Phylogenetic inference of changes in amino acid propensities with single-position resolution

Galya V. Klink[1], Georgii A. Bazykin[1,2]

[1]Institute for Information Transmission Problems (Kharkevich Institute) of the Russian Academy of Sciences, Moscow, Russia
[2]Skolkovo Institute of Science and Technology, Skolkovo, Russia

*Corresponding author: E-mail: g.bazykin@skoltech.ru.

Abstract

Fitness conferred by the same allele may differ between genotypes, and these differences shape variation and evolution. Changes in amino acid propensities at protein sites over the course of evolution have been inferred from sequence alignments statistically, but the existing methods are data-intensive and aggregate multiple sites. Here, we develop an approach to detect individual amino acids that confer different fitness in different groups of species from combined sequence and phylogenetic data. Our method is based on the mutual distribution of homoplasic (convergent or parallel) and divergent substitutions over a phylogeny, and uses the fact that a higher rate of homoplasic evolution within a clade signifies high fitness of the derived allele. We apply it to the *env* gene of two HIV-1 subtypes, A and B, and to the HA gene of two influenza A subtypes, H1 and H3, and show that the inferred fitness changes are consistent with the fitness differences observed in deep mutational scanning experiments. Changes in fitness do not always trigger episodes of positive selection, and therefore cannot be identified by existing methods for its detection.

Introduction

The same amino acid at a specific position of a protein may confer different fitness in different species. Changes in amino acid propensities can occur in the course of evolution due to substitutions at other epistatically interacting sites (Kondrashov et al. 2002; Jordan et al. 2015). Accounting for co-occurrence of amino acids improves predictions of the rates and patterns of substitutions (Kryazhimskiy et al. 2011) and polymorphism (Zaidi and Makova 2019), as well as the pathogenic potential of mutations (Kondrashov et al. 2002; Jordan et al. 2015) and their functional effects in experiments (Mann et al. 2014; Figliuzzi et al. 2016; Hopf et al. 2017; Riesselman et al. 2018). Changes in amino acid propensities can also happen due to shifts in protein fitness landscape caused by an altering environment, e.g. shifts in immune pressure on proteins of pathogens (Blackburne et al. 2008; Elena et al. 2009; Stray and Pittman 2012).

The variability of amino acid propensities has been inferred from statistical analyses of pooled data from numerous amino acid positions of large protein alignments (Bazykin 2015; Storz 2016). In particular, it has been postulated to bias the phylogenetic distribution of homoplasies – repeated substitutions giving rise to the same allele: parallelisms, convergencies (Povolotskaya and Kondrashov 2010; Goldstein et al. 2015; Zou and Zhang 2015; Klink et al. 2017; Klink and Bazykin 2017) and reversals (Naumenko et al. 2012; Soylemez and Kondrashov 2012). The idea behind this approach

is that the fixation probability of a mutation, and therefore the rate of substitution, giving rise to an allele increases with its conferred fitness (Kimura 1983). Therefore, substitutions producing the same allele are expected to be more frequent in groups of species where it confers high fitness. On a phylogeny, this can be observed as clustering of homoplasies in these groups of species, and such clustering has been interpreted as evidence for changes in fitness landscapes. To increase power, existing methods (Povolotskaya and Kondrashov 2010; Naumenko et al. 2012; Soylemez and Kondrashov 2012; Goldstein et al. 2015; Zou and Zhang 2015; Klink et al. 2017; Klink and Bazykin 2017) aggregate data across multiple sites, and therefore do not allow to track changes in amino acid propensities at individual sites.

Here, we develop an approach to estimate the probability, for a particular amino acid at a particular site, that it confers different fitness in the two given taxa. This approach is based on the phylogenetic distribution of substitutions at the analyzed site. Our method performs well on phylogenies that include hundreds of samples (species or strains) or more. Using this method, we infer sites in the gp160 protein (encoded by the *env* gene) of HIV-1 that have undergone a change in the relative fitness of different amino acids between two distal HIV-1 subtypes, and sites in the HA protein of influenza A that have undergone such a change between two distal influenza subtypes. We then validate these findings using the fitness differences measured directly in deep mutational scanning (DMS) experiments (Haddox et al. 2016; Haddox et al. 2018; Lee et al. 2018).

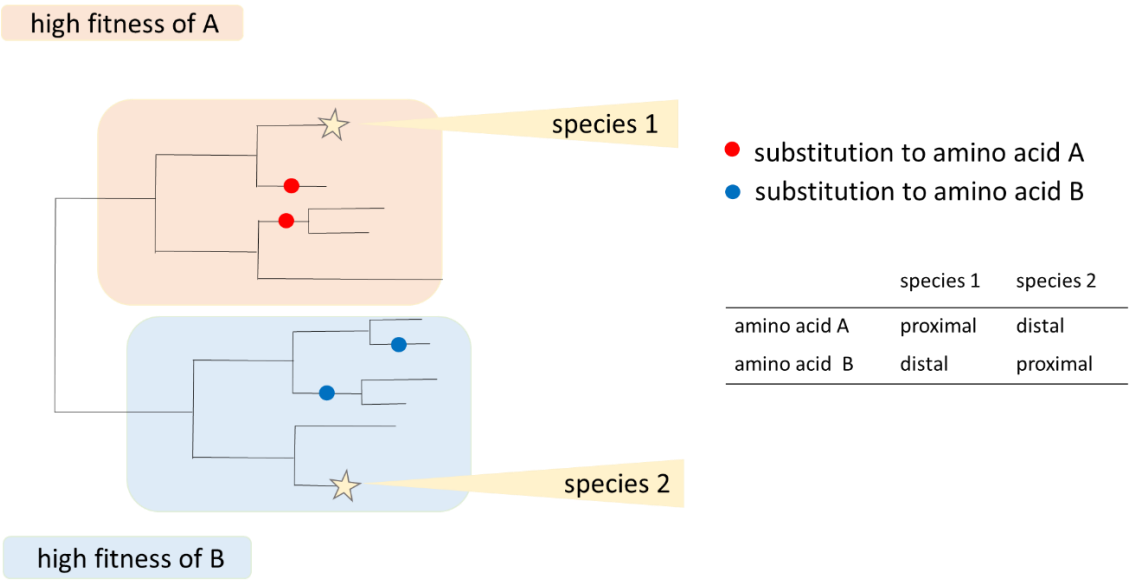
Results

New approach: inferring variable fitness amino acids

Using the reconstructed phylogenetic positions of substitutions, we analyze how homoplasic (parallel or convergent, i.e., giving rise to the same amino acid) and divergent (giving rise to different amino acids) substitutions at the same amino acid site are distributed relative to each other over the phylogeny. We test the hypothesis that aggregation of homoplasic substitutions at a particular fragment of the phylogenetic tree ("phylogenetic neighborhood") implies that the descendant allele confers higher fitness at this group of species, compared to the fitness conferred by it in other species (Klink and Bazykin 2017).

Specifically, we aim to infer individual amino acids at individual sites that undergo significant changes in frequency of homoplasies between the phylogenetic neighborhoods of two given species. For this, we start by considering the first of the two species as the "focal" species. For each amino acid at each site, we then test two hypotheses: that the phylogenetic positions of substitutions into this amino acid are biased towards ("proximal amino acids") or away from ("distal amino acids") the phylogenetic neighborhood of the focal species (Fig. 1). Next, we consider the second species as the focal species, and repeat the procedure. Finally, we assume that the fitness conferred by an amino acid has changed between the two focal species if it was classified as proximal

to one species and as distal from the other, and designate such amino acids as variable fitness amino acids.



	species 1	species 2
amino acid A	proximal	distal
amino acid B	distal	proximal

Figure 1. Schematic representation of the approach for inference of changes in fitness of individual amino acids. Both A and B are variable fitness amino acids because each of them is proximal to one species and distal from the other.

Method validation

To understand the power of our method to identify amino acids with variable fitness, we used SelVA (https://github.com/bazykinlab/SELVa) to simulate the evolution of individual amino acid sites under variable fitness landscape. The substitution rates between different amino acids were determined by their relative fitness, which could differ between tree branches. Specifically, we assumed that at a given site *a* and at each position of the tree, exactly one of the 20 amino acids, AA_k, has the scaled fitness value $F_a^{\mathrm{AA}_k} = 2N_e f_a^{\mathrm{AA}_k} = X$, while the remaining 19 amino acids have $F_a^{\mathbf{AA}_{i\neq k}} = 1$. The $\mathbf{AA}_k \longrightarrow \mathbf{AA}_{i\neq k}$ substitutions then occur at rate $\frac{1-X}{1-e^{X-1}}$, while the $\mathbf{AA}_{i\neq k} \longrightarrow \mathbf{AA}_k$ substitutions occur at rate $\frac{X-1}{1-e^{1-X}}$ (Yang and Nielsen 2008). To model fitness variability, we assumed that these values had changed exactly once at a predefined phylogenetic branch *b;* at this point, the new identity of the amino acid with the fitness *X*, AA_l, is chosen out of the 20 amino acids randomly, and the remaining amino acids get fitness 1. In the process, the fitness of amino acids AA_k and AA_l has changed, while the fitness of the remaining 18 variants has remained constant.

