Constrained mutational sampling of amino acids in HIV-1 protease evolution

Jeffrey I. Boucher,[1,*] Troy W. Whitfield,[2,3,*] Ann Dauphin,[4] Gily Nachum,[1] Konstantin B. Zeldovich,[3,†] Ronald Swanstrom,[5] Celia A. Schiffer,[1] Jeremy Luban,[1,4] and Daniel N. A. Bolon[1,‡]

[1]*Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605*

[2]*Department of Medicine, University of Massachusetts Medical School, Worcester, MA 01605*

[3]*Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01605*

[4]*Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA 01605*

[5]*Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599*

Abstract

HIV-1 samples a large number of sequence variants that help the virus to evade the immune system and escape from antiviral drugs. However, it is unclear how amino acid changes that require multiple nucleotide mutations contribute to the evolution of HIV-1. In a large database of HIV-1 protease sequences from circulating viruses we find that the most frequently observed amino acids are disproportionately accessible by single-base mutations from the consensus amino acid. Multiple-base mutations at a codon may be rarely sampled and/or exhibit fitness defects. To investigate the impact of fitness effects on mutant frequency, we quantified the experimental impacts of all individual amino acid changes in protease on the expansion of HIV-1 in tissue culture. Many amino acid changes requiring multiple nucleotide mutations that efficiently supported viral replication were observed rarely, or not at all, in the sequence database. Frequently observed amino acid changes that require multiple nucleotide mutations were accessible by single nucleotide mutational intermediates with higher fitness, on average, compared to the intermediates for unobserved amino acid changes. An additive quasispecies model that combined different mutational barriers for single- and multiple-base mutations with the experimental fitness effects showed an improved correspondence with clinical observations of mutant frequencies compared to fitness alone. Our results indicate that mutational sampling provides a larger barrier than fitness effects to amino acid changes requiring multiple base mutations in HIV-1 protease.

INTRODUCTION

Mutations are important to HIV-1 and many other viruses because they enable evasion of immune recognition[1] and escape from anti-viral drug treatments[2]. For example, when mutation rate was reduced by engineering a high-fidelity polymerase in polio, the resulting viruses had an impaired ability to infect and colonize in mice[3,4]. Similar observations of reduced infectivity for engineered high-fidelity viruses have been made in chikungunya virus[5] and human enterovirus 71[6]. In HIV, mutations that either increase or decrease the mutation rate show reduced replication efficiency in cell culture[7], suggesting that the mutation rate of HIV-1 has been subject to natural selection.

HIV-1 generates mutations as an inherent part of its infection cycle. HIV-1 is an RNA virus that replicates in host cells through a DNA intermediate. The process of reverse transcribing viral RNA into DNA is error prone and is a main contributor to the mutations that HIV-1 accumulates[8,9]. The error rate of HIV-1 replication in cell culture has been measured as 3×10^{-5} mutations per base per replication cycle[10]. This corresponds to 1 mutation for every 2-3 genomes of HIV-1. In addition to errors that occur during genome copying, additional mutations are caused after synthesis by host cytidine deaminases. Recent analyses of the frequency of null alleles of HIV-1 integrated as proviruses within host DNA[9] indicate that the *in vivo* error rate may be higher than estimates from cell culture, though these findings may be influenced by selection pressure that should disfavor functional proviruses that would kill the parental cell unless silenced. Because of the large number of virions in an infected individual, even conservative estimates of the error rate lead to tremendous genetic diversity. In a widely cited review[11], modeling of HIV-1 infection kinetics indicated that, "mutations will occur in every position in the genome multiple times each day and that a sizable fraction of all possible double mutations will also occur."

While mutations in HIV-1 occur at the nucleic acid level, many of the functional impacts are caused by changes in the encoded protein sequence[12]. The amino acid changes that are efficiently sampled by HIV-1 depend on multiple factors including the probabilities of different types of nucleic acid mutations and the genetic code. Single nucleotide substitutions are the most frequent errors generated when HIV-1 replicates[10,13]. Amino acids that are accessible by a one nucleotide mutation from the parental sequence should be more frequently sampled during HIV-1 infection and evolution. Within the genetic code, amino acids with

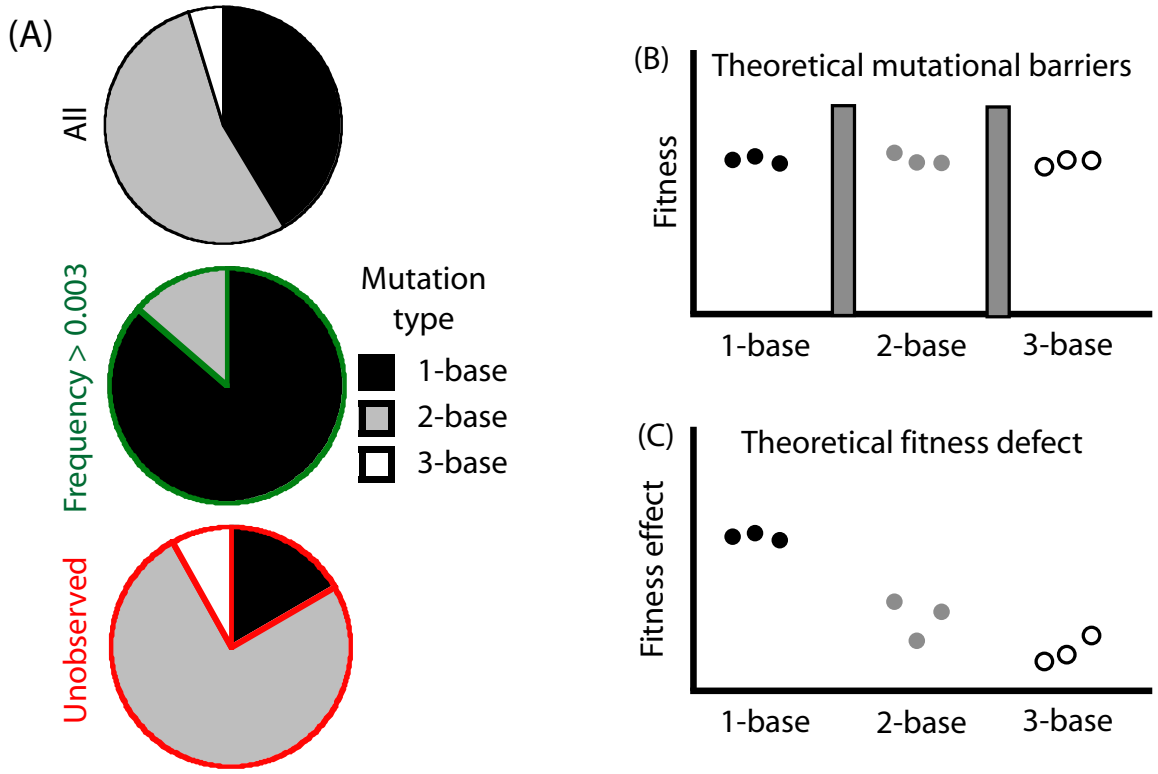

FIG. 1. Amino acid frequencies of patient derived HIV-1 protease suggest that mutational sampling and/or selection shapes amino acid preference. (A) Pie graphs illustrating the fraction of amino acid changes that require 1-, 2-, or 3-nucleotide conversions for all changes (top), commonly observed amino acids, and changes that were unobserved in patient isolates. (B) Depiction of a simple fitness landscape where all mutations exhibit similar fitness effects and mutational probabilities will strongly influence frequency. (C) A landscape where fitness effects provide a strong preference for amino acid changes accessible by single nucleotide conversions.

