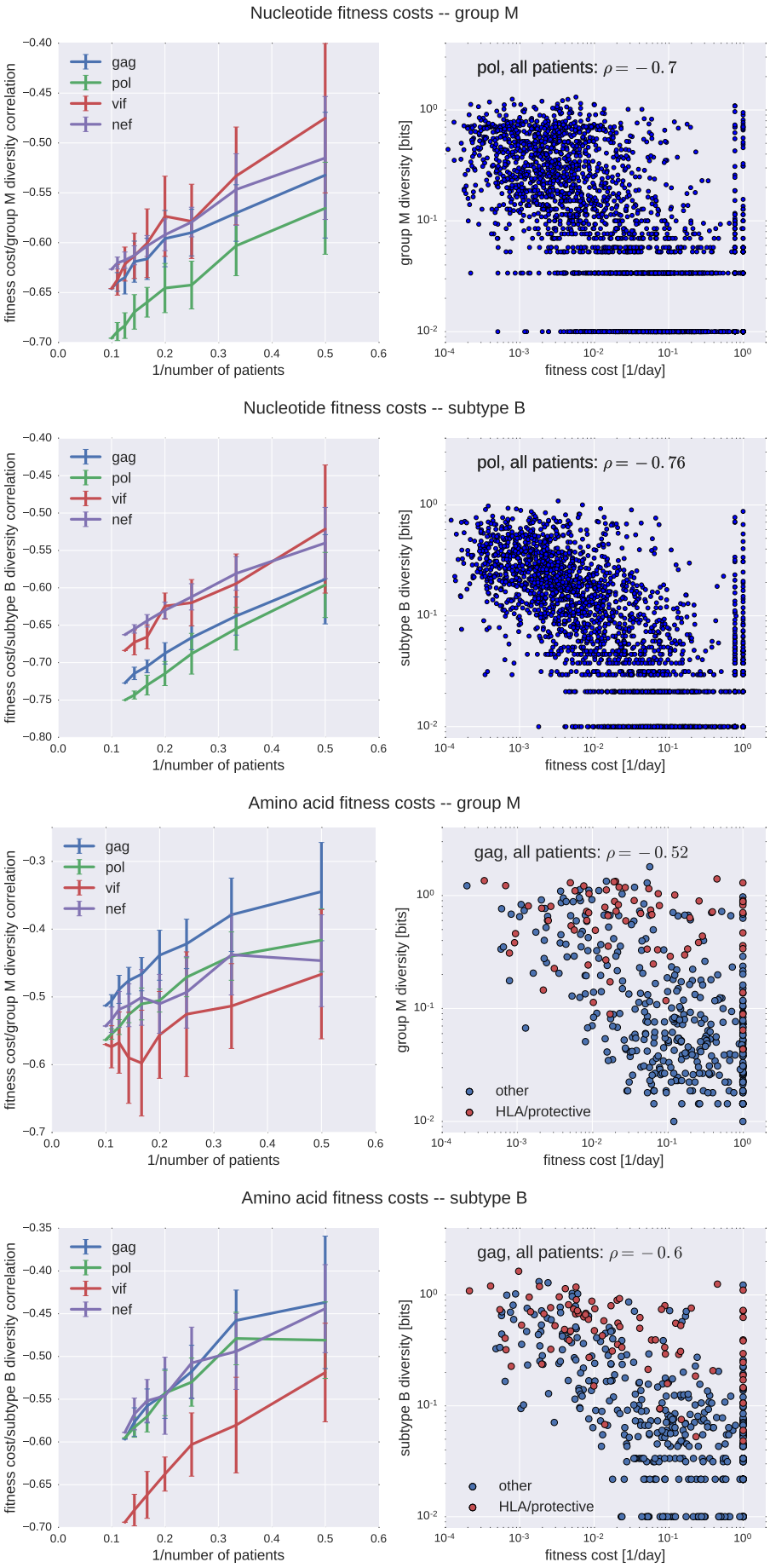

FIG. S4 **Correlation of fitness cost with global cross-sectional diversity.** The left panels show how correlation improves as fitness costs are estimates using data from more and more patients. The right panels show a scatter plot of fitness cost vs cross-sectional diversity using data from all patients for one of the proteins. The top panels show costs for nucleotide mutations, the bottom panels for amino acid mutations (and highlight HLA associated of protective sites, (Bartha *et al.*, 2013; Carlson *et al.*, 2012)).

