Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Evaluation of HIV-1 integrase variability by combining computational and probabilistic approaches

Davide Vergni[a,**], Daniele Santoni[b], Yagai Bouba[c,d], Saverio Lemme[d], Lavinia Fabeni[e], Luca Carioti[d], Ada Bertoli[d,f], William Gennari[g], Federica Forbici[e], Carlo Federico Perno[h], Roberta Gagliardini[i], Francesca Ceccherini-Silberstein[d], Maria Mercedes Santoro[d,*], on behalf of the HIV drug-resistance group

[a] *Istituto per le Applicazioni del Calcolo "Mauro Picone" - CNR, Rome, Italy*
[b] *Istituto di Analisi dei Sistemi ed. Informatica "Antonio Ruberti" - CNR, Rome, Italy*
[c] *Chantal BIYA International Reference Centre for research on HIV/AIDS prevention and management (CIRCB), Yaoundé, Cameroon*
[d] *Department of Experimental Medicine, University of Rome "Tor Vergata", Rome, Italy*
[e] *Laboratory of Virology, IRCCS, National Institute for Infectious Diseases "Lazzaro Spallanzani", Rome, Italy*
[f] *Laboratory of Virology, University Hospital "Tor Vergata", Rome, Italy*
[g] *Microbiology and Virology Unit, University Hospital, University of Modena and Reggio Emilia, Modena, Italy*
[h] *Multimodal Laboratory Research Department, Children Hospital Bambino Gesù, IRCCS, Rome, Italy*
[i] *HIV/AIDS Department, IRCCS, National Institute for Infectious Diseases "Lazzaro Spallanzani", Rome, Italy*

ARTICLE INFO

Keywords:
HIV-1
Integrase
Genetic variability
Bioinformatics
Mutational rate
Hellinger distance

ABSTRACT

This study aimed at updating previous data on HIV-1 integrase variability, by using effective bioinformatics methods combining different statistical instruments from simple entropy and mutation rate to more specific approaches such as Hellinger distance. A total of 2133 HIV-1 integrase sequences were analyzed in: i) 1460 samples from drug-naïve [DN] individuals; ii) 386 samples from drug-experienced but INI-naïve [IN] individuals; iii) 287 samples from INI-experienced [IE] individuals. Within the three groups, 76 amino acid positions were highly conserved ($\leq$0.2% variation, Hellinger distance: $<$0.25%), with 35 fully invariant positions; while, 80 positions were conserved ($>$0.2% to $<$1% variation, Hellinger distance: $<$1%). The H12-H16-C40-C43 and D64-D116-E152 motifs were all well conserved. Some residues were affected by dramatic changes in their mutation distributions, especially between DN and IE samples (Hellinger distance $\geq$1%). In particular, 15 positions (D6, S24, V31, S39, L74, A91, S119, T122, T124, T125, V126, K160, N222, S230, C280) showed a significant decrease of mutation rate in IN and/or IE samples compared to DN samples. Conversely, 8 positions showed significantly higher mutation rate in samples from treated individuals (IN and/or IE) compared to DN. Some of these positions, such as E92, T97, G140, Y143, Q148 and N155, were already known to be associated with resistance to integrase inhibitors; other positions including S24, M154, V165 and D270 are not yet documented to be associated with resistance. Our study confirms the high conservation of HIV-1 integrase and identified highly invariant positions using robust and innovative methods. The role of novel mutations located in the critical region of HIV-1 integrase deserves further investigation.

1. Introduction

The infection by human immunodeficiency virus (HIV) remains a major global public health issue. In recent years, the use of combined antiretroviral therapy (cART) has substantially decreased the AIDS related morbidity and mortality (de Machado et al., 2019) thanks to the constant improvement of the armamentarium of antiretroviral drugs (ARVs), which has transformed HIV/AIDS to a manageable chronic condition (Teeraananchai et al., 2017; Wandeler et al., 2016). Despite the availability of several regimens, the management of a subset of HIV

[*] Corresponding author at: Department of Experimental Medicine, University of Rome "Tor Vergata, Via Montpellier 1, Rome 00133, Italy.
[**] Corresponding author at. Istituto per le Applicazioni del Calcolo "Mauro Picone" - CNR, Rome, Italyva. Istituto per le Applicazioni del Calcolo "Mauro Picone" - CNR, Rome, Italy

E-mail addresses: davide.vergni@cnr.it (D. Vergni), maria.santoro@uniroma2.it (M.M. Santoro).

https://doi.org/10.1016/j.meegid.2022.105294
Received 16 December 2021; Received in revised form 24 March 2022; Accepted 28 April 2022
Available online 2 May 2022

infected individuals, especially those harbouring drug resistant strains and heavily-treatment-experienced individuals who have only limited treatment options, calls for the design of novel, safe and potent drugs with new mechanisms of action (Cihlar and Fordyce, 2016). Regarding this aspect, the HIV integrase represents an important target of clinical relevance for treating HIV infection and preventing evolution to AIDS (Brooks et al., 2019; Scarsi et al., 2020; Smith et al., 2021). The approval of integrase inhibitors (INIs), the last class of ARVs approved by the food and drug administration (FDA), and their introduction to clinical practice was an important event in the history of HIV treatment and has greatly strengthened cART (Brooks et al., 2019). This is because they have a remarkable efficacy and excellent safety and tolerability profiles (Brooks et al., 2019; Scarsi et al., 2020; Smith et al., 2021). So far, two waves of INIs were FDA-approved: the first generation INIs (raltegravir [RAL], elvitegravir [EVG]) and the second generation INIs (dolutegravir [DTG], bictegravir [BIC], and cabotegravir [CAB]) (Mbhele et al., 2021; Scarsi et al., 2020; Smith et al., 2021). Differently from first-generation INIs, second generation INIs show a very high genetic barrier to the development of resistance in both cART-naïve and cART-experienced individuals (Armenia et al., 2020; Marcelin et al., 2019; Mbhele et al., 2021; Smith et al., 2021; Yang et al., 2019).

The HIV-1 integrase is responsible for the chromosomal integration of newly synthesized double-stranded viral DNA into the host genomic DNA, an essential step for viral replication, enabling HIV-1 to establish a permanent genetic reservoir that can both initiate new virus production and replicate through cellular mitosis (Coffin et al., 1997; Rice et al., 1996). Following reverse transcription into the cytoplasm, within the pre-integration complex (PIC), the IN enzyme catalyzes the cleavage of two conserved nucleotides from the 3′ ends of both long terminal repeat (LTR) strands of the viral cDNA (3′ processing) (Engelman et al., 1991). After nuclear entry through the nuclear pore, the integrase catalyzes the integration of viral cDNA into the host genome (strand transfer) (Engelman et al., 1991). The integrase enzyme is a 32 kDa protein of 288 amino acids that is initially expressed and assembled into the virus particle as part of the large 160 kDa Gag–Pol precursor polyprotein, which contains other Gag (matrix, capsid, nucleocapsid and p6) and Pol [protease, reverse transcriptase and integrase] components (Swanstrom and Wills, 1997).

Looking at the HIV-1 integrase structure, it has three distinct domains, each playing a specific role (Chiu and Davies, 2005). The N-terminal domain (NTD) (residues 1–50) is highly conserved and contains a histidine-histidine cysteine-cysteine (H12-H16-C40-C43) motif coordinating the zinc binding and promotes protein multimerisation; the catalytic core domain (CCD) (residues 51–212) contains the catalytic triad (D64-D116-E152) and any mutation in these three positions leads to an abortive infection; and lastly the C-terminal domain (CTD) (residues 213–288) involved in DNA binding, is the least conserved of the three domains (Chiu and Davies, 2005). The reduction of INI susceptibility mainly occurs through the emergence of resistance mutations in the CCD or in the CTD (Collier et al., 2019; Jóźwik et al., 2020). In this regard, mutations at amino acidic positions 148 and 263 which enhance the viral DNA binding represent for example the main pathways to the development of resistance to second generation INIs (Chiu and Davies, 2005; Collier et al., 2019; Jóźwik et al., 2020; Mbhele et al., 2021; Smith et al., 2021; Thierry et al., 2017).

Another important aspect is the natural HIV integrase genetic variability. A study (Rhee et al., 2008) showed that polymorphism rates equal or above 0.5% were found for 34% of the CCD, 42% of the CTD and 50% of the NTD. Moreover, it has been previously documented that primary and secondary integrase associated mutations are generally absent or extremely rare in both cART-naïve individuals and cART-experienced INI-naïve individuals (Ceccherini-Silberstein et al., 2007, 2010; Rhee et al., 2008; Varghese et al., 2010).

The study of mutational landscape is essential for a better comprehension of the virus's genetic variability, in particular, the mechanisms that are at the basis of drug resistance. Simple statistics reporting the mutation rate for each amino acid position in a given data set were used to achieve this task, since the period when viral genomic sequences were made available (Coffin, 1995; Luciw et al., 1987). Novel instruments coming from information theory such as Shannon entropy were also used to study DNA/RNA sequences (Adami, 2004; Ohya and Sato, 2000); in order to take into consideration not only the overall fraction of amino acids, which are different from reference amino acid, but also how mutated residues are distributed (Lima de Lima et al., 2018; de Machado et al., 2019), Rhee and colleagues showed that integrase displayed a significantly decreased inter- and intra-subtype diversity and a lower Shannon's entropy than HIV-1 protease or reverse transcriptase (Li and De Clercq, 2016; Rhee et al., 2008).