similar physical and chemical properties are encoded by similar codons making it likely that single nucleotide mutations will cause conservative amino acid changes[14]. Detailed analyses indicate that strong selection acts on the genetic code[15] and that only one in a million random genetic codes is as effective at minimizing disruptive amino acid changes from single base mutations[16]. Thus, amino acids changes that require multiple-base mutations could be infrequent due in part to disruptive amino acids causing fitness defects.

Here, we sought to understand how the genetic code impacts protein evolution in HIV-1 where most if not all single nucleotide mutations are sampled within an infected individual. We chose to focus on protease because of a wealth of available sequence data and the

relatively small size of the 99 amino acid gene that facilitated experimental analyses. Many amino acids requiring multiple-base mutations were infrequently observed in clinical isolates yet were compatible with efficient viral replication. Despite the extensive genetic diversity of HIV-1 within an infected individual, our results indicate that inefficient sampling of amino acids requiring multiple nucleotide changes creates a mutational barrier such that many amino acids capable of supporting efficient viral replication are not readily available to the evolution of HIV-1 proteins.

RESULTS AND DISCUSSION

Analyses of protease sequences from clinical isolates

To explore mutation and selection acting on circulating HIV-1 in the absence of drug pressure, we analyzed the amino acid sequence of protease[17,18] from 38,781 protease-inhibitor-naïve people infected with subtype B HIV-1 (Fig. 1). The vast majority of commonly observed amino acids in protease (frequency > 0.003) from these sequences were accessible by single nucleotide changes from the consensus amino acid (Fig. 1A). In contrast, the preponderance of amino acids that were not observed required multiple nucleotide changes from the consensus amino acid. There are two main reasons that an amino acid may be at low frequency in a population: mutation and selection. If an amino acid causes a fitness defect, selection will drive it to low frequency. For neutral amino acids without any fitness effects, the probability of the amino acid will be proportional to the rate at which mutations cause the amino acid[19]. The infrequent observation of amino acids requiring multiple nucleotide changes in the compiled database could be due to fitness defects and/or mutational barriers (Fig. 1B).

Quantification of experimental fitness effects

To distinguish the influence of selection from mutation, we performed an EMPIRIC[20] mutational scan of the HIV-1 protease gene (Fig. 2). We individually randomized each amino acid position of protease in the NL4-3 strain of HIV-1 (which represents the consensus sequence from our sequence database) and quantified the experimental fitness of each possible mutation during infection of a T-cell line using a deep sequencing readout[21].

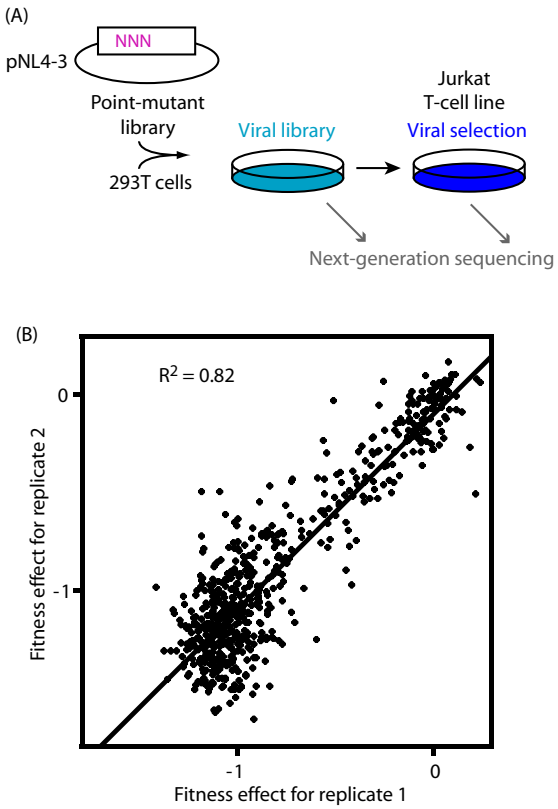

FIG. 2. EMPIRIC measurement of fitness effects. (A) Point mutations were generated in pNL4-3 and transfected into 293T cells to generate viral libraries that were expanded in Jurkat cells. Next-generation sequencing was used to quantify the change in frequency of mutations in the viral library before, during and after passage in Jurkat cells. (B) Comparison of fitness effects for amino acid changes from experimental replicates for three regions of protease (amino acids 20-29, 50-59, 80-89).

Fitness effects were scaled as selection coefficients where 0 corresponds to wildtype and -1 corresponds to a null allele (see Methods). Across the sequence of protease, synonymous mutations tended to exhibit fitness effects centered on 0 (Fig. S1), suggesting that fitness impacts were primarily determined by amino acid changes. Of note, windows of synonymous mutations showed modest yet statistically significant deviations from neutral expectations in three regions: at the beginning of protease and centered at amino acid positions 34 and 54 (Fig. S1). The fitness effects of synonymous mutations at the beginning of protease are consistent with selection on the amino acid sequence of the p6 reading frame that overlaps with this region of protease (Fig. S2). We do not have explanations for the observed fitness

effects of synonymous mutations at other regions of protease. As the observed fitness effects of synonymous mutations were modest, we chose to estimate the fitness effects of each amino acid change in protease by summing sequencing counts over all synonyms (Supplementary Table 1).

To assess the reliability of our fitness estimates, we repeated the viral competitions for three regions encompassing 30 of the 99 amino acid positions in protease (Fig. 2B). The fitness effects between repeated viral competitions show a strong linear correlation ($R^2 = 0.82$) indicating that the bulk competitions provide a reasonably precise estimation of fitness effects of each amino acid change under the conditions of the viral expansion. The fitness effects of amino acid changes tend to cluster around neutral and null, as has been almost universally observed for systematic or random mutation experimental evolution studies[22,23]. The neutral cluster and the null cluster are well distinguished in both experimental replicates.