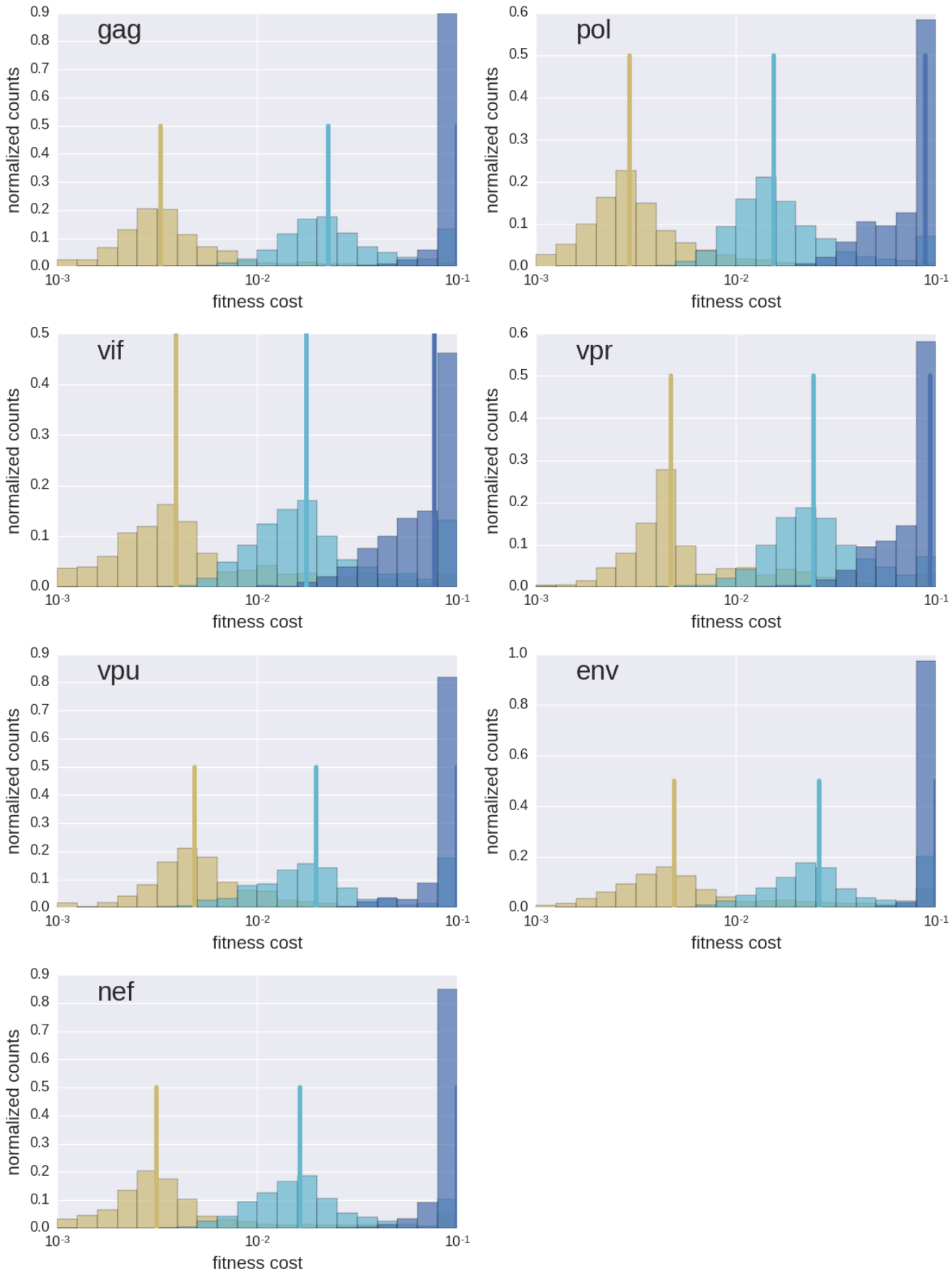

FIG. S5 Bootstrap confidence on fitness costs for various regions of the genome.

