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: March 21, 2016)

Mutation rates and fitness costs of deleterious mutations are difficult to measure *in vivo* but essential for a quantitative understanding of of evolution. Using whole genome deep sequencing data on longitudinal samples during untreated HIV-1 infection, we estimated mutation rates and the distribution of fitness costs in HIV-1 from the temporal dynamics of genetic variation. At approximately neutral sites, mutations accumulate with rates similar to those measured in cell cultures. Genetic diversity at other sites saturates and we estimated the fitness costs at those sites from the time and level of saturation. About half of all non-synonymous mutations have fitness costs greater than 10%, while half of synonymous mutations have costs below 1% such that they are essentially neutral over the course of a year. Fitness costs of mutations that are synonymous in one gene but affect proteins in other reading frame or important RNA structures are distributed similarly to non-synonymous mutations. Within patient fitness landscape explains a large fraction of global HIV-1 group M diversity.

Due to the error-prone HIV-1 reverse transcriptase and to a lesser degree human RNA polymerase II, mutations commonly occur during the replication cycle of HIV-1 (Abram *et al.*, 2010; Cuevas *et al.*, 2015; Mansky and Temin, 1995). These mutations are the source of genetic diversity, while sequence changes that accumulate and spread are filtered by selection. The key factors determining the rate and pattern of sequence evolution are (i) the mutation rate matrix, that is the rate at which the 12 nucleotide substitutions (e.g. $A \to G$) are generated per replication and (ii) the landscape of fitness effects of these mutations, i.e. the amounts by which individual mutations increase or decrease the replication capacity of the virus. Mutations can be divided in three classes depending on their effect on viral evolution: the majority of mutations are deleterious and impair virus replication, some mutations are neutral and have little or no effect, and a minority of mutations are beneficial.

While beneficial mutations are expected to increase in frequency and spread through the population, strongly deleterious single nucleotide polymorphisms (SNP) will settle into a noisy balance at low frequency: they are continuously generated by new, recurrent mutations, but they are also purged from the virus population by virtue of their negative effect on replication.

Deleterious mutations are expected to be largely shared across patients as they mainly impair general biological processes such as enzymatic activity and protein folding. This expectation is confirmed by deep sequencing data from untreated HIV-1 patients: minor SNPs at are found at similar frequencies in different patients, indicating a similar fitness cost independent of the host (Zanini *et al.*, 2016). In contrast, beneficial mutations are often patient-specific because they allow escape from adaptive immune responses, i.e. cytotoxic T-lymphocytes (CTL) and neutralizing antibodies, where the former depends on the HLA-type of the patient. Alternatively, if an escape mutation has spread in a previous host, the mutation might revert in the new host if that site is not targeted by the new immune system (Friedrich *et al.*, 2004; Leslie *et al.*, 2004; Li *et al.*, 2007; Zanini *et al.*, 2016).

The rate and spectrum of mutations during the replication cycle of HIV-1 has been measured using lacZ reporter assays (Abram *et al.*, 2010; Mansky and Temin, 1995), whereas fitness costs of individuals mutations are quantified by competing mutant and wild-type viruses (Martinez-Picado and Martinez, 2008; Parera *et al.*, 2007). Such measurements of replication capacity are done routinely for drug resistance testing (Petropoulos *et al.*, 2000) and have been used to infer fitness costs of mutations (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; Thyagarajan and Bloom, 2014).

Since mutation rates and fitness effects in cell culture systems might differ from their values *in vivo*, several approaches have been developed to estimate these quantities indirectly form diversity in large alignments of large global collections of HIV-1 sequences (Dahirel *et al.*, 2011; Ferguson *et al.*, 2013). Sites at which one amino acid predominates are inferred to be under strong purifying selection. A priori, it is unclear whether cross-sectional diversity reflects the intra-patient fitness landscape or whether it is influenced by transmission biases or by immune escape and distribution of HLA types. Furthermore, such methods can infer the relative fitness of different sequences, but do not allow to estimate absolute fitness costs.

Here, we use recent whole-genome deep sequencing data from longitudinal HIV samples (Zanini *et al.*, 2016) to infer in vivo mutation rates and the distribution of fitness costs. In that study, HIV RNA from 6-12 samples from 9 patients was amplified in six overlapping fragments and sequenced at 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%. The deep and longitudinal diversity data enable us to estimate the absolute values of the mutation rates and fitness costs.

Neutral mutation rate matrix

Mutations at neutral sites accumulate freely 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). Since the frequency of any particular mutation is subject to large stochastic effects due to genetic drift or physical linkage to other SNPs in the genome under selection, precise estimates of the divergence require averaging over many sites in the genome and ideally several independent evolutionary trajectories, e.g. HIV-1 evolution in different individuals.

To estimate the mutation rate, we defined an approximately neutral set of positions in the HIV-1 genome as those where mutations are synonymous and that are variable in a global sample of group M HIV-1 sequences. Fig. 1A and B show the average divergence from the approximate virus founder sequence in this neutral set across the nine patients, for all 12 nucleotide substitutions. The data confirm that divergence increases linearly and 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`$\rightarrow$`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 in the data. 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).

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), see Fig. S1. This quantitative agreement suggests that the average properties of mutations to HIV-1 depend little on the host cell. Because we measure the rate averaging across many sites, however, we cannot rule out that mutation rates depend on local sequence context (Abbotts *et al.*, 1993; Lewis *et al.*, 1999).

