Meixenberger *et al. Virology Journal* (2017) 14:223
DOI 10.1186/s12985-017-0887-1

Virology Journal

RESEARCH Open Access

Molecular evolution of HIV-1 integrase during the 20 years prior to the first approval of integrase inhibitors

Karolin Meixenberger[1*], Kaveh Pouran Yousef[2], Maureen Rebecca Smith[2], Sybille Somogyi[1], Stefan Fiedler[1], Barbara Bartmeyer[3], Osamah Hamouda[3], Norbert Bannert[1], Max von Kleist[2] and Claudia Kücherer[1]

Abstract

Background: Detailed knowledge of the evolutionary potential of polymorphic sites in a viral protein is important for understanding the development of drug resistance in the presence of an inhibitor. We therefore set out to analyse the molecular evolution of the HIV-1 subtype B integrase at the inter-patient level in Germany during a 20-year period prior to the first introduction of integrase strand inhibitors (INSTIs).

Methods: We determined 337 HIV-1 integrase subtype B sequences (amino acids 1–278) from stored plasma samples of antiretroviral treatment-naïve individuals newly diagnosed with HIV-1 between 1986 and 2006. Shannon entropy was calculated to determine the variability at each amino acid position. Time trends in the frequency of amino acid variants were identified by linear regression. Direct coupling analysis was applied to detect covarying sites.

Results: Twenty-two time trends in the frequency of amino acid variants demonstrated either single amino acid exchanges or variation in the degree of polymorphy. Covariation was observed for 17 amino acid variants with a temporal trend. Some minor INSTI resistance mutations (T124A, V151I, K156 N, T206S, S230 N) and some INSTI-selected mutations (M50I, L101I, T122I, T124 N, T125A, M154I, G193E, V201I) were identified at overall frequencies >5%. Among these, the frequencies of L101I, T122I, and V201I increased over time, whereas the frequency of M154I decreased. Moreover, L101I, T122I, T124A, T125A, M154I, and V201I covaried with non-resistance-associated variants.

Conclusions: Time-trending, covarying polymorphisms indicate that long-term evolutionary changes of the HIV-1 integrase involve defined clusters of possibly structurally or functionally associated sites independent of selective pressure through INSTIs at the inter-patient level. Linkage between polymorphic resistance- and non-resistance-associated sites can impact the selection of INSTI resistance mutations in complex ways. Identification of these sites can help in improving genotypic resistance assays, resistance prediction algorithms, and the development of new integrase inhibitors.

Keywords: HIV, Integrase, Drug resistance, Polymorphisms, Time trend, Covariation

Background

The HIV-1 integrase catalyses the integration of the reverse transcribed viral DNA into the host genomic DNA via a two-step process. In its active form the integrase forms a tetramer. The monomeric enzyme consists of 288 amino acids (aa) and contains three functional domains: the N-terminal zinc-binding domain (NTD, aa 1–49), the central catalytic core domain (CCD, aa 50–212), and the C-terminal DNA-binding domain (CTD, aa 213–288) [1–4]. Each region comprises motifs essential for the proper function of the enzyme, e.g. the zinc finger motif H12-H16-C40-C43 in the NTD, the active site D64-D116-E153 in the CCD, and the minimal nonspecific DNA-binding region ranging from I220 to D270 in the CTD [2–6].

Raltegravir was the first integrase strand inhibitor (INSTI) to be approved in Europe in 2007, followed by elvitegravir in 2012 and dolutegravir in 2014. Bictegravir [7] and cabotegravir [8] are in clinical trial development.

* Correspondence: meixenbergerk@rki.de
[1]HIV and other Retroviruses, Robert Koch Institute, Berlin, Germany
Full list of author information is available at the end of the article

INSTIs target the CCD, thereby inhibiting the strand transfer of the double-stranded viral DNA into the host genome [1]. Various allosteric inhibitors of integrase (ALLINIs), which modulate integrase multimerisation [9] and interfere with the cellular transcription factor LEDGF/p75 [10] are in development but have not so far made it further than Phase I clinical trial [11].

Due to its high rate of replication, mutation, and recombination, HIV is a virus of high genetic variability. The viability of virus variants in turn is limited by structural and functional constraints. At the same time, variants in the viral quasispecies can be selected by the pressure of the human immune system [12, 13] or antiretroviral treatment (ART) [14, 15]. In general, the HIV diversity at the intra-patient level increases during the course of infection [16] driven by both drift and selection [17]. During transmission to a new host, several stochastic and selective bottlenecks reduce the viral diversity to a few variants [18], and factors that contribute to shaping the HIV diversity at the inter-patient level are extensively discussed [19–24].

Naturally occurring polymorphisms can affect the genetic barrier to drug resistance by influencing the selection of resistance mutations, enzymatic activity, and replicative capacity [22, 25, 26]. Epistatic interactions between polymorphisms can further modulate viral fitness and the development of drug resistance in complex ways and have been shown to play an important role in the HIV-1 protease and reverse transcriptase [27–30]. Thus, to understand the selection of resistant variants in the presence of INSTIs, it is important to investigate the evolutionary dynamics of the polymorphic sites in the integrase.

The prevalence of HIV-1 integrase polymorphisms and INSTI resistance mutations has been investigated before in INSTI-naïve individuals [26, 31–35] and ART-naïve individuals [35–40]. However, time trends and covariation of complex mutation patterns preceding the availability of INSTIs have not so far been analysed. The aim of this study was to investigate covarying clusters of naturally occurring resistance- and non-resistance-associated amino acid variants and their frequencies over time at the inter-patient level to consider their potential relevance for INSTI resistance. To this end, HIV-1 integrase sequences were obtained from samples of ART-naïve individuals newly diagnosed with HIV-1 between 1986 and 2006, a 20-year period prior to the first approval of INSTI in Germany [41].

Methods

Study population

Plasma samples from individuals newly diagnosed with HIV-1 between 1986 and 1996 (N = 167) were archived at the former diagnostic unit of the National AIDS Centre in Germany and stored at −40 °C. The date of HIV-1 diagnosis is the same as the plasma sampling date for the HIV-1 integrase genotyping. Plasma samples from individuals newly diagnosed with HIV-1 between 1997 and 2006 (N = 170) were collected for the German HIV-1 Seroconverter Study [42–45] and stored at −70 ° C. These plasma samples were taken within 12 months after diagnosis. In total, the study population comprised 337 individuals (Table 1).

HIV-1 integrase genotyping

Viral RNA from 500 μl plasma was pelleted by centrifugation (20,800 g, 90 min, 4 °C) and then isolated using the Viral RNA Mini Kit (Qiagen, Germany) according to the manufacturer's instructions. Reverse transcription and polymerase chain reaction were performed using the OneStep RT-PCR kit (Qiagen, Germany) with primers 5′-INT (5′- ATT GGA GGA AAT GAA CAA GT -3′; nucleotides (nt) 4173–4192, Acc. K03455) and 3p31as (5′- ATC CTG TCT ACY TGC CAC ACA A -3′; nt 5066–5087, Acc. K03455) [37]. Amplicons were purified (QIAquick spin PCR purification kit, Qiagen, Germany) and sequenced by cycle-sequencing (ABI Big Dye 3.1, Gene Amp

Table 1 Number of HIV-1 integrase sequences per year/period of HIV-1 diagnosis

Year of diagnosis	No. of HIV-1 sequences/year	No. of HIV-1 sequences/period
1986	36	84
1987	3	
1988	17	
1989	28	
1990	0	58
1991	3	
1992	30	
1993	5	
1994	20	
1995	18	42
1996	7	
1997	7	
1998	10	
1999	7	49
2000	10	
2001	12	
2002	20	
2003	17	104
2004	22	
2005	24	
2006	41	