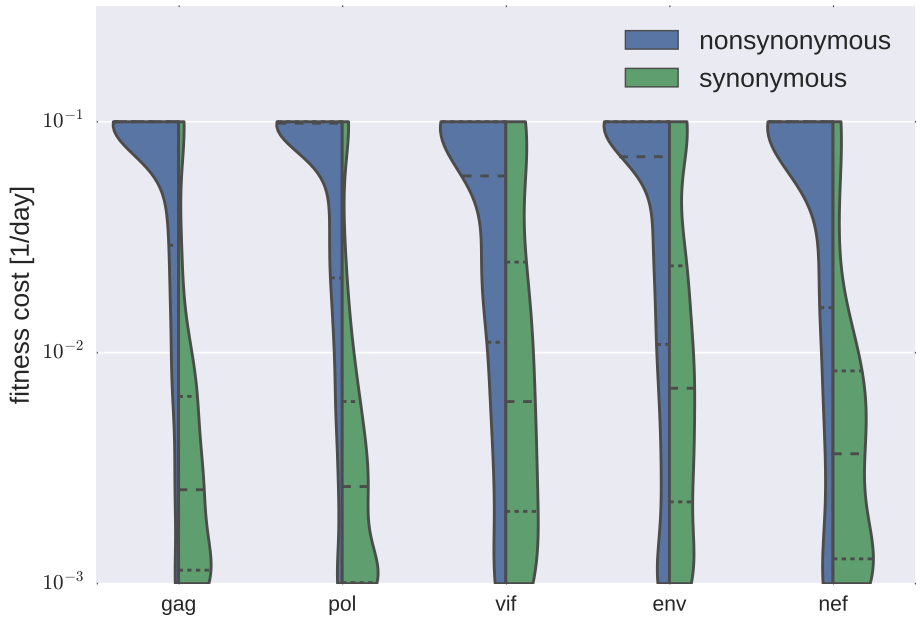

FIG. S6 **Fitness costs in different genes.** Distribution of fitness costs of synonymous and nonsynonymous mutations in different genes. Note that frequency estimates in gp120 are expected to be less accurate due to consistent difficulties amplifying this part of the genome.

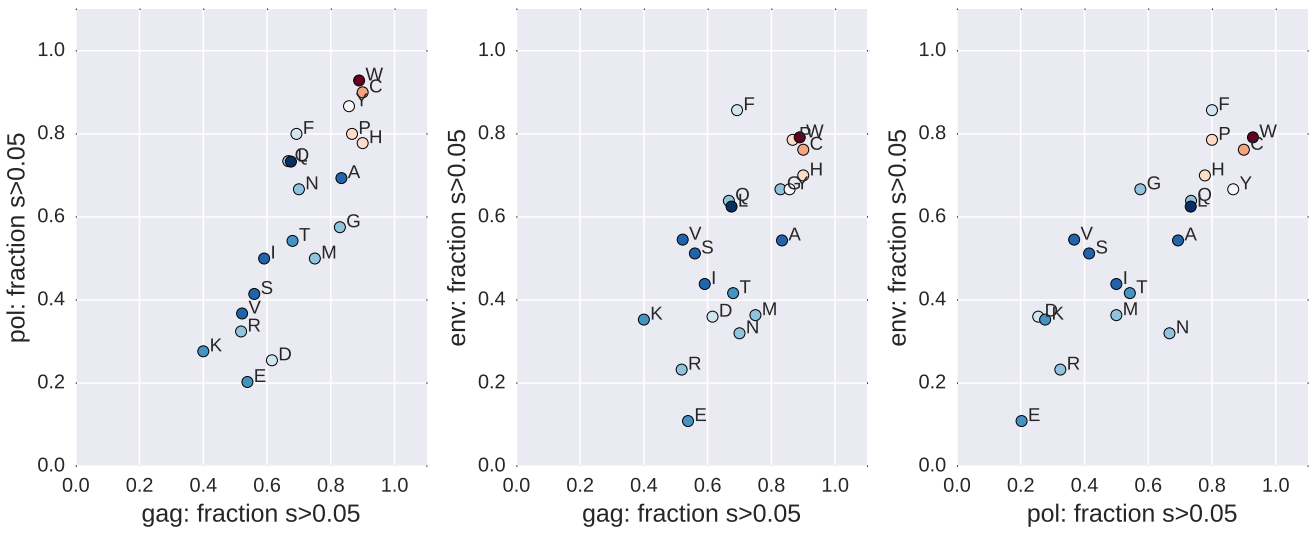

FIG. S7 **Fitness costs and consensus amino acid.** The fraction of sites with fitness costs > 0.05 per day depends consistently on the consensus amino acid. Mutations of cysteins (C), histidines (H), prolines (P), tryptophans (W), and tyrosines (Y) tend to be most costly. Points are colored according to the diagonal of the BLOSSUM80 matrix, from blue to white to red, indicating a fair degree of agreement, especially for the most costly residues.