Fitness effects compared to frequency in circulating variants

We compared experimental fitness effects with the frequency of amino acids in our inhibitor-naïve HIV-1 sequence database (Fig. 3A). We were cautious about the potential influence of epistasis on the interpretation of mutant effects measured in one genetic background compared to naturally occurring mutations that can appear in a wide array of genetic backgrounds. If the effect of most amino acid changes in protease were strongly dependent on other mutations, then we would not observe a clear relationship between fitness effects in NL4-3 and frequency in circulating variants. However, the amino acids most frequently observed in circulating variants (green dots in Fig. 3A) tend to exhibit small to no experimental fitness effect in NL4-3. Because strongly deleterious alleles are extremely unlikely to rise to high frequency[24], our results suggest that for commonly observed amino acids, experiments with NL4-3 provide a reasonable estimate of fitness effects in circulating HIV-1. Of note, this does not necessarily mean that epistasis is uncommon or unimportant in HIV-1 protease as epistasis is a general feature of protein evolution[25]. NL4-3 protease is the consensus of clinical sequences, and the common clinical amino acid changes in protease may frequently occur on a genetic background that is similar enough to NL4-3 to have closely related fitness effects.

Amino acids rarely observed in clinical isolates (grey dots in Fig. 3A) displayed a wide

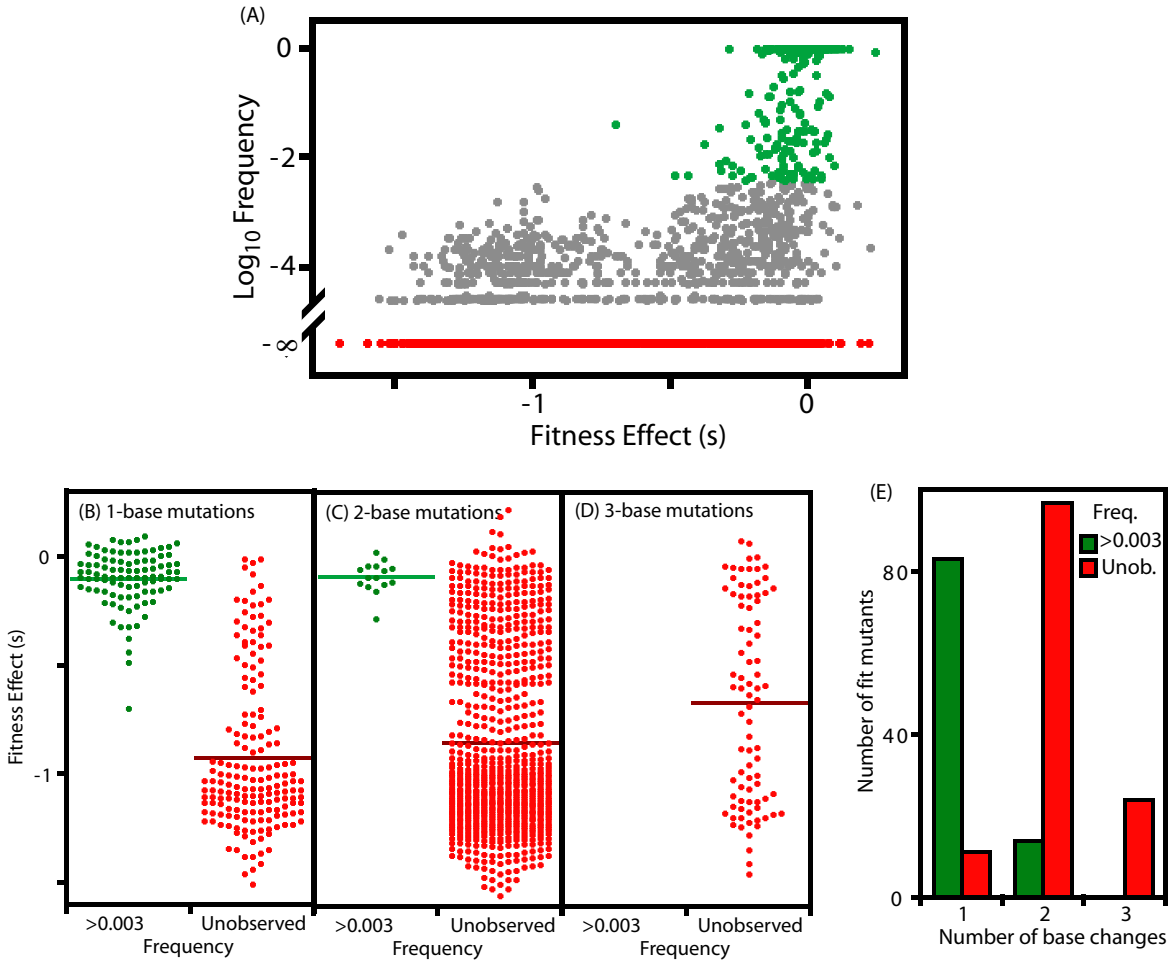

FIG. 3. Relationships between experimental fitness effects and the frequency of mutations in circulating viruses. (A) The frequency of mutations in sequenced viruses from drug-naïve patients compared to fitness effects. Mutations that occurred at a frequency greater than 0.003 are shown in green, unobserved mutations are shown in red, and all other frequencies are shown in grey. (B-D) Dot plot representations of the distribution of fitness effects for amino acid substitutions that were either observed at frequency greater than 0.003 or that were unobserved. (E) Bar graphs showing the number of observed (frequency $>$ 0.003, in green) and unobserved (red) 1-, 2- and 3-base mutations.

range of fitness effects in NL4-3. We considered different potential reasons for this range of fitness effect in NL4-3 that included many apparently null amino acid changes. We considered the possibility that viruses may expand more readily during infections of human hosts than in cell culture. However this seems unlikely as cell culture is generally a more permissive environment compared to hosts because there is no immune pressure[4]. Epistasis

may provide an explanation for the infrequently observed clinical amino acids that have a strong fitness defect in NL4-3, as the genomes where these amino acids occur could have accumulated secondary permissive mutations.

We also considered that null alleles may appear from sequencing errors in the database and/or from sequencing of non-infectious viral particles that could be generated from activated proviruses with genetic defects[26]. We cannot distinguish between these mechanisms because they can both cause null alleles to appear in the database. Consistent with these mechanisms, stop codons were present in the protease sequence database. To estimate the likelihood of observing a null protease allele in the database we compared the number of observed stop codons (197) to the number of single-base mutations from the consensus sequence that could lead to a stop codon (98). Based on this ratio we explored how many null amino acids we may expect to find in the clinical data. We defined amino acid changes as null if they had fitness effects within two standard deviations of the average stop codon in our experimental fitness scan (Supplementary Table 1). There were 1001 single nucleotide mutations that could lead to a null amino acid. Based on the 2 to 1 ratio of stop codons to single-base mutational pathways to stop codons, we expected roughly 2000 null amino acids accessible by single nucleotide mutations in the clinical database. We observed 2051, which closely matches the number expected based on the number of stop codons from the database. These findings indicate that termination codons and most experimentally null alleles seen in the database are likely generated by similar mechanisms, which could be errors in sequencing or non-viable viral lineages.