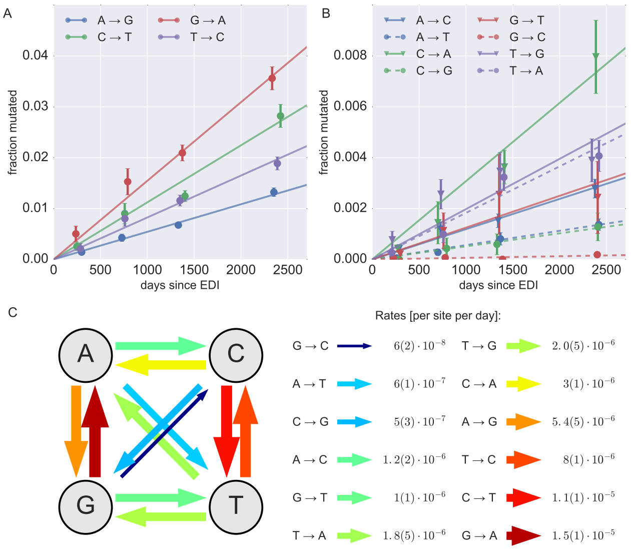

FIG. 1 **Accumulation of approximately neutral mutations.** Panels A&B show the accumulation of mutations at approximately neutral sites over time averaged over 9 patients, for transitions (A) and transversions (B). (C) The slope of the individual regression lines in panel A&B provide estimates of the 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.

Fitness costs of weakly conserved sites

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. The last term $\xi(x,t)$ describes stochastic effects including genetic drift and selection on linked SNPs at other loci in the genome. 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}. Linear accumulation of neutral mutations is recovered for $s \to 0$.

While the average trajectory is expected to follow this simple form, the trajectories of individual SNPs are noisy. For large s saturation is rapid and this noise can be overcome by averaging multiple samples. We will use this strategy below to obtain site specific estimates of s for most of the HIV-1 genome. To estimate typical fitness costs of weakly constrained sites, we average over sites with putatively similar properties.

Specifically, we group sites in the HIV-1 genome by global diversity in group M (i.e. we define quantiles of conservation) since sites with similar levels of conservation are expected to have similar fitness costs. 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. S2). We denote the combined frequency of all three mutations by x. For each conservation quantile, we average the frequencies x over all sites and patient samples in 7 time bins. These average diversities are indicated by dots in

Fig. 2A along with a nonlinear least square fit of Eq. (2) to the data of each quantile. We fit a single fit parameter per line, the fitness cost s, and set $\mu = 1.2 \cdot 10^{-5}$ per site per day consistent with our estimate of the neutral mutation rate. The resulting fitness costs and their error bars from 100 bootstraps over patients are shown in Fig. 2B as a blue line (”Sat”). To avoid confounding by CTL escape and reversions, we excluded a site if the major allele changed during the infection or if the initial allele of the patient did not agree with the HIV-1 group M consensus.

To extract additional information that is not accessible by looking at the average trajectory, we devised a method that accounts for correlations in diversity between time points. Similar to the initial saturation behavior, these correlations decay on a time scale s^{-1}. We parameterize the multivariant Gaussian distribution by the means and covariances calculated from Eq. (1) (see Methods). We estimate s and the noise parameter D by minimizing the Kullback-Leibler divergence between this distribution and an empirical distribution based averages and covariances inferred from the data (Konishi and Kitagawa, 2007). The average result and its standard deviation in 100 patient bootstraps is shown in Fig. 2B as a green line (“KL”).

Both methods yield similar estimates with average fitness costs increasing from about 10^{-3} or less per day for the most variable sites in the genome to above 0.01 for the most conserved half of the genome. Both methods effectively report an harmonic mean of fitness costs within each entropy quantile since SNP frequencies (proportional to the inverse selection coefficient) are averaged first and then used to calculate the average fitness cost. Harmonic averages put most emphasis on small selection coefficients, such that even in the most conserved regions the average is below 10%.

For strongly conserved sites, corresponding to fitness costs $s > 0.01$, saturation of diversity happens in less than 100 days, that is less than the typical interval between successive samples. We use a separate modelling approach for those conserved sites.

Fitness costs of strongly conserved sites

The rapid saturation of frequencies at sites where mutations carry a large cost implies that frequencies of minor variants at these sites are uncorrelated and that more accurate estimates of their frequencies can be obtained by averaging multiple samples. These more accurate frequency estimates from pooled samples allow direct estimation of s from the relation $\bar{x}_i = \mu_i / s_i$ at each site i.

From each patient, we calculated the average of SNP frequencies over all samples at least two year post infection weighted by the estimated template input. The average frequency of nucleotide or amino acid α at position k is then given by

$$\hat{x}_{k,\alpha} = \frac{1}{\sum_i w_i} \sum_i w_i x(t_i)_{k,\alpha} \qquad (3)$$

where $x(t_i)_{k,\alpha}$ is the frequency at time point t_i and w_i is the weight calculated from the template input T_i as $w_i = \frac{T_i}{1+T_i/500}$, where 500 is the inverse of the error rate.