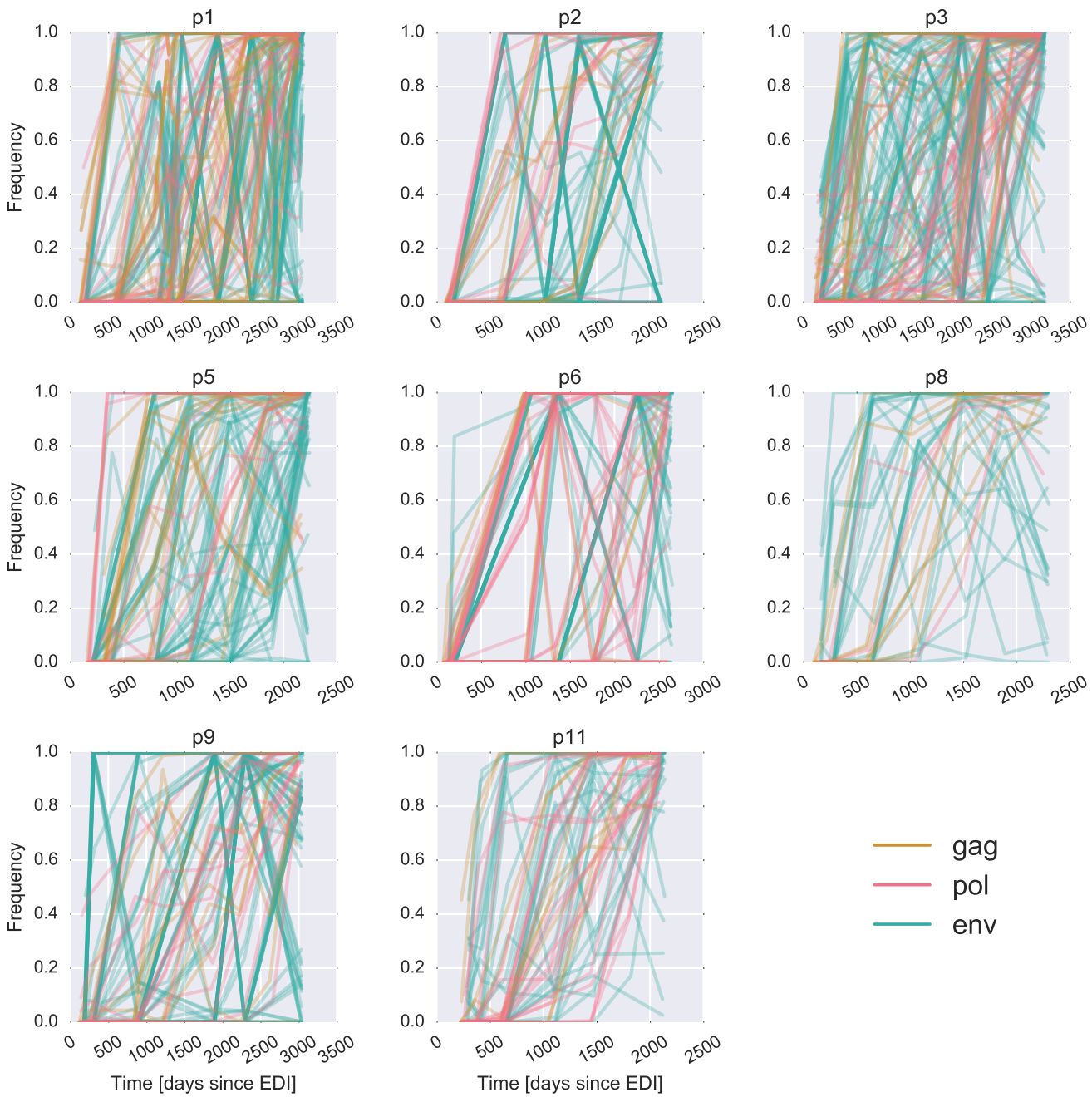

FIG. S8 Many mutations sweep across the viral population at the same time. Each panel shows the trajectories of putative selective sweeps in a study patient, i.e. mutations that reach 90% frequency at least at one time point. These trajectories include not only driver mutations, i.e. beneficially selected for, but also linked passenger mutations, e.g. synonymous mutations or mutations that carry little additional cost. The number of sweeps observed across a whole infection in our patients is as follows: p1, 145; p2, 95; p3, 147; p5, 95; p6, 94; p8, 41; p9, 111; p11, 63. p7 did not yield early samples and is not shown.