The appearance of null alleles in the database complicates estimates of fitness based on mutant frequency. Of note, experimentally null alleles were rare in the database and none were observed at a frequency greater than 0.003. By focusing on mutants with frequencies greater than 0.003 we excluded likely null alleles, providing a more reliable set of frequency-based estimates of fitness. We focused on the frequency of these mutants in all further analyses of sequenced variants from the database.

For understanding how HIV-1 samples amino acid changes requiring multiple-base mutations, we focused on amino acids commonly observed (frequency greater than 0.003) in circulating variants.

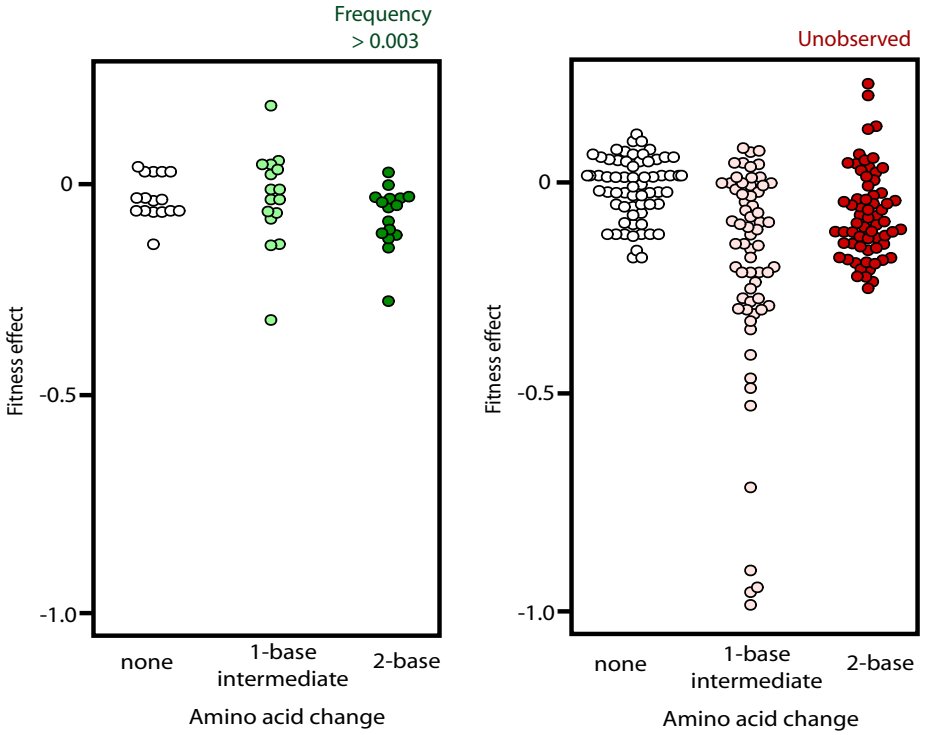

FIG. 4. Frequently observed amino acid changes requiring two nucleotide changes are accessible by fit intermediates compared to highly fit amino acid changes that were unobserved. (A) Dark green circles show fitness effects for amino acid changes requiring two nucleotide changes that were observed with a frequency greater than 0.003 in clinical samples. Light green circles show the most fit amino acid intermediate and white circles show on-pathway synonymous codons of the consensus amino acid. (B) Dark red circles represent a randomly selected set of highly fit amino acid changes requiring two nucleotide mutations that were not observed in the clinical database. Pink circles represent the most fit amino acid intermediate and white circles show on-pathway synonymous codons of the consensus amino acid. The intermediates for the frequently observed amino acid changes exhibit a higher average fitness ($p < 0.003$) than the intermediates for the unobserved amino acid changes.

Distinctions between single and multiple nucleotide mutations

We analyzed the number of nucleotide mutations required for each possible amino acid change from the consensus sequence. For amino acid changes accessible by changing a single base, the most common type of mutation made during HIV-1 replication[10], most experimentally fit amino acids were observed in circulating variants (Fig. 3B). From this observation,

we deduce the following: HIV-1 thoroughly samples single nucleotide mutations, consistent with previous modeling[11]; selection is a primary determinant of unobserved single base mutations in the clinical database; and experimental fitness effects in NL4-3 are predictive of common and unobserved amino acid changes in the absence of strong mutational barriers.

Two and three nucleotide codon changes (Figs. 3C and 3D) show a distinct pattern compared to the single base mutations (Fig. 3B). We used the fitness effects of amino acids that were common in circulating variants to define a range of naturally fit mutations. For multiple nucleotide mutations, there were many amino acid changes that supported efficient viral replication in culture but that were not observed in circulating variants (Fig. 3E). Mutational barriers that limit access to two and three base mutations provide a compelling rationale for these observations. According to well established population genetic theory[19], the probability of a neutral amino acid should be proportional to the rate of mutations leading to the amino acid. As single nucleotide changes are the most common mutation in HIV-1[10], amino acid changes requiring multiple base changes should often arise by the serial accumulation of individual independent mutations. The likelihood of this mechanism of multiple-base change is the product of the probabilities of each individual mutation and is less likely than the single-base mutations. The lower likelihood of multiple-base mutations compared to single-base mutations is a form of mutational barrier that should limit their frequency in populations of HIV-1.

Mutational walks

Because simultaneous replacement of two or more nucleotides in a codon is rare[10], mutational walks of separately occurring single base mutations may provide relevant access to these types of amino acid changes. The likelihood of a mutational walk depends on the fitness effects of intermediate steps[27,28]. If an intermediate step is strongly deleterious, then the pathway will be unlikely to occur even if the final state is neutral. We examined the intermediates for two nucleotide mutations (Fig. 4) where we identified a set of both common and unobserved clinical amino acids with similar high fitness. Among the two base mutations common in circulating variants (Fig. 4A), fitness effects for intermediate one base mutations exhibited a narrow and highly fit distribution. Intermediates for the unobserved two-base mutations (Fig. 4B) exhibited a broader range of effects with some fitness measurements