We asked if our method could retrieve the identity of AA_k and AA_l from simulated protein sequences. For simulation, we used the reconstructed phylogenetic tree of the gp160 protein encoded by the *env* gene of HIV-1 (see Methods). The tree has two clades, corresponding to HIV-1 subtypes A and B (Supplementary Figure 1). We assumed that the change in the relative fitness values had occurred in the last common ancestor (LCA) either of subtype A or of subtype B.

The results of these simulations are shown in Figure 2. We find that under moderate selection (*X* between ~4 and ~8), we correctly identify ~40% of the amino acids for which fitness has changed in the LCA of subtype A (Fig. 2A), and ~10% if it has changed in the LCA of subtype B (Supplementary Figure 2A). This sensitivity is achieved under high specificity, so that less than ~20% of the identified amino acids are false positives. For stronger selection (*X* ≈10), we correctly identify a similar fraction of amino acids that change fitness, with almost no false positives, among the amino acids with sufficient variability for our analysis ("potentially variable fitness amino acids", see Methods); however, under such strong selection, most sites are monomorphic, and our method cannot be applied. Conversely, under weaker selection (X<4), the power of the method drops, probably because changes in selection of such a low magnitude are likely masked by the noise associated with drift and tree topology.

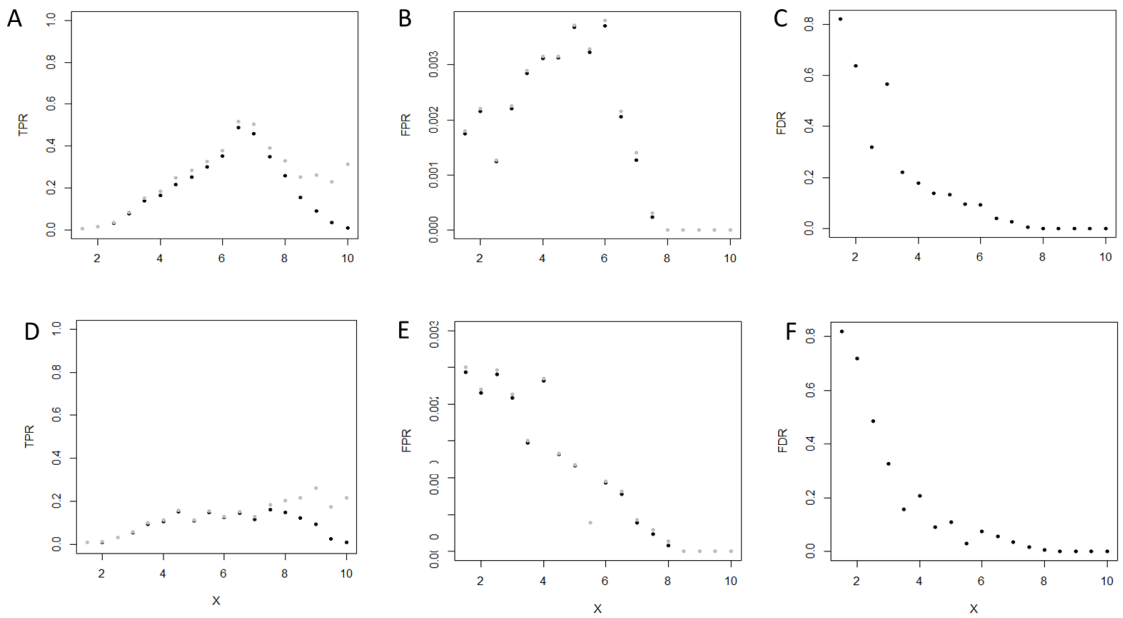


Figure 2. Performance of our approach in simulations for different scaled fitness values of the preferred variant *X;* fitness shift has occurred in the LCA of subtype A (A,B,C) or B (D,E,F); A,D, True Positive Rate: black, among all amino acids with variable fitness; grey, among amino acids with variable fitness that are also potentially variable fitness amino acids. B,E, False Positive Rate: black, among all amino acids with constant fitness; grey, among amino acids with constant fitness that are also potentially variable fitness amino acids; C,F, False Discovery Rate.

Many *env* sites change amino acid preferences in the course of evolution

As a test case, we applied our method to the evolution of the gp160 protein encoded by the *env* gene in HIV-1. This protein evolves under strong positive selection (Yang et al. 2003; Bazykin et al. 2006), suggesting that the site-specific amino acid propensities are likely to change between strains. As focal strains, we considered the two strains tested in DMS experiments (see below): BF520 strain of the subtype A and LAI strain of the subtype B.

Indeed, among the 3491 amino acids with sufficient variability for our analysis (potentally variable fitness amino acids), we detected 59 variable fitness amino acids (at FDR=11%; Supplementary figure 2; see Methods), distributed over 46 of the 724 sites of gp160 (Fig. 3). Since the FPR is below 0.003 for all scaled fitness values (Fig. 2C), we expect to falsely classify no more than 10 amino acids as variable fitness, so at least ~83% of the observed variable fitness amino acids must be real.

The substitutions giving rise to the variable fitness amino acids were positioned on the phylogeny up to 1.75 times closer to one of the focal species, and up to 1.9 times more distantly from the other, than expected randomly. Of these 59 amino acids, 36 amino acids at 34 sites were proximal to the BF520 strain and distal from the LAI strain; and 23 amino acids at 23 sites were proximal to the LAI strain and distal from the BF520 strain. At 11 sites, the changes in fitness were reciprocal: some amino acids were proximal to the BF520 strain and distal from the LAI strain, while others were proximal to the LAI strain and distal from the BF520 strain.

The sites carrying variable fitness amino acids were found at all domains and structural regions of gp120 and gp41, and were not biased towards or away from any functional domain (Fisher's exact test, mean two-sided p for functional domains > 0.6).

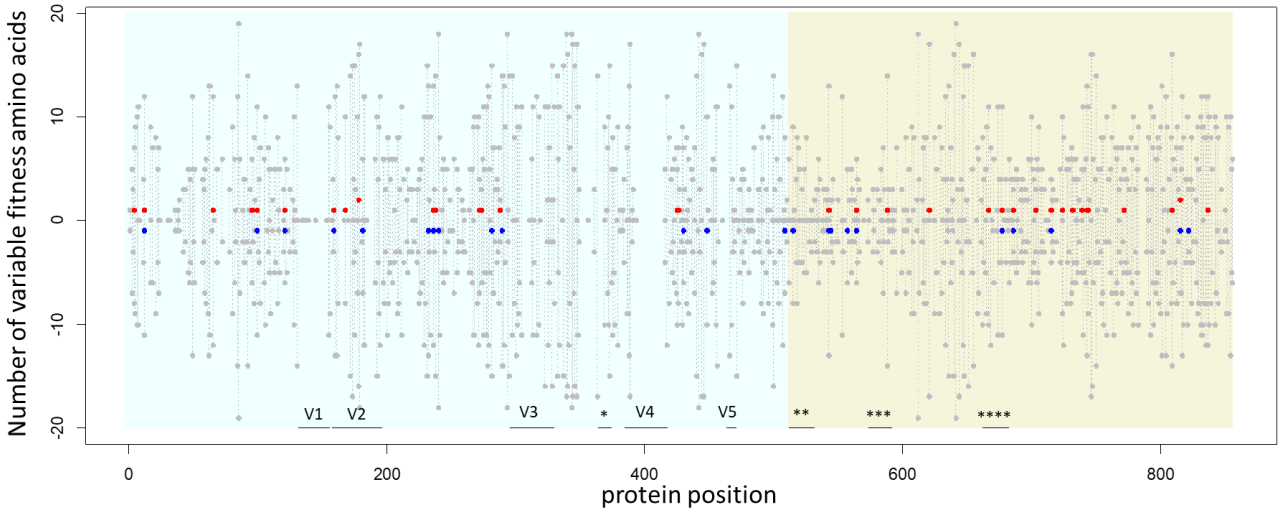


Figure 3. Amino acid sites carrying variable fitness amino acids in gp160 protein of HIV-1. For each position of the amino acid within the protein (horizontal axis), the red dot shows the number of amino acids that are simultaneously proximal to the BF520 strain and distal from the LAI strain; and the blue dot, the number of amino acids that are simultaneously proximal to the LAI strain and distal from the LBF520 strain, multiplied by -1. Grey dots represent the number of amino acids at a site that could potentially have significantly variable fitness (potentially variable fitness amino acids, see Methods). The positions without any dots are sites filtered out due to alignment ambiguity. Vertical grey lines connect dots for the same site. Blue background, gp120; green background, gp41. Variable loops V1-V5 and known functional domains are marked by horizontal black lines: *, CD4-binding loop; **, fusion peptide; ***, immunosuppressive region; ****, membrane proximal external region (MPER). Numbering of protein positions is hxb2-based.