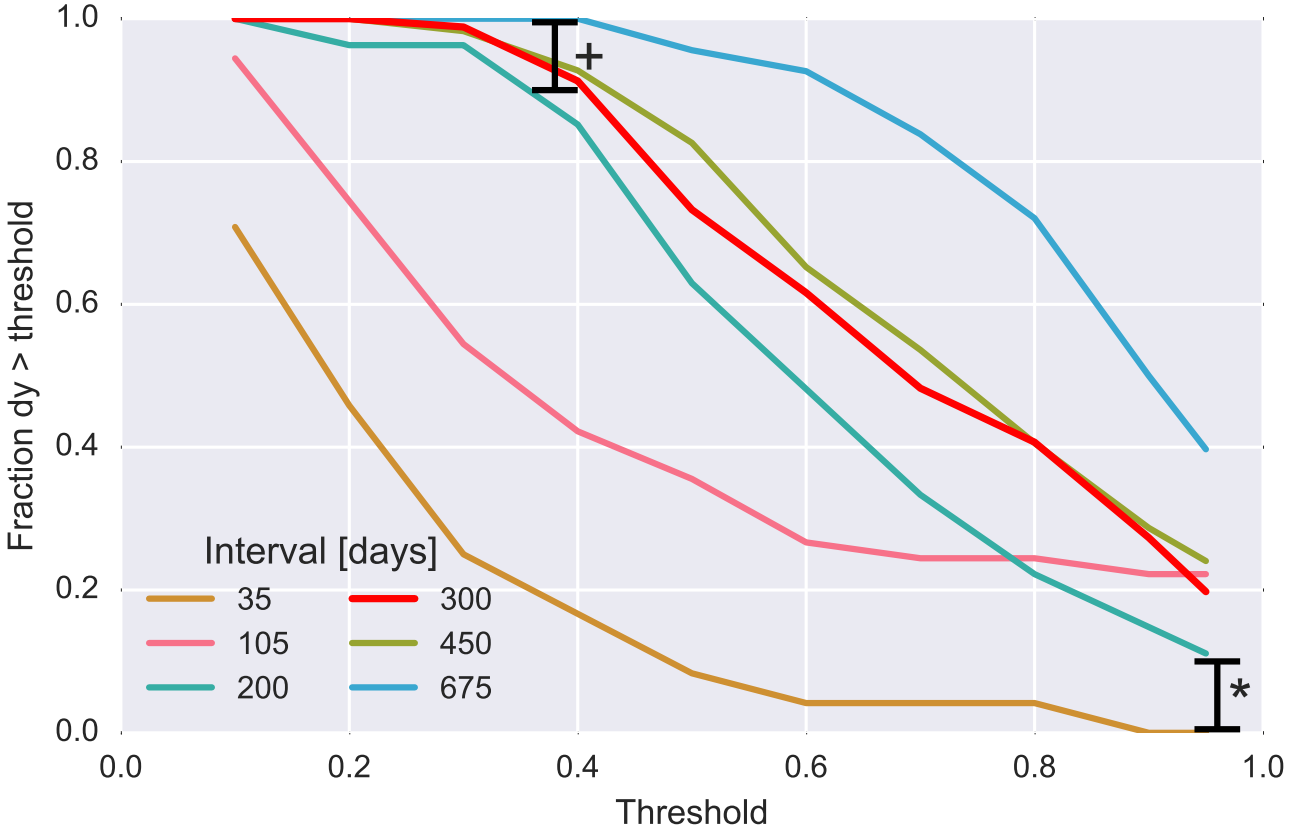

FIG. S9 Most sweeps suggest a fitness benefit around 1% per day. The increase in allele frequency between two consecutive samples is termed *dy*, and the fraction of sweeps with an increase larger than a threshold is shown for different thresholds (x axis) and for pairs of samples at different temporal distance (each line refers to one temporal distance category, as shown in the legend). Around 10% of sweeps happen in much faster than 200 days (fraction *), around 10% are much slower than 400 days (fraction +), and 80% of sweeps take around 300 days, which suggests a fitness benefit of around 1 to 2% per day, in agreement with previous estimates about chronic infection (Neher and Leitner, 2010). A similar result is obtained if only nonsynonymous changes are considered.