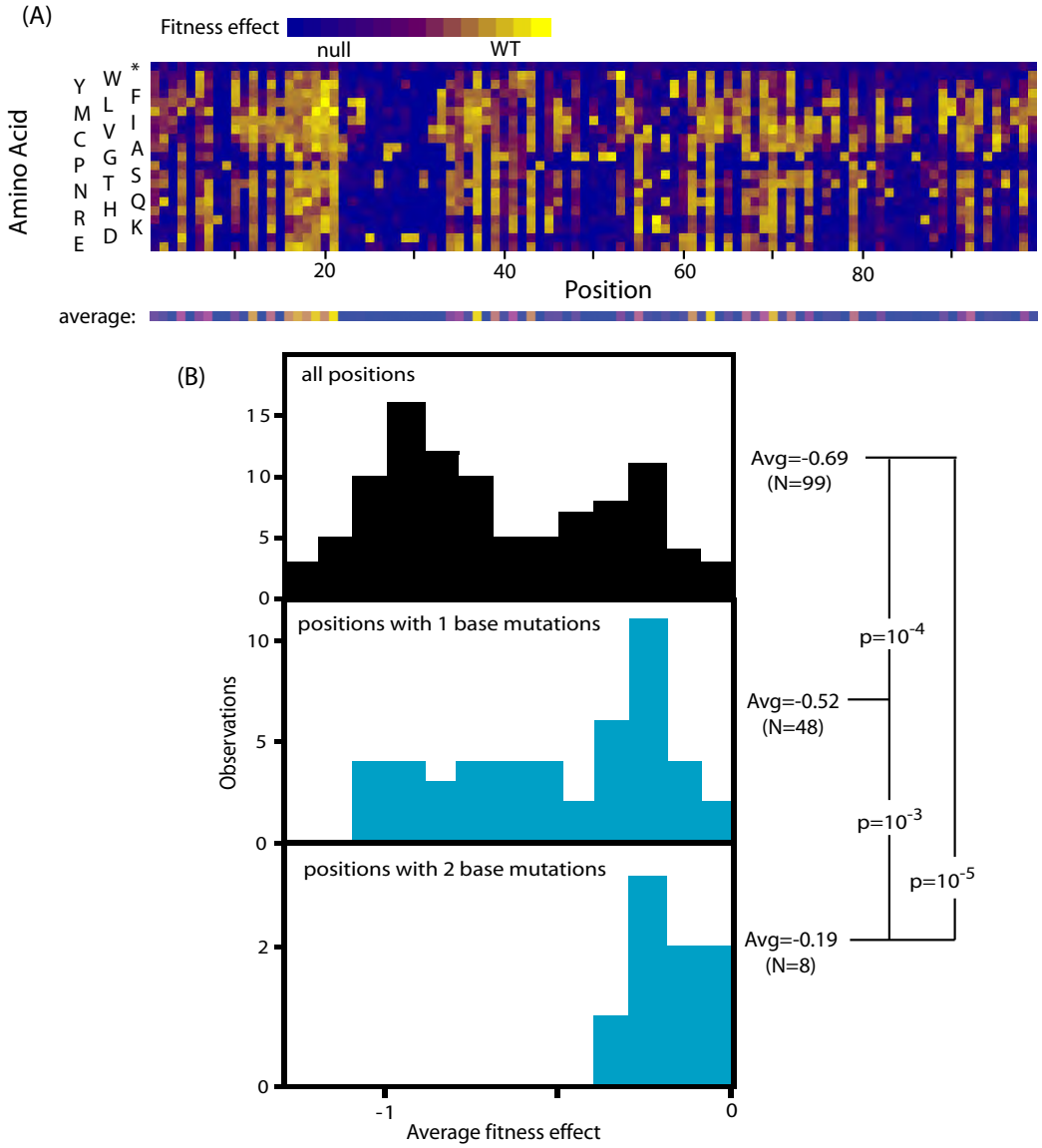

FIG. 5. Frequently observed mutations preferentially occur at positions in protease that are relatively tolerant to mutations. (A) Heatmap representation of the fitness effects of amino acid changes in protease from NL4-3 HIV. The average effect of all amino acid changes is shown on the bottom line and these values were used as a measure of the tolerance of a position to mutation in panel B. (B) The overall distribution of mutational tolerance (top panel) is bi-modal with a cluster of positions where the average fitness effect exhibited mild to no fitness effect and a cluster of positions where the average fitness effect is close to null (-1). Positions where mutations were observed in circulating viruses at frequencies above 0.003 were skewed towards higher tolerance (lower two panels). Statistical significance was assessed using one-tailed bootstrap procedure.

close to null. The average fitness effect of intermediates for unobserved mutations was lower than that for intermediates for common circulating mutations indicating that mutational walks contribute to multiple nucleotide codon changes in circulating viruses.

Multiple base mutations occurred predominantly at highly tolerant positions compared to both the overall distribution of sensitivity and the distribution of single base mutations (Fig. 5). Because selection is weaker at tolerant positions, they should more freely accumulate mutations during HIV-1 evolution. Such broad peaks in local fitness landscapes provide a greater opportunity for mutational walks to amino acids that involve multiple-base changes.

Quasispecies model

Quasispecies models[29,30] provide a framework for considering the likelihood of a genotype based on its fitness and those of its neighbors. We reasoned that the frequency of an amino acid might depend on the flatness of the fitness landscape as described by a quasispecies model. To examine this idea, we calculated amino acid probabilities using an additive quasispecies model, and experimental fitness measurements, and compared these to circulating frequency as a metric of evolution *in vivo* (Fig. 6).

The quasispecies model provides a stronger correlation with amino acid frequency in circulating variants than fitness measurements alone indicating that the model provides a useful way to consider both mutation and selection for codon evolution in HIV-1. Using a reported mutation rate for HIV-1 (3×10^{-5})[10], we observed an improved correlation coefficient compared to fitness effects alone. As expected, the equilibrium population from the quasispecies model is stratified by frequency into single-, double- and triple-base mutations separated by powers of the mutation rate (see Methods). This stratification, however, does not accurately recapitulate the distribution of amino acid frequencies in circulating variants. To account for the potential impact of multiple rounds of replication and mutation that can occur prior to host-to-host transmission, we varied the mutation rate in the quasispecies model and found that an apparent mutation rate $\sim$ 1000-fold higher than the per-generation rate provided the best fit with frequencies in circulating variants. The resulting modeled equilibrium population showed overlap in the frequencies for single-, double-, and triple-base mutations and showed a closer correspondence with the clinical frequencies of mutations compared to a model using the per generation mutation rate. These observations suggest that the qua-

