Virus Evolution, 2024, **10(1)**, veae036
DOI: https://doi.org/10.1093/ve/veae036
Advance Access Publication: 4 May 2024
Research Article

Changing selection on amino acid substitutions in Gag protein between major HIV-1 subtypes

Galya V. Klink,[1,2,†] Olga V. Kalinina,[3,4,5] and Georgii A. Bazykin[1,‡,*]

[1]Laboratory of Molecular Evolution, Institute for Information Transmission Problems (Kharkevich Institute) of the Russian Academy of Sciences, Bolshoy Karetny per. 19, build.1, Moscow 127051, Russia, [2]Center for Molecular and Cellular Biology, Skolkovo Institute of Science and Technology, Bolshoy Boulevard, 30, p.1, Skolkovo 121205, Russia, [3]Drug Bioinformatics, Helmholtz Institute for Pharmaceutical Research Saarland (HIPS)/Helmholtz Centre for Infection Research (HZI), Campus E8.1, Saarbrücken 66123, Germany, [4]Center for Bioinformatics, Saarland University, Campus E2.1, Saarbrücken 66123, Germany and [5]Medical Faculty, Saarland University, Kirrberger Str. 100, Homburg 66421, Germany
[†]https://orcid.org/0000-0001-8466-6958
[‡]https://orcid.org/0000-0003-2334-2751
*Corresponding author: E-mail: yegor.bazykin@gmail.com

Abstract

Amino acid preferences at a protein site depend on the role of this site in protein function and structure as well as on external constraints. All these factors can change in the course of evolution, making amino acid propensities of a site time-dependent. When viral subtypes divergently evolve in different host subpopulations, such changes may depend on genetic, medical, and sociocultural differences between these subpopulations. Here, using our previously developed phylogenetic approach, we describe sixty-nine amino acid sites of the Gag protein of human immunodeficiency virus type 1 (HIV-1) where amino acids have different impact on viral fitness in six major subtypes of the type M. These changes in preferences trigger adaptive evolution; indeed, 32 (46 per cent) of these sites experienced strong positive selection at least in one of the subtypes. At some of the sites, changes in amino acid preferences may be associated with differences in immune escape between subtypes. The prevalence of an amino acid in a protein site within a subtype is only a poor predictor for whether this amino acid is preferred in this subtype according to the phylogenetic analysis. Therefore, attempts to identify the factors of viral evolution from comparative genomics data should integrate across multiple sources of information.

Keywords: fitness landscape; changes in amino acid fitness; HIV-1 subtypes; Gag polyprotein; evolution of HIV-1; HIV-1 phylogenetic tree.

Introduction

HIV-1 is characterized by rapid evolution. It is capable of rapidly accumulating variability due to a combination of reasons, including a high mutation rate and a large population size within individuals (Fu 2001; Maldarelli et al. 2013; Cuevas et al. 2015; Zanini et al. 2015). These properties allow evolving HIV-1 populations to probe different combinations of multiple alleles, permitting it to extensively explore remote regions of its fitness landscape. An amino acid at a particular site can differentially impact viral fitness depending on which amino acids occupy other protein sites, a phenomenon known as epistasis; co-occurrence of multiple mutations within a single genome provides viruses with an opportunity to find new amino acid sequences that are similarly or even more fit under the current conditions than the wild-type variants (Rimmelzwaan et al. 2005; Kryazhimskiy et al. 2011; Ferretti et al. 2020; Zhang et al. 2020). Phenotypic diversity enabled by new combinations of mutations allows adaptation to new environments, such as acquisition of drug resistance or even expansion to a new host species (Neverov et al. 2015; Irwin et al. 2016; Biswas et al. 2019).

Fitness conferred by an amino acid at a site may differ between HIV-1 clades due to differences at epistatically interacting positions or in environmental factors. The latter include differences in prevalence of human leukocyte antigen (HLA) alleles or antiviral treatment regimes between host populations. Indeed, non-consensus amino acids at some sites may lead to escape from host-specific immune response or to resistance to antiviral drugs, while being neutral or even deleterious in the absence of these pressures (Kühnert et al. 2018; Avila-Rios et al. 2019). Differences in these factors between viral clades may therefore lead to differences in selection pressures.

In particular, such differences are likely between HIV-1 subtypes. HIV-1 subtypes of the M type can vary in severity of symptoms, response to treatment and effectiveness of immune response to infection, perhaps partially reflecting differences between subtypes in strategies of spreading through the host population (Taylor et al. 2008). Moreover, some subtypes are

more prevalent in some host populations than in others, and the differences between host populations can therefore contribute to differences in selection on the virus (McLaren and Carrington 2015).

Fitness conferred by an allele is reflected in the patterns of its evolution. First, selection favoring a variant increases the fraction of samples in which it is observed (Ferguson et al. 2013; Mann et al. 2014; Quadeer et al. 2020). In the limit, a site at which all but one of the alleles is lethal will be invariant. If this selection differs between groups, e.g. viral subtypes, different alleles may be more prevalent in different groups (Walter et al. 2009).

Second, more subtly, the fitness conferred by an allele also affects the frequency of substitutions involving it. When a variant is preferred, substitutions giving rise to it will be more frequent, and those replacing it, less frequent (Kimura 1983). If selection differs between groups, this can be observed as differences in the rates of substitutions to or from this allele between groups. Previously, we have developed a phylogenetic approach for inference of such differences in substitution rates (Klink, Kalinina, and Bazykin 2022).

Here, using the second of these two patterns, we ask whether some of the sites of the Gag protein have different amino acid preferences in different HIV-1 subtypes. Gag plays a crucial role in maturation of viral particles (Spearman 2015). It is thought to be a primary target for cellular immunity and was shown to evolve under selection pressure of cellular immune response within individual patients (Geldmacher et al. 2007; Piantadosi et al. 2009; Garcia-Knight et al. 2016). Different mutations allow escape from different HLA alleles, and in many cases, even escape from the same HLA allele is achieved by different mutations in different viral subtypes (Kinloch et al. 2019). Furthermore, Gag is considered a promising target for HIV-1 vaccines (Li et al. 2013) and antiretroviral drugs (Spearman 2015); one such drug, lenacapavir, has been recently approved by the food and drug administration (HIV-info.NIH.gov 2023). Moreover, mutations in Gag have been shown to play an important role in acquisition of drug resistance to protease inhibitors (Su, Koh, and Gan 2019). Finally, as Gag acts as an assembly machine with multiple interactions between sites of the polyprotein, epistatic interactions within it are expected to influence its evolution (Ganser-Pornillos, Yeager, and Sundquist 2008). All of this makes the existence of sites likely with between-subtype differences in amino acid preferences in this protein.