Applied Biosystems PCR System 9700, Thermo Fisher Scientific, Germany) with the primers listed above and the additional primers F2 s (5′- TAA GAC AGC AGT ACA AAT GGC AG -3′; nt 4745–4767, Acc. K03455) and F3as (5′- GCT GTC CCT GTA ATA AAC CCG -3′; nt 4899–4919, Acc. K03455). Sequencing was performed on an ABI Prism 310 capillary sequencer (Thermo Fisher Scientific, Germany), and SeqMan Pro (Lasergene v10.0.1, DNASTAR, USA) was used for sequence analysis. This genotyping assay had a detection limit of 10^3 copies/ml for HIV-1 subtype B and yielded a 915 bp amplicon spanning HIV-1 integrase bp 1–278. Only subtype B strains were included in the analyses. The HIV-1 subtype was determined using the REGA HIV Subtyping Tool (http://dbpartners.stanford.edu:8080/RegaSubtyping/stanford-hiv/typingtool/).

INSTI resistance mutations

Major INSTI resistance mutations (T66I, E92Q, F121Y, Y143CHR, S147G, Q148HKR, N155H) that confer substantial phenotypic resistance to at least one of the currently approved INSTI as well as minor INSTI resistance mutations (T66AK, L74 M, E92G, T97A, E138AK, G140AS, R263K) that increase INSTI resistance and/or viral replication capacity were identified according to the IAS-list [46]. In addition, the following minor INSTI resistance mutations according to the current definitions of the resistance prediction algorithm HIVdb (http://hivdb.stanford.edu, version March 2, 2017), ANRS (http://www.hivfrenchresistance.org, version no. 26, September 2016), HIV-GRADE (http://www.hiv-grade.de, version January 16, 2017), and Rega (https://rega.kuleuven.be/cev/avd/software/rega-algorithm, v9.0.1, October 29, 2013) were considered: A49G, H51Y, V54I, L68IV, L74I, E92V, Q95K, H114Y, G118R, S119R, T124A, A128T, E138T, G140C, Y143AGS, P145S, Q146IKLPR, Q148EG, V151AIL, S153FY, N155ST, K156 N, E157Q, G163KR, T206S, S230GNR, D232N, V260I. Moreover, INSTI-selected mutations that were observed in vitro or in vivo were investigated [47, 48]: M50I, G59E, I60L, I72A, Q95T, L101IY, T112S, F121Y, T122I, T125A, E138D, Y143K, Q148N, M154I, I162M, G163E, Q177R, G193E, V201I, I203M, I204T.

Assessment of phylogenetic bias

A unique anonymising code provided with the mandatory report of newly HIV-1 infected cases to the Robert Koch-Institute ensured that only one HIV-1 sequence per patient was included in our dataset. However, to investigate whether our results were biased by the overrepresentation of phylogenetically closely related sequences (i.e. sequences originating from direct transmission events) we identified clusters of sequences with a very small phylogenetic distance and replaced them with one representative sequence (the sequence from the patient who was diagnosed first). For clustering, a multiple sequence alignment (MSA) including an HIV-1 group N reference sequence (Acc. AJ006022) was generated with clustalW (BioEdit, v7.2.5, Tom Hall) and end-trimmed to nt 4230–5064 (Acc. K03455). We computed a maximum likelihood phylogeny with 500 times non-parametric bootstrap with replacement using the program RAxML [49]. The HIV-1 group N sequence was used to root the maximum likelihood tree. Applying the program Transmic [50], a set of phylogenetically closely related sequences was identified if the mean of all pairwise patristic distances did not exceed a threshold of 0.015 expected nucleotide substitutions per site and the most recent common ancestor node had a bootstrap support of 0.9 [50, 51]. We detected 11 clusters, each comprising two sequences. The final reduced dataset consisted of 326 sequences. All analyses were then performed with the full dataset and the reduced dataset to assess any potential bias. All reported results were confirmed on the basis of both datasets while all numbers reported in the results section are given for the full dataset.

Amino acid variability

MSAs for the full (337 sequences) and the reduced (326 sequences) dataset were generated with clustalW (BioEdit, v7.2.5, Tom Hall) and end-trimmed to nt 4230–5064 (Acc. K03455). The nucleotide sequences were then translated into amino acid sequences. Nucleotide sequence ambiguities of codons were resolved during translation. An X was assigned if multiple amino acids resulted from the translation of codons containing nucleotide sequence ambiguities to avoid consideration of amino acid variants only present in mixtures due to PCR and sequencing errors. The frequencies of amino acids at positions 1 to 278 of the HIV-1 integrase were calculated. A position was defined to be polymorphic if, on the basis of the full and the reduced dataset at this position, a total of 1% or more amino acid variants were present when compared to the consensus B sequence (http://hivdb.stanford.edu/pages/documentPage/consensus_amino_acid_sequences.html). All other positions were defined as conserved.

We also analysed the Shannon entropy E_i, which quantifies the degree of variability at a single amino acid position i [1,278], according to

$$E_i = -\sum\nolimits_X P(X_i = x)\log_2 P(X_i = x)$$

where $P(X_i = x$) denotes the probability of observing the particular amino acid x at position i [52]. A high value of entropy E_i indicates high amino acid variability at position i.

Amino acid time trends

The samples were grouped into five time periods (1986–1989, 1990–1994, 1995–1998, 1999–2002, and 2003–2006) to obtain a more homogenous distribution and to render the time trend analyses more robust (Additional file 1: Table S1). Next, we calculated the frequencies of all amino acids at positions 1 to 278 of the HIV-1 integrase within the periods. We then fitted a linear regression by minimizing the sum of least square deviation between detected and predicted frequencies y and $f(a, b)$ for each period i according to

$$\{a, b\} = argmin \sum_i \left(f(a, b, x_i) - y(x_i)\right)^2$$

with $f_i(a, b, x_i) = a \cdot x_i + b$ where a denotes the slope and b the intersect of the linear function. The variable x_i denotes the distance of the centre of period i [in years] to the centre of the first period. In order to detect significant time trends, we generated 10,000 bootstrap samples by drawing sequences from the original sequence set with replacement. For each resampled set, we computed mutation frequencies and performed the linear regression as described above. Raw P values were then computed from the bootstrap distribution of fitted slopes a in analogy to Katchanov et al. [53], i.e. we computed $P^+ = \frac{\#a \le 0.001}{10000}$ to test whether the frequency of that mutation is significantly increasing by at least 0.1% per year ($H_0 : a \le 0.001$ vs. $H_1 : a > 0.001$) and conversely $P^- = \frac{\#a \ge -0.001}{10000}$ to assess whether it is significantly decreasing by at least 0.1% per year ($H_0 : a \ge -0.001$ vs. $H_1 : a < -0.001$). P values ($P = \min(P^-, P^+)$) were subsequently corrected for multiple testing using the false discovery rate (FDR) method by Benjamini-Hochberg [54]. Time trends were considered significant if they were identified on the basis of the full and the reduced dataset with a FDR corrected $P < 0.05$.