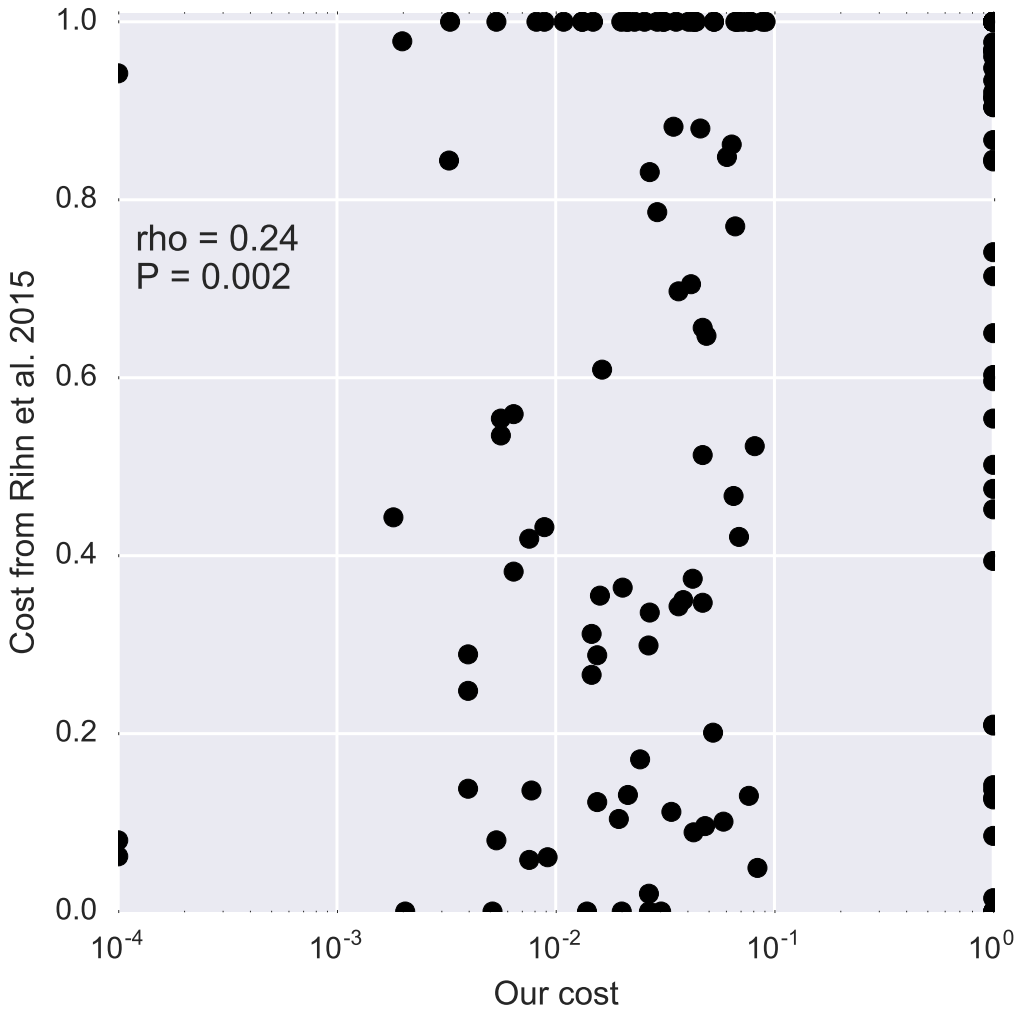

FIG. S10 Fitness costs in integrase are correlated with published *in vitro* experiments (Rihn *et al.*, 2015). The rank correlation coefficient is 0.24 (P-value = 0.002), which indicates a partial agreement between our results and Rihn et al. (2015). There are three reasons why no perfect correlation is expected. First, cell culture fitness determinations are sensitive to costs above 3-5% whereas our *in vivo* method is accurate between 0.1% and 10% approximately. This makes the two approaches nicely complementary in scope. Second, cell cultures are not perfect models of the viral dynamics in a patient, hence some selective pressures might differ. Third, one limitation of our study is that for each site we do not test for a specific mutation, so a few discrepancies might be due to this methodological difference. In cases when Rihn *et al.* (2015) tested more than one mutation at a site, the same cost from our table was reused. To further test the significance of the correlation, we repeated the correlation analysis several times after reshuffling sites and costs and found no significant correlation in those cases.