In this study, we aimed at updating previous data on HIV-1 integrase variability in a large group of samples from drug-naïve and drug-experienced (both INI-naïve and INI-treated) individuals, all infected by HIV-1 B subtype, by using effective bioinformatics methods combining different statistical instruments from simple entropy and mutation rate to more specific approaches such as Hellinger distance, in order to evaluate differences between residue distributions in the different samples. In particular, we provided insights on the molecular response of HIV-1 in terms of differential mutational events occurring in treated and untreated HIV-1 infected individuals. The reliability of the analysis was supported by a non-parametric statistical test.

2. Materials and methods

2.1. Dataset

This study included 2133 HIV-1 integrase sequences obtained for clinical purposes over the period August 2004–October 2019 period. Genotyping was performed on plasma samples from HIV-1 B subtype-infected patients by using the ViroSeq HIV-1 Integrase Genotyping System (Celera Diagnostics, Alameda, CA, USA) or an in-house assay, as previously described (Armenia et al., 2014). Sequences having a mixture of wild type and mutant residues at single positions were considered to have the mutant(s) at that position. Individuals included in the study were followed in various clinical centers in Central and North Italy; 1460 were drug-naïve [DN], 386 drug-experienced but INI-naïve [IN], and 287 INI-experienced [IE]. Regarding this last group of individuals, overall, they were exposed to an average of one INI (262 with one INI; 22 with two INIs; 3 with three INIs, (median [interquartile range, IQR] exposure 26 [13–53] months). In detail, 217 (75.6%) individuals were exposed to RAL (median [IQR] exposure: 28 [12–56] months), 53 (18.5%) individuals were exposed to DTG (median [IQR] exposure: 14 [10–25] months), and 44 (15.3%) to EVG (median [IQR] exposure 15 [6–24] months).

2.2. Ethics

All data used in the study were previously anonymized, according to the requirements set by the EU Regulation 2016/679 and by the Italian Data Protection Code. The research was conducted on anonymous samples in accordance with the principles of the Declaration of Helsinki and the Italian Ministry of Health. All information, including sequences, virological and therapeutic data, was recorded in an anonymized database.

2.3. Sequences and mutation indices in HIV-1 integrase

As reported above, the dataset under investigation was composed of three different groups of HIV-1 integrase subtype B sequences, related to samples from DN, IN and IE individuals. We defined as A the set of the 20 canonical amino acids (alphabet) and as $S_{\mathrm{DN}} = \{s^{i}_{\mathrm{DN}}\}$ the set of integrase sequences from DN samples, with $\mathrm{i} = 1, 2, \ldots, \mathrm{N}_{\mathrm{DN}}$, and N_{DN} the sample size of DN. s^{i}_{DN} indicated the *i*-th sequence of the sample, $s^{i}_{\mathrm{DN}} = \mathrm{a}^{i_1}a^{i_2}\ldots a^{i_j}\ldots a^{i_n}$, where $a^{i_j} \in A$ is the amino acid at the $\mathrm{j} - \mathrm{th}$ position of the i-th sample sequence and n = 288 was the number of amino acids of the integrase. In the same way we defined S_{IE} for the IE samples and S_{IN} for the IN samples. In the following, for the introduction of the statistical indicators, we omitted, when not needed, to specify the reference sample (whether DN, IN or IE).

For every amino acid position of the sequence, j, we defined the distribution frequency $p_j(a)$ for amino acid a as the ratio between the number of occurrences of a in position j and the total number of sequences in the considered dataset. Based on $p_j(a)$ it was possible to introduce two different indicators of the degree of mutability of each residue at position j.

The first was simply based on the rate of mutation with respect to the reference

$$M(j) = \sum_{a \neq a_{\mathrm{rj}}} p_j(a) = 1 - p_j\left(a_{\mathrm{rj}}\right)$$

where a_{rj} indicates the reference amino acid for the integrase in the position j, and $M(j)$ measures the percentage of mutations that can be found at position j. To summarize the mutability characteristics of a residue, we introduced mutability levels as follows:

L = 0 if $M < 0.01$ ————— Not mutable or very poorly mutable.
L = 1 if $0.01 \leq M < 0.05$ ————— Poorly mutable.
L = 2 if $0.05 \leq M < 0.2$ ————— Mutable.
L = 3 if $M \geq 0.2$ ————— Highly mutable.

Using $p_j(a)$ another interesting quantity can be introduced, i.e. the number of different amino acids that can be found at position j,

$$N(j) = \sum_{a \in A} \theta\left(p_j(a)\right)$$

where $\Theta(p_j(a)) = 1$ when $p_j(a) > 0$ and $\Theta(p_j(a)) = 0$ when $p_j(a) = 0$, i.e. $\Theta(p_j(a)) = 1$ if and only if there exists at least one sequence in the sample in which at position j the amino acid a is present. Values of N(j) greater than 1 mean that residue j has mutations. The greater the value of N(j) the greater the number of occurring amino acid mutations. Although they may seem similar, the M and N measures are not so closely linked because a site j might present a high percentage of mutations, i.e. a high M(j), but the mutations might be associated with only one amino acid, i.e. N(j) = 2. On the other hand, a position could present many different amino acids, and therefore a large N(j), but with a very low frequency, and therefore a not too large M(j).

The second measure of mutability was based on the entropy of the frequency distribution of the amino acids at position j and can be defined as:

$$E(j) = -\sum_{a \in A} p_j(a) log_{20}\left(p_j(a)\right)$$

The entropy values, between 0 and 1, measure the degree of randomness of the frequency distribution $p_j(a)$: if the distribution is uniform, i.e. if each amino acid occurs with the same probability (therefore the residue at position j is extremely mutable) the entropy is equal to 1. The entropy value decreases as the distribution becomes more and more heterogeneous until it reaches 0 in the limiting case in which a single amino acid is present in that position, i.e. there are no mutations for the residue j.

2.4. Hellinger distance

To assess differences of mutability for a given residue with respect to the three different sample groups, the distribution frequency, $p_j(a)$, for amino acids a in residue j, were exploited. In particular, in order to evaluate the differential mutability profile with respect to two different samples, one of the various distance measures between probability distributions could be used. One of most used distances between probability distributions is the Hellinger distance defined as

$$H_j(\mathrm{S}, \mathrm{R}) = \frac{1}{\sqrt{2}} \sqrt{\sum_{a \in A} \left(\sqrt{p_j(\mathrm{a}; \mathrm{S})} - \sqrt{p_j(\mathrm{a}; \mathrm{R})} \right)^2}.$$

The Hellinger distance, with values between 0 and 1, assigns the maximum value 1 when the two distributions have no amino acids in common and assigns the value 0 when the two distributions are identical. The values of Hellinger distance, $H_j(\mathrm{S}, \mathrm{R})$, can be interpreted as the percentage difference between the two distributions of amino acids at position j related to samples S and R and it was chosen because it highlights both changes leading to new mutations and to the absence of previously occurring mutations, which is very important from the drug resistance perspective.

2.5. Statistical test

In order to assess whether the observed differences in the amino acid distributions for a given residue in different samples (for example the difference in the amino acid distribution at position 151 in DN and IE) are real biological facts or simple random fluctuations, a solid non parametric test, such as the Mann-Whitney test is needed. Unfortunately, given the considerable difference in sample sizes, and given the presence of multiple events associated with the quasi-species, the reliability of the Mann-Whitney test, especially in the case of quite similar distributions of small samples could not be guaranteed.

Therefore, we have chosen to implement an empirical statistical technique able to compute a p-value by measuring the reliability of the values obtained for quantities of interest, by randomly shuffling the original amino acid sets. For example, let us consider the Hellinger distance associated with a certain residue position, j, between DN and IE. The goal is to assess whether the distance obtained for the original samples, $H_j(\mathrm{DN}, \mathrm{IE})$, is a reliable value or can be considered a random fluctuation. In this case the null hypothesis, H_0, is that the value of $H_j(\mathrm{DN}, \mathrm{IE})$ is due to a random fluctuation. In order to measure the p-value for H_0, the two samples associated with the position j of DN and IE are merged and a joined set, O_j, is obtained. The probability distribution of a random Hellinger distance, associated to the shuffling of the original sequences, can be obtained by randomly extracting, many times, two samples of amino acids, DN^R and IE^R, from the joined set, O_j, each of which with the same number of elements of the original samples, and computing the distance between the random amino acids distribution, $H_j(\mathrm{DN}^R, \mathrm{IE}^R)$. By doing the above procedure repeatedly, the probability distribution of the random Hellinger distance, $P(H_j(\mathrm{DN}^R, \mathrm{IE}^R))$, can be computed and the p-value for H_0 can be obtained by looking at the quantile of the non-random value $H_j(\mathrm{DN}, \mathrm{IE})$, or by summing up the right tail of the probability of $P(H_j(\mathrm{DN}^R, \mathrm{IE}^R))$ that exceeds the non-random value $H_j(\mathrm{DN}, \mathrm{IE})$. In other terms, the probability that the random Hellinger distance exceeds the non-random value gives the p-value of $H_j(\mathrm{DN}, \mathrm{IE})$. P values <0.05 obtained by using this empirical statistical technique can be considered as significant.

3. Results

The mutational landscape of HIV-1 integrase was investigated comparing mutation events in the three analyzed datasets, DN, IN and IE. Firstly, the three sample sets were analyzed in terms of number of different residues and percentage of mutated residues in different positions, providing a global view of the mutational landscape of the three datasets. Secondly, the entropy and the mutation level of amino acid for each position were analyzed in order to investigate which positions had no mutations between the three samples and which were subject to drug resistance mutation. Finally, significant amino acid positions were selected and analyzed clustering them into three classes according to their behavior (positions conserved in all datasets, positions with higher wild type frequency in samples from treated individuals with respect to those from drug-naïve individuals, and positions with lower wild type frequency in samples from treated individuals with respect to those from drug-naïve individuals).