Amino acid covariation

In order to determine if amino acid positions are covarying, we applied direct coupling analysis where direct correlations are disentangled from transitive correlations. We computed evolutionary coupling terms ec$_{ij}$ using the recently developed plmc tool (https://github.com/debbiemarkslab/plmc) [55]. This tool infers couplings by fitting a Potts model to the MSA using a pseudo likelihood approach with L$_2$ regularisation. We used default regularisation parameters of $\lambda_1 = 0.01$ and $\lambda_2 = 100$ for single site contributions and direct couplings terms $e_{ij}(\alpha,\beta)$, respectively. To correct for a phylogenetic bias by reweighting neighbouring sequences, we chose the default parameter $\theta = 0.01$. Ambiguous amino acids (X) were discarded during inference. As expected, the inferred coupling terms are approximately normally distributed with mean 0 and standard deviation $1/\lambda_2$.

The direct coupling terms $e_{ij}(\alpha,\beta)$ describe the direct correlation of two amino acids α and β at positions i and j. The overall interaction of two positions i and j are given by the evolutionary coupling term ec_{ij}, which is the Frobenius norm of the direct coupling terms

$$\|e_{ij}\|_2 = \sqrt{\sum\nolimits_{\alpha,\beta} e_{ij}(\alpha,\beta)^2}$$

with average product correction (APC), to suppress phylogenetic bias effects [56]:

$$ec_{ij} = \|e_{ij}\|_2 - \frac{\|e_{i*}\|_2 \cdot \|e_{*j}\|_2}{\|e_{**}\|_2}.$$

To identify significant coupling terms, z-scores were computed according to

$$Zscore_{ij} = \frac{e_{ij} - \mu(e_{**})}{\sigma(e_{**})}$$

where $\mu(e_{**})$ and $\sigma(e_{**})$ denote the mean and standard deviation of the estimated evolutionary coupling terms between all position pairs $i < j$. Z-scores were then converted into P values and corrected for multiple testing using the Benjamini-Hochberg method [54]. We used $P < 0.005$ to detect significant couplings [57], which corresponds to a Bayes factor of 14 typically used in Bayesian inference [58].

To ensure the robustness of the statistical inference, Halsey et al. [59] recently proposed to combine power analysis with P value based statistical analysis. In line with these recommendations, we generated 1000 MSAs by sampling sequences with replacement from the original alignment. For each resampled MSA, we inferred the coupling terms and performed the statistical tests as outlined above. In line with Halsey's recommendations, we only reported coupling terms that were significant in at least 95% of the resamplings, i.e. where the respective test power was >0.95. An interactive plot to explore the detailed results of the direct coupling analysis was generated using the EVZoom tool (https://github.com/debbiemarkslab/EVzoom) and can be accessed through http://page.mi.fu-berlin.de/msmith/couplings_integrase.html.

Results

Characteristics of the study population

All 337 individuals were newly diagnosed with HIV-1, ART-naïve, and infected with an HIV-1 subtype B strain. The median age at HIV-1 diagnosis was 32 (inter-quartile range 27–38) years. Most individuals were male (297/337, 88.1%), 11.0% (37/337) were female, and for 0.9% (3/337) the gender was not documented. The most common transmission group was men having sex with men (220/337, 65.3%), followed by 14.8% heterosexual transmission (50/337) and 14.2% injection drug users (48/337). For 5.6% (19/337) of the individuals the risk factor for transmission was unknown.

Frequency of INSTI resistance mutations

As expected, no major INSTI resistance mutations were detected in this ART-naïve study population from the period prior to INSTI release. Moreover, all major INSTI resistance-associated positions were fully conserved, and amino acids at these positions corresponded to the respective INSTI-sensitive consensus B amino acids. However, some minor INSTI resistance mutations and INSTI-selected mutations were identified (Table 2).

Table 2 Overall frequency (%) of minor INSTI resistance mutations and INSTI-selected mutations

	Overall frequency (%) of	
Variant	minor INSTI resistance mutation	INSTI-selected mutation
M50I		12.8
V54I	0.9	
G59E		2.1
L68V	1.2	
L74I	2.7	
L74 M	0.6	
L101I		36.8
S119R	3.0	
T122I		24.3
T124A	25.8	
T124 N		
T125A		17.8
A128T	0.9	
E138D		1.5
V151A	0.3	
V151I	7.4	
M154I		18.4
K156 N	5.3	
E157Q	2.4	
G163E		2.4
G163R	0.3	
V165I		
H171Q		
G193E		10.1
V201I		36.8
I203M		3.3
T206S	6.8	
N222 K		
S230G	0.6	
S230 N	8.0	
V260I	0.3	

Amino acid variability

Eighty-six polymorphic sites (30.9%) were observed within the 278 amino acid positions examined in the HIV-1 integrase. Proportionately, most polymorphic positions were found in the NTD (20/49, 40.8%). There were equal proportions of polymorphic sites within the CCD (47/163, 28.8%) and the CTD (19/66, 28.8%) (Fig. 1). The highest amino acid variability with entropies >0.75 was found in the NTD at positions 11 and 17 as well as in the CCD at positions 72, 101, 119, 122, 124, 125, 154, and 201 (Fig. 2).

The zinc finger motif (H12-H16-C40-C43), the active site (D64-D116-E153), the multimerisation motif (K186-R187-K188), and most sites interacting with the cellular cofactor LEDGF/p75 or within the minimal nonspecific DNA binding region were conserved. Nevertheless, positions 125, 165, 167, 171, 172, and 173, involved in binding LEDGF/p75, were polymorphic with entropies up to 0.96. Likewise, within the minimal nonspecific DNA binding region positions 220–222, 227, 230, 232, 234, 251–256, and 265 exhibited entropies up to 0.66 (Fig. 2).

Amino acid time trends

Twenty-two significant time trends in amino acid frequencies were discovered at 13 polymorphic positions. Proportionally, most of these sites were located in the CCD (9/163, 5.5%), followed by location in the NTD (2/49, 4.1%) and the CTD (2/66, 3.0%) (Fig. 3).

At nine positions (6, 11, 72, 101, 135, 154, 165, 201, 265) an amino acid exchange between two variants was identified. In contrast, at sites 119 and 124 increasing polymorphy was observed since variants decreased in frequency without being replaced by another variant. Finally, variants 122I and 256E increased in frequency without replacing another variant, although a decrease of T122 and D256 was observed that was close to statistical significance (Fig. 3).

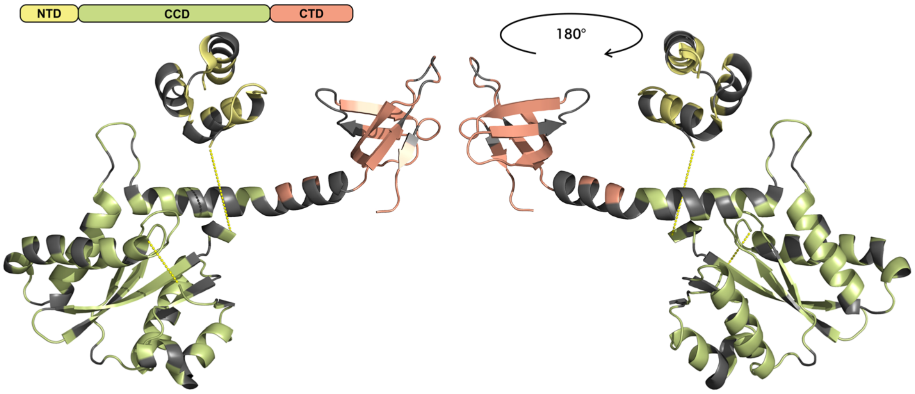

Fig. 1 Ribbon model of the HIV-1 integrase monomer from two viewpoints. The model was constructed in PyMol [67] as an overlay of protein database structures 1ex4 [3] and 1k6y [4]. The NTD (aa 1–49) is coloured yellow, the CCD (aa 50–212) green, and the CTD (aa 213–288) red. Polymorphic sites are coloured grey

Variants L101I, T122I, and V201I with increasing frequency as well as M154I with decreasing frequency were associated with INSTI resistance (Fig. 3).