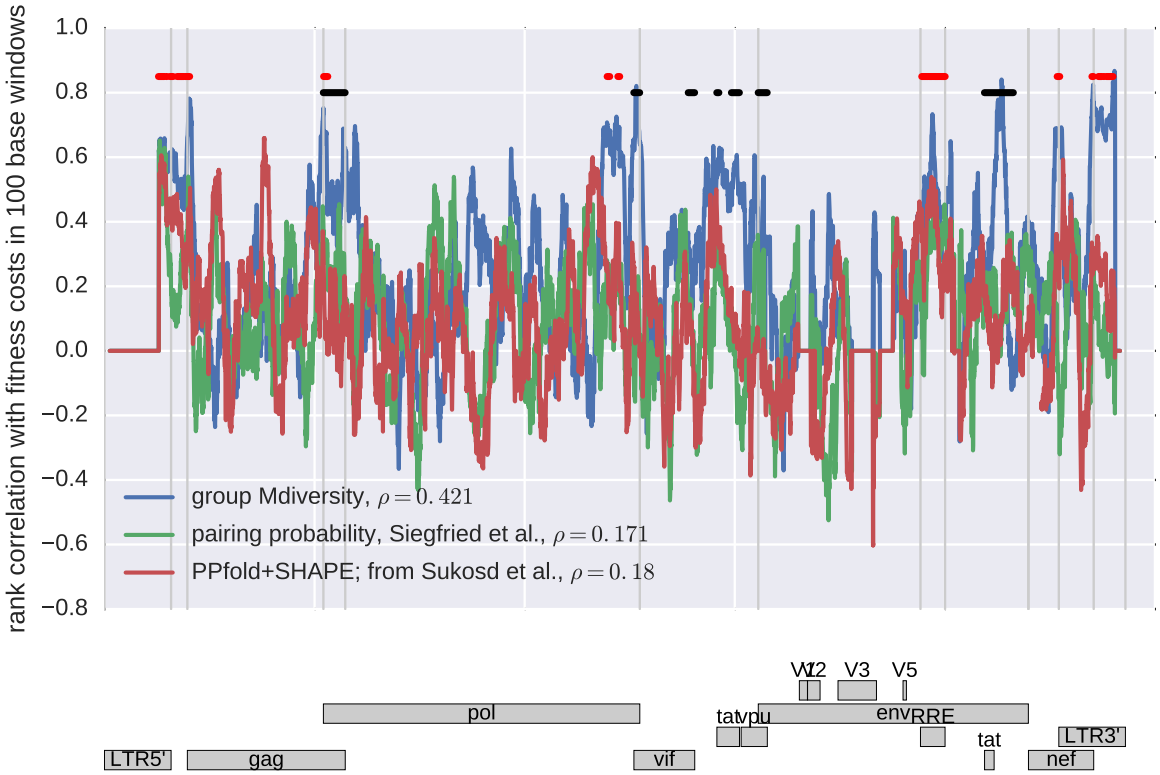

FIG. S11 Fitness estimates are at synonymous* sites are well correlated (in sliding 100 bp windows) with group M diversity, but correlation with RNA structure prediction by Siegfried *et al.* (2014) and Sükösd *et al.* (2015) is weaker and limited to a few regions. Pronounced peaks of the correlation between diversity and fitness costs at synonymous positions coincide with overlapping reading frames (marked in black in the top part of the figure) and known regulatory elements (marked in red). The strongest correlation is observed in the central and 3' poly purine tracts, around the overlap of gag and pol, and in the 3' LTR. The genome wide correlation (given in the legend) is highly significant in all cases but low for RNA structure predictions. * synonymous sites are defined here as those at least the transition does not result in in an amino acid change in *gag*, *pol*, *vif*, *vpu*, *env*, and *nef*.

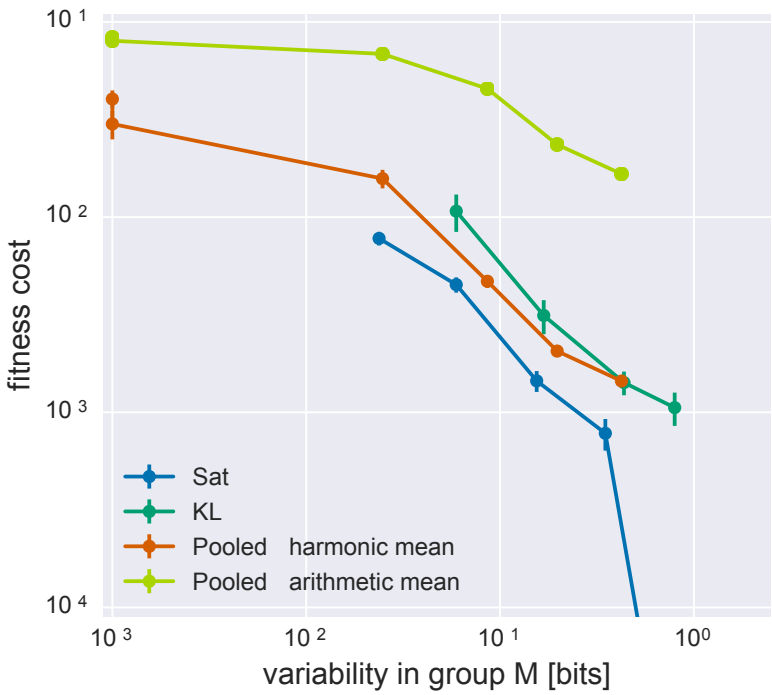

FIG. S12 Fitness cost estimates based on temporal correlations of allele frequencies are consistent with the saturation and the pooled estimates. The ”Sat” and ”Pooled” curves are like in Fig. 2, the ”KL” curve uses the estimate method based on minimization of Kullback-Leibler divergence (see below). The arithmetic mean of the ”Pooled” estimate is higher than the harmonic mean, as expected; this quantity basically describes the fraction of polymorphisms within each group of sites.