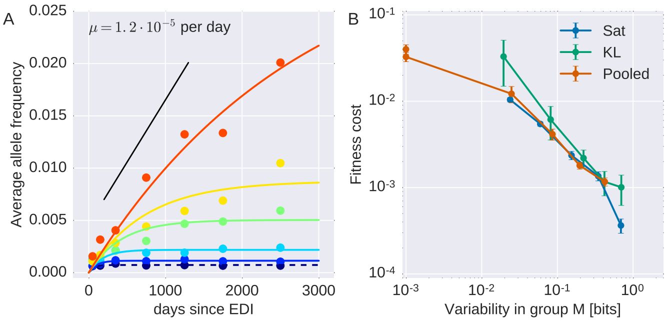

FIG. 2 **Average intra-patient fitness cost in six quantiles of global HIV-1 group M variability.** (A) Average derived SNP frequencies (1 - frequency of the ancestral state) saturate fast at positions in the conserved quantiles (blue), while intra-patient diversity keeps increasing in variable quantiles (yellow and 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 “saturation”. KL refers to the probabilistic Kullback-Leibler inference method, while “pooled” refers to harmonic averages of site specific cost estimates (see main text for details). Error bars indicate 100 bootstraps over patients. The Sat and KL method are not applicable in the most conserved third of the genome. 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.

To further reduce the noise in the minor SNP frequency estimates, we combine data from different patients. 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. We will denote the minor frequency of non-consensus nucleotides or amino acids simply by $\hat{x}_i$.

Minor diversity within patients correlates strongly with global HIV-1 group M diversity (rank correlation $\rho \approx 0.7$ for per site diversity measured by entropy), even though each of these measurements of minor SNP frequency is conditioned on the majority variant being equal to the consensus variant. The correlation increases as intra-patient variation is estimated using more patients (see Fig. S4), suggesting that fitness costs at individual sites is largely conserved between patients. After pooling samples from different patients, the estimates of minor variation are accurate enough to calculate fitness costs for individual sites via $s_i \approx \mu_i / \hat{x}_i$, where μ_i is the total mutation rate at site i. Since sequencing depth and the number of available samples is substantially lower in env, we will base this analysis mostly on positions in the genes gag, pol, vif, vpr, vpu, and nef and include env only when looking at variation of constraint along the genome.

Fig. 3A and C show the distributions of estimated selection coefficients for synonymous and non-synonymous positions. We observe marked differences between these distributions: about half of all non-synonymous 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 or overlap with RNA structures (e.g. RNA stems at the beginning of gag), resembles that of non-synonymous mutations Fig. 3B. The harmonic average of selection coefficients in different quantiles of global diversity reproduces the above estimates, see "pooled" in Fig. 2B. To assess the accuracy of our estimates, in Fig. 3D we show the variation in the fitness cost estimate after bootstrapping of patients. The variation is approximately 2-fold in each direction, so fitness costs above 5% are clearly separated from costs of 1% or less.

Fig. 4A shows fitness costs of mutations at most positions along the HIV-1 genome (including env) separately for synonymous and non-synonymous mutations. Fig. 4B shows their distribution for different genes. In gag and pol, the contrast between synonymous and non-synonymous mutations is greatest. Synonymous mutations are costly in the RNA stem at very beginning of gag and in overlaps between genes, consistent with Fig. 3B. Overall, synonymous mutations are estimated to be more costly in env, where the overlap with the tat/rev exon and the rev responsive element (RRE) constrain synonymous mutations.

Non-synonymous mutations are strongly enriched among sites that are globally variable (entropy above 0.1) but monomorphic within patients (odds ratio 5). This enrichment is most pronounced in pol, gag and nef (odds ratios > 20). This observation is consistent with host-specific selection pressures (CTL selection) that result in costly adaptations that 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.

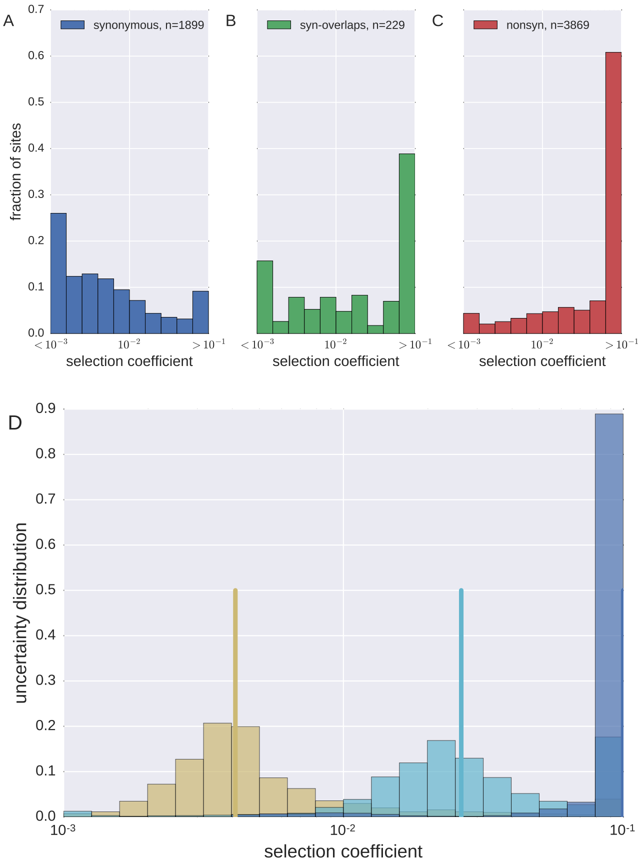

FIG. 3 **Distribution of fitness costs.** (A-C) Distributions of (A) synonymous mutations, (B) mutations that are synonymous in one gene but affect another protein in a different reading frame or known RNA structures and (C) non-synonymous mutations. The extremal bins include all points larger or smaller than the axis boundary. (D) Variation in our estimates of selection coefficients with the same median (indicated by the vertical lines) after bootstrapping patients. The upper graph includes sites in gag, pol, vif, vpu, vpr, the lower graph uses data for gag only (see Fig. S3 for other genomic regions).