Amino acid covariation

Overall, 42 significant couplings were identified (Table 3). 10 couplings appeared within the NTD, 10 within the CCD, and two within the CTD. 13 couplings were observed between the NTD and the CCD, five between the CCD and the CTD, and two between the NTD and the CTD. In total, 25 and 28 couplings involved at least one position within the NTD and the CCD, respectively (Fig. 4).

The LEDGF/p75-binding sites 125 and 165 were observed in one coupling each. Nine couplings involved sites 222, 234, 253, 256, or 265 of the minimal nonspecific DNA-binding region (Table 3).

17 amino acid variants with temporal trend were found to covary with other time-trending variants. The individual time trends were generally compatible, i.e. both time trends were concordant if the coupling terms were positive and discordant if the coupling terms were negative (Table 4).

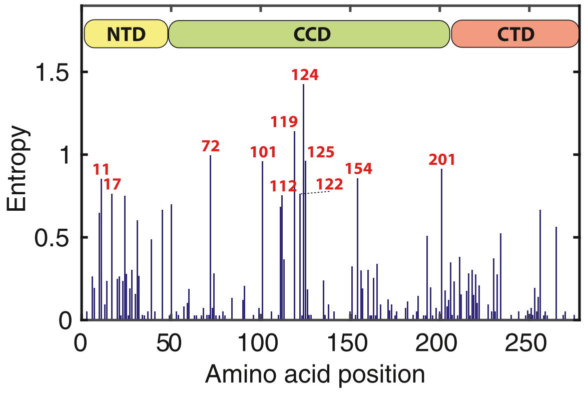

Fig. 2 Levels of Shannon entropy at amino acid positions 1–278 of the HIV-1 integrase. Positions with entropy levels >0.75 and the three functional domains of the integrase are indicated

Two minor INSTI resistance mutations (T124A, E157Q) and five INSTI-selected mutations (L101I, T122I, T125A, M154I, V201I) covaried with non-resistance-associated variants (Table 4).

Discussion

Our analyses could confirm and considerably extend previously published results based on INSTI-naïve [26, 31–35] or ART-naïve [35–40] study populations that either focused on HIV-1 subtype B [31, 36, 38] or included various HIV-1 group M subtypes [26, 32–35, 37, 39, 40]. We restricted our analyses to HIV-1 subtype B strains because these are predominant in Germany [43–45] and because different HIV-1 subtypes have different consensus amino acids at some sites, which can bias the degree of variability when compared to consensus B [33, 35].

We found the highest amino acid variability determined by entropy, time trends, and direct coupling analysis within the CCD, the NTD, and between CCD and NTD. Sites important for enzymatic activity were in general conserved, however, some positions involved in binding the cellular cofactor LEDGF/p75 (sites 125, 165) and within the minimal nonspecific DNA binding region (sites 220, 230, 232, 234, 256, 265) were polymorphic with an overall variability ≥5%. Covariation between positions 125, 165, 256, or 265 and other sites was observed, and the DNA-binding site 234 covaried with the DNA-binding site 253. The most frequent substitutions were T125A, V165I, I220L, S230 N, D232E, L234I, D256E, and A265V. All of them occurred within the same biochemical class of amino acids, with the exception of T125A that represents a switch from a hydrophilic to a hydrophobic amino acid.

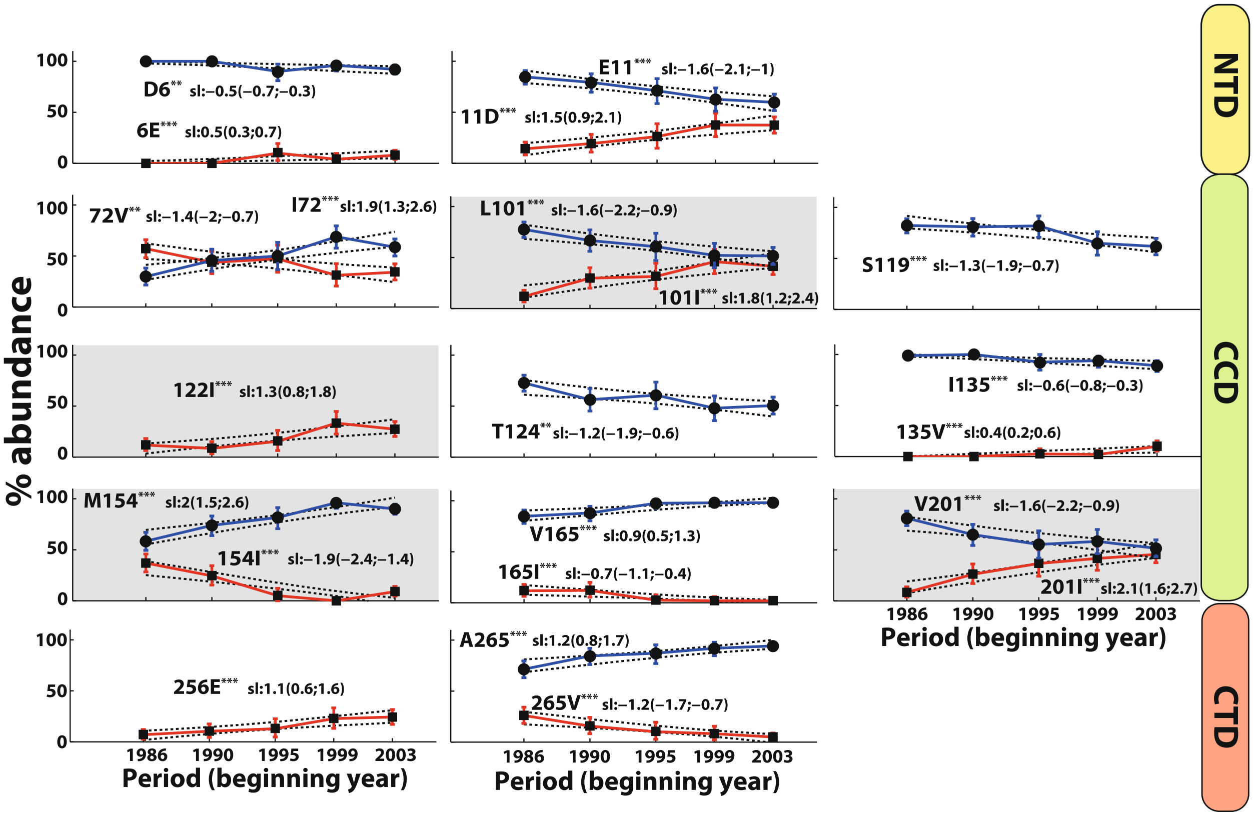

Fig. 3 Amino acid variants with significant trends in frequency over time. The median frequencies of variants corresponding to the consensus B reference sequence within each period are indicated by filled circles with blue error bars indicating the 5th–95th percentile ranges. The median frequencies of variants differing from the consensus B reference sequence within each period are indicated by filled squares with red error bars indicating the 5th–95th percentile ranges. The median predicted slopes are depicted in blue or red solid lines. The 5th–95th percentile ranges of the model-predicted trends are indicated by dotted lines. Predicted slopes sl are given in units %/year. Positions are grouped by the three functional domains of the integrase. Positions showing trends in INSTI resistance-associated variants are shaded grey. *** time trend is significant at the $P < 0.01$ level, ** time trend is significant at the $P < 0.05$ level

The effect of this switch is not known and should be investigated experimentally. A time trend in frequency was observed for variants T125A, D256E, and A265V. The knowledge about the variability of the integrase should be taken into account for the design of genotypic resistance assays.