3.1. Global mutational landscape and comparison between current and expected distributions

Besides the percentage of mutated amino acids, we evaluated the amount and the type, other than wild type, of amino acids occurring in our datasets. Details about statistical and mathematical measures taken into account are reported in the Section 2.3.

In Fig. 1 the sample probability distribution, $P(N)$, of the number of different amino acids that can occur in each residue position, all along the integrase protein, is shown. A significantly different behavior between DN, showing a larger number of mutations, and both IN and IE samples, has been revealed. In fact, apart from $N = 1$ (no mutations, i.e. wild type) and $N = 2$ (associated to residues with only one possible amino acid mutation), the curve associated to DN is always above the other curves, i.e., a greater number of mutated amino acid for DN sample are allowed. In particular, for $\mathrm{N} = 1$ the percentage for DN is around 18% while it is around 32% for both IN and IE, so that IN and IE show a higher number of positions where the amino acid associated with the wild type is always present with respect to DN. Moreover, for $N \geq 3$ the number of different amino acids occurring in mutated positions of DN is higher than the same number for IN and IE. so, once again, DN shows more mutations. It can be reasonably hypothesized, as discussed in the next sections, that this behavior can be associated with the pressure of drugs imposing a constraint on mutation events. In the inset, the same plot is shown in logarithmic scale also reporting the Poisson approximation (obtained by assuming independence of each individual amino acid mutation, with a fitted average rate of mutation for each residue of $\mathrm{l} = 2.24$ for DN and $\mathrm{l} = 1.42$ for both IN and IE). As can be observed, the independent approximation fits a limited number of amino acid positions, those with the lowest number of amino acids mutations; however, the analysis provides a quantitative measure of the higher mutability of DN.

The percentage of mutated residues with respect to the wild type for each residue was also investigated through mutation levels, $\mathrm{L} = 0\ldots3$, and the related sample distributions are reported in Fig. 2. DN samples showed both a large number of not mutable or very poorly mutable ($\mathrm{L} = 0$) and highly mutable ($\mathrm{L} = 3$) residues with respect to the other two sample sets, while both IN and IE showed a large number of poorly mutable ($\mathrm{L} = 1$) residues, and only IE showed a large number of mutable residues ($\mathrm{L} = 2$) with respect to DN. Again, the percentage (i.e., fraction of mutated sequences for a given position) averaged out over all the position in the DN sequences (3.7%) was always higher than in IN (3.0%) and in IE (3.2%), meaning that in the DN sample there are many poorly ($\mathrm{L} = 0$) and highly ($\mathrm{L} = 3$) mutated sites with respect to that in the IN and IE samples.

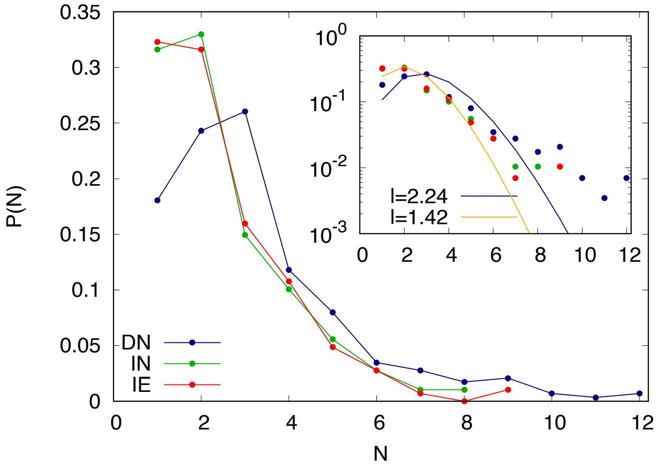

Fig. 1. Number of different residues in the HIV-1 integrase occurring per positions. The sample probability distribution of positions ($P(N)$ - y-axis) for a given number of different amino acids (N- x-axis) is reported for the three data sets analyzed (drug-naïve [DN] – blue points, drug-experienced but INI-naïve [IN] – green points, INI-experienced [IE] – red points). $\mathrm{N} = 1$ means no mutations, while greater values of N are associated to the presence of different amino acids mutation for a given position. In the inset, the same plots are reported in log scale (y-axis) with superimposed (continuous lines) the Poisson approximation for DN (blue line) and for both IN and IE (yellow line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The inset in Fig. 2 reports, in a logarithmic scale, the sample probability distribution for each percentage of mutation rate. Unlike the previous analysis, related to the number of different amino acids occurring per positions, the three sample datasets showed a very similar shape compatible with a Pareto heavy tail distribution with exponent − 1. Therefore, although the number of residues belonging to the various mutability classes is different in the various samples, the probability distributions have a similar heavy tail behavior that is usually associated to phenomena in which a typical representative value is not observed, i. e. a large portion of residues present few mutations while a small portion of residues has a very large number of mutations.

3.2. Local mutational landscape: mutation rate and entropy of single residue positions

In the previous paragraphs, we have shown the statistics of mutability properties of the integrase protein, here we begin to study the mutability characteristics of single residues. The percentage of mutated amino acids M(j) for each residue position for the three datasets was computed and reported in Fig. 3. The behavior of the three datasets seems to be quite similar, but, as shown in the inset, a zoom in a smaller region reveals that in some positions the percentage of mutation can be very different (see for example the region near residue 147 and near residue 153).

Similarly, Fig. 4 shows the entropy values E(j), measured for each residue position, j, for the three datasets. Even, in this case the behavior of the entropy for each residue seems quite similar among the three samples, but, as shown in the inset, a zoom in a central region of the

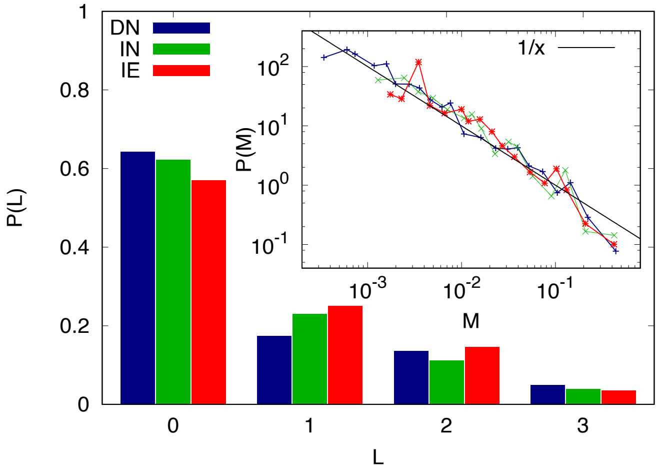

Fig. 2. HIV-1 integrase mutation levels distribution. The percentage of residues ($P(L)$ - y-axis) related to a given mutation level (L - x-axis) is reported for the three datasets analyzed (drug-naïve [DN], drug-experienced but INI-naïve [IN], INI-experienced [IE] samples). In the inset the sample probability distribution ($P(M)$ - y-axis) of the mutation rate (M- x-axis) in log-log scale is shown. The black line (1/x) represents the Pareto distribution, with exponent − 1, fitting the data.

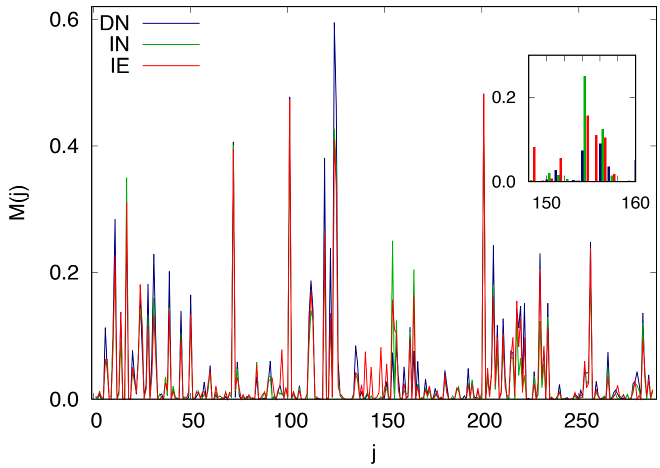

Fig. 3. Mutation rate for each position of HIV-1 integrase. Overview of the mutational landscape measured via the Mutation rate - M(j) - (y-axis), reported for each position (x-axis) for the three datasets analyzed (drug-naïve [DN], drug-experienced but INI-naïve [IN], INI-experienced [IE] samples). In the inset a zoom (from position 148 to position 160) of the same plot is shown.

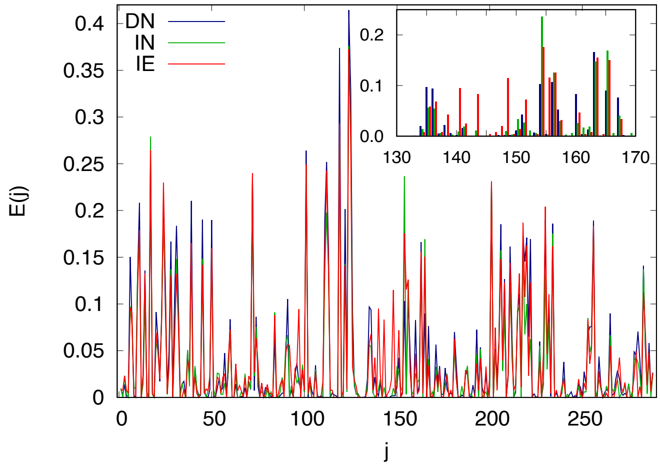

Fig. 4. Entropy for each position of HIV-1 integrase. Overview of the mutational landscape measured via the Entropy - E(j) - (y-axis), reported for each position (x-axis) for the three datasets analyzed (drug-naïve [DN], drug-experienced but INI-naïve [IN], INI-experienced [IE] samples). In the inset a zoom of the same plot (from position 130 to position 170) is shown.

protein reveals that in some positions the percentage of mutations can be very different (also in this case the most interesting regions are those near residue 147 and near residue 153).