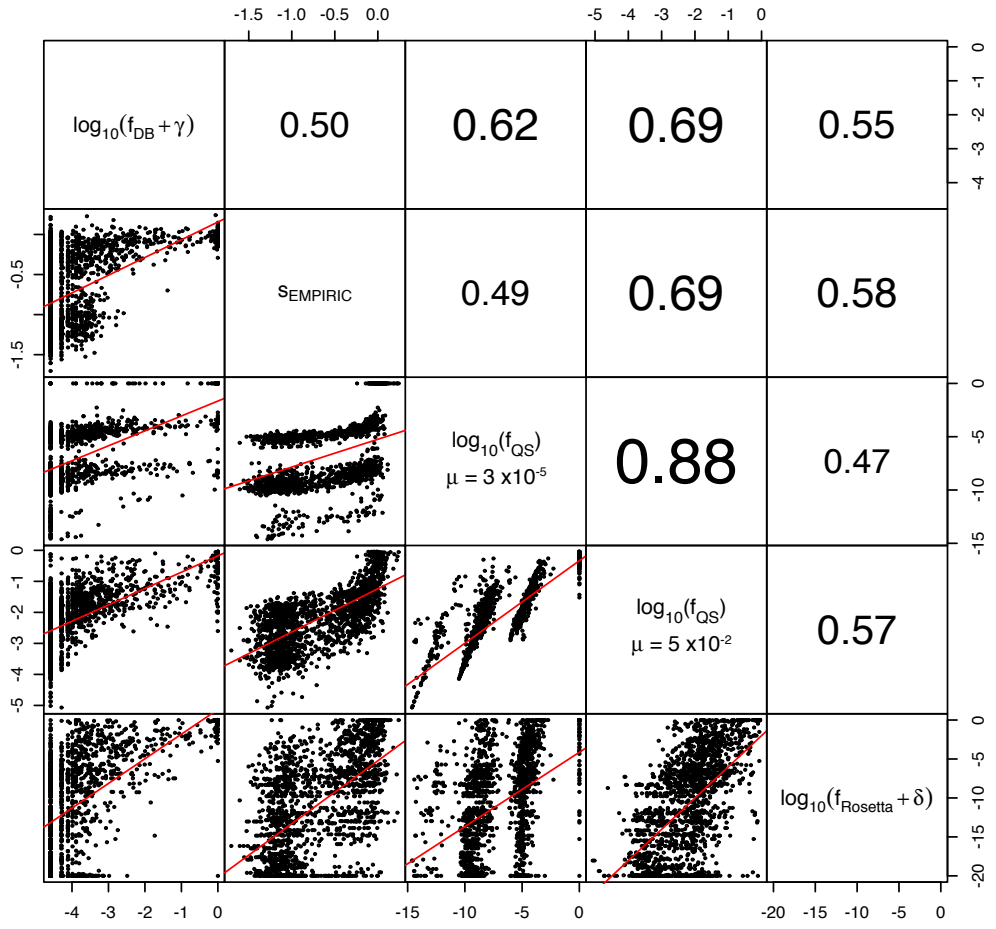

FIG. 6. Fitness effects and a quasispecies model provide improved correlation with mutant frequencies in clinical isolates. Comparison of the frequency of amino acids in clinical samples (f_{DB}) with the following: fitness effects without a mutational model ($s_{EMPIRIC}$); projected frequencies based on an additive quasispecies model (f_{QS}) with either the experimentally measured mutation rate of HIV-1 (3×10^{-5}) or the mutation rate (5×10^{-2}) that provided the strongest correlation with clinical frequencies; and projected frequencies based on previously reported[31] structural simulations with a specialized mutational model ($f_{Rosetta}$). Small offsets ($\gamma = 1/38,781$ and $\delta = 10^{-25}$) were included where necessary to facilitate logarithmic transformations. Spearman rank correlation coefficients are indicated in the upper right portion of the figure.

sispecies model captured key features of the population structure of HIV-1 among different infected hosts.

We compared the measured fitness effects and the quasispecies model with a previously

published biophysical model[31] of protease function (Fig. 6). Kortemme and co-workers used Rosetta[32] to simulate the effects of all possible amino acid changes on critical features of protease function including the folding and dimerization of protease and binding to peptide substrates[31]. Of note, this biophysical study focused exclusively on amino acid substitutions resulting from single-base mutations. The Rosetta predictions more closely correlate with clinical frequency than our experimental fitness measurements without a mutational model, but not as well as the quasispecies model that uses both fitness measurements and a mutational model. These comparisons indicate the importance of both mutation and selection to the sequence evolution of HIV-1. These comparisons also highlight the predictive potential of biophysical models[33,34] when structures of most of the biologically relevant states are available, and the additional value of experimental fitness landscapes.

While the quasispecies model improves the fit with frequencies in circulating variants, it is important to note that roughly half of the variation in frequency remains unexplained. This unexplained variation likely has many contributing factors, including noise in the experimental fitness measurements, and epistasis. The variation between replicate competitions was on the order of 20%, which would have about a four-fold impact on predicted frequency based on our best-fit quasispecies model[35]. While experimental noise is clearly meaningful, it does not fully explain discrepancies with the frequencies in circulating variants. As noted earlier, the sequence database of circulating variants contains null alleles of protease, which makes this frequency an imperfect proxy for *in vivo* fitness. In addition, epistasis should also contribute to discrepancies between fitness measurements made in one genetic background with estimates of clinical samples with diverse genetic backgrounds. Despite these caveats, the fit of the quasispecies model indicates that the fitness effects of mutational neighbors contributes to HIV-1 evolution in circulating viruses.

CONCLUSIONS

Combined analyses of the sequence of circulating variants and an experimental protein fitness landscape of HIV-1 protease indicate that sampling of multiple base substitutions in the same codon is limited during HIV-1 evolution. Likely for this reason, the distribution of amino acid changes in circulating variants is skewed towards amino acid changes accessible by single nucleotide mutations. The errors that HIV-1 makes during genome copying are

predominantly single nucleotide mutations that are unlikely to simultaneously occur in the same codon. Therefore, multiple nucleotide changes to codons are likely to depend on the fitness of single-base-change intermediates. Because the majority of amino acid changes require multiple mutations, this mechanism of mutational sampling can have a large influence on protein sequence evolution, even for viruses with high genetic diversity such as HIV-1.

METHODS

Library construction

To facilitate the initial introduction of mutations, protease plus 50 bases of upstream and downstream flanking sequence bracketed by KpnI sites was cloned from pNL4-3 into pRNDM[36]. Each codon of protease in the pRNDM plasmid was individually subjected to site saturation mutagenesis using a cassette ligation strategy[36]. A pNL4-3Δprotease plasmid was generated to efficiently accept protease mutants from the pRNDM construct. The pNL4-3Δprotease plasmid was constructed with a unique AatII restriction site. The pNL4-3Δprotease plasmid was treated with AatII enzyme followed by T4 DNA polymerase without nucleotides to remove the 3′ overhang. Protease mutant libraries in pRNDM were excised with KpnI and treated with T4 DNA polymerase without nucleotides in order to remove 3′ overhangs. The protease mutant libraries and treated pNL4-3Δprotease samples contained 25 bases of complementarity at either end and this facilitated efficient assembly using Gibson Assembly®. All enzymes were from New England Biolabs. Mutants at 9-10 consecutive positions in protease were pooled to generate 10 library samples that together include mutations at each of the 99 positions in protease.