Here, we provide a list of amino acids with variable fitness across the six subtypes of HIV-1. These data can be informative in evaluating potential targets for antiviral treatments. We also describe associations between amino acid preferences and HLA allele escape in the corresponding subtypes, highlighting the biological relevance of the observed phylogenetic patterns. Finally, we show that changes in fitness are not necessarily associated with positive selection, highlighting the distinctness of these phenomena.

Results

Many sites of Gag contain variable fitness amino acids

To describe the role of changes in amino acid preferences in the evolution of Gag, for each of the six HIV-1 subtypes: A, B, C, D, F, and G, we first identified those amino acids that were gained significantly more or significantly less frequently in the lineages more closely related to this subtype than expected by chance.

For this, we use our previously developed approach, referred to as the *d*-test. It is based on the *d*-statistic, which is a measure of whether the substitutions towards a particular amino acid variant preferentially occur in the strains that are closely related to the group of interest. By comparing *d* with its expected distribution in the absence of any changes in preferences between subtypes, we can identify the amino acids that are significantly favored in some subtypes and disfavored in others (Klink, Kalinina, and Bazykin 2022; see 'Methods' section).

If an amino acid at a certain site was gained unexpectedly frequently in the strains closely related to a particular subtype (including those within this subtype), we refer to this amino acid as 'proximal' for this subtype. Conversely, an amino acid that was preferentially gained in the less related strains is referred to as 'distal' for this subtype. If an amino acid was proximal for at least one subtype, and distal for at least one other subtype, this implies that the fitness conferred by it has changed over the course of evolution; we refer to such amino acids as 'variable fitness amino acids'.

We filtered out 68 of 500 sites of Gag protein due to a high proportion of gaps in the alignment. In the remaining 432 sites, we detected 131 variable fitness amino acids, positioned at 69 sites (Fig. 1). Forty of these sites carried more than one variable fitness amino acid. To measure the strength of the bias in the *d*-statistic, we calculated the number of standard deviations between the mean of the null distribution of *d* and the observed *d* (z-score). In case of variable fitness amino acids, the best proximal z-score is the leftmost (i.e. the lowest one), and the best distal z-score is the rightmost (i.e. the highest one) among subtypes (Fig. 2). The mean proximal and distal z-scores were –4.3 and 3.8, respectively, for variable fitness amino acids, while their difference was much lower (–1.5 and 1.3, respectively) for all amino acids (Supplementary Fig. S1).

An increase in the frequency at which a variant is gained within a subtype, as reflected by the *d*-statistic, does not necessarily lead to a high prevalence of this variant within this subtype, e.g. if the gains tend to occur on terminal branches. Conversely, a low frequency of gains can be associated with high prevalence, e.g. if the site is invariant within a subtype. To better understand the link between phylogenetic preference for a variant and its prevalence, we measured the total fraction of samples carrying an amino acid in subtypes for which it is proximal and those for which it is distal.

Variable fitness amino acids had on average higher prevalence in subtypes for which they were proximal than in subtypes for which they were distal (two-sided Wilcoxon's rank sum test, $P < 2.2e\text{-}16$; Supplementary Fig. S2). To better understand this, we compared the two measures of variable fitness amino acids: the one based on frequency of substitutions (substitution-based effect size, Es = (best distal z-score—best proximal z-score)) and the one based on prevalence (prevalence-based effect size, Ep = (total proximal prevalence—total distal prevalence)). These measures correlated significantly, but the correlation was far from perfect (Spearman's test, P-value < 0.001, rho = 0.36; Fig. 3). This is as expected, because these metrics reflect two different aspects of relative fitness of an amino acid.

The variable fitness amino acid with the strongest substitution-based effect is I487 (Fig. 2). 487I substitutions occur almost exclusively (in 209 of 214 cases) in the two subtypes for which I487 is proximal, namely, A or C; and just three times in the two subtypes for which it is distal, namely, B or D. The Ep of this amino acid is also rather high, although not extreme, with 35 per cent and 20 per cent of strains in subtypes A and C but just 0.12 per cent

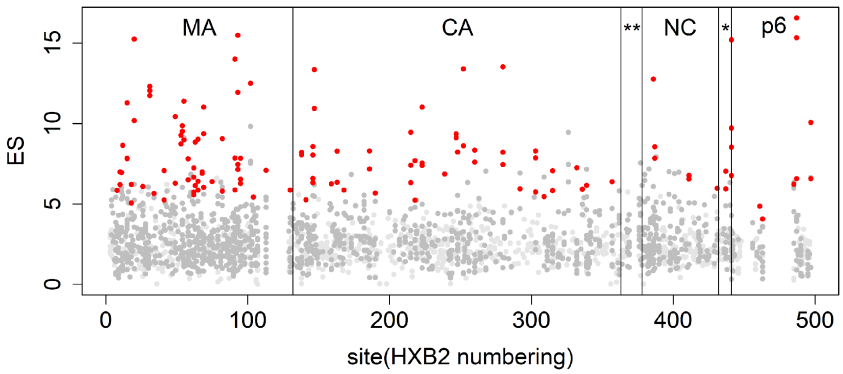

Figure 1. Variable fitness amino acids in the Gag protein. Red dots, variable fitness amino acids; dark gray dots, constant fitness amino acids; light gray dots, amino acids with insufficient number of gains for the *d*-test to be applicable (see 'Methods' section). MA, matrix protein; CA, capsid protein; **, spacer peptide 1 (sp1); NC, nucleocapsid protein; *, spacer peptide 2 (sp2); p6, p6-gag. Coordinates of peptides are taken from Su, Koh, and Gan (2019). ES (Y-axis), substitution-based effect size (best distal z-score—best proximal z-score, see text).