Figs. 3 and 4 are visual overviews of the mutational landscape (measured with percentage of mutation and entropies) of the various residues in the HIV-1 integrase in the three different samples. They are not supposed to be quantitative but just to give an immediate view of which positions are the most mutable and which are the most conserved, highlighting differences and overlapping in the mutational landscape of the three samples.

In order to quantitatively highlight the most interesting regions and identify those residue positions that undergo the greatest selective pressure by the drugs, we used the previously introduced mutation level and the Hellinger distance, which is a comprehensive measure of difference between different distributions. In particular, from a biological point of view, it is interesting to look for those differences in mutation between the different samples that have a more dramatic impact on specific residues, i.e. it is interesting to find those residues which, under the pressure of the drug, are forced to remain fixed or, vice versa, are forced to mutate. In Figs. 5, 6 and 7 the Hellinger distance (see Section 2.4) for each residue in the different samples are displayed, highlighting those positions for which there is a difference in the mutation level (see Section 2.3).

Some residues clearly appeared to be affected by a dramatic change in their mutation distributions. In particular, the most notable differences were related to IE samples, in particular the differences in the mutation distributions were wider between the samples of DN and IE, Fig. 6, specifically in the position 97, 140, 143, 148 and 155, passing from a mutability level L = 0 (less than 1%) to a mutability level L = 2 (between 5 and 20%). The graphical representation allows an immediate overview of the differential mutation profile of the different samples. In the next section we will present a detailed analysis of the individual positions, again based on the mutability percentage and the Hellinger distance, which allows us to highlight the most relevant positions associated to the differential mutation percentages in the different samples.

3.3. *Analysis of significant amino acid positions according to their mutations in the three datasets*

By analyzing the differential mutation percentages in the various sequences, significant amino acid positions were partitioned into three different classes according to their mutations among the three groups: i) the first class made of 156 amino acid positions whose residues were well-conserved among all three groups; ii) the second class made of 15 amino acid positions with fewer mutations in samples from treated individuals (IN and IE) with respect to those from DN individuals; iii) the third class made of 8 amino acid positions with a higher number of mutations in samples from treated individuals (IN and IE) with respect to those from DN individuals. The remaining 109 amino acid positions were associated to residues whose mutation level is not relevant or significantly different among the three groups of samples analyzed and can be associated to simple random mutations due to the high mutability of the HIV-1.

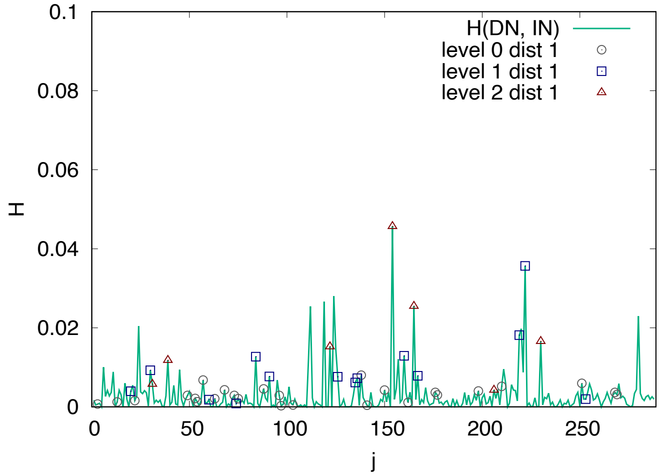

Fig. 5. Hellinger distance for each HIV-1 integrase amino acid position between drug-naïve (DN) and INI-naïve (IN) groups. Hellinger distance between DN and IN (y-axis) is reported for each position (x-axis). Positions showing different mutation levels in the two datasets are marked with circles (passing from mutability level L = 0, i.e. less than 1%, to mutability level L = 1, between 1 and 5%), squares (passing from mutability level L = 1 to a mutability level L = 0 or L = 2, i.e. between 5 and 20%) or triangles (passing from mutability level L = 2 to a mutability level L = 1 or L = 3, i.e. above 20%) according to the top right legend.

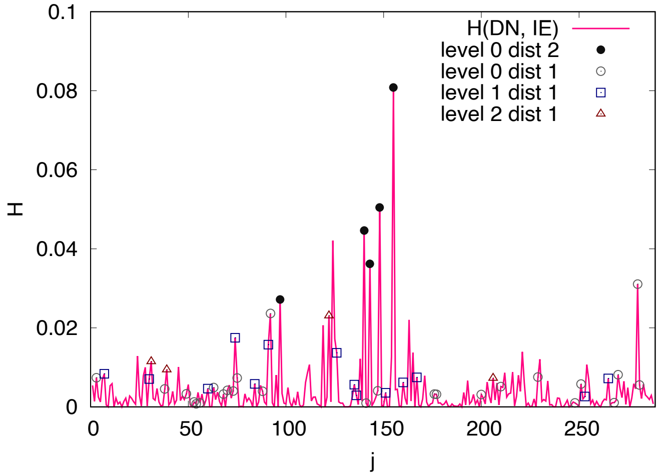

Fig. 6. Hellinger distance for each HIV-1 integrase amino acid position between drug-naïve (DN) and INI-naïve (IN) groups INI-experienced (IE) groups. Hellinger distance between DN and IE (y-axis) is reported for each position (x-axis). Positions showing different mutation levels in the two datasets are marked according to the top right legend. In particular, the black circle (passing from mutability level L = 0, i.e. less than 1%, to mutability level L = 2, i.e. between 5 and 10%) has been added to the symbols already present in Fig. 5, indicating a considerable variation in mutability.

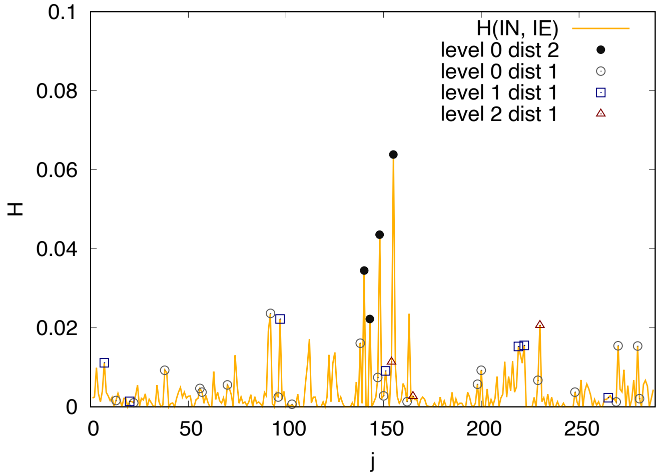

Fig. 7. Hellinger distance for each amino acid position between drug-experienced but INI-naïve (IN) and INI-experienced (IE) groups. Hellinger distance between IN and IE (y-axis) is reported for each position (x-axis). Positions showing different mutation levels in the two datasets are marked with circles, squares or triangles (the same as in Fig. 6) according to the top right legend.

3.3.1. *Conserved amino acid positions*

A total of 156 conserved amino acid positions among the three groups were found (rate of variation <1% within each group). Of these positions, 76 were highly conserved (rate of variation ≤0.2% within each group, with a Hellinger distance <0.25%; Table 1; Supplementary Fig. 1), while 80 amino acid positions were conserved (rate of variation >0.2% and < 1% within each group, with a Hellinger distance <1%; Supplementary Table 1). Among the highly conserved positions, no amino acid mutation was found in 35 of them. Consecutive highly conserved positions were observed in groups of two (9 pairs), three (4 triplets: 107–109, 116–118, 241–243 and 245–247), four (one quadruplet: 223–226) and five (one quintuplet: 129–133) amino acids (Table S1). The amino acids of histidine-histidine-cysteine-cysteine (H12-H16-C40-C43) motif, coordinating the zinc binding and promoting protein multimerization, were all found to be highly conserved. Regarding the residues D64, D116 and E152 of the catalytic triad, the residue D116 was highly conserved in all groups. Despite the fact that a good conservation was found for the residues D64 and E152, a small variation (0.3% and 0.5%, respectively) was found in IN individuals (Supplementary Table 1). It is worth noticing that amino acid positions 118, 142, 145 and 149, which have been reported to reduce susceptibility to INIs, fell in the highly conserved class.

3.3.2. *Amino acid positions with a decreased mutation rate in samples from treated individuals*

We found a total of 15 amino acid positions (D6, S24, V31, S39, L74, A91, S119, T122, T124, T125, V126, K160, N222, S230, C280) showing a significant decrease in the overall mutation rate in samples from treated individuals (IN and/or IE) compared to those from drug-naïve individuals (Table 2). In this table, for the sake of completeness, we also reported the differences between IN and IE mutation profiles. The decrease in mutation rate in each of the two-by-two group-comparisons was considered as significant when the following conditions were satisfied at the same time: i) having a Hellinger distance ≥1%; ii) having a *p* value <0.05 (Table 2; Supplementary Fig. 2). Of note, we found a significant decrease in the overall mutation rate from INI-naïve to INI-exposed samples for the residues L74, A91, T122, T124, T125, N222, S230, and C280. Specifically, a significant decrease in the overall mutation rate from both DN to IE and from IN to IE individuals was found for the residues L74 and A91, suggesting that the INI-pressure might have a role in the mutation variation of these positions. A significant decrease in the mutation rate of the other positions (T122, T124, T125, N222, S230, and C280) was found in each of the two-by-two group-comparisons groups, thus suggesting that drug-pressure in general (not only INI-pressure) might have a role in the mutation rate of these positions.