Most time trends were based on an exchange of two amino acids, however, a general diversification was observed at sites 119 and 124. 17 out of 22 amino acid variants with increasing or decreasing frequency covaried among each other. In general, we observed a concordant time trend for pairs with a positive direct coupling term and a discordant time trend for pairs with a negative direct coupling term. Exceptions to this rule were couplings between positions 154–265 and 201–256 (Table 4). The time trends for the individual variants 154I–A265, M154-265 V, and V201-256E were discordant despite positive direct coupling terms. The reason for this may be unidirectional coupling, i.e. 154I requires the presence of A265, but not vice versa. Likewise, the time trends of M154-A265, 154I-265 V, and 201I-256E were concordant despite negative direct coupling terms.

The prevailing concordance of significant time trends and significant couplings in our study suggests the selection of coevolving epistatic clusters. However, due to the transmission bottlenecks [18], genetic drift may be another viable explanation for the observed time trends in the frequency of certain amino acid variants. The role of genetic drift in HIV evolution is debated and has been quantified to some extent at the level of intra-patient evolution and for known transmission pairs rather than on inter-patient level [60, 61]. Genetic drift on inter-patient level requires inheritance of selectively neutral substitutions. Large parts of the integrase may be under negative selection to maintain enzymatic functionality; nevertheless, particular positions and certain substitutions of the integrase may be selectively neutral. Therefore, we considered the possibility of genetic drift for all time-trending substitutions by assessing whether there

Table 3 Evolutionary coupling terms for covarying positions within the HIV-1 integrase

position *i*	position *j*	ec*ij*
6	10	0.0823
10	17	0.0623
10	72	0.1550
10	122	0.0994
11	17	0.0668
11	21	0.0847
11	24	0.0924
11	31	0.0883
11	111	0.0858
11	119	0.0995
11	154	0.0500
17	31	0.0516
17	119	0.0562
17	124	0.0755
23	201	0.0721
24	25	0.0717
24	45	0.0543
24	72	0.0554
28	39	0.0826
31	154	0.0362
39	154	0.0449
39	201	0.1201
39	256	0.0559
39	265	0.0411
45	160	0.0757
72	265	0.0897
101	111	0.0738
101	112	0.0677
101	154	0.0740
112	124	0.0567
112	201	0.0612
119	122	0.2792
124	125	0.0466
124	211	0.0589
154	165	0.1792
154	256	0.0488
154	265	0.0647
157	160	0.1177
201	256	0.0781
201	265	0.0641
219	222	0.0730
234	253	0.0679

Evolutionary coupling terms ec*ij* between position *i* and *j* with $P < 0.005$ and power > 0.95 are given

was evidence for (i) inheritance of the time-trending substitutions and (ii) whether the time-trending substitutions were selectively neutral. Ad (i): We statistically compared the mean patristic distance of random sequences versus the mean patristic distance of sequences carrying a specific time-trending substitution to investigate if time-trending substitutions appeared more frequently in phylogenetically closely related sequences (Additional file 1: Table S1). Sequences carrying the time-trending substitutions 72 V, 154I, 165I, and 265 V had significantly smaller mean patristic distances than random, by this indicating inheritance. Interestingly, all of these substitutions decreased in frequency over time. Ad (ii): First, we performed Tajima's D test [60, 61] with a result of D = −1.44, by this indicating negative selection. Next, we calculated the ratio of nonsynonymous over synonymous mutations (dN/dS ratio) [60, 61], finding that most regions of the integrase were under strong negative selection, including sites 72, 154, 165, and 265 (Additional file 2: Figure S1). In summary, we could not observe a clear contribution of genetic drift to the time trends of the examined substitutions.

By using Sanger sequencing of bulk RT-PCR products with a sensitivity of approximately 30% [62, 63] and excluding ambiguous amino acids in our analyses we could only investigate the major virus variant from each patient sample. Minor variants and linkage between minor variants can only be investigated by using more sensitive techniques like single genome sequencing (SGS) or next generation sequencing (NGS) [63–65]. To minimize the probability that technical errors during RT-PCR and Sanger sequencing lead to false positive predictions with regard to coupling terms, we combined our direct coupling analysis with a power analysis, which essentially requires that an amino acid pair has to be present in multiple sequences to be repeatedly identified by direct coupling analysis. Recently, the use of covariation methods as a measure of coevolution has been questioned by Talavera et al. [66]. Based on a computational study, the authors point out that a strong covariation signal is caused by a low evolutionary rate. We therefore assessed our results accordingly but could not find a relation between the rarity of pairwise substitutions and high coupling terms or the occurrence of single substitutions in couplings (Additional file 3: Figure S2).

16 minor INSTI resistance mutations and 11 INSTI-selected mutations were observed as naturally occurring in our ART-naïve study population, which originated from the time prior to INSTI approval. Among these resistance-associated variants, three increased in frequency over time and seven covaried with non-resistance-associated variants. The complex

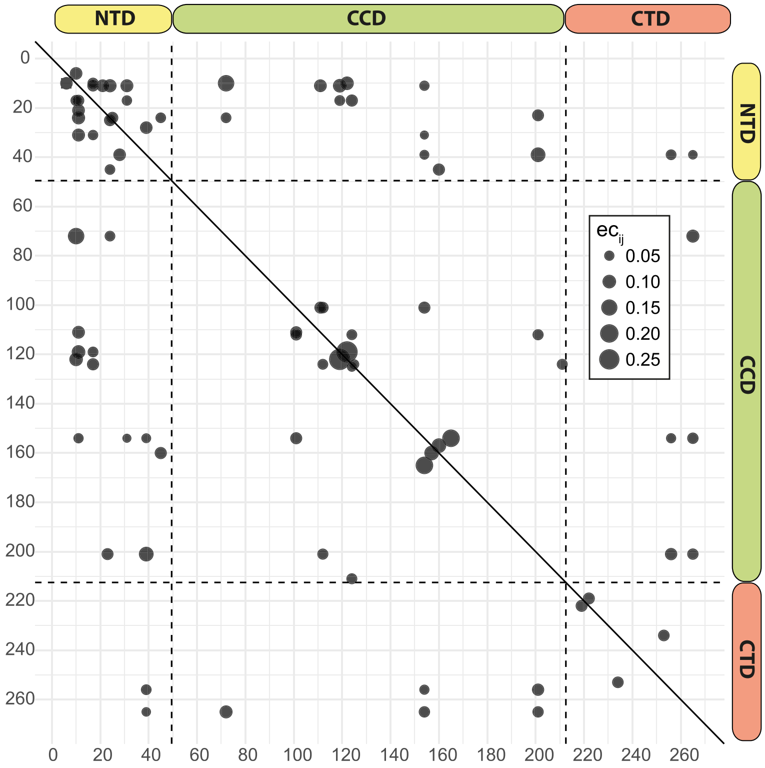

Fig. 4 Significant couplings within the HIV-1 integrase. The size of the dots indicate the values of the evolutionary coupling terms ec*ij* and are plotted mirrored for position *i* against position *j*. The dashed horizontal and vertical lines seperate the three functional domains of the integrase

interdependent evolution of these mutations might control enzymatic activity and replication capacity independent of selective pressure through INSTIs at the inter-patient level. Indeed, accessory drug resistance mutations that compensate viral fitness are often already polymorphic in drug-sensitive HIV-1, suggesting that these mutations may naturally enhance viral fitness and virulence with progression of the HIV-1 epidemic [21, 22]. INSTI-independent linkage between non-resistance-associated sites and resistance-associated sites or sites targeted by INSTIs can affect the selection of resistance mutations in the presence of INSTIs. This knowledge should be taken into account for the improvement of resistance prediction algorithms as well as for the development and preclinical evaluation of new INSTIs and ALLINIs. Deeper analyses of the observed resistance-associated variants are needed to evaluate their clinical relevance. In particular, those with naturally increasing frequencies that were linked to covariation should be investigated, i.e. L101I, T122I, and V201I. The absence of major INSTI resistance mutations in our ART-naïve study population underscores the suitability of INSTIs for first-line antiretroviral regimens.