The degree to which the substitutions giving rise to an amino acid may be clustered on the phylogeny should depend on the number of such substitutions. To compare the clustering between sites, we therefore introduced an unbiased measure of its extent. For this, we measured Z - the standard score (z-score) for the tendency of substitutions giving

rise to the considered amino acid to occur in the neighborhood of the first species, and away from the second (see Methods). For the variable fitness amino acids, the absolute value of Z ranged between 3 and 10 (Supplementary figure 3).

Phylogenetically inferred changes in fitness between strains match those observed in DMS experiments

We compared the changes in phylogenetically inferred amino acid preferences between two strains to those inferred experimentally by DMS (Haddox et al. 2016; Haddox et al. 2018). To estimate the changes in DMS preferences, for each amino acid, we measured the ratio $r = \pi_x/\pi_y$, where π_x is the DMS preference for this amino acid in the strain for which it is phylogenetically proximal, and π_y is the preference for it in the strain for which it is distal. As a control, we analyzed the distribution of r between the two strains for invariable fitness amino acids.

The values of r were significantly higher for the variable fitness amino acids than in the control (one-tailed Wilcoxon's test, p = 0.005; Fig. 4), indicating that the phylogenetically and experimentally inferred variable fitness amino acids match. This trend is pronounced both for amino acids that conferred higher fitness in BF520 and in LAI strains (Fig. 4), and is mainly driven by amino acids with modest preferences ($\pi <$ 0.2; Fig. 4B).

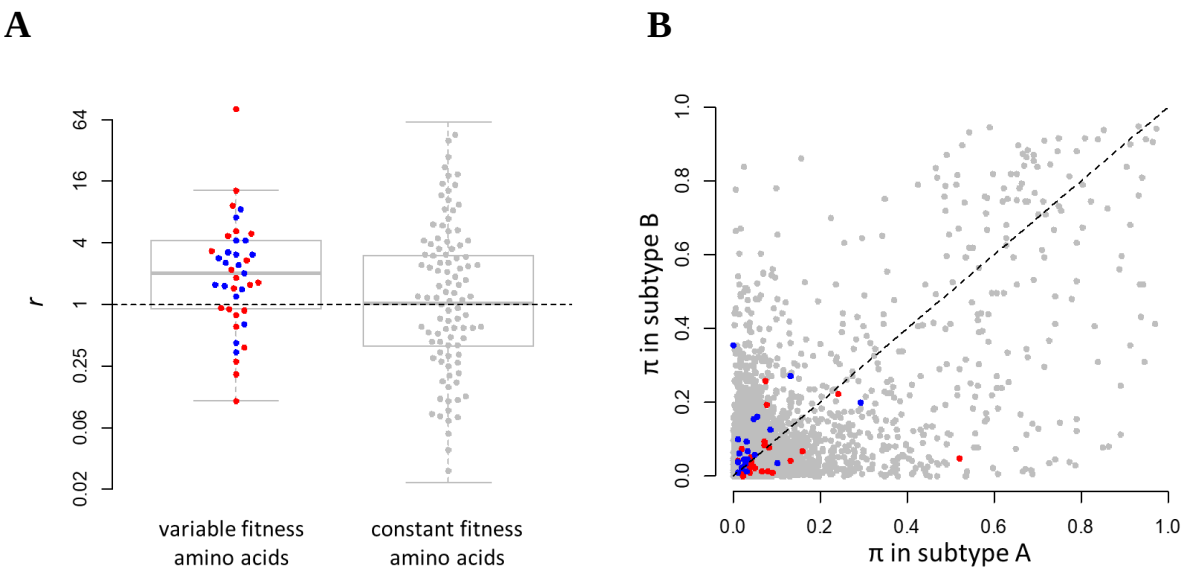


Figure 4. *A*, The ratio of experimentally measured preferences for an amino acid between the strain for which this amino acid is phylogenetically proximal and the strain for which it is distal. Amino acids with significant (left, p<0.01 for both strains) or insignificant (right, p>0.1 for both strains) changes in conferred fitness between strains are shown; bold line is the median, box borders represent inter quartile range (IQR) between first (Q1) and third (Q3) quartiles, whiskers are Q1-1.5*IQR and Q3 + 1.5 * IQR intervals, outliers are not shown; boxplots are made using all data; each variable amino acid close to subtype A or subtype B and a subsample of 100 other amino acids are shown as red, blue and grey dots respectively. Dashed line indicates no change in preferences (r=1). *B*, Experimentally measured fitness of each amino acid in two strains of HIV1. Grey dots, amino acids that persist at least in one strain in our phylogenetic tree; red, variable fitness amino acids

that are proximal to strain on X axis; blue, variable fitness amino acids that are proximal to strain on Y axis; dashed line - mean line.

The majority of the sites with variable fitness amino acids show no evidence of positive selection

Changes in fitness conferred by an allele can trigger episodes of positive selection. To study the association between these processes, we identified sites under positive selection, as evidenced by an above-neutral ratio of nonsynonymous to synonymous substitutions (dN/dS) estimated by codeml ($\omega > 1$), in subtype A, subtype B or both. Among the 46 sites carrying variable fitness amino acids, only 20 (43%) have experienced positive selection in subtypes A or B, despite the fact that the power of both the site-specific dN/dS test for positive selection and our test increases with the number of substitutions. An even lower overlap was observed with the sets of positively selected sites described in previous studies (Yang et al. 2003: 17%; Wood et al. 2009: 9%). Although there was an excess of positive selection among sites with variable fitness amino acids (Methods; Fisher's exact test, two-sided $p < 0.001$, Table 1), the fact that no positive selection was detected in over a half of variable sites suggests that these two phenomena are frequently uncoupled. Furthermore, r was not higher for variable fitness amino acids at positively selected sites than at sites without any signal of positive selection (one-tailed Wilcoxon's test, $p = 0.09$).

Table 1. Sites with variable fitness amino acids and positively selected sites.

	Sites with variable fitness amino acids	Sites without variable fitness amino acids	p-value*
Our analysis			
Positively selected sites	20	71	
Other sites	26	422	<0.001
(Wood et al. 2009)			
Positively selected sites	4	21	
Other sites	42	472	0.26
(Yang et al. 2003)			
Positively selected sites	8	37	
Other sites	38	456	0.04

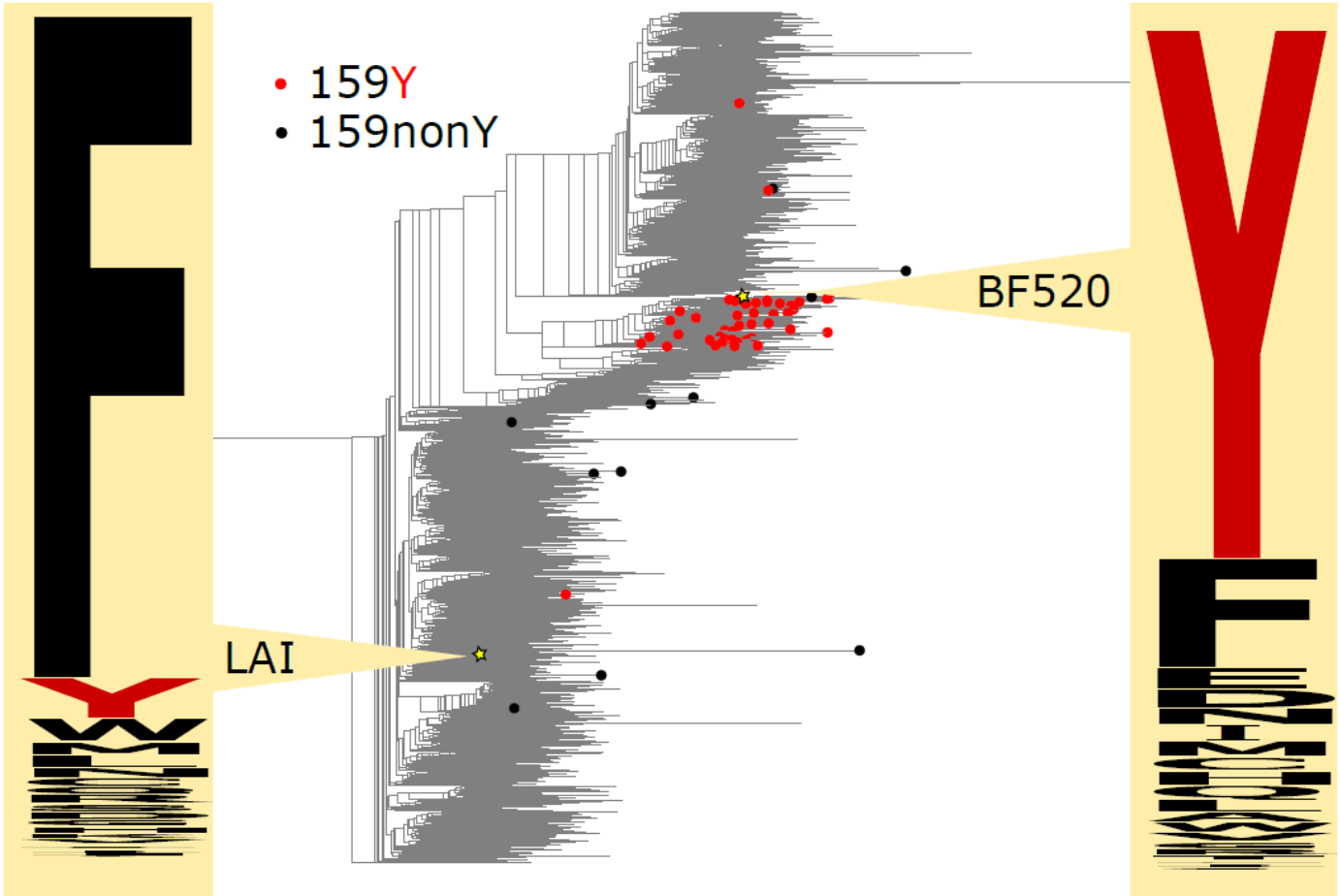
Only sites with potentially variable fitness amino acids are included. *2-sided Fisher's exact test