For each position, several mutations contributed to the amino acid variation. For example, in the case of position 74, the overall mutation rate was 5.9% in drug-naïve samples, 4.8% in IN samples and 3.5% in IE samples, with a wild type amino-acid rate of 94.1%, 95.2% and 96.5%, respectively. By evaluating the mutations that contributed to the amino acid variation, mutations 74 M/I/V/Q were found. However, the mutation variation was mainly due to the accessory mutations 74I and 74 M. In fact, the prevalence of polymorphic mutation 74I was 4.8% in drug-naïve samples, 3.7% in IN samples and 0.9% in IE samples, while the prevalence of 74 M was 0.3%, 0.3% and 1.9%, respectively. Concerning this last mutation, our results confirm that it is related to INI-exposure; in fact, it is known that it is selected in patients receiving RAL and EVG.

3.3.3. *Amino acid positions with an increased mutation rate in samples from treated individuals*

We found a total of 8 amino acid positions (E92, T97, G140, Y143, Q148, M154, N155, V165) with a significantly higher mutation rate in samples from treated individuals (IN and/or IE) compared to those who were drug-naïve (Table 3). In this table, for the sake of completeness, we also report the differences between IN and IE mutation profiles. The increase in mutation rate in each of the two-by-two group-comparisons was considered as significant when the following conditions were satisfied at the same time: i) having a Hellinger distance ≥1%; ii) having a *p* value <0.05 (Table 3; Supplementary Fig. 3). For seven positions (E92, T97, G140, Y143, Q148, M154 and N155), the wild type amino acid was less conserved in the group of IE individuals compared to both IN and drug-naïve individuals. Noteworthy, six of these positions (E92, T97, G140, Y143, Q148, and N155) are known to be associated with resistance to INIs; in fact, several mutations related to these positions are selected under INI-pressure (https://hivdb.stanford.edu/dr-summary/resistance-notes/INSTI/). For all these positions, the wild type amino acid was less conserved in the group of IE individuals. Concerning the positions Y143 and N155 for instance, the wild type amino acids

Table 1
Prevalence of the highly conserved amino acid positions.

AA position	Wild type[a]	Variation prevalence, %		
		Drug naïve (*N* = 1460)	INI naïve (*N* = 386)	INI experienced (*N* = 287)
5	I	0.0	0.0	0.2
12	**H**	0.0	0.0	0.0
16	**H**	0.0	0.0	0.0
18	N	0.1	0.1	0.2
19	**W**	0.0	0.0	0.0
29	P	0.0	0.0	0.0
36	I	0.0	0.0	0.0
40	C	0.1	0.0	0.0
43	C	0.1	0.0	0.0
44	Q	0.1	0.1	0.2
52	**G**	0.0	0.0	0.0
58	P	0.0	0.1	0.0
61	**W**	0.0	0.0	0.0
62	Q	0.0	0.0	0.0
67	H	0.0	0.1	0.0
69	**E**	0.0	0.0	0.0
76	A	0.0	0.1	0.0
78	H	0.0	0.1	0.0
81	S	0.1	0.0	0.2
83	**Y**	0.0	0.0	0.0
85	E	0.0	0.1	0.0
86	A	0.0	0.0	0.2
94	**G**	0.0	0.0	0.0
102	L	0.0	0.1	0.2
105	A	0.1	0.0	0.0
107	R	0.1	0.0	0.0
108	**W**	0.0	0.0	0.0
109	P	0.1	0.0	0.0
116	D	0.1	0.0	0.0
117	N	0.1	0.0	0.0
118	**G**	0.0	0.0	0.0
120	**N**	0.0	0.0	0.0
129	A	0.1	0.1	0.0
130	**C**	0.0	0.0	0.0
131	**W**	0.0	0.0	0.0
132	**W**	0.0	0.0	0.0
133	A	0.0	0.0	0.0
142	**P**	0.0	0.0	0.0
144	N	0.0	0.0	0.0
145	P	0.0	0.0	0.2
149	G	0.1	0.1	0.2
158	L	0.0	0.1	0.0
164	Q	0.0	0.0	0.2
168	Q	0.1	0.0	0.0
175	A	0.1	0.0	0.0
178	M	0.1	0.0	0.0
180	V	0.1	0.0	0.0
183	H	0.0	0.0	0.0
184	**N**	0.0	0.0	0.0
186	K	0.0	0.0	0.0
189	G	0.1	0.1	0.0
190	G	0.0	0.1	0.0
192	G	0.1	0.0	0.0
197	G	0.1	0.1	0.0
199	**R**	0.0	0.0	0.0
202	D	0.1	0.0	0.0
209	Q	0.1	0.0	0.0
223	F	0.0	0.0	0.0
224	R	0.1	0.0	0.0
225	**V**	0.0	0.0	0.0
226	Y	0.0	0.0	0.0
228	**R**	0.0	0.0	0.0
235	W	0.1	0.0	0.0
237	**G**	0.0	0.0	0.0
238	**P**	0.0	0.0	0.0
241	**L**	0.0	0.0	0.0
242	L	0.0	0.1	0.0
243	W	0.1	0.0	0.0
245	**G**	0.0	0.0	0.0
246	E	0.1	0.0	0.0
247	G	0.0	0.1	0.2
250	V	0.0	0.0	0.0
261	**P**	0.0	0.0	0.0
262	R	0.2	0.0	0.0
274	Q	0.1	0.0	0.0
276	**A**	0.0	0.0	0.0

In the table are reported all the amino acid positions highly conserved among the three groups: with a variation ≤0.2% within each group and a with a Hellinger distance <0.25% in the two-by-two group comparisons.
[a]Wild type amino acid according to the consensus B reference sequence.
In red: amino acid positions that are consecutive.
In bold: positions without any variation.
Underlined and in italic: position associated with a major resistance mutation.
In italic: positions associated with accessory resistance mutations.
In box: amino acid positions involved in the HHCC (12, 16, 40 and 43) and DDE (116) motifs.
AA: amino acid. INI: integrase inhibitor.

were conserved in 100% and 99.9% respectively in DN individuals, against 94.9% and 89.0% respectively in those who had an exposure to INIs. Among the other residues, a significant increase in the mutation rate between DN and IE individuals was found for M154 (92.6% and 84.3%, respectively) and V165 (92.4% and 83.6%, respectively). The following mutations were found to mainly contribute to the variations of these positions: 154I (DN: 5.0%; IN: 19.2%; IE: 11.5%) and 154 L (DN: 2.4%; IN: 5.6%; IE: 4.2%); 165I (DN: 7.6%; IN: 20.5%; IE: 16.3%).

4. Discussion

This study aimed at evaluating the rate of HIV-1 integrase mutation/conservation and at identifying new integrase amino acid positions potentially associated to resistance to INIs, analyzing a large number of HIV-1 B subtype integrase sequences, through an innovative bioinformatics and statistics method based on computational and probabilistic strategy. In particular, in order to identify those residue positions showing significant differences among different datasets (drug-naïve, INI-naive, INI-experienced), besides the classical instruments such as entropy and mutation rate, some other instruments were exploited. The concept of mutation level seems to be an effective and reliable solution to provide an at-a-glance picture of the differential mutational landscape and to identify significant residue positions. A more comprehensive instrument such as the Hellinger distance was introduced to take into consideration not only the fraction of mutated residues but also what kind of different residues occur as mutations in order to trace the effects of drug constraints. This new method showed that the HIV-1 integrase gene is well conserved, further confirming the previous data regarding its little propensity to show variations with respect to the entire HIV proteome (Ceccherini-Silberstein et al., 2010, 2009; Li and De Clercq, 2016; Rhee et al., 2008). It is interesting to observe that, according to our method, more than 50% of amino acid positions distributed across all three domains of integrase was conserved (≤1% mutation prevalence), regardless of drug exposure. In a previous study focusing only on INI-naïve patients (either drug-naïve or drug-experienced) (Ceccherini-Silberstein et al., 2009), a proportion of 65% of residues conservation (corresponding to 187/288 amino acids) was observed; this proportion was also confirmed in our new analysis, analyzing only drug-naïve and INI-naive individuals (186/288 amino acids) (data not shown). In particular, in both studies some of the amino acids were found in invariant regions of consecutive amino acid positions. Furthermore, among these conserved amino acid positions, our study identified several positions which are highly conserved among the three groups (Table 1), independently of drug pressure; these could be a potential target of new drugs. As expected, amino acid residues in the H12-H16-C40-C43 and the D64-D116-E152 motifs which are indispensable for

Table 2
Amino acid positions with a decreased mutation rate in samples from treated individuals.