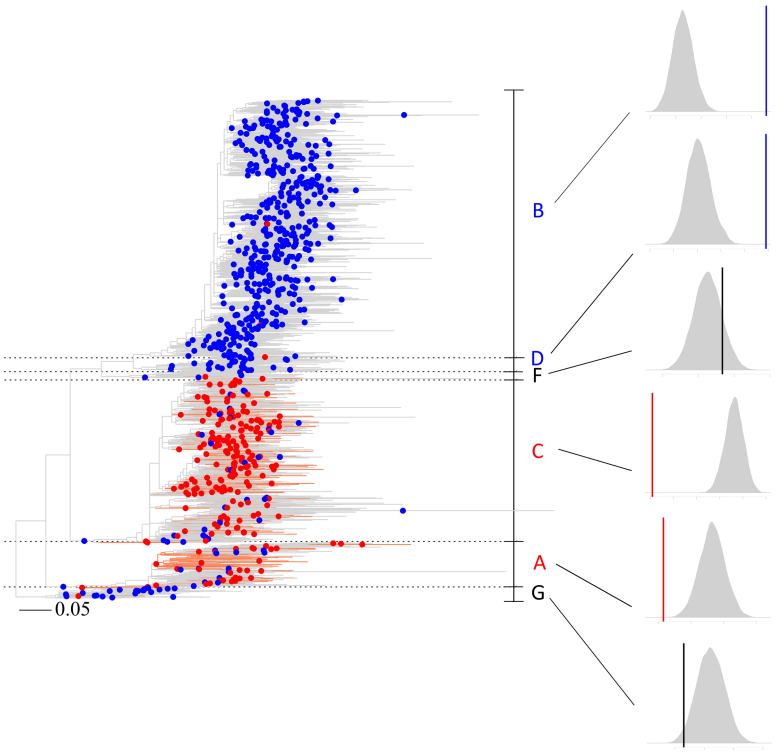

Figure 2. Detection of variable fitness amino acids from the best proximal and distal effect sizes, based on variable fitness amino acid I in site 487. The six plots represent the null distributions of *d* (mean phylogenetic distance between substitutions to an amino acid and the focal strain, see Methods section) for each subtype. Solid line represents the actual *d*, colored blue for subtypes for which amino acid I is distal, red for subtypes for which amino acid I is proximal, and black otherwise. Red circles, 487I substitutions; blue circles, 487X substitutions (where X is any non-I amino acid) from the same ancestral amino acids. Coral branches are those with descendant nodes carrying 487I. Letters mark corresponding subtypes (red, subtypes for which I is proximal; blue, subtypes for which I is distal). Branch lengths are in substitutions per site.

and 1.1 per cent in subtypes B and D carrying I487. I487 was proposed to allow escape from HLA alleles B*42:01 and C*17 in subtype C (Carlson et al. 2014). Therefore, its variable fitness between the subtypes possibly reflects differences in prevalence of the corresponding HLA alleles in human populations that are preferentially affected by different subtypes of HIV-1. Specifically, we propose that the high frequency of 487I substitutions in subtype C is due to the high frequencies of HLA alleles B*42:01 and C*17 in those populations, where subtype C is widespread.

Variable fitness amino acids with high substitution-based effect size, but relatively low prevalence-based effect size, may indicate that most substitutions to this amino acid occur on terminal or near-terminal branches of its proximal subtype(s), suggesting that this amino acid may play a more important role in within-host than in global evolution of the subtype. For example, 91Q is among the ten amino acids with the highest substitution-based effect size, but its prevalence-based effect size is very modest (Fig. 4). It is proximal for subtype A and distal for subtypes B, D, and F. All 23 substitutions 91Q in subtype A occurred on terminal branches, resulting in its 8.3 per cent prevalence, which is still higher than its prevalence in other subtypes which does not exceed 4.7 per cent (Fig. 4).

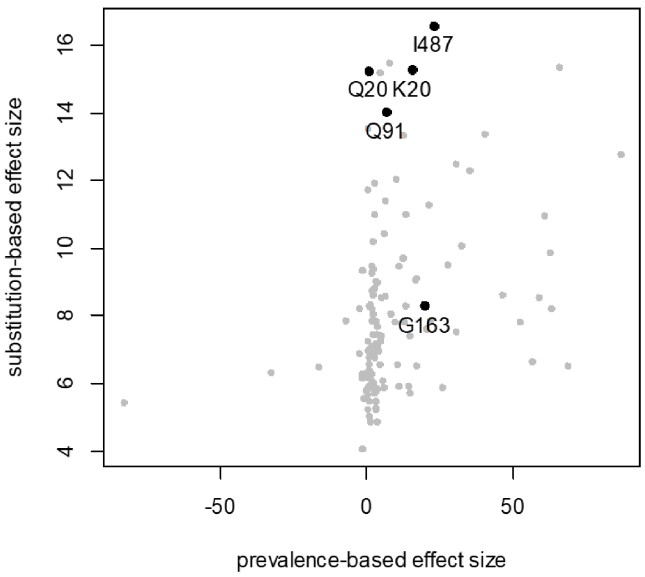

Figure 3. Prevalence-based and substitution-based effect sizes for differences in variable fitness amino acid preferences. Dots correspond to individual variable fitness amino acids at specific sites; marked dots are those discussed in the text.

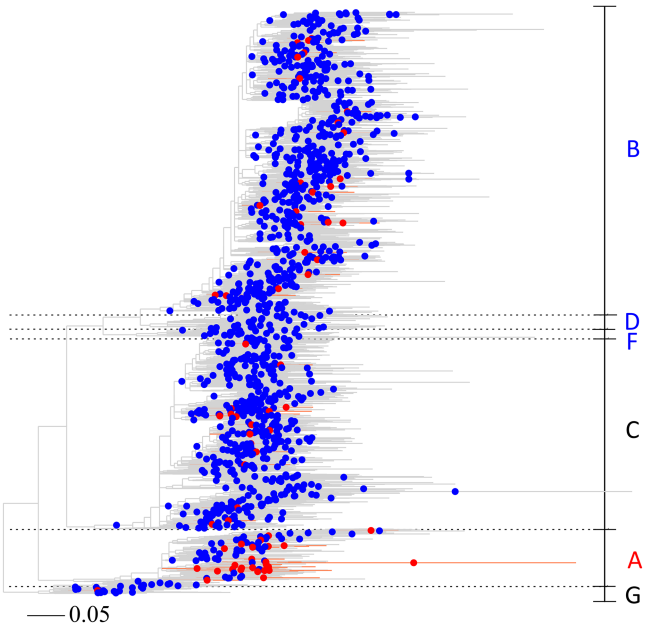

Figure 4. Site 91 of Gag. Red, 91Q substitutions; blue, 91X substitutions (where X is any non-Q amino acid) from the same ancestral amino acids. Coral, branches with descendant nodes carrying 91Q. Substitutions 91Q are concentrated in subtype A, occurring on terminal branches. Letters mark corresponding subtypes (red, subtypes for which Q is proximal; blue, subtypes for which Q is distal). Branch lengths are in substitutions per site.