A closer examination of the data helps understand the interplay between the dN/dS ratio and fitness variability. For example, according to the DMS data, Y159 is strongly preferred within subtype A, while F159 is strongly preferred in subtype B. In our analysis, Y159 is a variable fitness amino acid proximal to the subtype A. At site 159, codeml has detected positive selection in subtype A ($\omega = 3.47$), but not in subtype B ($\omega = 0.10$). Within subtype A, F159Y substitutions constitute nearly all (81%) observed substitutions at site 159, so the signal of positive selection picked up by codeml has been created by them, implying recurrent adaptation (Fig. 5A).

By contrast, while S273 was designated a variable fitness amino acid in our analysis, no positive selection has been detected at site 273 by codeml. R273S substitutions constitute

only 5.7% of all substitutions at this site in subtype B, while they account for 46.5% of substitutions in the remaining subtypes (including subtype A). Consistently, S273 confers higher fitness in DMS experiments in subtype A than in subtype B. Still, even within subtype A, S273 is just one of the several variants with near-equal fitness (Fig. 5B). Therefore, the increased frequency of R273S substitutions in subtype A is due to relaxation of negative selection against 273S, rather than positive selection in its favour.

A



B

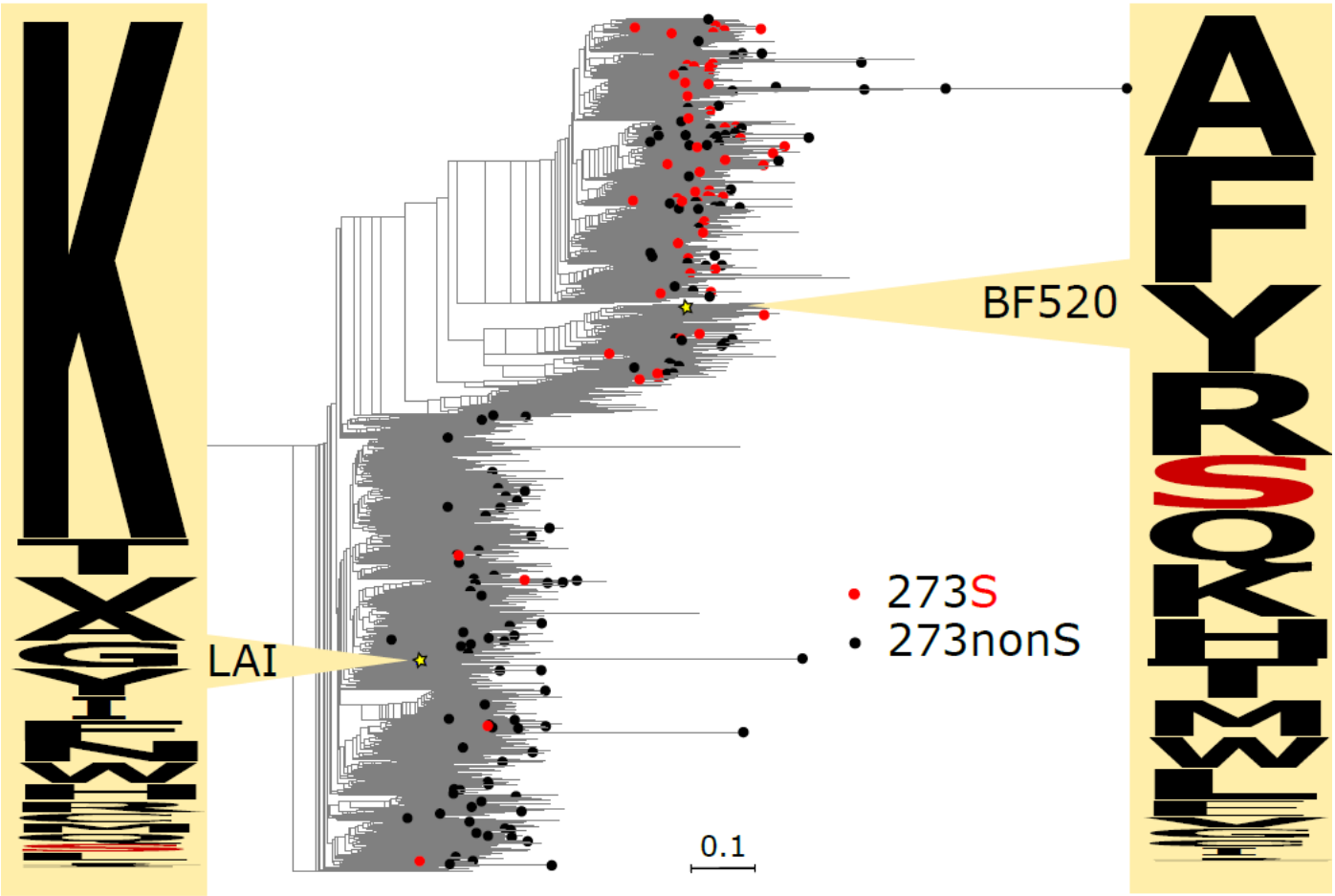


Figure 5. Substitutions to variable fitness amino acids (shown in red) and to other amino acids (shown in black) at sites 159(A) and 273(B) of gp-160. Phylogenetic distances are in numbers of amino acid substitutions per site. Numeration of protein positions is hxb2-based. Vectors of experimentally derived amino acid propensities were built with "Weblogo" online tool (https://weblogo.berkeley.edu/logo.cgi).

Differences in fitness constraints between 3 subtypes

Our approach can be extended to detect amino acids with fitness differing in a predetermined pattern between multiple subtypes. To illustrate this, we study the amino acids that change the conferred fitness between the three major subtypes of HIV-1, A, B and C, which are the most prevalent subtypes in Asia, Northern America/Europe and Africa, respectively. Besides the 59 variable fitness amino acids detected in the A-B pairwise analysis (see above), we also found 268 additional amino acids that had changed fitness between C and at least one of the subtypes A and B. These amino acids are listed in Table S1; the web script that lists the variable fitness amino acids for an arbitrarily chosen pattern of preferences between subtypes is available at http://makarich.fbb.msu.ru/galkaklink/hiv_landscape/.

Variable fitness amino acids in HA of Influenza A

We applied our method to the haemagglutinin gene of H1 and H3 subtypes of human influenza A, and detected 9 variable fitness amino acids in 7 sites (FDR = 39%; Fig. 6).