Fitness costs are weakly correlated with disorder and solvent accessibility

Perturbations to protein structure are expected to reduce virus fitness. Hence mutations that increase the folding energy, 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 in (Li *et al.*, 2015). We correlated these in silico derived scores with intra-patient 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 and be far the best correlate of intra-patient diversity (and hence fitness cost estimates) is cross-sectional conservation, see Table I.

Frequencies and fitness costs of drug resistance mutations

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), and protease inhibitors (PI). 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 JZ *et al.*, 2011) Some earlier 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 JZ *et al.*, 2011). Fig. 5 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 one or two patients (pooled across all time points of each patient). Only the protease mutation M46I and RT G190ASEQ are observed consistently across several patients.

The frequency of drug resistance mutations is expected to

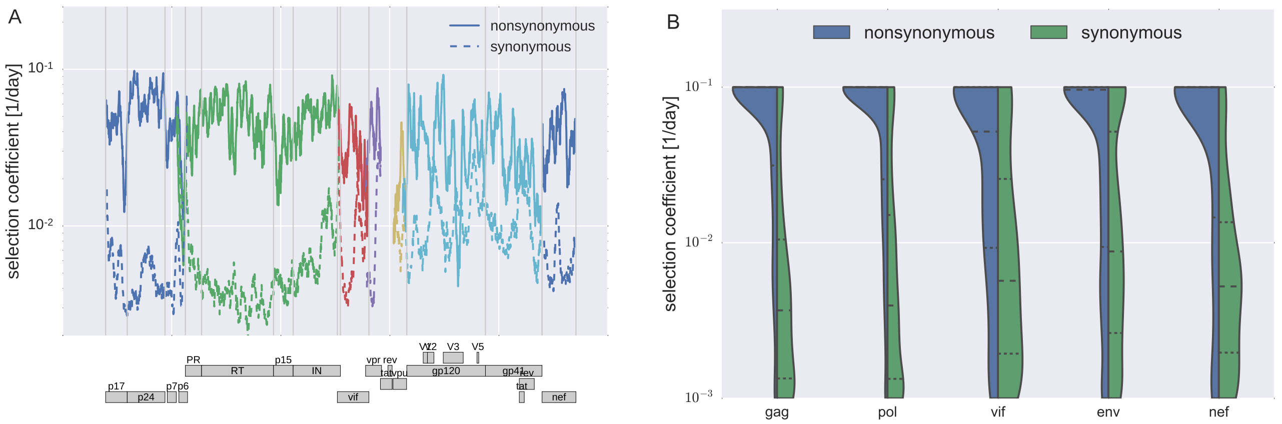

FIG. 4 **Fitness costs along the HIV-1 genome**. Panel (A) shows fitness costs of synonymous and non-synonymous mutations in gag, pol, vif, vpu, env, and nef as a geometric sliding average with window size 30. Panel (B) shows the corresponding distributions of estimated effects in each gene. Note that frequency estimates in gp120 are expected to be less accurate due to consistent difficulties amplifying this part of the genome.

gene	group M	subtype B	disorder	accessibility
gag	0.50	0.61	0.27	0.27
pol	0.52	0.59	0.08	0.26
nef	0.52	0.57	0.35	0.25
env	0.42	0.41	0.06	-0.07
vif	0.65	0.73	0.15	0.10

TABLE I Spearman's correlation coefficients of pooled intra-patient diversity with cross-sectional diversity (measured as entropy in group M and subtype B alignments) and disorder scores and solvent accessibility values obtained from (Li *et al.*, 2015). Fig. S4 shows how intra-patient/global diversity correlations improve when basing intra-patient estimates on larger numbers of patients.

be inversely proportional to their fitness cost in absense 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 10% (sites where minor variation is not or only sporadically observed).

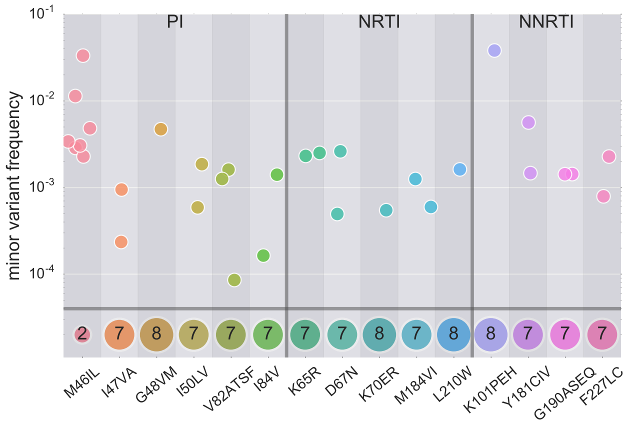

FIG. 5 **Pre-existing drug resistance mutations.** Each point shows the time averaged frequency of minor amino acids in individual patients. The bottom row indicates in how many out of 9 patients each mutation is not observed. Most mutations are observed only in a minority of patients suggesting high fitness costs. The following mutations were never found at frequencies above 0.1% in any sample: PI: L24I, V32I, I54VTAM, L76V, N88S, L90M; NRTI: M41L, K70ER, L74VI, Y115F, T215YF, K219QE; NNRTI: L100I, K103N, V106AM, E138K, V179DEF, Y188LCH, M230L.