Sites with variable fitness amino acids may participate in viral immune escape

Gag is a highly immunogenic protein, and the dynamics of infection is affected by the combination of viral epitopes and host HLA alleles (Kiepiela et al. 2007; Kinloch et al. 2019). Several experimental works measured HLA binding with different alleles in Gag epitopes (Tomiyama et al. 1999; Yang et al. 2003; Sanchez-Merino et al. 2008), and some Gag variants that provide HLA escape during infection were found in case studies (Feeney et al. 2004; Chopera et al. 2017). Systematically, numerous escape variants can be found by association studies that compare how frequently particular HLA and Gag alleles co-occur in the same patient. As the results of such approaches might be influenced by phylogenetic dependencies, accounting for the evolutionary relationships between studied viral sequences is important (Carlson et al. 2008). One popular method for inference of amino acids that contribute to increase or decrease in fitness of the virus in presence of particular HLA alleles is association search method based on a phylogenetic dependency network (PDN) model (Carlson et al. 2008). It was used to find amino acids that are associated with different host HLA alleles in subtypes A, B, C, and D in several human populations (Carlson et al. 2012, 2014; Chikata et al. 2014; Soto-Nava et al. 2018; Kinloch et al. 2019). To evaluate the relationship of our results with differential HLA escape, we compared our results with these findings.

Among the sixty-nine variable fitness sites, sixty-one (88 per cent) were previously described as affecting the HLA-I-binding affinity at least in one of the subtypes A, B, C, and D (Supplementary Table S1). Seventeen such sites were found as epitopic in one subtype, twenty-six in two subtypes, thirteen in three subtypes, and five in all four subtypes. Overall, there was no tendency of variable fitness amino acids that were proximal to a subtype to be associated with HLA-driven adaptations previously detected in this subtype.

We counted the variable fitness amino acids that were previously detected as adapted/non-adapted to any HLA allele in any subtype by PDN model (Carlson et al. 2012, 2014; Chikata et al. 2014; Soto-Nava et al. 2018; Kinloch et al. 2019), and were simultaneously proximal/distal for this subtype according to the *d*-test. Forty-one amino acids were simultaneously adapted and proximal; twenty, adapted and distal; twenty-six, non-adapted and proximal; and thirteen, non-adapted and distal. Thus, in fifty-four cases *d*-test matched and in fort-six cases mismatched PDN results. This is not surprising, as PDN and *d*-test answer different questions. PDN reflects associations of specific viral alleles with high or low immunogenicity, irrespective of the role of such associations in inter-patient evolution of the virus. It might be difficult for PDN to find an escape amino acid variant in a subtype if it is already too frequent, since it may frequently remain unchanged in a host with other HLA variants (Chikata et al. 2014), weakening association. On the other hand, detecting amino acids as non-adapted to a specific HLA allele might be difficult when they are too rare. In support of this, only 23 per cent of adapted amino acids represented subtype consensus, but this fraction was 64 per cent for non-adapted amino acids. Alternatively, *d*-test reflects the overall increase or decrease of relative fitness effect of an amino acid in a subtype, irrespective of the reason. Thus, the results of these methods can be considered together to better understand selective constraints of Gag.

Our results can help in extrapolation of PDN results for one subtype to other subtypes. For example, a well-known example of differential HLA escape between subtypes is epitope KF11 for HLA-B*57:03. A163G is an escape mutation in subtype C, but not in subtype B, due to higher fitness cost connected with the G163 variant in subtype B (Payne et al. 2014). According to the d-test, G163 is a variable fitness amino acid which is proximal to subtypes A and G and distal from subtypes B and D (Fig. 5). Its prevalence is 2.9 per cent and 24 per cent in G and A, but 0 per cent and 0.24 per cent in D and B. Thus, although PDN did not find 163 G as 'adapted' in subtype A, possibly due to the high prevalence of this amino acid reducing the power of PDN due to high probability of transmission

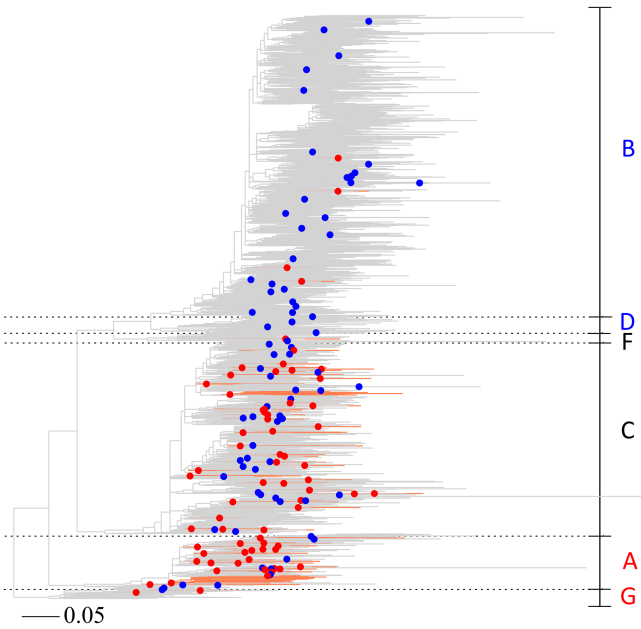

Figure 5. Site 163 of Gag. Red, A163G substitutions; blue, A163X substitutions (where X is any non-G amino acid). Coral branches are those with descendant nodes carrying G163. Letters mark corresponding subtypes (red, subtypes for which G is proximal; blue, subtypes for which G is distal). Branch lengths are in substitutions per site.

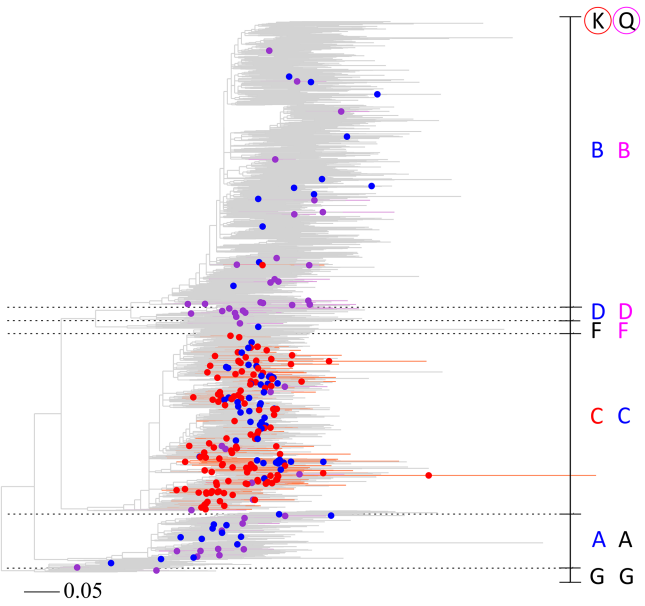

Figure 6. Substitutions from R in site 20 of Gag. Red, R20K; purple, R20Q; blue, R20X (where X is any non-K and non-Q amino acid). Coral, branches with descendant nodes carrying 20 K; purple, 20Q; grey, other. Letters mark corresponding subtypes (red/purple, subtypes for which K/Q is proximal; blue, subtypes for which K/Q is distal). Branch lengths are in substitutions per site.

to people without HLA-B*57:03 allele, this variant is likely to play an important role in adaptation to HLA-B*57:03 allele in subtype A. Major variant 163A was found as 'nonadapted' to HLA-B*57:03 in this subtype by PDN.