AA position	AA	Mutation prevalence, %			Hellinger distance, % (p-value)		
		Drug naïve (*N* = 1460)	INI naïve (N = 386)	INI experienced (N = 287)	Drug-naïve vs. INI-naïve	Drug-naive vs. INI-experienced	INI-naive vs. INI-experienced
6	**All**[a]	11.3	6.3	6.3	**1.0 (0.029)**	0.7 (0.298)	0.4 (0.826)
	V	0.1	0.0	0.0			
	A	0.3	0.0	0.3			
	T	0.4	0.5	0.3			
	S	1.1	0.3	0.3			
	N	0.7	1.0	0.3			
	E	8.7	4.5	4.9			
	D[b]	88.7	93.7	93.7			
24	**All**[a]	15.9	14.3	18.1	**2.0 (<0.001)**	1.3 (0.068)	0.5 (0.570)
	A	0.8	0.3	0.7			
	G	2.4	6.5	6.4			
	T	0.1	0.0	0.2			
	N	10.6	5.9	8.5			
	D	1.1	1.6	2.1			
	E	0.3	0.0	0.2			
	H	0.4	0.0	0.0			
	K	0.1	0.0	0.0			
	S[b]	84.1	85.7	81.9			
31	**All**[a]	22.9	15.9	13.2	0.6 (0.003)	**1.1 (<0.001)**	0.1 (0.577)
	M	0.2	0.1	0.1			
	I	22.7	15.9	13.2			
	V[b]	77.1	84.1	86.8			
39	**All**[a]	20.2	14.4	14.0	**1.2 (0.019)**	0.9 (0.105)	0.9 (0.413)
	C	17.0	13.3	11.5			
	M	1.0	0.3	0.3			
	Y	0.0	0.1	0.0			
	G	0.0	0.1	0.0			
	T	0.1	0.0	0.0			
	N	1.6	0.5	1.6			
	Q	0.1	0.0	0.0			
	H	0.1	0.0	0.0			
	R	0.2	0.0	0.6			
	S[b]	79.8	85.6	86			
74	**All**[a]	5.9	4.8	3.5	0.1 (0.821)	**1.7 (<0.001)**	**1.3 (0.026)**
	M	0.3	0.3	1.9			
	I	4.8	3.7	0.9			
	V	0.7	0.8	0.5			
	Q	0.0	0.0	0.2			
	L[b]	94.1	95.2	96.5			
91	**All**[a]	6.0	3.6	3.0	0.8 (0.092)	**1.6 (0.011)**	**1.9 (0.028)**
	G	0.3	0.0	0.3			
	T	1.4	0.8	1.2			
	S	2.5	0.8	1.3			
	Q	0.5	0.3	0.0			
	D	0.1	0.0	0.0			
	E	1.2	1.8	0.0			
	P	0.0	0.0	0.2			
	A[b]	94.0	96.4	97.0			
119	**All**[a]	38.1	23.9	26.2	**2.7 (<0.001)**	**2.1 (<0.001)**	0.1 (0.941)
	A	0.4	0.3	0.3			
	G	9.9	2.8	3.0			
	T	4.4	5.1	4.9			
	R	3.2	2.9	2.5			
	P	20.2	12.8	15.5			
	S[b]	61.9	76.1	73.8			
122	**All**[a]	23.8	13.0	13.6	**1.5 (<0.001)**	**2.3 (<0.001)**	**1.3 (0.026)**
	I	22.5	11.7	13.1			
	V	1.3	1.3	0.0			
	A	0.0	0.0	0.3			
	S	0.0	0.0	0.2			
	T[b]	76.2	87.0	86.4			
124	**All**[a]	59.5	42.7	40.9	**2.8 (<0.001)**	**4.2 (<0.001)**	**1.1 (0.672)**
	A	23.0	19.8	20.7			
	G	0.3	0.0	0.0			
	S	3.1	3.5	2.1			
	N	32.3	18.3	16.3			
	Q	0.6	0.8	0.3			
	D	0.1	0.1	0.0			
	E	0.0	0.0	0.3			
	H	0.1	0.3	0.3			
	R	0.0	0.0	0.3			
	K	0.1	0.0	0.4			

(*continued on next page*)

Table 2 (*continued*)

AA position	AA	Mutation prevalence, %			Hellinger distance, % (p-value)		
		Drug naïve (*N* = 1460)	INI naïve (N = 386)	INI experienced (N = 287)	Drug-naïve vs. INI-naïve	Drug-naive vs. INI-experienced	INI-naive vs. INI-experienced
	T[b]	40.5	57.3	59.1			
125	**All[a]**	46.3	36.7	33.5	**1.5 (0.005)**	**1.7 (0.006)**	**1.4 (0.087)**
	C	0.1	0.0	0.0			
	M	0.7	1.0	0.0			
	I	0.3	0.4	0.0			
	V	6.1	3.6	5.5			
	A	38.3	31.5	27.4			
	S	0.3	0.0	0.2			
	N	0.1	0.0	0.0			
	Q	0.1	0.0	0.0			
	K	0.1	0.3	0.0			
	P	0.3	0.0	0.3			
	T[b]	53.7	63.3	66.5			
126	**All[a]**	6.0	3.2	1.9	0.8 (0.066)	**1.4 (0.002)**	0.6 (0.582)
	M	0.7	0.3	0.0			
	F	0.3	0.0	0.0			
	I	0.0	0.3	0.0			
	L	4.7	2.5	1.9			
	A	0.3	0.3	0.0			
	T	0.1	0.0	0.0			
	V[b]	94.0	96.8	98.1			
160	**All[a]**	5.1	1.3	2.4	**1.3 (0.002)**	0.6 (0.448)	0.6 (0.491)
	G	0.3	0.0	0.3			
	T	0.2	0.0	0.3			
	N	0.5	0.4	0.2			
	Q	3.9	0.9	1.6			
	E	0.1	0.0	0.0			
	R	0.1	0.0	0.0			
	K[b]	94.9	98.7	97.6			
222	**All[a]**	15.1	3.8	9.8	**3.6 (<0.001)**	0.6 (0.153)	**1.6 (0.012)**
	I	0.0	0.3	0.0			
	T	0.2	0.0	0.0			
	H	1.7	0.6	1.0			
	R	0.3	0.0	0.3			
	K	12.9	2.8	8.4			
	N[b]	84.9	96.2	90.2			
230	**All[a]**	23.0	12.3	20.6	**1.7 (<0.001)**	**1.2 (0.003)**	**2.1 (0.003)**
	G	0.3	0.5	0.7			
	T	0.0	0.0	0.2			
	N	22.5	11.8	17.8			
	H	0.1	0.0	0.0			
	R	0.1	0.0	1.9			
	S[b]	77.0	87.7	79.4			
280	**All[a]**	4.3	2.2	0.0	**2.3 (0.001)**	**3.1 (0.001)**	**1.5 (0.072)**
	L	0.1	0.0	0.0			
	G	2.8	0.0	0.0			
	S	1.4	1.9	0.0			
	R	0.0	0.3	0.0			
	C[b]	95.7	97.8	100.0			

In the table are reported all the amino acid positions that showed a significant decrease of the overall mutation rate in samples from treated individuals (INI-naïve and/or INI-experienced) compared to those from drug-naïve individuals. The decrease in mutation rate in each of the two-by-two group-comparisons was considered significant when the Hellinger distance was ≥1% and the p value was <0.05 (in bold). [a] Overall mutated amino acid prevalence per position. [b] Wild type amino acid according to the consensus B reference sequence. Underlined and in italic: positions associated with major resistance mutations. In italic: positions associated with accessory resistance mutations. AA: amino acid. INI: integrase inhibitor.

the integrase function, were found to be well conserved within the three groups. Moreover, among the conserved amino acid positions involved in contact with the cellular cofactor LEDGF/p75 that were described to be conserved (variability ≤0.25%) (Ceccherini-Silberstein et al., 2009), we found that eight of them (H12, L102, A129, C130, W131, W132, Q168, M178) were indeed highly conserved within the three groups (drug-naïve, INI-naive and INI-experienced) (Table 1), while the remaining (A128, I161, R166, E170, T174, Q214) showed a small variation (between >0.2% and ≤ 1.0% among the three groups, and with a distance <1%) (Supplementary Table 1). All these observations are confirmed by the low Shannon's entropy between the three groups, especially around the regions that determine the integrase functions (Supplementary Fig. 4).

The data obtained through our metric distance-based method showed that some amino acid positions displayed a significantly higher mutation rate in INI-experienced patients compared to drug-naïve and/or INI-naïve drug-experienced patients. At positions 92, 140, 143, 148 and 155 for instance, the main contributing mutations were respectively the 92Q/G, 140S/A/R, 143R/C/G, 148H/K/R and N155H, which are known to be associated with major INI resistance. At these positions, the wild type amino acid was found to be highly conserved among drug-naive and INI-naïve groups (Table 3). The positions 97, 140, 143, 148 and 155 went from a mutability index 0 (i.e. less than 1%) to a mutability index 2 (i.e. between 10 and 20%, see Fig. 6).

It is important to note that positions 154 and 165, not previously reported to be associated with resistance to INIs, showed a significantly increased mutation rate among INI-exposed samples as well as in INI-naïve drug-experienced patients. Interestingly, similar to the resistance

Table 3
Amino acid positions with an increased mutation rate in samples from treated individuals.