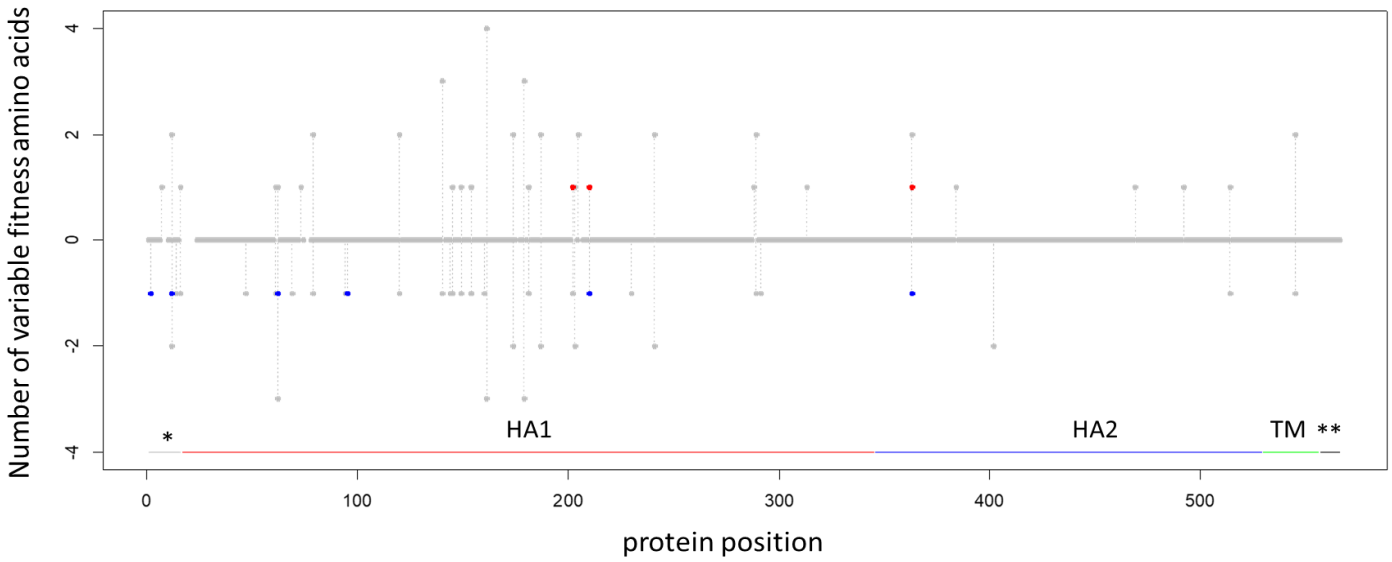


Figure 6. Amino acid sites carrying variable fitness amino acids in hemagglutinin of influenza virus. For each protein position (horizontal axis), the red dot shows the number of amino acids that are simultaneously proximal to the H1 strain and distal from the H3 strain, and the blue dot, the number of amino acids that are simultaneously proximal to the H3 strain and distal from the H1 strain, multiplied by -1. Grey dots represent the number of amino acids at a site that could potentially have significantly variable fitness (see Methods). The positions without any dots are sites filtered out due to alignment ambiguity. Vertical grey lines connect dots for the same site. Protein domains are marked by horizontal lines: *, signal peptide; HA1, HA1 ectodomain; HA2, HA2 ectodomain; TM, transmembrane domain; **, cytoplasmic tail; numeration of protein positions is A/Perth/16/2009 based.

The results for the hemagglutinin were again consistent with the DMS experiment: amino acids with an excess of substitutions into them in the phylogenetic neighborhood of one of the strains conferred higher fitness at this strain than in the other strain (one-tailed Wilcoxon's test, p = 0.02; Fig. 7).

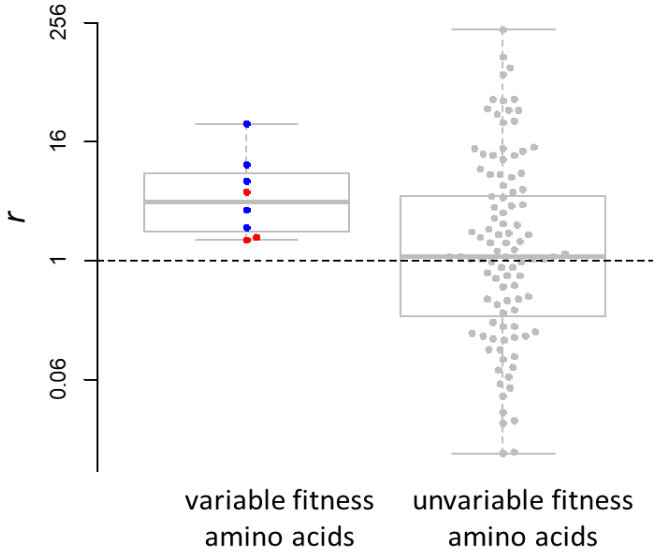


Figure 7. The ratio between experimentally measured amino acid preferences of the strain for which an amino acid is proximal and the strain for which it is distal, influenza hemagglutinin. Amino acids with significant (left, $p<0.05$ for both strains) or insignificant (right, $p>0.1$ for both strains) changes in the phylogenetic analysis are shown; notations are the same as in Figure 4.

For some of the sites, changes in fitness could be interpreted from the functional perspective. Site 202 (in A/Perth/16/2009 numbering) is one of the two sites that were previously shown to be antigenic in H3, but not in the H1 subtype (Stray and Pittman 2012). In site 202, S is the ancestral amino acid for both H1 and H3 clades. Both our analysis and the DMS experiment suggest that P202 confers higher relative fitness in H1 than in H3. Four S202P substitutions have occurred in the H1 clade, and 90% of the H1 isolates contain P202. Conversely, no substitutions into P have occurred in the H3 clade, and 95% of H3 isolates carry G202, which has originated from S in a total of 3 substitutions. The observed differences in the fitness landscapes between subtypes might reflect the differences in the function of site 202 between the two viral subtypes.

Discussion

Substitution patterns in the course of evolution of a protein are shaped by the selection pressure experienced by it, and changes in this pressure may lead to non-uniformity of substitution rates between evolving lineages. While the overall non-stationarity of evolutionary rate (heterotachy, Lopez et al. 2002; Yang and Nielsen 2002; Murrell et al. 2012) as well as of individual substitution types (heteropecilly, Tamuri et al. 2009; Roure and Philippe 2011; Klink et al. 2017; Klink and Bazykin 2017) have been used to infer the changes in underlying selection, these approaches require aggregating multiple sites to obtain the statistical signal. Therefore, they are unable to pick up changes in amino acid propensities within individual sites, especially if the directions of these changes were discordant between sites.

Large sequence datasets with inferable phylogenetic relationships can be harvested to increase the resolution of these methods. Here, we develop an approach to infer, for individual amino acids at individual sites, the changes in conferred relative fitness between two related lineages.

Our method makes use of the changes in the relative frequencies of convergent (or parallel) and divergent substitutions between evolving lineages. While the high frequency of substitutions giving rise to the same amino acid suggests a high fitness of that amino acid, usage of divergent substitutions provides an internal control, making the method robust and insensitive to other types of non-stationarity. In particular, first, it is not sensitive to differences in the overall substitution rate (across sites and mutation types) between lineages. Second, as it considers each site individually, it is not sensitive to differences in substitution rates or patterns between sites. Third, as it controls for the identity of the ancestral amino acids, it is not sensitive to differences in rates of different

types of substitutions (e.g., transitions vs. transversions) within a site that are conserved throughout the considered phylogeny. Fourth, for the same reason, it is not sensitive to changes in the rate of the substitutions of the same type that are caused by the changes in the prevalence of a particular ancestral amino acid (Klink and Bazykin 2017).

Still, our method has some limitations (Klink and Bazykin 2017). First, as it is based on the analysis of phylogenetic distribution of amino acid substitutions, its power is limited by the number of substitutions observed, which in turn depends on the tree shape and the selection regime. In particular, it is inadequate for detection of changes in amino acid fitness between closely related species or at slowly evolving sites. Here, we explicitly study the power of our method for different selection strengths. Simulations show that our method works best within a window of selection coefficients corresponding to moderate selection. This is to be expected. If selection is too weak, both weakly advantageous and weakly deleterious substitutions occur throughout the phylogeny, making our approach impotent. Conversely, if selection is very strong, the number of variable sites becomes low: the favored amino acid is usually nearly fixed in one of the clades, so that there are no substitutions giving rise to it to analyze. In particular, our method is inadequate for detection of selection in favour of a highly conserved ancestral variant, since multiple substitutions into this variant are unlikely. We can identify the amino acids with sufficient variability for our approach a priori; among those, the power of our method is high, so that just a few of the variable fitness amino acids identified here can be false positives.

Second, our test detects any change in the frequency of an individual substitution at a site independent of its cause. While at the amino acid level, such changes are probably always caused by changes in amino acid propensities, if a similar approach is applied to nucleotide substitutions, rate changes can also occur due to differences in mutation biases or in non-amino-acid level selection, e.g. selection for nucleotide composition. In theory, synonymous sites can be used to control for non-selective changes in substitution biases.

To illustrate and experimentally validate our approach, we make use of two DMS datasets for viral surface proteins, in each of which the amino acid preferences have been measured experimentally for two different strains: the gp160 protein in two HIV-1 strains, and the HA protein in two influenza A strains. Overall, we find that the results of the phylogenetic test are consistent with experimental results. If a particular amino acid is gained more frequently by a species closely related to a particular strain, it also tends to confer higher fitness in experimental assays in that same strain.

Still, the concordance between the two tests is imperfect. The observed discordances are illustrative of both the strengths and the limitations of our approach. We find that the inferred variable fitness amino acids mostly have modest experimentally measured preference π even in the strain for which they are proximal. There can be multiple reasons for this. First, our phylogenetic test is not guaranteed to assign high preference to the

fittest amino acid at a site. In particular, it is only able to assign differential fitness to derived, but not ancestral, variants. For example, 159F confers high experimentally measured fitness in the gp160 of the LAI strain, but its fitness is lower in the BF520 strain (Fig. 5a); still, it cannot be picked up by the phylogenetic test, as it is the ancestral variant, and the substitutions giving rise to it are rare in both strains (Fig. 5a). As the variant that is much more fit than all other variants is likely to be conserved throughout the clade, the phylogenetic method is less likely to pick it up, compared to a variant with fitness comparable to that of at least one other variant. Thus, π may be high at only one of the strains even for those amino acids for which no change in fitness is inferred phylogenetically.