Because the analysed dataset was rather small (n = 337), our results may require further validation from the analysis of larger, independent datasets. Due to the relatively small number of samples, some of our results might not have reached statistical significance, e.g. the temporal trend of T122 and D256. Generally, given the small sample size, overrepresentation of almost identical sequences (i.e. from transmission chains) may profoundly bias any downstream analysis of time trends and covariation patterns. To assess whether our analyses were affected by such sampling bias, we additionally performed them using a reduced dataset in which clusters of closely related sequences were replaced by one representative only. The results obtained from the reduced dataset confirmed all results obtained from the full dataset.

Conclusions

Our aim was to analyse the molecular evolution of the HIV-1 integrase prior to the approval of INSTIs and, thus, INSTI selective pressure at the inter-patient level. We found significant time trends in the frequency of certain amino acid variants, suggesting ongoing adaptation of the enzyme. Upon closer inspection, we found that amino acid variants with significant time trends covaried with other time-trending variants. Such a linkage may impose constraints that determine the evolutionary

Table 4 Direct coupling terms for specific amino acid variants at covarying positions within the HIV-1 integrase

Position *i*	Amino acid α	Position *j*	Amino acid β	*eij* (α, β)	Time trend	Minor INSTI resistance mutation	INSTI-selected mutation
10	D	122	I	0.0683			*
10	E	122	I	−0.0753			*
11	D	119	S	−0.0940	*		
11	D	154	I	−0.0527	*		*
11	D	154	M	0.0335	*		
11	E	119	S	0.0909	*		
11	E	154	I	0.0577	*		*
11	E	154	M	−0.0368	*		
17	N	124	A	0.0690		*	
17	S	124	A	−0.0598		*	
23	A	201	I	−0.0490			*
23	A	201	V	0.0525			*
23	V	201	I	0.0432			*
23	V	201	V	−0.0443			*
31	I	154	I	−0.0438			*
31	V	154	I	0.0352			*
39	C	154	I	−0.0329			*
39	C	201	I	0.0776			*
39	C	201	V	−0.0725			*
39	S	154	I	0.0384			*
39	S	201	I	−0.0859			*
39	S	201	V	0.0828			*
72	I	265	A	0.0587	*		
72	I	265	V	−0.0612	*		
72	V	265	A	−0.0622	*		
72	V	265	V	0.0676	*		
101	I	111	K	−0.0743			*
101	I	111	T	0.0334			*
101	I	154	I	−0.0462	*		*
101	I	154	M	0.0645	*		*
101	L	154	I	0.0606	*		*
101	L	154	M	−0.0695	*		
112	A	124	A	0.0205		*	
112	I	124	A	0.0205		*	
112	I	201	I	0.0482			*
112	I	201	V	−0.0538			*
112	T	124	A	−0.0447		*	
112	T	201	I	−0.0483			*
112	T	201	V	0.0490			*
119	G	122	I	0.0576			*
119	P	122	I	0.1303			*
119	S	122	I	−0.1695	*		*
119	T	122	I	−0.0194			*

Table 4 Direct coupling terms for specific amino acid variants at covarying positions within the HIV-1 integrase *(Continued)*

Position *i*	Amino acid α	Position *j*	Amino acid β	*eij* (α, β)	Time trend	Minor INSTI resistance mutation	INSTI-selected mutation
124	A	125	T	0.0266		*	
124	A	125	V	−0.0288		*	
124	A	211	K	−0.0568		*	
124	A	211	R	0.0263		*	
124	A	211	T	0.0294		*	
124	S	125	A	0.0244			*
124	T	125	A	−0.0205			*
154	I	165	I	0.0893	*		*
154	I	165	V	−0.0976	*		*
154	I	256	D	0.0333			*
154	I	256	E	−0.0369	*		*
154	I	265	A	0.0520	*		*
154	I	265	V	−0.0484	*		*
154	M	165	I	−0.1011	*		
154	M	165	V	0.1118	*		
154	M	256	E	0.0525	*		
154	M	265	A	−0.0496	*		
154	M	265	V	0.0482	*		
157	Q	160	K	−0.0632		*	
157	Q	160	Q	0.0649		*	
201	I	256	D	0.0623			*
201	I	256	E	−0.0611	*		*
201	I	265	A	0.0484	*		*
201	I	265	V	−0.0466	*		*
201	V	256	E	0.0604	*		
201	V	265	A	−0.0535	*		
201	V	265	V	0.0547	*		

Couplings with time-trending amino acid variants, minor INSTI resistance mutations, and INSTI-selected mutations are indicated by a star

trajectory of the integrase and that influence the selection of INSTI resistance mutations. Our results can help in evaluating the resistance potential of naturally occurring polymorphisms and in understanding the development of resistance in the presence of INSTIs.

Additional files

Additional file 1: Table S1. Statistics assessing whether time-trending substitution could have resulted from inheritance. (DOCX 11 kb)

Additional file 2: Figure S1. dN/dS ratio for different regions of the HIV-1 integrase. The red line depicts the median and the grey shaded area the 25 to 75 percentile range. Black dots mark time-trending positions, i.e. codons 6, 11, 72, 101, 119, 122, 124, 135, 154, 165, 201, 256, 265. A sliding window with window size 5 was applied. (EPS 542 kb)

Additional file 3: Figure S2. Relation between significant coupling and frequency of single/pairwise substitutions. a) Relation between the relative frequencies of substitution pairs within the MSA and the strength of their coupling values (only significant couplings are considered). b) Relation between the relative frequencies of single substitutions within the MSA and their absolute frequency in significant couplings. (ZIP 133 kb)

Acknowledgments
We are very grateful to all of the study patients and medical doctors who participate in the German HIV-1 Seroconverter Study. We thank Angelika Zach for study assistance, Steve Norley for proofreading of the manuscript, and the sequencing facility of the RKI.

Funding
Funding of the German HIV-1 Seroconverter Study by the BMG (Federal Ministry of Health) and BMBF (Federal Ministry of Education and Research) in the framework of various research projects between 1998 and 2012 is greatly acknowledged. KPY, MRS, and MvK acknowledge funding from the BMBF through project 031A307 and the Einstein centre ECMATH, project CH20.

Availability of data and materials
The sequence data was deposited at GenBank and assigned the accession numbers KM677336 to KM677672.