Matches can also be found. For example, R20K was shown to be an adaptation to A74* HLA type only in subtype C. In agreement, it is proximal to subtype C (and distal from subtypes A, B, and D) in our test. It comprises 16 per cent of sequences in this subtype and 0.059 per cent of sequences in other subtypes, and 121 of 122 R20K substitutions occurred in subtype C (Fig. 6). Meanwhile, R20Q was shown as adaptation to the same HLA type (in particular, A*74:01) in subtype D, and it is proximal to this subtype (as well as to subtypes B and F; and distal from C) according to our test. Its prevalence in subtypes B, D, and F is 1.05 per cent, 16 per cent, and 1.9 per cent, and 0.82 per cent in subtype C.

Amino acid I in position 250 is an escape mutation from HLA-B58 supertype alleles. It was shown experimentally that it reduces viral replication both in subtypes B and C, but replication capacity is more easily restored by compensatory mutations in C than in B (Chopera et al. 2012). According to our data, the prevalence of I is 0.82 per cent in subtype B and 3.4 per cent in subtype C. Due to the high conservation of the site, it contains no variable fitness amino acids with significance threshold 0.01 that is used throughout this paper. However, increasing this threshold to 0.05 makes I a variable fitness amino acid which is more fit in subtype C and less fit in subtypes B and F.

Interplay between changes in amino acid preferences and action of positive selection

We used codeml (Yang 2007) to identify sites under positive selection for each subtype separately. In total, we have found forty-five sites, 44 per cent of which were positively selected in only one of the six subtypes. Subtype B had the smallest group of positively selected sites (thirteen sites) and the smallest percentage of subtype-specific selected sites (8 per cent), while subtype A had the largest group (twenty-five sites) with the highest percentage of subtype-specific sites (24 per cent). On average, 53 per cent of positively selected sites were common between any two subtypes, and this percentage was independent of phylogenetic distance between the lowest common ancestors of two subtypes.

Among the sixty-nine sites with variable fitness amino acids, thirty-two (46 per cent) experienced positive selection at least in one subtype, including twenty-eight that experienced positive selection in subtypes from either proximal or distal groups for at least one variable fitness amino acid. This suggests intensive selection in favor of proximal amino acids or against distal amino acids in 41 per cent (28 of 69) of variable fitness sites.

Increased evolutionary rate detectable by codeml might be a signal of either directional selection to an amino acid that became more preferred in a site, or diversifying selection against the ancestral variant. To study the relationship of the fitness shifts that we have found with directional selection to particular amino acids, we performed the search for positively selected amino acids in each subtype separately with FADE method from HyPhy tool (Pond, Frost, and Muse 2005). The 52.3 per cent of variable fitness amino acids were positively selected in at least one subtype, where they were proximal ('proximal group') and were not selected in any subtype where they were distal ('distal group'). A 35.4 per cent were not selected in any group, 10 per cent were selected in at least one subtype from both groups, and 2.3 per cent were not selected in any subtypes from the proximal group, but were selected in at least one subtype from the distal group.

Thus, a considerable fraction of changes in amino acid preferences that we detected caused episodes of positive selection.

Discussion

We described differences in amino acid preferences in Gag protein between six HIV-1 subtypes. We also assessed the extent of these differences with substitution-based and prevalence-based effects

and showed that considering both these values is important as they reveal different manifestations of viral fitness. We showed that our results might add information about evolution of epitopic sites of Gag, in particular, sites 163 and 250. We also estimated that about 40 per cent of fitness variability that we detected might be associated with positive selection. To the best of our knowledge, our study is the first work that systematically finds changes in amino acid fitness in Gag polyprotein that can help understand differences in functional constraints between subtypes. All variable fitness amino acids with effect sizes and evidence of positive selection are listed in Supplementary Table S2.

Our study has caveats. One common caveat for all studies that use phylogenetic trees is possible dependence of the results on mistakes in topology and ancestral states reconstruction. The topology of deep nodes representing LCA of subtypes in our tree is consistent with the topology reconstructed using full genomes (Bletsa et al. 2019). Mistakes in topological reconstruction of multiple nodes and in ancestral states reconstruction can influence the results by creating false homoplasies.

Another thing that can influence the results is the difference in sampling strategies in different subtypes (for example, a bias between untreated and treated infections or early and late infections). In such cases, our findings may reflect differences associated not with the biology of the viral subtype but with a confounding variable. Nevertheless, such differences would represent consequences of the environment of the virus in the broad sense.

The third caveat is associated with phylogenetic interpretation of differences we detect. Although we believe that our results reflect true differences between subtypes, strictly speaking, we found amino acids with variable fitness between phylogenetic neighborhoods of particular strains that we take to estimate test statistic d, not between the subtypes themselves (see 'Methods' section). Theoretically, shifts in amino acid preferences may occur inside a subtype, especially in some HIV-1 subtypes that are known to be genetically subdivided (Désiré et al. 2018). Ironically, there are relatively few sequenced samples of these subtypes in our data. But when we applied the d-test to two distinct strains from subtype B, which is the most prevalent in our dataset (51 per cent of sequences), we have found no amino acids that change fitness between the phylogenetic neighborhoods of them.

Overall, our results were not explained by adaptation (maladaptation) to HLA in proximal (distal) subtypes. We used information about the adaptation of Gag amino acids to HLA alleles from associative studies. These studies had been performed on empirical data from particular human populations, and even significant associations that had been found may not be reflected in our phylogenetic tree which was built based on the global data. Moreover, some variable fitness amino acids were marked as 'adapted' to one HLA allele while 'non-adapted' to the other in the same subtype, suggesting that the global influence of an amino acid variant in Gag epitope on the effectiveness of host immune response may be complex.

Evidence of the role of host HLA alleles for HIV-1 within-host evolution are numerous, and there are also many suggestions about their importance in global HIV evolution (Matthews et al. 2009; Goulder and Walker 2012; Kløverpris, Leslie, and Goulder 2016). However, genome-wide association studies were able to describe less than 23 per cent of variation in viral load by host genetic polymorphisms (Carrington and Walker 2012), leaving room for a potential role of viral genetics. Moreover, the importance of HLA alleles in disease outcome and viral evolution may differ between host populations due to different availability of antiretroviral therapies (Kløverpris, Leslie, and Goulder 2016). Therefore, shifts in amino acid preferences that we have found in HLA epitopes may easily be associated with selection pressures other than HLA binding.