Discussion

Sequence evolution of HIV-1 is the determined by the rate and spectrum of mutations as well as their phenotypic effects. Mutations that increase the replication rate of the virus (at least transiently) spread through the population: their effect on replication can be estimated from the speed at which they rise in frequency (Asquith *et al.*, 2006; Ganusov *et al.*, 2011; Kessinger *et al.*, 2013; Neher and Leitner, 2010). The majority of mutations, however, are deleterious and stays at low frequencies within hosts. Their contribution to sequence evolution can nevertheless be substantial 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. Selection is constantly pruning deleterious variation from the population to maintain a functional genome. 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 concordance suggests that the mutation rate of HIV-1, which is the joint rate of the HIV-1 RT 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 $G \to A$ with transitions being about 5-fold faster than transversions; the lowest rates are between base pairing partners, e.g. $G \leftrightarrow C$, see Fig. 1. While consistent with cell culture estimates, the rates we estimate differ from those reported by (Cuevas *et al.*, 2015). Cuevas *et al.* (2015) counted mutations in proviral DNA integrated into host cell genomes and estimated that the combined mutation rate due to reverse transcription errors and hypermutation by enzymes of the APOBEC family (Malim, 2009) is 4×10^{-3} per site and replication – more than 100 times higher than our estimate. The primary reason for this discrepancy is the fact that our estimate effectively excludes the APOBEC contribution to mutation which almost always results in non-functional virus (Armitage *et al.*, 2012). Since virus production typically results in cell death on the order of a day, proviral DNA is enriched for hypermutated or otherwise deactivated viruses that accumulates in long lived CD4+ cells. A mutation estimate based on proviral DNA therefore does not reflect mutations in the typical replication cycle.

In agreement with our earlier results, we found that diversity within patients is well predicted by cross-sectional diversity (Zanini *et al.*, 2016). Here, we report that average fitness costs increase from 10^{-3} or less for non-conserved sites to above 0.1 for the most conserved third of the genome. Intra-patient diversity explains to about half of the diversity in global alignments of HIV-1 sequences. A subset of positions is diverse globally, but shows little diversity within patients. Consistent with the hypothesis that these sites are globally diverse because of costly adaptation to host-specific selection pressures, this subset is strongly enriched for non-synonymous mutations. Such host-specific adaptations have the potential to confound estimation of fitness landscapes from cross-sectional data (Dahirel *et al.*, 2011; Ferguson *et al.*, 2013).

Our approach based on accumulation and saturation of minor variation within patients is complementary to cell culture based experiments(Martinez-Picado and Martinez, 2008; Thyagarajan and Bloom, 2014). 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. The longitudinal nature of the data allowed us to compare different methods of estimating fitness costs, either from the saturation time or the saturation frequency of deleterious mutations. By pooled SNP frequency data from over 60 samples, we obtained accurate estimates of minor variation, that enabled us to estimate explicit fitness costs for most positions in the HIV genome. These estimates are based on the assumption of a balance between mutations and selection at the level of individual positions. This assumption is justified for sites with fitness costs above 0.002, where SNP frequencies are expected to equilibrate in about one year. While the data on frequencies of rare minor variants are noisy even after averaging many samples, bootstrap resampling indicated that the accuracy of per site fitness cost estimates is about 2-fold in either direction, i.e. we can clearly tell apart fitness effects of 1% and 5%. Furthermore, estimates between different types of mutations (e.g. synonymous, non-synonymous, synonymous in reading frame overlaps) are clearly distinct.

We find that a minority of synonymous mutations are strongly constrained by selection (mostly in overlaps between reading frames), about 50% have intermediate selection coefficients around 1%, while the remainder is free to vary on the time scale of a few years. These observations are consistent with comparative analysis of RNA secondary structure that concluded that pairing patterns evolve rapidly in most of the genome (Pollom *et al.*, 2013) but conserved in isolated regions (Lavender *et al.*, 2015). While the sampling depth and hence the accuracy of the inferences is lower in env, our results nevertheless suggest that constraint on synonymous mutations is 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). About half of non-synonymous mutation have deleterious effects large enough that we rarely or never observed these mutations above 0.2% frequencies in about 60 late, deeply sequenced samples. Our mutation rate estimates imply that these positions have selection coefficients in excess of 10%.

Variations in fitness costs and evolutionary rates across sites in proteins are partly explained by characteristics such as solvent accessible area, intrinsic disorder scores, or estimates of changes in folding free energy (Echave *et al.*, 2016). While we find strong correlation between fitness costs and conservation in global sample of HIV-1 sequences, disorder scores of solvent accessible area only explain a small fraction of the variation, consistent with (Meyer and Wilke, 2015).

Understanding the fitness landscape of HIV-1 is an important part of understanding how the virus evolves under changing selection by the immune system or drug treatment. The almost perfectly linear accumulation of diversity at approximately neutral sites (synonymous and globally diverse) can be used estimate the time since infection from diversity in individual samples (Kouyos *et al.*, 2011). The pol region in particular accumulates diversity without much interference by selective sweeps (Zanini *et al.*, 2016).

As whole genome deep sequencing becomes more common, this type of analysis could be extended to a much large number of samples, giving more accurate estimates of minor SNP frequencies and their associated fitness costs.

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 (see supplementary material). 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). These scores were mapped to homologous positions in the virus populations via alignments to the reference sequence NL4-3. Positions without scores were discarded.