Authors' contributions

KM: designed the study, determined the sequences, performed the analyses, wrote the manuscript. KPY: performed the analyses, contributed to writing of the manuscript. MRS: performed the analyses, contributed to writing of the manuscript. SS: validated the genotyping assay. SF: determined the sequences, performed the analyses. BB: collected the epidemiological data. OH: collected the epidemiological data. NB: contributed reagents, contributed to writing of the manuscript. MvK: performed the analyses, contributed to writing of the manuscript. CK: designed the study, contributed to writing of the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

The study protocol and the informed consent procedure of the German HIV-1 Seroconverter Study was initially approved in 2005 by the Ethics Committee of the Charité, University Medicine Berlin (EA2/105/05), with approval amended and confirmed in 2013.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]HIV and other Retroviruses, Robert Koch Institute, Berlin, Germany. [2]Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany. [3]HIV/AIDS, STI and Blood-borne Infections, Robert Koch Institute, Berlin, Germany.

Received: 12 April 2017 Accepted: 31 October 2017
Published online: 14 November 2017

References

1. Ceccherini-Silberstein F, Malet I, D'Arrigo R, Antinori A, Marcelin AG, Perno CF. Characterization and structural analysis of HIV-1 integrase conservation. AIDS Rev. 2009;11:17–29.
2. Bushman FD, Engelman A, Palmer I, Wingfield P, Craigie R. Domains of the integrase protein of human immunodeficiency virus type 1 responsible for polynucleotidyl transfer and zinc binding. Proc Natl Acad Sci U S A. 1993;90:3428–32.
3. Chen JC, Krucinski J, Miercke LJ, Finer-Moore JS, Tang AH, Leavitt AD, et al. Crystal structure of the HIV-1 integrase catalytic core and C-terminal domains: a model for viral DNA binding. Proc Natl Acad Sci U S A. 2000;97:8233–8.
4. Wang JY, Ling H, Yang W, Craigie R. Structure of a two-domain fragment of HIV-1 integrase: implications for domain organization in the intact protein. EMBO J. 2001;20:7333–43.
5. Lutzke RA, Plasterk RH. Structure-based mutational analysis of the C-terminal DNA-binding domain of human immunodeficiency virus type 1 integrase: critical residues for protein oligomerization and DNA binding. J Virol. 1998;72:4841–8.
6. Cherepanov P, Ambrosio AL, Rahman S, Ellenberger T, Engelman A. Structural basis for the recognition between HIV-1 integrase and transcriptional coactivator p75. Proc Natl Acad Sci U S A. 2005;102:17308–13.
7. Sax PE, DeJesus E, Crofoot G, Ward D, Benson P, Dretler R, et al. Bictegravir versus dolutegravir, each with emtricitabine and tenofovir alafenamide, for initial treatment of HIV-1 infection: a randomised, double-blind, phase 2 trial. Lancet HIV. 2017;4:e154–60.
8. Markowitz M, Frank I, Grant RM, Mayer KH, Elion R, Goldstein D, et al. Safety and tolerability of long-acting cabotegravir injections in HIV-uninfected men (ECLAIR): a multicentre, double-blind, randomised, placebo-controlled, phase 2a trial. Lancet HIV. 2017;4(8):e331–e340.
9. Desimmie BA, Schrijvers R, Demeulemeester J, Borrenberghs D, Weydert C, Thys W, et al. LEDGINs inhibit late stage HIV-1 replication by modulating integrase multimerization in the virions. Retrovirology. 2013;10:57.
10. Christ F, Voet A, Marchand A, Nicolet S, Desimmie BA, Marchand D, et al. Rational design of small-molecule inhibitors of the LEDGF/p75-integrase interaction and HIV replication. Nat Chem Biol. 2010;6:442–8.
11. Zhang FH, Debnath B, Xu ZL, Yang LM, Song LR, Zheng YT, et al. Discovery of novel 3-hydroxypicolinamides as selective inhibitors of HIV-1 integrase-LEDGF/p75 interaction. Eur J Med Chem. 2017;125:1051–63.
12. Leslie A, Kavanagh D, Honeyborne I, Pfafferott K, Edwards C, Pillay T, et al. Transmission and accumulation of CTL escape variants drive negative associations between HIV polymorphisms and HLA. J Exp Med. 2005;201:891–902.
13. Allen TM, Altfeld M, Geer SC, Kalife ET, Moore C, O'Sullivan KM, et al. Selective escape from CD8+ T-cell responses represents a major driving force of human immunodeficiency virus type 1 (HIV-1) sequence diversity and reveals constraints on HIV-1 evolution. J Virol. 2005;79: 13239–49.
14. Nijhuis M, van Maarseveen NM, Boucher CA. HIV protease resistance and viral fitness. Curr Opin HIV AIDS. 2007;2:108–15.
15. Chen L, Perlina A, Lee CJ. Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. J Virol. 2004;78:3722–32.
16. Meixenberger K, Hauser A, Jansen K, Yousef KP, Fiedler S, von Kleist M, et al. Assessment of ambiguous base calls in HIV-1 pol population sequences as a biomarker for identification of recent infections in HIV-1 incidence studies. J Clin Microbiol. 2014;52:2977–83.
17. Maldarelli F, Kearney M, Palmer S, Stephens R, Mican J, Polis MA, et al. HIV populations are large and accumulate high genetic diversity in a nonlinear fashion. J Virol. 2013;87:10313–23.
18. Joseph SB, Swanstrom R, Kashuba AD, Cohen MS. Bottlenecks in HIV-1 transmission: insights from the study of founder viruses. Nat Rev Microbiol. 2015;13:414–25.
19. Alizon S, von Wyl V, Stadler T, Kouyos RD, Yerly S, Hirschel B, et al. Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load. PLoS Pathog. 2010;6:e1001123.
20. Alizon S, Fraser C. Within-host and between-host evolutionary rates across the HIV-1 genome. Retrovirology. 2013;10:49.
21. Theys K, Deforche K, Vercauteren J, Libin P, van de Vijver DA, Albert J, et al. Treatment-associated polymorphisms in protease are significantly associated with higher viral load and lower CD4 count in newly diagnosed drug-naive HIV-1 infected patients. Retrovirology. 2012;9:81.
22. Theys K, Abecasis AB, Vandamme AM. HIV-1 drug resistance: where do polymorphisms fit in? Future Microbiol. 2013;8:303–6.
23. Li G, Piampongsant S, Faria NR, Voet A, Pineda-Pena AC, Khouri R, et al. An integrated map of HIV genome-wide variation from a population perspective. Retrovirology. 2015;12:18.
24. Vrancken B, Rambaut A, Suchard MA, Drummond A, Baele G, Derdelinckx I, et al. The genealogical population dynamics of HIV-1 in a large transmission chain: bridging within and among host evolutionary rates. PLoS Comput Biol. 2014;10:e1003505.
25. Brenner BG, Lowe M, Moisi D, Hardy I, Gagnon S, Charest H, et al. Subtype diversity associated with the development of HIV-1 resistance to integrase inhibitors. J Med Virol. 2011;83:751–9.
26. Garrido C, Geretti AM, Zahonero N, Booth C, Strang A, Soriano V, et al. Integrase variability and susceptibility to HIV integrase inhibitors: impact of subtypes, antiretroviral experience and duration of HIV infection. J Antimicrob Chemother. 2010;65:320–6.
27. von Kleist M, Metzner P, Marquet R, Schutte C. HIV-1 polymerase inhibition by nucleoside analogs: cellular- and kinetic parameters of efficacy, susceptibility and resistance selection. PLoS Comput Biol. 2012;8:e1002359.
28. Martinez JP, Bocharov G, Ignatovich A, Reiter J, Dittmar MT, Wain-Hobson S, et al. Fitness ranking of individual mutants drives patterns of epistatic interactions in HIV-1. PLoS One. 2011;6:e18375.
29. Hinkley T, Martins J, Chappey C, Haddad M, Stawiski E, Whitcomb JM, et al. A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. Nat Genet. 2011;43:487–9.
30. Theys K, Deforche K, Beheydt G, Moreau Y, van Laethem K, Lemey P, et al. Estimating the individualized HIV-1 genetic barrier to resistance using a nelfinavir fitness landscape. BMC Bioinformatics. 2010;11:409.
31. Lataillade M, Chiarella J, Kozal MJ. Natural polymorphism of the HIV-1 integrase gene and mutations associated with integrase inhibitor resistance. Antivir Ther. 2007;12:563–70.