Antiretroviral therapies can also contribute to differences in selective constraints between subtypes. Subtypes can have different constraints in acquiring drug resistance (Soares et al. 2007). Moreover, epistasis may play an important role in the evolution of such sites as it was shown that the mutational landscape of sites involved in drug resistance cannot be explained by a site-independent model (Biswas, Haldane, and Levy 2022). Two of the 130 variable fitness amino acids, V437 and L497, were previously shown to originate in response to treatment with protease inhibitors (PIs) and to contribute to PI resistance (Myint et al. 2004). Variable fitness amino acid L in site 497 is proximal to subtype A where its prevalence is 33 per cent, and distal to subtypes B, D, and F where it does not exceed 1 per cent. Therefore, variability in preference of L497 in viral subtypes might be caused by differences in treatment strategies among countries or by differences in genomic context, but this question needs further research.

We also considered the relationship between changes in amino acid preferences and action of positive selection by identifying sites that are under positive selection at least in one of the six subtypes. Forty-six per cent of variable fitness sites experienced positive selection, and 71 per cent of positively selected sites were also variable fitness sites. Thus, positive selection and changes of amino acid preferences are likely to be associated in Gag sites. In addition, nearly half of variable fitness amino acids experienced directional selection in proximal, but not in distal, subtypes. Meanwhile, many variable fitness amino acids did not show evidence of directional selection in proximal subtypes. Thus, considering both fitness variability and action of positive selection can help in distinguishing between several scenarios of evolution of a protein site.

In making conclusions about one viral subtype on a base of other subtypes, it is important to take into account that amino acid preferences in some sites can differ between subtypes. Such variability can be caused by genetic divergence as well as by environmental differences between subtypes (such as treatment and common immune alleles in host populations). Our approach can divide protein sites into variable fitness sites and constant fitness sites (Fig. 1). We expect constant sites to behave more similarly in different subtypes, for example, in response to drugs, while variable sites are more likely to respond differently to the same treatment.

There are other methods that may help to find variability of amino acid preferences in protein sites, both experimental (Da Silva et al. 2010; Fowler and Fields 2014) and analytical (Tamuri et al. 2009; Kryazhimskiy et al. 2011; Louie et al. 2018; Laine, Karami, and Carbone 2019; Bloom and Neher 2023). All these methods have their own advantages and disadvantages, and it is important to study viral proteins with several approaches simultaneously to make the most complete picture of their fitness landscapes which may help find effective antiviral treatments.

Materials and methods

Data

We downloaded 3,596 gag DNA sequences belonging to nine subtypes of HIV-1 from the Los Alamos HIV sequence database (Kuiken, Korber, and Shafer 2003) (http://www.hiv.lanl.gov/context/index, accessed 17 January 2023) using the following

Table 1. Prevalence of HIV-1 subtypes in the dataset.

Subtype	Number of strains	Focal strain
A	303	A1.KE.95.clone_936.GQ430800
B	1710	B.US.86.AD87_ADA.AF004394
C	1104	C.BR.04.04BR038.AY727524
D	91	D.CD.85.Z2Z6_Z2_CDC_Z34.M22639
F	53	F1.ES.11.VA0053_nfl.KJ883138
G	68	G.NG.10.10NG020134.KX389619

settings: alignment type = filtered web, organism = HIV-1/SIVcpz, region = GAG, subtype = M group without recombinants, year: 2018. We discarded sequences with lengths not divisible by three and internal stop codons, as well as sequences from subtypes H, J, and K, because these subtypes were represented by only fourteen sequences in sum. This left us with 2,388 sequences from six subtypes (Table 1). We translated them using standard genetic code and obtained amino acid alignments and codon-informed nucleotide sequence alignments using Pal2Nal (Suyama et al. 2006) and Mafft (Katoh 2002). Sites with more than 1 per cent of gaps were excluded from the analysis, after that 432 of 500 Gag sites remained. We used HXB2 site numbering throughout the paper.

Using nucleotide alignments, we built a maximum likelihood phylogenetic tree using RAxML (version 8.0.0) under the GTRGAMMA model (Stamatakis 2014). The tree was rooted according to Bletsa et al. (2019). We then optimized branch lengths on the basis of the amino acid alignment under the PROTGAMMAGTR model and reconstructed ancestral states with codeml program of the PAML package (version 4.6) (Yang 2007). The analysis was performed on amino acid alignments.

Functional regions were defined as annotated in UniProt (entry P04591) (The UniProt Consortium 2019)

We have determined the structural class for each such site with respect to its position in the Gag three-dimensional structure using StructMAn (Gress et al. 2016).

Detecting variable fitness amino acids

To find amino acids in the Gag protein that change their fitness between any two of six HIV-1 subtypes A, B, C, D, F, G, we applied our previously developed method called '*d*-test' (Klink, Kalinina, and Bazykin 2022). The test is based on computing for each amino acid A the so-called *d*-statistic and comparing it with the expectation. The *d*-statistic is the mean phylogenetic distance *d* from the focal point on a phylogenetic tree to every substitution to A. The expected distribution of *d* is calculated individually for each amino acid A accounting for ancestral amino acids. If *d* is significantly smaller than it is expected, we call amino acid A 'proximal' for the focal point, and if it is larger than expected, we call it 'distal'. If amino acid A is proximal for one focal point and distal for the other one, it is considered to have higher fitness in the phylogenetic vicinity of the first than of the second point on the phylogeny.

We randomly designated one strain in each of the six subtypes as 'focal' (Table 1). For each amino acid in every site, for each focal strain we tested a hypothesis that the mean phylogenetic distance *d* between substitutions to this amino acid and the focal strain is smaller than that expected from the overall phylogenetic distribution of substitutions from the same ancestral amino acids at this site, using significance threshold of 0.01.

Identifying sites under positive selection

Sites of Gag were tested for positive selection using the codeml program of the PAML package (version 4.6) with site model (Yang 2007). We separately analyzed subtypes A, B, C, D, F, and G, using corresponding subtrees of the initial phylogenetic tree. The presence of positive selection was tested by comparison of models M7 and M8 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.

For finding evidence of directional selection, FADE option of the command line version of HyPhy (release 2.5.11) was used with default parameters (Pond, Frost, and Muse 2005). Subtrees for each subtype were analyzed separately.

Visualization

Plots were made with R language (version 4.0.3) (R Core Team 2020), using basic R and package 'ggpubr' (Kassambara 2020). Phylogenetic trees were visualized with Python package ete3 (Huerta-Cepas, Serra, and Bork 2016).

Data availability

The data underlying this article are available in the article and in its Online Supplementary Material.

Supplementary data

Supplementary data are available at *VEVOLU* online.

Funding

This work was supported by the Russian Science Foundation grant No 21–74-20160.

Conflict of interest: None declared.