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. $A \rightarrow G$) 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 ± 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 three different approaches, called "'Sat", "KL", 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. See the supplementary script `fitness_cost_saturation.py` for the implementation. The resulting fits are shown in Fig. 2A and the fitness costs s in Fig. 2B.

Kullback-Leibler divergence minimization

The "KL" estimates in Fig. 2B result from a different modeling approach. The basic idea is to exploit the correlations in SNP frequency between samples at short temporal distance. 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 via a probabilistic model. We combine all SNP trajectories (summed minor derived states) of all sites within one conservation quantile into $\mathbf{x}$, separately for each patient. We model 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 simplified constant noise term $\eta(t)$ to make it mathematically tractable. The solution is a multivariate Gaussian distribution: and covariances is a Gaussian distribution of the form

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

where $K_{i,j} = K(t_i, t_j) = \langle x(t_i)\, x(t_j) \rangle - \langle x(t_i) \rangle \langle x(t_j) \rangle$ is the expected covariance matrix of SNP frequency trajectories that for the simplified Eq. (1) is given by

$$K(t, t') = \frac{D}{s} \left[e^{-s|t-t'|} - e^{-s(t+t')} \right], \tag{5}$$

where parameter D defines the noise intensity. The obtain estimates of the parameters of the model (s and D while fixing μ at the values estimated above), we need to compare this distribution to the data.

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:

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

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.

A convenient measure of the divergence between the two distributions is so-called Kullback-Leibler divergence, defined as

$$KL = \int ... \int P(\mathbf{x}) \log \left[\frac{P(\mathbf{x})}{W(\mathbf{x})} \right] dx_1 ... dx_N, \qquad (7)$$

where $P(\mathbf{x})$ and $W(\mathbf{x})$ are respectively the empirical and the theoretical distributions. We minimize the Kullback-Leibler divergence with respect to parameters of the theoretical model: s, μ and D.

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

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

Eq.(8) has to be minimized with respect to the parameters of the theoretical distribution: s, μ and D. This procedure can be performed numerically and allows straightforward generalization to several site categories with different fitness parameters s: the Kullback-Leibler divergences for these categories are additive.

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. The weight saturates as $T_i/(T_i + 500)$ as sequencing and PCR errors dominate start to become relevant at frequencies of about 0.2%. The weighted average is performed within 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 using the scripts `combined_af.py` and `combined_af_aa.py`.

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. 3D 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 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.

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.

Echave, J., S. J. Spielman, and C. O. Wilke, 2016, Nat Rev Genet **17**(2), 109, ISSN 1471-0056.

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, **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.

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

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.

Konishi, S., and G. Kitagawa, 2007, *Information Criteria and Statistical Modeling* (Springer Publishing Company, Incorporated), 1st edition, ISBN 0387718869, 9780387718866.

Kouyos, R. D., V. von Wyl, S. Yerly, J. Bni, P. Rieder, B. Joos, P. Taff, C. Shah, P. Brgisser, T. Klimkait, R. Weber, B. Hirschel, *et al.*, 2011, Clin Infect Dis **52**(4), 532, ISSN 1058-4838.

Lavender, C. A., R. J. Gorelick, and K. M. Weeks, 2015, PLOS Comput Biol **11**(5), e1004230, ISSN 1553-7358, URL `http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004230`.

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 JZ, Paredes R, Ribaudo HJ, and et al, 2011, JAMA **305**(13), 1327, ISSN 0098-7484.

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.

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.

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

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.

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

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

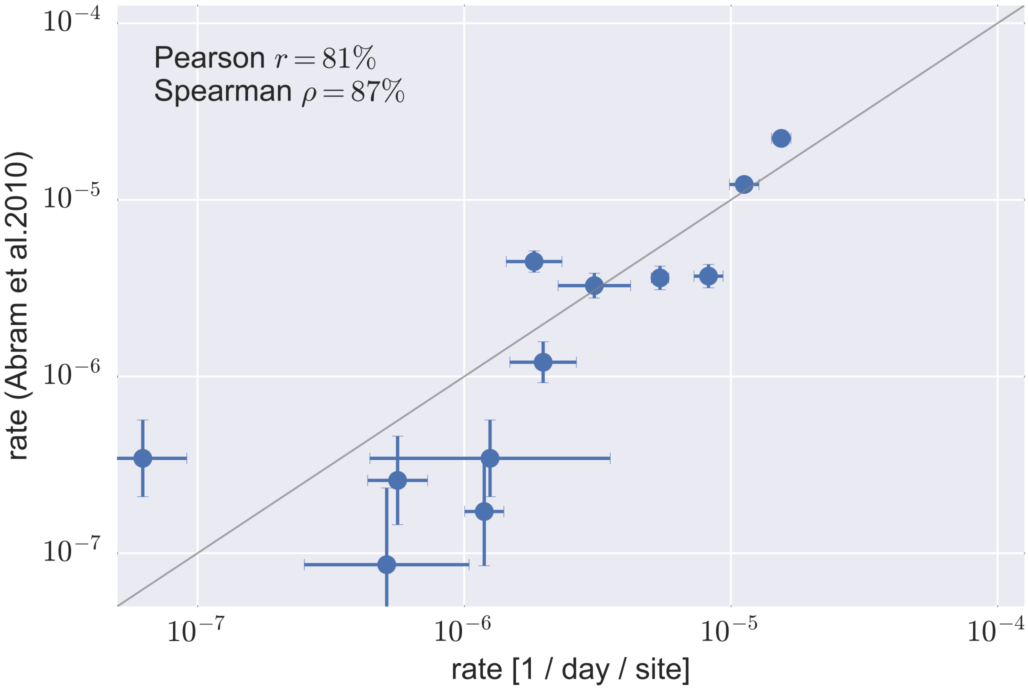

FIG. S1 Comparison to previously published in vitro measurements of the mutation rate matrix by Abram *et al.* (2010). Error bars for the estimates are standard deviations over 100 patient bootstraps. Error bars for the values from Abram et al. (2010) are standard deviations of binomial sampling noise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