Second, many of the fitness shifts between A and B subtypes of HIV-1 probably did not affect the fittest amino acid. Indeed, among the gp160 sites with variable fitness amino acids between LAI and BF520, only 30% carry different amino acids in these two strains, which is no more than among remaining sites (Fisher's exact test, $p = 0.08$).

It has been shown that an experimentally informed codon model (ExpCM; Hilton et al. 2017) that incorporates the results of the DMS scan into evolutionary models outperforms other substitution models, because it takes into account selective constraints at individual sites (Haddox et al. 2018; Lee et al. 2018). Thus, ExpCM was interpreted as the product of negative selection. Here, we show that the changes in the DMS matrices can also reveal episodes of positive selection. Less than a half of sites where we detected variable fitness amino acids also had signatures of positive selection that could be detected by a standard dN/dS approach. Therefore, changes in site-specific fitness landscape are not necessarily coupled with strong positive selection for new optimal variants. For example, this can be the case when the fittest amino acid remains the same in the course of evolution, but some other amino acids experience changes in their acceptability. Such changes may affect subsequent changes in the fitness landscape that can be dramatic; for example, an amino acid can become acceptable at a site due to permissive substitution at another site and then become the most prevalent under positive selection.

Our method can be applied to more than two focal species to find amino acids that have the same relative fitness in some lineages but change it in other lineages. We illustrate this using the three subtypes of HIV-1. As the number of strains substantially differs between these three clades (223, 2035 and 1265 for A, B and C correspondingly), the power of our method differs between them. Still, our results show that with enough data we are able to compare fitness of an amino acid between more than two clades. It is important to note that it is not necessary for the focal lineages to belong to well separated clades to be suitable for our test, unlike some other methods (Tamuri et al. 2009). Moreover, our method makes no assumption regarding model of epistatic interactions or fitness shift (Tamuri et al. 2009; Louie et al. 2018).

Predicting the effects of genetic variation on fitness is often based on the degree of position conservation (Sim et al. 2012; Adzhubei et al. 2013; Kircher et al. 2014), but functionally important residues are not necessarily conserved (Wu et al. 2015). Our method can be used to distinguish between rapidly evolving sites with relaxed and variable constraints. For example, in gp160 there are 10 sites with 18 different amino acids that aroused across the phylogenetic tree. Among them, we have found 5 sites with variable fitness amino acids. It possibly indicates that the fitness landscape is variable in these sites. As our current method is limited to relatively rapidly evolving sites, it cannot answer questions such as on what properties of the site the probability of switching its amino acid propensities depends most, and this is an area for future research.

Despite the recent surge in DMS experiments that allow high-throughput fitness measurements, such experiments can still only be performed for a relatively few model systems, and are still impractical for large numbers of strains. Moreover, the fitness measurements in experimental conditions are not necessarily reflective of the selective pressure experienced by a lineage over the course of its evolution. By contrast, comparative genomics allows to assess the fitness differences that have shaped evolution for a large number of non-model species simultaneously.

Materials and Methods

Inference of phylogenetic biases

We define the phylogenetic distance d between the focal species and each substitution into *A* as the sum of branch lengths along the path between them on the phylogeny, and $\bar{d}$ as the mean of those distances. If the substitutions preferentially occur in the phylogenetic neighborhood of the focal species, we expect $\bar{d}$ to be downward biased, compared to the null expectation. By contrast, if the substitutions are underrepresented in the phylogenetic neighborhood of the focal species, we expect $\bar{d}$ to be upward biased. To obtain the null distribution of $\bar{d}$,we made use of the substitutions into non-*A* amino acids. $\bar{d}$ depends both on the phylogenetic distribution of the amino acids ancestral to *A*, and on the number of substitutions into *A* (Klink and Bazykin 2017). To control for this, we subsampled the substitutions from the same ancestral amino acids as follows. We listed all "ancestral" amino acids that were at least once substituted by *A*. For each such ancestral amino acid, we then considered all substitutions into any (*A* or non-*A*) amino acid, and subsampled, out of these substitutions, the number of substitutions equal to that into *A*. We then pooled these substitutions across all ancestral amino acids together, and measured the mean phylogenetic distance $\tilde{d}$ for them. This procedure was repeated 10,000 times to obtain the null distribution of $\bar{d}$. We estimated the extent of bias of the substitutions as $E = \bar{d}/\text{mean}(\tilde{d})$; $E < 1$ if the substitutions into *A* are clustered near the focal species, and $E > 1$ if they preferentially happened far from it. The p-value for the

deviation of E from 1 was estimated as the percentile of $\bar{d}$ in this distribution, and for its bias away from the focal species, as 1 minus this value.

We required p < T for proximal amino acids, and p > (1-T) for distal amino acids, where T is a significance threshold. An amino acid was considered "potentially variable fitness" if the minimal attainable p-value (i.e., the fraction of the values of $\bar{d}$ in the most extreme bin) was lower than T. For each significance threshold T, we calculated the expected fraction of false positives (false discovery rate, FDR) as argmin ((T*PotVar)/VarFit, 1), where T is the significance threshold, PotVar is the number of potentially variable fitness amino acids at this T, and VarFit is the number of variable fitness amino acids at this T. We used T=0.01 for gp160 throughout; the percent of expected false positives at this p-value is less than 11% (Supplementary figure 2). For HA, were the substitution rate is lower, we used T 0.05, yielding the FDR<39%.

Finally, for a pair of species 1 and 2, we defined, for an amino acid A at this site, the effect size E_1/E_2, where E_1 and E_2 are the values of the bias in each species. To make the effect size of different amino acids comparable, for each amino acid, we measured its standardized score (z-score) Z using the null distribution of E_1/E_2 that was obtained in the same way as the null distribution of $\bar{d}$.

Data

gp-160

We downloaded 4241 gp160 DNA and protein sequences belonging to nine subtypes of HIV-1 from the Los Alamos HIV sequence database (http://www.hiv.lanl.gov/) using the following settings: alignment type=filtered web, organism=HIV-1/SIVcpz, region=Env, subtype=M group without recombinants, year: 2016. The subtypes were represented unequally, with the majority of sequences from subtypes B (>2000 isolates) and C (>1000 isolates). We discarded sequences with lengths not divisible by 3 and internal stop codons, leaving 3789 sequences. We then added two experimentally studied sequences LAI and BF520, and an outgroup type O (isolate_ANT70), for a total of 3792 sequences. Using Pal2Nal (Suyama et al. 2006) and Mafft (Katoh et al. 2002), we obtained amino acid alignments and codon-informed nucleotide sequence alignments. Using nucleotide alignments, we built a maximum likelihood phylogenetic tree using RAxML under the GTRGAMMA model. We then optimized branch lengths on the basis of the amino acid alignment under the PROTGAMMAGTR model (Stamatakis 2014) and reconstructed ancestral states with codeml program of the PAML package (Yang 2007).

The analysis was performed on amino acid alignments, using LAI or BF520 as focal species. To ensure that our results are robust to uncertainty of phylogenetic reconstruction, for all analyses involving one "proximal-distal" comparison, we constructed maximum likelihood phylogenetic trees on the basis of the same alignment in 10 independent runs, identified the variable fitness amino acids for each tree, and

performed downstream analyses with just those amino acids that had significant shifts in fitness in all 10 replicates.

Functional regions were defined as annotated in UniProt (The UniProt Consortium 2017; entry P04578). We obtained solvent accessibility data for the LAI strain from (Haddox et al. 2016). To define buried and exposed residues, we used the threshold relative solvent accessibility = 0.25.

Hemagglutinin

We obtained 2238 and 50109 DNA sequences of hemagglutinins from seasonal H1N1 and H3N2 influenza A subtypes correspondingly from GISAID database (Shu and McCauley 2017). We discarded sequences shorter than 1680 nucleotides, with lengths not divisible by 3 and internal stop codons, leaving 2142 sequences of H1 and 49543 sequences of H3. We then used cd-hit program (Li and Godzik 2006) to discard identical sequences from the dataset which leaved 1557 sequences of H1 an 20191 sequences of H3. Finally, we randomly chose 1557 H3 sequences to make the H1 and H3 datasets equal in size. Protein sequences were obtained from DNA sequences with BioPerl. All subsequent procedures were conducted as for the HIV-1 dataset. We obtained experimentally derived amino acid propensities for A/WSN/1933 (H1N1) and A/Perth/16/2009 (H3N2) strains from (Lee et al. 2018) .