Estimation of selection coefficients by Kullback-Leibler divergence minimization

In addition to the two modelling methods presented in Fig. 2, "Sat" and "Pooled", we tested a third approach that exploits the time information of samples (like the "Sat" method) but also models the temporal correlations of SNP frequencies (see Fig. S12). These correlations are not accounted for in the "Sat" fitting procedure which simply fits average values for each bin.

We capture the correlation structure of the SNP frequency trajectories by modelling the full probability distribution $P(\mathbf{x})$ of observing all SNPs from all times at a certain combination of frequencies:

$$\mathbf{x} = (x_{t,i} \dots),$$

where t indicates each time point and i each conservation group. We combine all SNP trajectories (summed minor derived states) of all sites within one conservation quantile into $\mathbf{x}$, separately for each patient. We approximate the joint probability distribution $P(\mathbf{x})$ by a theoretical distribution $W(\mathbf{x})$ that is the solution of the stochastic equation (1) with a constant diffusive noise term $\eta(t)$ to make it mathematically tractable

$$\left\langle \eta^2(t) \right\rangle \propto Dt.$$

where D defines the noise intensity. The solution of eq. (1) under these simplifying assumptions is a multivariate Gaussian distribution:

$$W(\mathbf{x}) = \frac{\exp\left[-\frac{1}{2}(\mathbf{x} - \langle \mathbf{x} \rangle)^T K^{-1} (\mathbf{x} - \langle \mathbf{x} \rangle)\right]}{\sqrt{(2\pi)^N \det K}}, \tag{4}$$

where K is the covariance matrix of SNP frequencies. Mean and covariance of $W(\mathbf{x})$ are given respectively by

$$\begin{aligned} \langle x(t) \rangle &= \frac{\mu}{s}\left(1 - e^{-st}\right), \\ K(t,t') &= \frac{D}{s}\left[e^{-s|t-t'|} - e^{-s(t+t')}\right], \end{aligned} \tag{5}$$

We now want to estimate the parameters s and D from the data while keeping μ, the mutation rate, fixed at the measured value $1.2 \cdot 10^{-5}$ per day per site. To this end, we construct an empirical distribution of SNP frequency trajectories as a multivariate Gaussian with mean and covariances obtained by averaging the data across sites:

$$\begin{aligned} \hat{x}(t) &= \frac{1}{L}\sum_k x_k(t), \\ \kappa(t_i, t_j) &= \frac{1}{L-1}\sum_k \left[x_k(t_i) - \hat{x}(t_i)\right]\left[x_k(t_j) - \hat{x}(t_j)\right]. \end{aligned} \tag{6}$$

Here k is the site/position index, the $\hat{x}$ designates average minor SNP frequency in the conservation quantile analysed, t_i and t_j are time points along the trajectory, and L is the number of sites used in the average.

Mean and covariance fully determine the empirical Gaussian distribution, so we can extract the best model parameters by minimizing the distance of this distribution and the theoretical one. A convenient measure of the divergence between the two distributions is so-called Kullback-Leibler divergence, defined as

$$KL = \int P(\mathbf{x}) \log\left[\frac{P(\mathbf{x})}{W(\mathbf{x})}\right] d\mathbf{x}. \tag{7}$$

Averaging over the empirical distribution $P(\mathbf{x})$ is now equivalent to averaging over sites, which allows us to write the Kullback-Leibler divergence (KL) as

$$\begin{aligned} KL &= C - \frac{1}{L}\log W(\mathbf{x}) = C + \log\sqrt{(2\pi)^N \det K} \\ &+ \frac{1}{2}\sum_{i,j}\left\{\left[\hat{x}(t_i) - \langle x(t_i)\rangle\right](K^{-1})_{ij}\left[\hat{x}(t_j) - \langle x(t_j)\rangle\right] + (K^{-1})_{ij}\kappa_{ji}\right\}. \end{aligned} \tag{8}$$

Finally, we notice that for different conservation groups, the KL is additive. We can thus sum over all conservation groups to esitmate all s and D parameters simultaneously (one s and one D per group). The resulting values for s are shown in Fig. S12 as the "KL" curve and is in good agreement with the two previous methods used to estimate average fitness costs.