Growth competition

Viral recovery and competitions were performed similar to previous descriptions[21]. Briefly, 2.5 μg of plasmid DNA encoding full length HIV-1 NL4-3 was transfected into 293T cells using calcium phosphate. Supernatant of recovered P0 viral libraries was harvested after 48 hours, clarified by filtration through 0.45 μm filters and stored at -20 C. We used an RT assay, which quantifies reverse transcriptase activity by real-time PCR, to normalize virion production[37]. Viral infections were performed using 5×10^8 RT units of

virus (P0) and 3×10^6 Jurkat T cells in 500 μL RPMI complete media for 2 hours. Cells were washed twice with sterile PBS and seeded in 1.5 mL RPMI complete media in a 24-well plate. P1 viral supernatant was collected and cells were split on days 2, 4, 8, 11, 14, and 16. Fresh media was replaced after each collection time-point. Viral RNA was isolated from P1 viral supernatant collected on days 4 and 8 using Qiagen's QIAamp MinElute Virus Spin kit. Ultracentrifugation of 300 μL of viral supernatant was performed to pellet virions using a Beckman Optima Max XP ultracentrifuge with a TLA-55 rotor at 25,000 rpm for 1 hour at 4 C. Supernatant was removed and virion-associated RNA in the pellet was isolated according to the kit protocol, and eluted in 20 μL.

DNA preparation and sequencing

Samples were prepared for sequencing essentially as previously described[21]. HIV genomic RNA was extracted from supernatants containing virions using High Pure Viral RNA kit (Roche Inc.). SuperScript III and a primer downstream of the randomized regions were used to reverse transcribe viral RNA to cDNA. The cDNA samples were processed as previously described[36] to add barcodes to distinguish pre- and post-selection samples as well as to add base sequences needed for Illumina sequencing.

Sequence analysis

Viral RNA samples drawn from transfected cells at days 0 (initial transfection from 293T cells to Jurkat T-cells), 4 and 8 were processed using Illumina 36-bp single read sequencing on a Genome Analyzer II. Reads with a Phred score of 20 or above (>99% confidence) were used for time-dependent analysis. Using these counts, for each species, i, the selection-rate constant[38], relative to the mean Malthusian fitness of stop codons can be defined as the slope of a normalized logarithmic abundance versus time

$$\widetilde{m}_i = \frac{d}{dt} \ln(N_i(t)/\lambda(t)) = m_i - m_\lambda,$$

where $N_i(t)$ is counts at time t and m_i is the standard (i.e. un-normalized) Malthusian fitness. Defining the normalization factor

$$m_\lambda = \frac{d}{dt} \ln \lambda(t) = \frac{1}{n_{\text{stop}}} \frac{d}{dt} \ln \left(\prod_{k=1}^{n_{\text{stop}}} N_k(t) \right),$$

where n_{stop} is the number of stop codons in the library, ensures that the mean fitness of stop codons is zero. Defining the relative (i.e. Wrightian) fitness with respect to the wild-type (NL4-3 here) sequence

$$w_i = \frac{\widetilde{m}_i}{\widetilde{m}_{WT}},$$

ensures that $w_{WT} = 1$. The selection coefficient is $s_i = w_i - 1$. In order to collect measurements across the entire protease sequence, a set of 10 non-overlapping EMPIRIC libraries was prepared, each one spanning 9 or 10 contiguous amino acid positions. Each library was normalized separately, as described above, to generate the fitness landscape depicted in Fig. 5A. For several of these libraries, biological replicate experiments were performed (see Figs. 2B and S1), yielding $R^2 = 0.72$ at the codon level or $R^2 = 0.82$ at the amino acid level, after averaging among synonymous codons.

Analysis of sequences from circulating isolates

Sequence information for HIV-1 protease from clinical isolates was collected from the Stanford University HIV drug resistance database[17,18] in April, 2017 using the Genotype-Rx protease data-set. The protease sequences were restricted to include only subtype B strains that were annotated "None" for treatment with protease inhibitors and had a complete nucleotide sequence (i.e. "NASeq"). These requirements resulted in 38,781 sequences, of which 13,958 contained at least one position with an annotated mixture of amino acids (e.g. "KR" represents a mixture of lysine and arginine). At positions annotated with mixtures, each amino acid in the reported mixture was counted in our analyses. We did not count stop codons because they represent known null alleles, of which there were 197 in the database. We examined the impacts of excluding mixtures on the relationship between amino acid frequency and fitness (Fig. S3). Excluding sequences that contained mixtures did not dramatically alter the correspondence between frequency and experimental fitness. We examined the location of amino acid mixtures (Fig. S4) and note that they tend to occur at positions that exhibit the most amino acid divergence between different isolates. This observation is consistent with many of the mixtures representing intra-host variation and was a motivation for including reported amino acid mixtures in our analyses.

Modeled populations

While data collected via EMPIRIC provide a direct readout of relative fitness for the various species in the library, under certain population dynamics assumptions these fitness measurements can also be used to extrapolate initial (constructed) populations to those at later times.

The quasispecies model[29,39] has previously been used to describe the evolutionary dynamics of HIV[40]. Since our measurements were done on libraries of single codon substitutions, we consider an additive discrete-time quasispecies model where at time t, the codon $(1 \leq j \leq 64)$ frequencies, $f_j^{(i)}(t)$, at each position, i, in the protease are related to those at later times by

$$f_j^{(i)}(t+1) = \frac{\sum_k W_{jk}^{(i)} f_k^{(i)}(t)}{F^{(i)}(t)}, \tag{1}$$

where $W_{jk}^{(i)} = w_j^{(i)} Q_{jk}$ is the mutation-selection matrix at position i. This matrix is the product of a site-specific fitness landscape, $w_j^{(i)}$ (characterized by fitting the EMPIRIC data) and a single-site mutational landscape[41],

$$Q_{jk} = \left(\frac{\mu}{(l-1)(1-\mu)} \right)^{d_{jk}} (1-\mu)^N,$$

where each residue position implies a (single codon) sequence length $N = 3$, there are $l = 4$ possible nucleotides at any position in this sequence, μ is the probability of a mutation occurring during an iteration of Eq. 1 and d_{jk} is the Hamming distance between states (i.e. codons) j and k. The average fitness of the population at time t is

$$F^{(i)}(t) = \sum_j \sum_k W_{jk}^{(i)} f_k^{(i)}(t).$$

The steady state solution to Eq. 1 is given by the dominant eigenvector of $W_{jk}^{(i)}$. In order to predict steady state frequency distributions under this model, the mutation rate was fitted by minimizing the mean per-residue Kullback-Leibler divergence between $f_j^{(i)}$ and the corresponding distribution found in multiple sequence alignment of patient-derived protease sequences.

ACKNOWLEDGMENTS

We are thankful to L. Jiang for guidance and helpful advice related to the EMPIRIC experiments. We are also thankful to Tanya Kortemme for sharing data from reference[31] and for helpful discussions. This work was supported by grant P01GM109767 from the National Institutes of Health.

---

* J.I.B. (Jeffrey I. Boucher) and T.W.W. (Troy W. Whitfield) contributed equally to this work.

† Current address: Sanofi Pasteur, Cambridge, MA 02139.

‡ To whom correspondence should be addressed. Dan.Bolon@umassmed.edu

[1] Nowak MA, et al. (1991) Antigenic diversity thresholds and the development of AIDS. *Science* 254(5034):963–969.