Simulated evolution

We simulated evolution of the protein across the phylogeny of two HIV-1 subtypes, A and B. The phylogeny has been constructed as above, but only sequences from A and B subtypes were used. Simulations were performed with the Simulator of Evolution with Landscape Variation (SELVa) program (https://github.com/bazykinlab/SELVa). SELVa allows to specify the branches of the phylogenetic tree where fitness shift occurs, and the vectors of amino acid fitness before and after the shift. It simulates substitutions between amino acids according to the model described in (Yang and Nielsen 2008), and assumes that mutation probability is the same among all pairs of amino acids.

We simulated evolution of 700 sites along the phylogenetic tree with the amino acid fitness vector {X,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1}, for each value of *X* between 1.5 and 11 with a 0.5 step. In simulations with variable fitness landscape, the identity of the amino acid with fitness *X* differed between subtypes A and B. Fitness shift has occurred in the last common ancestor of subtype A or B.

For each X, we calculated three parameters: true positive rate (TPR), or sensitivity - the proportion of correctly identified actual positives; false positive rate (FPR) – the proportion of actual negatives that were incorrectly identified as positives (1-FPR is called specificity); false discovery rate (FDR) – the proportion of false positives among all positives. The results are shown on Figure 2.

Comparison with DMS data

To test whether our results match the results of the DMS experiments, for each amino acid, we measured the ratio r between the DMS preference for this amino acid in the strain for which it is phylogenetically more proximal ($??_x$), and the preference for it in the strain for which it is more distal ($??_y$) according to obtained p-values. We then compared the mean r for variable fitness and other amino acids with Wilcoxon's rank sum test.

Revealing sites under positive selection

Sites of gp160 were tested for positive selection using codeml program of the PAML package with site model. The presence of positive selection was tested by comparison of models M1a and M2a with the likelihood ratio test. Probabilities that positive selection acts at each site were calculated by BEB (Bayes Empirical Bayes) method as implemented in codeml. We separately performed the analysis for subtype A subtree, subtype B subtree and A+B tree (all trees were built as described in the Data section but only with sequences from A and B subtypes).

Acknowledgments

This work was supported by the Russian Foundation for Basic Research grant no. 18-34-00358\18. We thank Juhye Lee, Hugh Haddox and Jesse Bloom for sharing their DMS data pre-publication and for valuable comments. We also thank Alexey Neverov for helpful discussions.

References

Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. Chapter 7:Unit7.20.

Bazykin GA. 2015. Changing preferences: deformation of single position amino acid fitness landscapes and evolution of proteins. Biol. Lett. [Internet] 11:20150315. Available from: http://rsbl.royalsocietypublishing.org/lookup/doi/10.1098/rsbl.2015.0315

Bazykin GA, Dushoff J, Levin SA, Kondrashov AS. 2006. Bursts of nonsynonymous substitutions in HIV-1 evolution reveal instances of positive selection at conservative protein sites. Proc. Natl. Acad. Sci. U. S. A. 103:19396–19401.

Blackburne BP, Hay AJ, Goldstein RA. 2008. Changing selective pressure during antigenic changes in human influenza H3. PLoS Pathog. 4:e1000058.

Elena SF, Agudelo-Romero P, Lalić J. 2009. The evolution of viruses in multi-host fitness landscapes. Open Virol. J. 3:1–6.

Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. 2016. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. Mol. Biol. Evol. 33:268–280.

Goldstein RA, Pollard ST, Shah SD, Pollock DD. 2015. Nonadaptive Amino Acid Convergence Rates Decrease over Time. Mol. Biol. Evol. [Internet] 32:1373–1381. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msv041

Haddox HK, Dingens AS, Bloom JD. 2016. Experimental Estimation of the Effects of All Amino-Acid Mutations to HIV's Envelope Protein on Viral Replication in Cell Culture.Swanstrom R, editor. PLOS Pathog. [Internet] 12:e1006114. Available from: http://dx.plos.org/10.1371/journal.ppat.1006114

Haddox HK, Dingens AS, Hilton SK, Overbaugh J, Bloom JD. 2018. Mapping mutational effects along the evolutionary landscape of HIV envelope. eLife [Internet] 7. Available from: https://elifesciences.org/articles/34420

Hilton SK, Doud MB, Bloom JD. 2017. phydms: software for phylogenetic analyses informed by deep mutational scanning. PeerJ 5:e3657.

Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, Marks DS. 2017. Mutation effects predicted from sequence co-variation. Nat. Biotechnol. 35:128–135.

Jordan DM, Frangakis SG, Golzio C, Cassa CA, Kurtzberg J, Task Force for Neonatal Genomics, Davis EE, Sunyaev SR, Katsanis N. 2015. Identification of cis-suppression of human disease mutations by comparative genomics. Nature 524:225–229.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.

Kimura, M. 1983. The Neutral Theory of Molecular Evolution. Cambridge: Cambridge University Press

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46:310–315.

Klink GV, Bazykin GA. 2017. Parallel Evolution of Metazoan Mitochondrial Proteins. Genome Biol. Evol. [Internet] 9:1341–1350. Available from: https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evx025

Klink GV, Golovin AV, Bazykin GA. 2017. Substitutions into amino acids that are pathogenic in human mitochondrial proteins are more frequent in lineages closely related to human than in distant lineages. PeerJ [Internet] 5:e4143. Available from: https://peerj.com/articles/4143

Kondrashov AS, Sunyaev S, Kondrashov FA. 2002. Dobzhansky-Muller incompatibilities in protein evolution. Proc. Natl. Acad. Sci. U. S. A. 99:14878–14883.

Kryazhimskiy S, Dushoff J, Bazykin GA, Plotkin JB. 2011. Prevalence of epistasis in the evolution of influenza A surface proteins. PLoS Genet. 7:e1001301.

Lee JM, Huddleston J, Doud MB, Hooper KA, Wu NC, Bedford T, Bloom JD. 2018. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. Proc. Natl. Acad. Sci. U. S. A. 115:E8276–E8285.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics [Internet] 22:1658–1659. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl158

Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol. Biol. Evol. 19:1–7.

Louie RHY, Kaczorowski KJ, Barton JP, Chakraborty AK, McKay MR. 2018. Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. Proc. Natl. Acad. Sci. U. S. A. 115:E564–E573.

Mann JK, Barton JP, Ferguson AL, Omarjee S, Walker BD, Chakraborty A, Ndung'u T. 2014. The fitness landscape of HIV-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing. PLoS Comput. Biol. 10:e1003776.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 8:e1002764.

Naumenko SA, Kondrashov AS, Bazykin GA. 2012. Fitness conferred by replaced amino acids declines with time. Biol. Lett. [Internet] 8:825–828. Available from: http://rsbl.royalsocietypublishing.org/cgi/doi/10.1098/rsbl.2012.0356

Povolotskaya IS, Kondrashov FA. 2010. Sequence space and the ongoing expansion of the protein universe. Nature [Internet] 465:922–926. Available from: http://www.nature.com/articles/nature09105

Riesselman AJ, Ingraham JB, Marks DS. 2018. Deep generative models of genetic variation capture the effects of mutations. Nat. Methods 15:816–822.

Roure B, Philippe H. 2011. Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. BMC Evol. Biol. 11:17.

Shu Y, McCauley J. 2017. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance [Internet] 22. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=22750

Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 40:W452-457.

Soylemez O, Kondrashov FA. 2012. Estimating the rate of irreversibility in protein evolution. Genome Biol. Evol. 4:1213–1222.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics [Internet] 30:1312–1313. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu033

Storz JF. 2016. Causes of molecular convergence and parallelism in protein evolution. Nat. Rev. Genet. [Internet] 17:239–250. Available from: http://www.nature.com/doifinder/10.1038/nrg.2016.11

Stray SJ, Pittman LB. 2012. Subtype- and antigenic site-specific differences in biophysical influences on evolution of influenza virus hemagglutinin. Virol. J. 9:91.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34:W609-612.

Tamuri AU, Dos Reis M, Hay AJ, Goldstein RA. 2009. Identifying changes in selective constraints: host shifts in influenza. PLoS Comput. Biol. 5:e1000564.

The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45:D158–D169.

Wu NC, Olson CA, Du Y, Le S, Tran K, Remenyi R, Gong D, Al-Mawsawi LQ, Qi H, Wu T-T, et al. 2015. Functional Constraint Profiling of a Viral Protein Reveals Discordance of Evolutionary Conservation and Functionality. PLoS Genet. 11:e1005310.