AA position	AA[a]	Mutation prevalence, %			Hellinger distance, % (p-value)		
		Drug naïve (N = 1460)	INI naïve (N = 386)	INI experienced (N = 287)	Drug-naïve vs. INI-naïve	Drug-naive vs. INI-experienced	INI-naive vs. INI-experienced
92	**All**[a]	0.0	0.0	3.3	0.0 (1.000)	**2.4 (<0.001)**	**2.4 (<0.001)**
	A	0.0	0.0	0.1			
	G	0.0	0.0	0.3			
	Q	0.0	0.0	2.8			
	R	0.0	0.0	0.1			
	P	0.0	0.0	0.1			
	E[b]	100.0	100.0	96.7			
97	**All**[a]	0.8	1.2	7.8	0.0 (0.443)	**2.7 (<0.001)**	**2.2 (<0.001)**
	A	0.8	1.2	7.7			
	E	0.0	0.0	0.2			
	T[b]	99.2	98.8	92.2			
140	**All**[a]	0.1	0.4	7.5	0.1 (0.231)	**4.5 (<0.001)**	**3.5 (<0.001)**
	A	0.0	0.1	0.3			
	S	0.1	0.3	7.1			
	R	0.0	0.0	0.1			
	G[b]	99.9	99.6	92.5			
143	**All**[a]	0.0	0.5	5.1	0.4 (<0.001)	**3.6 (<0.001)**	**2.2 (0.001)**
	C	0.0	0.0	1.6			
	G	0.0	0.0	0.3			
	H	0.0	0.0	0.2			
	R	0.0	0.5	3.0			
	Y[b]	100.0	99.5	94.9			
148	**All**[a]	0.1	0.4	8.2	0.2 (0.117)	**5.0 (<0.001)**	**4.4 (<0.001)**
	N	0.0	0.1	0.0			
	H	0.1	0.1	6.1			
	R	0.0	0.0	0.9			
	K	0.0	0.1	1.2			
	Q[b]	99.9	99.6	91.8			
154	**All**[a]	7.4	25.0	15.7	**4.6 (<0.001)**	**1.3 (<0.001)**	**1.1 (0.013)**
	I	5.0	19.2	11.5			
	L	2.4	5.6	4.2			
	V	0.0	0.3	0.0			
	M[b]	92.6	75.0	84.3			
155	**All**[a]	0.1	0.1	11.0	0.2 (0.145)	**8.1 (<0.001)**	**6.4 (<0.001)**
	T	0.1	0.0	0.0			
	H	0.0	0.1	11.0			
	N[b]	99.9	99.9	89.0			
165	**All**[a]	7.6	20.5	16.4	**2.5 (<0.001)**	**1.4 (<0.001)**	0.3 (0.164)
	M	0.0	0.0	0.1			
	I	7.6	20.5	16.3			
	V[b]	92.4	79.5	83.6			
270	**All**[a]	0.8	0.0	2.2	0.6 (0.219)	0.8 (0.357)	**1.5 (<0.001)**
	G	0.0	0.0	0.4			
	N	0.3	0.0	0.9			
	Q	0.1	0.0	0.0			
	E	0.2	0.0	0.0			
	H	0.2	0.0	0.9			
	K	0.1	0.0	0.0			
	D[b]	99.2	100.0	97.8			

In the table are reported all the amino acid positions that showed a significant increase of the overall mutation rate in samples from treated individuals (INI-naïve and/or INI-exposed) compared to those who were drug-naïve. The increase in mutation rate in each of the two-by-two group-comparisons was considered significant when the Hellinger distance was ≥1% and the p value was <0.05 (in bold). a Overall mutated amino acid prevalence per position. b Wild type amino acid according to the consensus B reference sequence. Underlined and in italic: positions associated with major resistance mutations. In italic: positions associated with accessory resistance mutations.
AA: amino acid. INI: integrase inhibitor.

positions 140, 145, 148 and 155, positions 154 and 165 are found in a flexible loop and an amphipathic alpha-helix (α4), a region which is involved in the interaction between integrase and viral DNA (Rhee et al., 2008). One of our previous studies showed that there was a positive association of some mutations at positions 154 and 165 with specific reverse transcriptase resistance mutations (Ceccherini-Silberstein et al., 2010); in particular, 154L (occuring in 2.4% of drug-naïve, 5.6% of INI-naïve and 4.2% of INI-experienced sequences) positively associated with 227L and 215Y, while 165I (occuring in 7.6% of drug-naïve, 20.5% of INI-naïve and 16.3% of INI-expericeenced sequences) positively associated with mutation 227L. The same study also showed that 154L and 165I significantly and positively co-occurred together (co-variation frequency: 44.4%). Moreover, the mutation 154I was previously documented to be involved in some mutational pathways (Anstett et al., 2017; Meixenberger et al., 2017; Seki et al., 2015).

Our method also evidenced that the exposure to ART (in both INI-naïve and/or INI-experienced individuals) is associated with a reduced variability at some integrase amino acid positions in HIV-1 integrase (Fig. 2). Among these amino acid positions, position 74 is known to be associated with accessory resistance mutations (Kobayashi et al., 2008; Low et al., 2009). However, looking specifically at mutation 74M, a mutation reported to be selected in patients receiving RAL or EVG (Kobayashi et al., 2008; Low et al., 2009), a significantly higher prevalence of 1.9% was observed among INI-exposed samples, compared to only 0.3% in drug-naïve samples. Other residues already reported in literature are the hypervariable residues S119, T124, and T125 that have been shown to interact with the tDNA; in particular, the mutation S119R could have a substantial impact on how the integrase binds to the tDNA (de Machado et al., 2019).

Moreover, this specific mutation enhances primary INI resistance (Hachiya et al., 2015). In this regard, an experiment showed that the occurrence of mutation A91E (another position in this group) restored the lost viral fitness due to the presence of S119R (Brockman et al., 2012).

On the other hand, the amino acid position 280, showed no variation in INI-exposed samples, compared to a variation of 4.3% and 2.2% in samples from drug-naïve and INI-naïve individuals, respectively. A mutation that contributed to the variation of this position is the 280S in samples from both drug-naïve (1.4%) and INI-naïve (1.9%) individuals. Mutants with 280S substitution is believed to be relatively more resistant to oxidation when compared with integrase with a wild type amino acid (Jenkins et al., 1996). The role of other amino acid positions found in this group with less variability in drug-exposed samples might deserve further investigation.

This study has some limitations. In fact, the number of analyzed sequences in the group of INI-treated patients, especially those receiving EVG and DTG, was few. Thus, we could not evaluate whether differences in the variability of specific amino acid position were related to a different type of INI usage. Further studies on a larger set of sequences are needed to answer this specific question. Moreover, it would have been more interesting to also include additional sequences from patients treated with second generation INIs currently used in clinical settings such as DTG and BIC. This analysis considered only HIV-1 B subtype; however, such in-depth analysis could also provide insights into what happens among HIV-1 non-B subtypes, which are reported to show a higher variability and rate of accessory resistance mutations (Semengue et al., 2021).

In conclusion, in the present work, a novel approach to study mutational events of biological sequences, both on a global and local scale, has been introduced by exploiting solid probabilistic and computational techniques. In particular, the focus was on drug-induced differential mutations in HIV-1 integrase, but the same approach could be applied in different contexts. The study confirmed an overall high conservation of the integrase with respect to other HIV proteins, and identified the presence of known interesting sites associated to drug-resistance. In addition, previously unknown positions highly varying between INI-exposed and drug-naïve individuals have been found and they surely deserve further investigation. This type of analysis constituted a first statistical attempt to deeply investigate rules and constraints that drive mutational events, and can be followed by several other steps in line with this approach. For example, chemical-physical properties of amino acids could be taken into account in order to highlight the most relevant mutations, or co-mutation events of couples/cluster of amino acids in different proteins of HIV or finally the complete RNA structure of HIV could be considered in order to analyze, via mutual information or conditional mutation probability.

Credit authorship contribution statement

Davide Vergni: Conceptualization; Formal analysis; Methodology; Writing – original draft. Daniele Santoni: Conceptualization; Formal analysis; Writing – original draft. Yagai Bouba: Investigation; Writing – original draft. Saverio Lemme: Investigation; Formal analysis; revising. Lavinia Fabeni, Luca Carioti: Data curation; Revising. Ada Bertoli, William Gennari, Federica Forbici, Carlo Federico Perno, Roberta Gagliardini: Revising. Francesca Ceccherini: Silberstein: Conceptualization; Writing – review. Maria Mercedes Santoro: Conceptualization; Investigation; Supervision; Writing – original draft.

Funding

The funding for this study was provided by an unrestricted grant from the AVIRALIA Foundation, and from the CNR project DIT. AD021.161.001 / Analisi probabilistica di dataset biologici e network dynamics.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationship that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Debra Mandatori for having revised and edited the manuscript, and Andrea Biddittu and Massimiliano Bruni for the data management.

HIV-drug resistance group.

University of Rome “Tor Vergata”, Rome, Italy: F. Ceccherini-Silberstein, M.M. Santoro; Policlinic of Rome Tor Vergata, Rome: M. Andreoni, M. Iannetta, A. Bertoli, L. Sarmati, V. Malagnino; E. Teti; Saint Camillus International University of Health Sciences, Rome, Italy: D. Armenia; National Institute for Infectious Diseases “L. Spallanzani”, IRCCS, Rome, Italy: A. Antinori, F. Baldini, R. Bellagamba, G. Berno, M. Camici, S. Cicalini, F. De Zottis, R. Esvan, L. Fabeni, F. Forbici, M. Fusto, R. Gagliardini, S. Gebremeskel, F. Gili; E. Girardi, E. Grilli, S. Grisetti, S. Lanini, I. Mastrorosa, V. Mazzotta, A. Mondi, N. Orchi; S. Ottou, C. Pinnetti, S. Pittalis; D. Pizzi, M. Plazzi, A. Vergori; San Gallicano Hospital, IRCCS, Rome, Italy: A.R. Buonomini, M. Giuliani, A. Latini, A. Pacifici; Bambino Gesù Children’s Hospital, Rome, Italy: C.F. Perno; La Sapienza University, Rome, Italy: V. Belvisi, C. Del Borgo, A. Carraro, M. Lichtner, R. Marocco; University Hospital, University of Modena and Reggio Emilia, Modena, Italy: V. Borghi, C. Mussini, W. Gennari.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2022.105294.

References

Adami, C., 2004. Information theory in molecular biology. Phys Life Rev 1, 3–22. https://doi.org/10.1016/j.plrev.2004.01.002.

Anstett, K., Brenner, B., Mesplede, T., et al., 2017. HIV drug resistance against strand transfer integrase inhibitors. Retrovirology 14, 36. https://doi.org/10.1186/s12977-017-0360-7.

Armenia, D., Fabeni, L., Alteri, C., et al., 2014. HIV-1 integrase genotyping is reliable and reproducible for routine clinical detection of integrase resistance mutations even in patients with low-level viraemia. J. Antimicrob. Chemother. 70, 1865–1873. https://doi.org/10.1093/jac/dkv029.