References

Avila-Rios, S. et al. (2019) 'Clinical and Evolutionary Consequences of HIV Adaptation to HLA: Implications for Vaccine and Cure', *Current Opinion in HIV & AIDS*, 14: 194–204.

Biswas, A. et al. (2019) 'Epistasis and Entrenchment of Drug Resistance in HIV-1 Subtype B', *eLife*, 8: e50524.

Biswas, A., Haldane, A., and Levy, R. M. (2022) 'Limits to Detecting Epistasis in the Fitness Landscape of HIV', *PloS One*, 17: e0262314.

Bletsa, M. et al. (2019) 'Divergence Dating Using Mixed Effects Clock Modelling: An Application to HIV-1', *Virus Evolution*, 5: vez036.

Bloom, J. D., and Neher, R. A. (2023) 'Fitness Effects of Mutations to SARS-CoV-2 Proteins', *Virus Evolution*, 9: vead055.

Carlson, J. M. et al. (2008) 'Phylogenetic Dependency Networks: Inferring Patterns of CTL Escape and Codon Covariation in HIV-1 Gag', *PLoS Computational Biology*, 4: e1000225.

Carlson, J. M. et al. (2012) 'Correlates of Protective Cellular Immunity Revealed by Analysis of Population-Level Immune Escape Pathways in HIV-1', *Journal of Virology*, 86: 13202–16.

Carlson, J. M. et al. (2014) 'Selection Bias at the Heterosexual HIV-1 Transmission Bottleneck', *Science*, 345: 1254031.

Carrington, M., and Walker, B. D. (2012) 'Immunogenetics of Spontaneous Control of HIV', *Annual Review of Medicine*, 63: 131–45.

Chikata, T. et al. (2014) 'Host-Specific Adaptation of HIV-1 Subtype B in the Japanese Population', *Journal of Virology*, 88: 4764–75.

Chopera, D. R. et al. (2012) 'Intersubtype Differences in the Effect of a Rare P24 Gag Mutation on HIV-1 Replicative Fitness', *Journal of Virology*, 86: 13423–33.

Chopera, D. R. et al. (2017) 'Early Evolution of Human Leucocyte Antigen-associated Escape Mutations in Variable Gag Proteins Predicts CD4+ Decline in HIV-1 Subtype C-infected Women', *AIDS*, 31: 191–7.

Cuevas, J. M. et al. (2015) 'Extremely High Mutation Rate of HIV-1 in Vivo', *PLOS Biology*, 13: e1002251.

Da Silva, J. et al. (2010) 'Fitness Epistasis and Constraints on Adaptation in a Human Immunodeficiency Virus Type 1 Protein Region', *Genetics*, 185: 293–303.

Désiré, N. et al. (2018) 'Characterization Update of HIV-1 M Subtypes Diversity and Proposal for Subtypes A and D Sub-subtypes Reclassification', *Retrovirology*, 15: 80.

Feeney, M. E. et al. (2004) 'Immune Escape Precedes Breakthrough Human Immunodeficiency Virus Type 1 Viremia and Broadening of the Cytotoxic T-Lymphocyte Response in an HLA-B27-Positive Long-Term-Nonprogressing Child', *Journal of Virology*, 78: 8927–30.

Ferguson, A. L. et al. (2013) 'Translating HIV Sequences into Quantitative Fitness Landscapes Predicts Viral Vulnerabilities for Rational Immunogen Design', *Immunity*, 38: 606–17.

Ferretti, L. et al. (2020) 'Pervasive Within-host Recombination and Epistasis as Major Determinants of the Molecular Evolution of the Foot-and-mouth Disease Virus Capsid', *PLOS Pathogens*, 16: e1008235.

Fowler, D. M., and Fields, S. (2014) 'Deep Mutational Scanning: A New Style of Protein Science', *Nature Methods*, 11: 801–7.

Fu, Y.-X. (2001) 'Estimating Mutation Rate and Generation Time from Longitudinal Samples of DNA Sequences', *Molecular Biology and Evolution*, 18: 620–6.

Ganser-Pornillos, B. K., Yeager, M., and Sundquist, W. I. (2008) 'The Structural Biology of HIV Assembly', *Current Opinion in Structural Biology*, 18: 203–17.

Garcia-Knight, M. A. et al. (2016) 'Viral Evolution and Cytotoxic T Cell Restricted Selection in Acute Infant HIV-1 Infection', *Scientific Reports*, 6: 29536.

Geldmacher, C. et al. (2007) 'CD8 T-cell Recognition of Multiple Epitopes within Specific Gag Regions Is Associated with Maintenance of a Low Steady-state Viremia in Human Immunodeficiency Virus Type 1-seropositive Patients', *Journal of Virology*, 81: 2440–8.

Goulder, P. J. R., and Walker, B. D. (2012) 'HIV and HLA Class I: An Evolving Relationship', *Immunity*, 37: 426–40.

Gress, A. et al. (2016) 'StructMAn: Annotation of Single-nucleotide Polymorphisms in the Structural Context', *Nucleic Acids Research*, 44: W463–8.

HIVinfo.NIH.gov. (2023) *FDA-Approved HIV Medicines* <https://hivinfo.nih.gov/understanding-hiv/fact-sheets/fda-approved-hiv-medicines> Accessed 26 Dec 2023.

Huerta-Cepas, J., Serra, F., and Bork, P. (2016) 'ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data', *Molecular Biology and Evolution*, 33: 1635–8.

Irwin, K. K. et al. (2016) 'Antiviral Drug Resistance as an Adaptive Process', *Virus Evolution*, 2: vew014.

Kassambara, A. (2020) *Ggpubr: 'Ggplot2' Based Publication Ready Plots* <https://CRAN.R-project.org/package=ggpubr> Accessed 17 Jan 2023.

Katoh, K. (2002) 'MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform', *Nucleic Acids Research*, 30: 3059–66.

Kiepiela, P. et al. (2007) 'CD8+ T-cell Responses to Different HIV Proteins Have Discordant Associations with Viral Load', *Nature Medicine*, 13: 46–53.

Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge: CUP.

Kinloch, N. N. et al. (2019) 'Genotypic and Mechanistic Characterization of Subtype-Specific HIV Adaptation to Host Cellular Immunity', *Journal of Virology*, 93: e01502–18.

Klink, G. V., Kalinina, O. V., and Bazykin, G. A. (2022) 'Phylogenetic Inference of Changes in Amino Acid Propensities with Single-position Resolution', *PLOS Computational Biology*, 18: e1009878.

Kløverpris, H. N., Leslie, A., and Goulder, P. (2016) 'Role of HLA Adaptation in HIV Evolution', *Frontiers in Immunology*, 6: 665.