Yang W, Bielawski JP, Yang Z. 2003. Widespread Adaptive Evolution in the Human Immunodeficiency Virus Type 1 Genome. J. Mol. Evol. [Internet] 57:212–221. Available from: http://link.springer.com/10.1007/s00239-003-2467-9

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

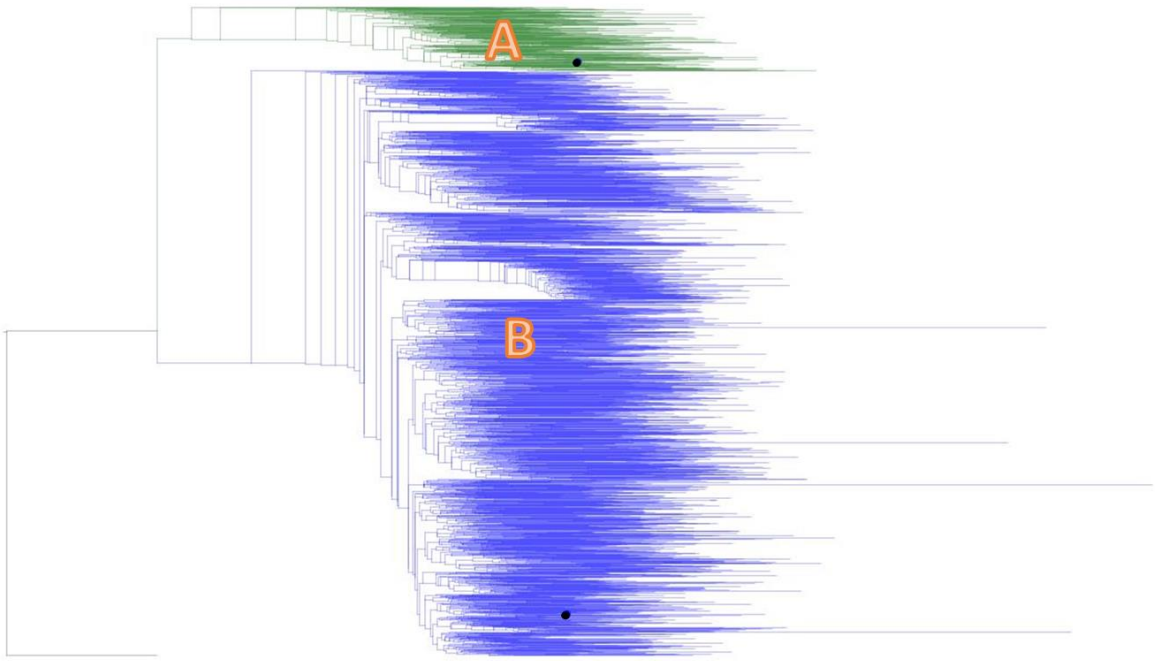
Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19:908–917.

Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol. Biol. Evol. 25:568–579.

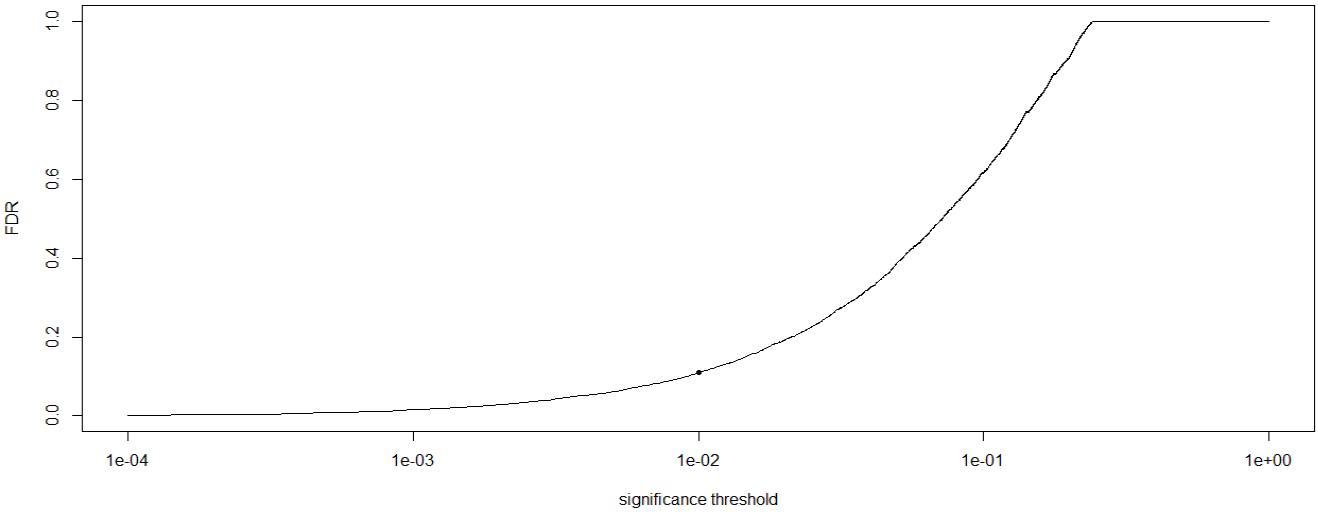
Zaidi AA, Makova KD. 2019. Investigating mitonuclear interactions in human admixed populations. Nat. Ecol. Evol. 3:213–222.

Zou Z, Zhang J. 2015. Are Convergent and Parallel Amino Acid Substitutions in Protein Evolution More Prevalent Than Neutral Expectations? Mol. Biol. Evol. [Internet] 32:2085–2096. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msv091

Supplementary materials

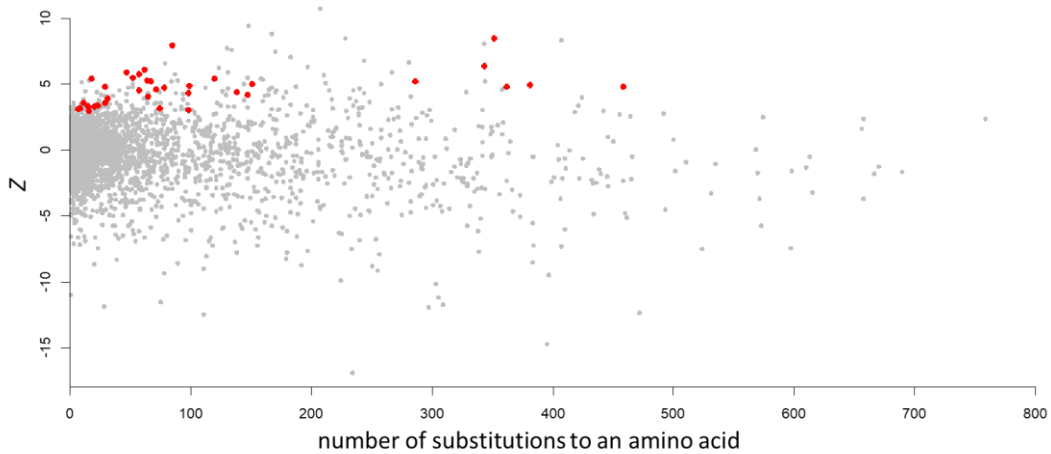


Supplementary figure 1. Phylogenetic tree of A and B HIV-1 subtypes that was used in simulation analysis and in search for positive selection. Green clade is subtype A; blue clade is subtype B; black dots are strains for that we applied our test in simulation analysis (LAI in subtype A, BF520 in subtype B).

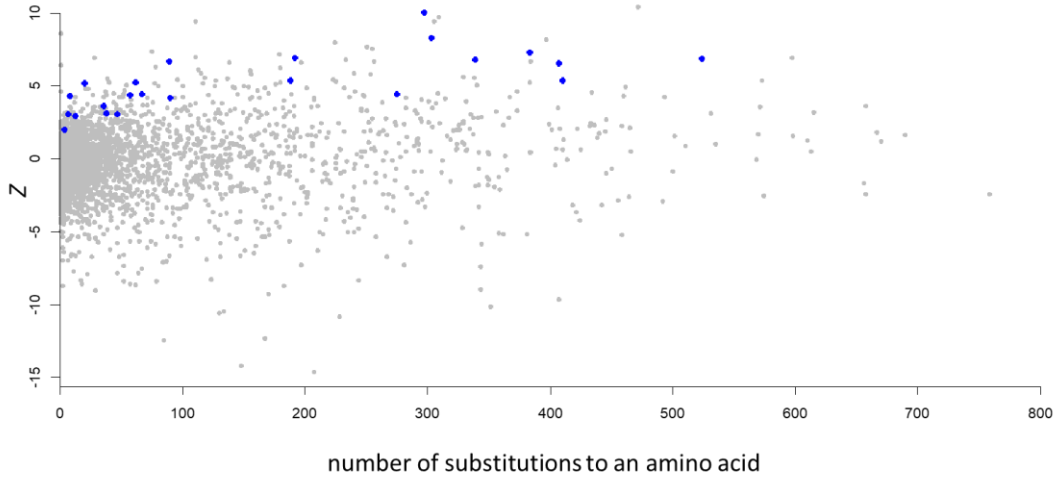


Supplementary figure 2. FDR for different significance thresholds obtained as argmin ((T*PotVar)/VarFit, 1), where T is the significance threshold, PotVar is the number of potentially variable fitness amino acids at this T, and VarFit is the number of significantly variable fitness amino acids at this T, in testing of the hypothesis that an amino acid is proximal to BF520 strain.

A



B



Supplementary figure 3. The effect size *Z* for amino acids relative to BF520 (A) or LAI (B). Red (*A*) or blue (*B*) dots – variable fitness amino acids that are proximal to BF520 (*A*) or LAI (*B*); grey dots – all other amino acids.