Armenia, D., Bouba, Y., Gagliardini, R., et al., 2020. Evaluation of virological response and resistance profile in HIV-1 infected patients starting a first-line integrase inhibitor-based regimen in clinical settings. J. Clin. Virol. 130, 104534 https://doi.org/10.1016/j.jcv.2020.104534.

Brockman, M.A., Chopera, D.R., Olvera, A., et al., 2012. Uncommon pathways of immune escape attenuate HIV-1 Integrase replication capacity. J. Virol. 86, 6913–6923. https://doi.org/10.1128/jvi.07133-11.

Brooks, K.M., Sherman, E.M., Egelund, E.F., et al., 2019. Integrase inhibitors: after 10 years of experience, is the best yet to come? Pharmacotherapy. Pharmacotherapy Publications Inc. https://doi.org/10.1002/phar.2246.

Ceccherini-Silberstein, F., Svicher, V., Sing, et al., 2007. Characterization and structural analysis of novel mutations in human immunodeficiency virus type 1 reverse transcriptase involved in the regulation of resistance to nonnucleoside inhibitors. J. Virol. 81, 11507–11519. https://doi.org/10.1128/JVI.00303-07.

Ceccherini-Silberstein, F., Malet, I., D’Arrigo, R., et al., 2009. Characterization and structural analysis of HIV-1 integrase conservation. AIDS Rev. 11, 17–29. Available at. https://pubmed.ncbi.nlm.nih.gov/19290031/.

Ceccherini-Silberstein, F., Malet, I., Fabeni, L., et al., 2010. Specific HIV-1 integrase polymorphisms change their prevalence in untreated versus antiretroviral-treated HIV-1-infected patients, all naive to integrase inhibitors. J. Antimicrob. Chemother. 65, 2305–2318. https://doi.org/10.1093/jac/dkq326.

Chiu, T., Davies, D., 2005. Structure and function of HIV-1 Integrase. Curr. Top. Med. Chem. 4, 965–977. https://doi.org/10.2174/1568026043388547.

Cihlar, T., Fordyce, M., 2016. Current status and prospects of HIV treatment. Curr. Opin. Virol. 18, 50–56. https://doi.org/10.1016/j.coviro.2016.03.004.

Coffin, J.M., 1995. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science (80-.) 267, 483–489. https://doi.org/10.1126/science.7824947.

Coffin, J.M., Hughes, S.H., Varmus, H.E., 1997. Retroviruses, Retroviruses. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY). Available at. https://www.ncbi.nlm.nih.gov/books/NBK19376/.

Collier, D.A., Monit, C., Gupta, R.K., 2019. The impact of HIV-1 drug escape on the global treatment landscape. Cell Host Microbe. https://doi.org/10.1016/j.chom.2019.06.010.

de Lima, E.N.C., Piqueira, J.R.C., Camargo, M., et al., 2018. Impact of antiretroviral resistance and virological failure on HIV-1 informational entropy. J. Antimicrob. Chemother. 73, 1054–1059. https://doi.org/10.1093/jac/dkx508.

de Machado, L.A., da Gomes, M.F.C., Guimarães, A.C.R., 2019. Raltegravir-induced adaptations of the HIV-1 Integrase: analysis of structure, variability, and mutation co-occurrence. Front. Microbiol. 10, 1–12. https://doi.org/10.3389/fmicb.2019.01981.

Engelman, A., Mizuuchi, K., Craigie, R., 1991. HIV-1 DNA integration: mechanism of viral DNA cleavage and DNA strand transfer. Cell 67, 1211–1221. https://doi.org/10.1016/0092-8674(91)90297-C.

Hachiya, A., Ode, H., Matsuda, M., et al., 2015. Natural polymorphism S119R of HIV-1 integrase enhances primary INSTI resistance. Antivir. Res. 119, 84–88. https://doi.org/10.1016/j.antiviral.2015.04.014.

Jenkins, T.M., Engelman, A., Ghirlando, R., et al., 1996. A soluble active mutant of HIV-1 integrase: involvement of both the core and carboxyl-terminal domains in multimerization. J. Biol. Chem. 271, 7712–7718. https://doi.org/10.1074/jbc.271.13.7712.

Jóźwik, I.K., Passos, D.O., Lyumkis, D., 2020. Structural biology of HIV Integrase Strand transfer inhibitors. Trends Pharmacol. Sci. 41, 611–626. https://doi.org/10.1016/j.tips.2020.06.003.

Kobayashi, M., Nakahara, K., Seki, T., et al., 2008. Selection of diverse and clinically relevant integrase inhibitor-resistant human immunodeficiency virus type 1 mutants. Antivir. Res. 80, 213–222. https://doi.org/10.1016/j.antiviral.2008.06.012.

Li, G., De Clercq, E., 2016. HIV genome-wide protein associations: a review of 30 years of research. Microbiol. Mol. Biol. Rev. 80, 679–731. https://doi.org/10.1128/mmbr.00065-15.

Low, A., Prada, N., Topper, M., et al., 2009. Natural polymorphisms of human immunodeficiency virus type 1 Integrase and inherent susceptibilities to a panel of Integrase inhibitors. Antimicrob. Agents Chemother. 53, 4275–4282. https://doi.org/10.1128/AAC.00397-09.

Luciw, P.A., Cheng-Mayer, C., Levy, J.A., 1987. Mutational analysis of the human immunodeficiency virus: the orf-B region down-regulates virus replication. Proc. Natl. Acad. Sci. 84, 1434–1438. https://doi.org/10.1073/pnas.84.5.1434.

Marcelin, A.-G., Grude, M., Charpentier, C., et al., 2019. Resistance to integrase inhibitors: a national study in HIV-1-infected treatment-naive and -experienced patients. J. Antimicrob. Chemother. 74, 1368–1375. https://doi.org/10.1093/jac/dkz021.

Mbhele, N., Chimukangara, B., Gordon, M., 2021. HIV-1 integrase strand transfer inhibitors: a review of current drugs, recent advances and drug resistance. Int. J. Antimicrob. Agents 57, 106343. https://doi.org/10.1016/j.ijantimicag.2021.106343.

Meixenberger, K., Yousef, K.P., Smith, M.R., et al., 2017. Molecular evolution of HIV-1 integrase during the 20 years prior to the first approval of integrase inhibitors. Virol. J. 14, 223. https://doi.org/10.1186/s12985-017-0887-1.

Ohya, M., Sato, K., 2000. Use of information theory to study genome sequences. Reports Math. Phys. 46, 419–428. https://doi.org/10.1016/S0034-4877(00)90010-7.

Rhee, S.-Y., Liu, T.F., Kiuchi, M., et al., 2008. Natural variation of HIV-1 group M integrase: implications for a new class of antiretroviral inhibitors. Retrovirology 5, 74. https://doi.org/10.1186/1742-4690-5-74.

Rice, P., Craigie, R., Davies, D.R., 1996. Retroviral integrases and their cousins. Curr. Opin. Struct. Biol. 6, 76–83. https://doi.org/10.1016/S0959-440X(96)80098-4.

Scarsi, K.K., Havens, J.P., Podany, A.T., et al., 2020. HIV-1 Integrase inhibitors: a comparative review of efficacy and safety. Drugs. https://doi.org/10.1007/s40265-020-01379-9.

Seki, T., Suyama-Kagitani, A., Kawauchi-Miki, et al., 2015. Effects of Raltegravir or Elvitegravir resistance signature mutations on the barrier to Dolutegravir resistance in vitro. Antimicrob. Agents Chemother. 59, 2596–2606. https://doi.org/10.1128/AAC.04844-14.

Semengue, E.N.J., Armenia, D., Inzaule, S., et al., 2021. Baseline integrase drug resistance mutations and conserved regions across HIV-1 clades in Cameroon: implications for transition to dolutegravir in resource-limited settings. J. Antimicrob. Chemother. 13, 1277–1285. https://doi.org/10.1093/jac/dkab004.

Smith, S.J., Zhao, X.Z., Passos, D.O., et al., 2021. Integrase Strand transfer inhibitors are effective anti-HIV drugs. Viruses 13, 205. https://doi.org/10.3390/v13020205.

Swanstrom, R., Wills, J., 1997. Synthesis, Assembly, and Processing of Viral Proteins. Retroviruses. Available at. https://pubmed.ncbi.nlm.nih.gov/21433349/.

Teeraananchai, S., Kerr, S.J., Amin, J., et al., 2017. Life expectancy of HIV-positive people after starting combination antiretroviral therapy: a meta-analysis. HIV Med. 18, 256–266. https://doi.org/10.1111/hiv.12421.

Thierry, E., Deprez, E., Delelis, O., 2017. Different pathways leading to integrase inhibitors resistance. Front. Microbiol. 7, 1–13. https://doi.org/10.3389/fmicb.2016.02165.

Varghese, V., Liu, T.F., Rhee, S.Y., et al., 2010. HIV-1 integrase sequence variability in antiretroviral naïve patients and in triple-class experienced patients subsequently treated with raltegravir. AIDS Res. Hum. Retrovir. 26, 1323–1326. https://doi.org/10.1089/aid.2010.0123.

Wandeler, G., Johnson, L.F., Egger, M., 2016. Trends in life expectancy of HIV-positive adults on antiretroviral therapy across the globe: comparisons with general population. Curr. Opin. HIV AIDS. https://doi.org/10.1097/COH.0000000000000298.

Yang, L.-L.L., Li, Q., Zhou, L.-B.B., et al., 2019. Meta-analysis and systematic review of the efficacy and resistance for human immunodeficiency virus type 1 integrase strand transfer inhibitors. Int. J. Antimicrob. Agents. https://doi.org/10.1016/j.ijantimicag.2019.08.008.