32. Myers RE, Pillay D. Analysis of natural sequence variation and covariation in human immunodeficiency virus type 1 integrase. J Virol. 2008;82:9228–35.
33. Rhee SY, Liu TF, Kiuchi M, Zioni R, Gifford RJ, Holmes SP, et al. Natural variation of HIV-1 group M integrase: implications for a new class of antiretroviral inhibitors. Retrovirology. 2008;5:74.
34. Varghese V, Liu TF, Rhee SY, Libiran P, Trevino C, Fessel WJ, et al. HIV-1 integrase sequence variability in antiretroviral naive patients and in triple-class experienced patients subsequently treated with raltegravir. AIDS Res Hum Retrovir. 2010;26:1323–6.
35. Rhee SY, Sankaran K, Varghese V, Winters MA, Hurt CB, Eron JJ, et al. HIV-1 protease, reverse transcriptase, and Integrase variation. J Virol. 2016;90:6058–70.
36. Low A, Prada N, Topper M, Vaida F, Castor D, Mohri H, et al. Natural polymorphisms of human immunodeficiency virus type 1 integrase and inherent susceptibilities to a panel of integrase inhibitors. Antimicrob Agents Chemother. 2009;53:4275–82.
37. Sichtig N, Sierra S, Kaiser R, Daumer M, Reuter S, Schulter E, et al. Evolution of raltegravir resistance during therapy. J Antimicrob Chemother. 2009;64:25–32.
38. Ceccherini-Silberstein F, Malet I, Fabeni L, Dimonte S, Svicher V, D'Arrigo R, et al. 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. 2010;65:2305–18.
39. Sierra S, Lubke N, Walter H, Schulter E, Reuter S, Fatkenheuer G, et al. The SnoB study: frequency of baseline raltegravir resistance mutations prevalence in different non-B subtypes. Med Microbiol Immunol. 2011; 200:225–32.
40. Piralla A, Paolucci S, Gulminetti R, Comolli G, Baldanti F. HIV integrase variability and genetic barrier in antiretroviral naive and experienced patients. Virol J. 2011;8:149.
41. Meixenberger K, Yousef KP, Somogyi S, Fiedler S, Bartmeyer B, von Kleist M, et al. Characterization of natural polymorphic sites of the HIV-1 integrase before the introduction of HIV-1 integrase inhibitors in Germany. J Int AIDS Soc. 2014;17:19746.
42. Duwe S, Brunn M, Altmann D, Hamouda O, Schmidt B, Walter H, et al. Frequency of genotypic and phenotypic drug-resistant HIV-1 among therapy-naive patients of the German Seroconverter study. J Acquir Immune Defic Syndr. 2001;26:266–73.
43. Poggensee G, Kucherer C, Werning J, Somogyi S, Bieniek B, Dupke S, et al. Impact of transmission of drug-resistant HIV on the course of infection and the treatment success. Data from the German HIV-1 Seroconverter study. HIV Med. 2007;8:511–9.
44. Bartmeyer B, Kuecherer C, Houareau C, Werning J, Keeren K, Somogyi S, et al. Prevalence of transmitted drug resistance and impact of transmitted resistance on treatment success in the German HIV-1 Seroconverter cohort. PLoS One. 2010;5:e12718.
45. Zu Knyphausen F, Scheufele R, Kucherer C, Jansen K, Somogyi S, Dupke S, et al. First line treatment response in patients with transmitted HIV drug resistance and well defined time point of HIV infection: updated results from the German HIV-1 seroconverter study. PLoS One. 2014;9:e95956.
46. Pineda-Pena AC, Faria NR, Imbrechts S, Libin P, Abecasis AB, Deforche K, et al. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. Infect Genet Evol. 2013;19:337–48.
47. Anstett K, Brenner B, Mesplede T, Wainberg MA. HIV drug resistance against strand transfer integrase inhibitors. Retrovirology. 2017;14:36.
48. Liu TF, Shafer RW. Web resources for HIV type 1 genotypic-resistance test interpretation. Clin Infect Dis. 2006;42:1608–18.
49. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22: 2688–90.
50. Pouran Yousef K, Meixenberger K, Smith MR, Somogyi S, Gromoller S, Schmidt D, et al. Inferring HIV-1 transmission dynamics in Germany from recently transmitted viruses. J Acquir Immune Defic Syndr. 2016; 73:356–63.
51. Hassan AS, Pybus OG, Sanders EJ, Albert J, Esbjornsson J. Defining HIV-1 transmission clusters based on sequence data. AIDS. 2017;31:1211–22.
52. Buslje CM, Santos J, Delfino JM, Nielsen M. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. Bioinformatics. 2009; 25:1125–31.
53. Katchanov J, von Kleist M, Arasteh K, Stocker H. 'Time-to-amphotericin B' in cryptococcal meningitis in a European low-prevalence setting: analysis of diagnostic delays. QJM. 2014;107:799–803.
54. Benjamini Y, Hochberg Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B-Methodological. 1995;57:289–300.
55. Hopf TA, Ingraham JB, Poelwijk FJ, Scharfe CP, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nat Biotechnol. 2017;35:128–35.
56. Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2012;28:184–90.
57. Altman N, Krzywinski M. Points of significance: interpreting P values. Nat Methods. 2017;14:213–4.
58. Sellke T, Bayarri MJ, Berger JO. Calibration of p values for testing precise null hypotheses. Am Stat. 2001;55:62–71.
59. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods. 2015;12:179–85.
60. Shriner D, Shankarappa R, Jensen MA, Nickle DC, Mittler JE, Margolick JB, et al. Influence of random genetic drift on human immunodeficiency virus type 1 env evolution during chronic infection. Genetics. 2004;166:1155–64.
61. Edwards CT, Holmes EC, Pybus OG, Wilson DJ, Viscidi RP, Abrams EJ, et al. Evolution of the human immunodeficiency virus envelope gene is dominated by purifying selection. Genetics. 2006;174:1441–53.
62. Gunthard HF, Wong JK, Ignacio CC, Havlir DV, Richman DD. Comparative performance of high-density oligonucleotide sequencing and dideoxynucleotide sequencing of HIV type 1 pol from clinical samples. AIDS Res Hum Retrovir. 1998;14:869–76.
63. Palmer S, Kearney M, Maldarelli F, Halvas EK, Bixby CJ, Bazmi H, et al. Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. J Clin Microbiol. 2005;43:406–13.
64. Beerenwinkel N, Gunthard HF, Roth V, Metzner KJ. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front Microbiol. 2012;3:329.
65. Schirmer M, Sloan WT, Quince C. Benchmarking of viral haplotype reconstruction programmes: an overview of the capacities and limitations of currently available programmes. Brief Bioinform. 2014;15:431–42.
66. Talavera D, Lovell SC, Whelan S. Covariation is a poor measure of molecular Coevolution. Mol Biol Evol. 2015;32:2456–68.
67. Schrodinger LLC: The PyMOL molecular graphics system, version 1.8. 2015.