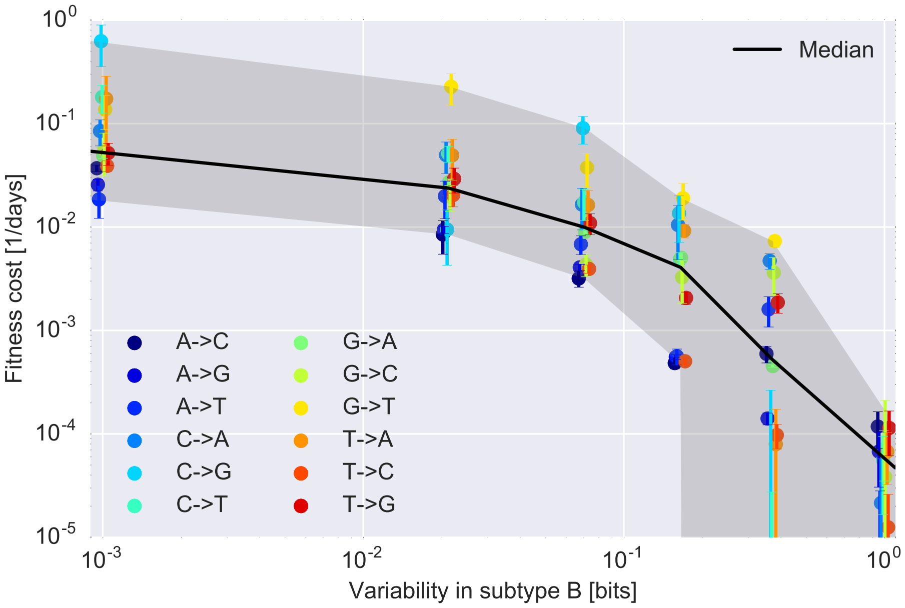

FIG. S2 Fitness cost estimates from saturation curves similar to Fig. 2 but separate for each of the 12 mutations. The general picture is the same like shown in Fig. 2, but some mutations appear to be slightly more or less suppressed than the average.