[2] Kantor R, et al. (2004) Evolution of resistance to drugs in HIV-1-infected patients failing antiretroviral therapy. *AIDS* 18(11):1503–1511.

[3] Pfeiffer JK, Kirkegaard K (2005) Increased fidelity reduces poliovirus fitness and virulence under selective pressure in mice. *PLoS Pathog.* 1(2):e11.

[4] Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R (2006) Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* 439(7074):344–348.

[5] Coffey LL, Beeharry Y, Bordería AV, Blanc H, Vignuzzi M (2011) Arbovirus high fidelity variant loses fitness in mosquitoes and mice. *Proc. Natl. Acad. Sci. USA* 108(38):16038–16043.

[6] Meng T, Kwang J (2014) Attenuation of human enterovirus 71 high-replication-fidelity variants in AG129 mice. *J. Virol.* 88(10):5803–5815.

[7] Dapp MJ, Heineman RH, Mansky LM (2013) Interrelationship between HIV-1 fitness and mutation rate. *J. Mol. Biol.* 425(1):41–53.

[8] Preston BD, Poiesz BJ, Loeb LA (1988) Fidelity of HIV-1 reverse transcriptase. *Science* 242(4882):1168–1171.

[9] Cuevas JM, Geller R, Garijo R, López-Aldeguer J, Sanjuán R (2015) Extremely high mutation rate of HIV-1 in vivo. *PLoS Biol.* 13(9):e1002251.

[10] Mansky LM (1996) Forward mutation rate of human immunodeficiency virus type 1 in a T

lymphoid cell line. *AIDS Res. Hum. Retroviruses* 12(4):307–314.

[11] Perelson AS (2002) Modelling viral and immune system dynamics. *Nat. Rev. Immunol.* 2(1):28–36.

[12] Zanini F, Puller V, Brodin J, Albert J, Neher RA (2017) In vivo mutation rates and the landscape of fitness costs of HIV-1. *Virus Evol.* 3(1):vex003.

[13] Abram ME, Ferris AL, Shao W, Alvord WG, Hughes SH (2010) Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. *J. Virol.* 84(19):9864–9878.

[14] Woese CR (1965) On the evolution of the genetic code. *Proc. Natl. Acad. Sci. USA* 54(6):1546–1552.

[15] Sengupta S, Higgs PG (2015) Pathways of genetic code evolution in ancient and modern organisms. *J. Mol. Evol.* 80(5-6):229–243.

[16] Freeland SJ, Hurst LD (1998) The genetic code is one in a million. *J. Mol. Evol.* 47(3):238–248.

[17] Rhee SY, et al. (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* 31(1):298–303.

[18] Shafer RW (2006) Rationale and uses of a public HIV drug-resistance database. *J. Infect. Dis.* 194 Suppl 1:S51–8.

[19] Kimura M (1983) *The Neutral Theory of Molecular Evolution.* (Cambridge University Press, Cambridge, UK).

[20] Hietpas RT, Jensen JD, Bolon DNA (2011) Experimental illumination of a fitness landscape. *Proc. Natl. Acad. Sci. USA* 108(19):7896–7901.

[21] Duenas-Decamp M, Jiang L, Bolon D, Clapham PR (2016) Saturation Mutagenesis of the HIV-1 Envelope CD4 Binding Loop Reveals Residues Controlling Distinct Trimer Conformations. *PLoS Pathog.* 12(11):e1005988.

[22] Jiang L, Mishra P, Hietpas RT, Zeldovich KB, Bolon DNA (2013) Latent effects of Hsp90 mutants revealed at reduced expression levels. *PLoS Genet.* 9(6):e1003600.

[23] Canale AS, Cote-Hammarlof PA, Flynn JM, Bolon DN (2018) Evolutionary mechanisms studied through protein fitness landscapes. *Curr. Opin. Struc. Biol.* 48:141–148.

[24] Ohta T (1973) Slightly deleterious mutant substitutions in evolution. *Nature* 246(5428):96–98.

[25] Pollock DD, Thiltgen G, Goldstein RA (2012) Amino acid coevolution induces an evolutionary Stokes shift. *Proc. Natl. Acad. Sci. USA* 109(21):E1352–9.

[26] Maldarelli F, et al. (2014) HIV latency. Specific HIV integration sites are linked to clonal

expansion and persistence of infected cells. *Science* 345(6193):179–183.

[27] Gillespie JH (1983) A simple stochastic gene substitution model. *Theor. Popul. Biol.* 23(2):202–215.

[28] Elena SF, Lenski RE (2003) Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat. Rev. Genet.* 4(6):457–469.

[29] Eigen M (1971) Self-organization of matter and the evolution of biological macromolecules. *Naturwissenschaften* 58:465–523.

[30] Wilke CO (2005) Quasispecies theory in the context of population genetics. *BMC Evol. Biol.* 5:44.

[31] Humphris-Narayanan E, Akiva E, Varela R, Ó Conchúir S, Kortemme T (2012) Prediction of mutational tolerance in HIV-1 protease and reverse transcriptase using flexible backbone protein design. *PLoS Comput. Biol.* 8(8):e1002639.

[32] Kuhlman B, et al. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* 302(5649):1364–1368.

[33] Chi PB, Liberles DA (2016) Selection on protein structure, interaction, and sequence. *Protein Sci.* 25(7):1168–1178.

[34] Bershtein S, Serohijos AW, Shakhnovich EI (2017) Bridging the physical scales in evolutionary biology: from protein sequence space to fitness of organisms and populations. *Curr. Opin. Struc. Biol.* 42:31–40.

[35] Ribeiro RM, Bonhoeffer S, Nowak MA (1998) The frequency of resistant mutant virus before antiviral therapy. *AIDS* 12(5):461–465.

[36] Hietpas R, Roscoe B, Jiang L, Bolon DNA (2012) Fitness analyses of all possible point mutations for regions of genes in yeast. *Nat. Protoc..* 7(7):1382–1396.

[37] Vermeire J, et al. (2012) Quantification of reverse transcriptase activity by real-time PCR as a fast and accurate method for titration of HIV, lenti- and retroviral vectors. *PloS one* 7(12):e50859.

[38] Lenski RE, Rose MR, Simpson SC (1991) Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2,000 generations. *American Naturalist* 138(6):1315–1341.

[39] Eigen M, McCaskill J, Schuster P (1988) Molecular Quasi-Species. *J. Phys. Chem.* 92(24):6881–6891.

[40] Nowak MA, May RM, Anderson RM (1990) The evolutionary dynamics of HIV-1 quasispecies

and the development of immunodeficiency disease. *AIDS* 4(11):1095–1103.

[41] Jain K, Krug J (2007) Adaptation in simple and complex fitness landscapes in *Structural Approaches to Sequence Evolution*, eds. Bastolla U, Porto M, Roman H, Vendruscolo M. (Springer, Berlin), pp. 299–339.