Kryazhimskiy, S. et al. (2011) 'Prevalence of Epistasis in the Evolution of Influenza A Surface Proteins', *PLoS Genetics*, 7: e1001301.

Kühnert, D. et al. (2018) 'Quantifying the Fitness Cost of HIV-1 Drug Resistance Mutations through Phylodynamics', *PLOS Pathogens*, 14: e1006895.

Kuiken, C., Korber, B., and Shafer, R. W. (2003) 'HIV Sequence Databases', *AIDS Reviews*, 5: 52–61.

Laine, E., Karami, Y., and Carbone, A. (2019) 'GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects', *Molecular Biology and Evolution*, 36: 2604–19.

Li, G. et al. (2013) 'Functional Conservation of HIV-1 Gag: Implications for Rational Drug Design', *Retrovirology*, 10: 126.

Louie, R. H. Y. et al. (2018) 'Fitness Landscape of the Human Immunodeficiency Virus Envelope Protein that Is Targeted by Antibodies', *Proceedings Of the National Academy Of Sciences*, 115: E564–73.

Maldarelli, F. et al. (2013) 'HIV Populations are Large and Accumulate High Genetic Diversity in a Nonlinear Fashion', *Journal of Virology*, 87: 10313–23.

Mann, J. K. et al. (2014) 'The Fitness Landscape of HIV-1 Gag: Advanced Modeling Approaches and Validation of Model Predictions by in Vitro Testing', *PLoS Computational Biology*, 10: e1003776.

Matthews, P. C. et al. (2009) 'HLA Footprints on Human Immunodeficiency Virus Type 1 are Associated with Interclade Polymorphisms and Intraclade Phylogenetic Clustering', *Journal of Virology*, 83: 4605–15.

McLaren, P. J., and Carrington, M. (2015) 'The Impact of Host Genetic Variation on Infection with HIV-1', *Nature Immunology*, 16: 577–83.

Myint, L. et al. (2004) 'Gag Non-cleavage Site Mutations Contribute to Full Recovery of Viral Fitness in Protease Inhibitor-resistant Human Immunodeficiency Virus Type 1', *Antimicrobial Agents and Chemotherapy*, 48: 444–52.

Neverov, A. D. et al. (2015) 'Coordinated Evolution of Influenza A Surface Proteins', *PLOS Genetics*, 11: e1005404.

Payne R P. et al. (2014) 'Differential Escape Patterns within the Dominant HLA-B*57:03-Restricted HIV Gag Epitope Reflect Distinct Clade-Specific Functional Constraints. *J Virol*, 88: 4668–78.

Piantadosi, A. et al. (2009) 'HIV-1 Evolution in Gag and Env Is Highly Correlated but Exhibits Different Relationships with Viral Load and the Immune Response', *AIDS London England*, 23: 579–87.

Pond, S. L. K., Frost, S. D. W., and Muse, S. V. (2005) 'HyPhy: Hypothesis Testing Using Phylogenies', *Bioinformatics*, 21: 676–9.

Quadeer, A. A. et al. (2020) 'Deconvolving Mutational Patterns of Poliovirus Outbreaks Reveals Its Intrinsic Fitness Landscape', *Nature Communications*, 11: 377.

R Core Team. (2020) *R: A Language and Environment for Statistical Computing* <https://www.R-project.org/> Accessed 17 Jan 2023.

Rimmelzwaan, G. F. et al. (2005) 'Full Restoration of Viral Fitness by Multiple Compensatory Co-mutations in the Nucleoprotein of Influenza A Virus Cytotoxic T-lymphocyte Escape Mutants', *Journal of General Virology*, 86: 1801–5.

Sanchez-Merino, V. et al. (2008) 'Identification and Characterization of HIV-1 CD8 + T Cell Escape Variants with Impaired Fitness', *The Journal of Infectious Diseases*, 197: 300–8.

Soares, E. A. J. M. et al. (2007) 'Differential Drug Resistance Acquisition in HIV-1 of Subtypes B and C', *PLoS ONE*, 2: e730.

Soto-Nava, M. et al. (2018) 'Weaker HLA Footprints on HIV in the Unique and Highly Genetically Admixed Host Population of Mexico', *Journal of Virology*, 92: e01128–17.

Spearman, P. (2015) 'HIV-1 Gag as an Antiviral Target: Development of Assembly and Maturation Inhibitors', *Current Topics in Medicinal Chemistry*, 16: 1154–66.

Stamatakis, A. (2014) 'RAxML Version 8: A Tool for Phylogenetic Analysis and Post-analysis of Large Phylogenies', *Bioinformatics*, 30: 1312–3.

Su, C.-T.-T., Koh, D. W.-S., and Gan, S. K.-E. (2019) 'Reviewing HIV-1 Gag Mutations in Protease Inhibitors Resistance: Insights for Possible Novel Gag Inhibitor Designs', *Molecules*, 24: 3243.

Suyama M, Torrents D and Bork P. (2006) 'PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments', *Nucleic Acids Research*, 34: W609–12.

Tamuri, A. U. et al. (2009) 'Identifying Changes in Selective Constraints: Host Shifts in Influenza', *PLoS Computational Biology*, 5: e1000564.

Taylor, B. S. et al. (2008) 'The Challenge of HIV-1 Subtype Diversity', *New England Journal of Medicine*, 358: 1590–602.

Tomiyama, H. et al. (1999) 'Identification of Multiple HIV-1 CTL Epitopes Presented by HLA-B*5101 Molecules', *Human Immunology*, 60: 177–86.

The UniProt Consortium. (2019) 'UniProt: A Worldwide Hub of Protein Knowledge', *Nucleic Acids Research*, 47: D506–15.

Walter, B. L. et al. (2009) 'Functional Characteristics of HIV-1 Subtype C Compatible with Increased Heterosexual Transmissibility', *AIDS*, 23: 1047–57.

Yang, Z. (2007) 'PAML 4: Phylogenetic Analysis by Maximum Likelihood', *Molecular Biology and Evolution*, 24: 1586–91.

Yang, O. O. et al. (2003) 'Determinants of HIV-1 Mutational Escape from Cytotoxic T Lymphocytes', *The Journal of Experimental Medicine*, 197: 1365–75.

Zanini, F. et al. (2015) 'Population Genomics of Intrapatient HIV-1 Evolution', *eLife*, 4: e11282.

Zhang, T. et al. (2020) 'Predominance of Positive Epistasis among Drug Resistance-associated Mutations in HIV-1 Protease', *PLoS Genetics*, 16: e1009009.

Virus Evolution, 2024, **10(1)**, veae036
DOI: https://doi.org/10.1093/ve/veae036
Advance Access Publication: 4 May 2024
Research Article