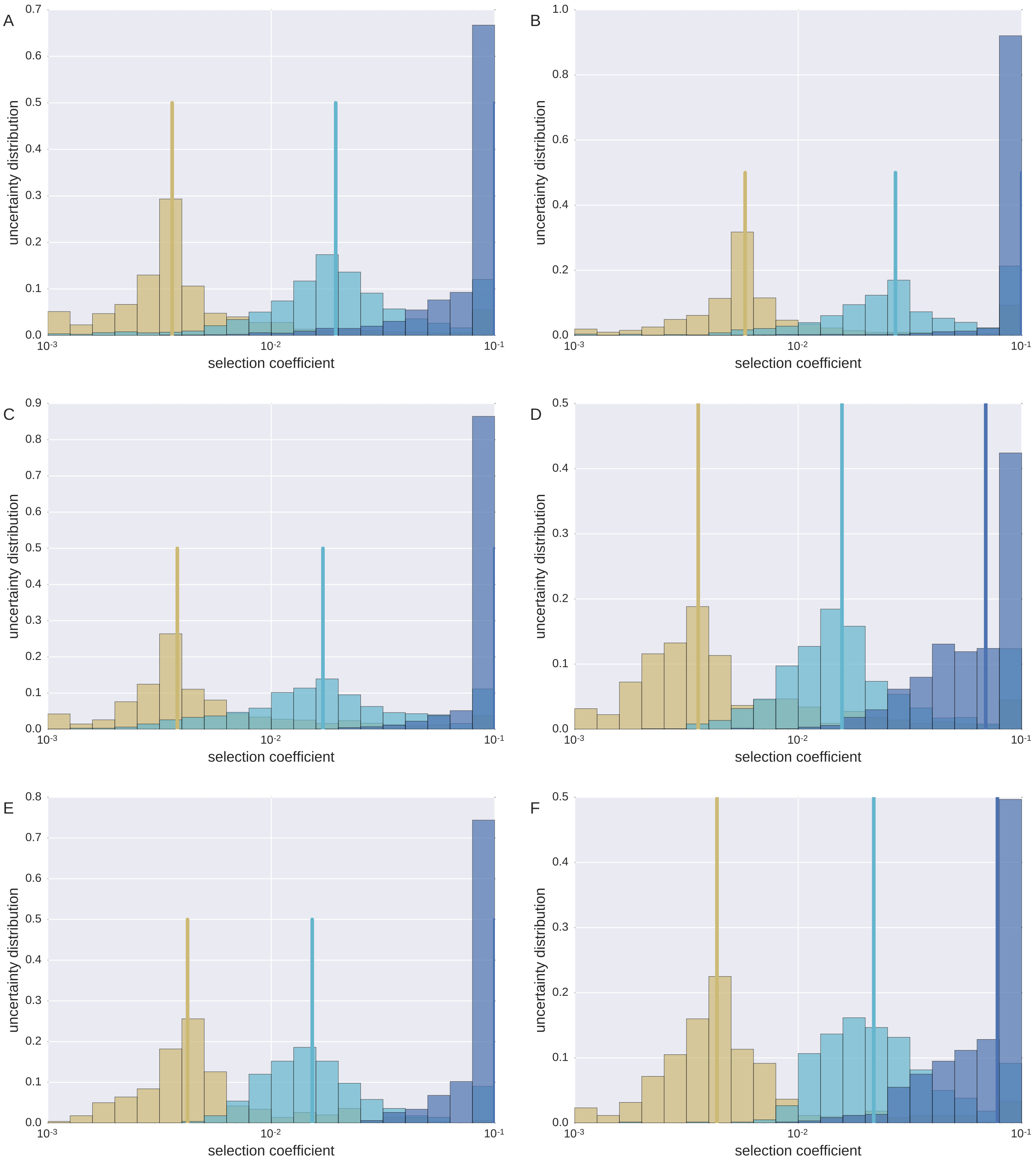

FIG. S3 Bootstrap confidence on fitness costs, like Fig. 3D but for other regions of the genome. Panel A: pol, B: env, C: nef, D: vif, E: vpu, F: vpr.

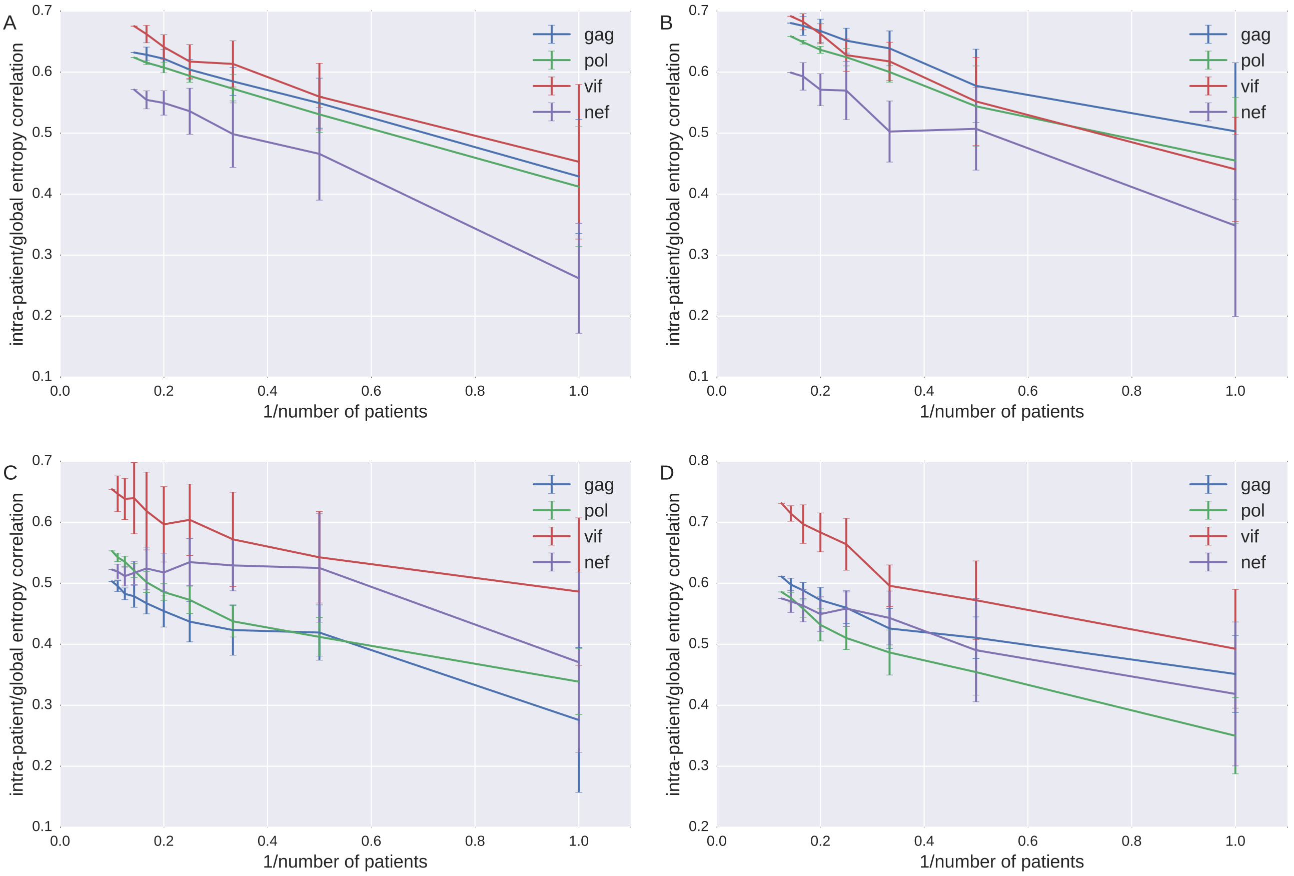

FIG. S4 Correlation of intra-patient diversity with global cross-sectional diversity vs the number of patients from which virus populations are used to estimate typical intra-patient variability (only at sites without sweeps where majority state agrees with the consensus state). Panels A and B show correlation at the nucleotide level, while panels C and D show correlations at the amino acid level. A and C include patients from all subtypes and compare against diversity in group M, while panels B and D are restricted to subtype B.