Global analysis of sequence diversity within HIV-1 subtypes across geographic regions

Austin Huang[1,2], Joseph W Hogan[3], Sorin Istrail[2], Allison DeLong[3], David A Katzenstein[4] & Rami Kantor*[1]

[1]Division of Infectious Diseases, Brown University, Providence, RI, USA
[2]Center for Computational Molecular Biology, Brown University, Providence, RI, USA
[3]Center for Statistical Sciences, Brown University, Providence, RI, USA
[4]Division of Infectious Diseases, Stanford University, Stanford, CA, USA
*Author for correspondence: The Miriam Hospital, RISE 154, 164 Summit Avenue, Providence, RI 02906, USA
■ Tel.: +1 401 793 4997 ■ Fax: +1 401 793 4709 ■ rkantor@brown.edu

Aims: HIV-1 sequence diversity can affect host immune responses and phenotypic characteristics such as antiretroviral drug resistance. Current HIV-1 sequence diversity classification uses phylogeny-based methods to identify subtypes and recombinants, which may overlook distinct subpopulations within subtypes. While local epidemic studies have characterized sequence-level clustering within subtypes using phylogeny, identification of new genotype–phenotype associations are based on mutational correlations at individual sequence positions. We perform a systematic, global analysis of position-specific *pol* gene sequence variation across geographic regions within HIV-1 subtypes to characterize subpopulation differences that may be missed by standard subtyping methods and sequence-level phylogenetic clustering analyses. **Materials & methods:** Analysis was performed on a large, globally diverse, cross-sectional *pol* sequence dataset. Sequences were partitioned into subtypes and geographic subpopulations within subtypes. For each subtype, we identified positions that varied according to geography using VESPA (viral epidemiology signature pattern analysis) to identify sequence signature differences and a likelihood ratio test adjusted for multiple comparisons to characterize differences in amino acid (AA) frequencies, including minority mutations. Synonymous nonsynonymous analysis program (SNAP) was used to explore the role of evolutionary selection witihin subtype C. **Results:** In 7693 protease (PR) and reverse transcriptase (RT) sequences from untreated patients in multiple geographic regions, 11 PR and 11 RT positions exhibited sequence signature differences within subtypes. Thirty six PR and 80 RT positions exhibited within-subtype geography-dependent differences in AA distributions, including minority mutations, at both conserved and variable loci. Among subtype C samples from India and South Africa, nine PR and nine RT positions had significantly different AA distributions, including one PR and five RT positions that differed in consensus AA between regions. A selection analysis of subtype C using SNAP demonstrated that estimated rates of nonsynonymous and synonymous mutations are consistent with the possibility of positive selection across geographic subpopulations within subtypes. **Conclusion:** We characterized systematic genotypic *pol* differences across geographic regions within subtypes that are not captured by the subtyping nomenclature. Awareness of such differences may improve the interpretation of future studies determining the phenotypic consequences of genetic backgrounds.

The characterization of genetic diversity is central to epidemiological tracking of the expanding HIV epidemic [1–4]. HIV genotypes are organized into clades using the subtyping nomenclature [5], which partitions them into a phylogenetic hierarchy [5]. Subtyping is often used for sequence stratification prior to analysis or as part of the inclusion criteria for sequences in a study [1,5–9].

HIV-1 subtypes are strongly associated with specific geographic regions [4,10]. For example, the globally predominant HIV-1 subtype C has been identified mainly in southern Africa, Ethiopia, Latin America, India and regions in China. However, molecular epidemiological studies have also described sequence clustering within subtypes [11–21]. For example, clustered *env* sequences within subtypes B in Thailand and C in Ethiopia and India have been designated as Thai B/B′, Ethiopia C′ and C-IN, respectively [11–14,17–19,21]. With some exceptions [22–25], these characterizations have largely focused on the *env* gene for its high degree of diversity and implications for

Keywords
■ geography ■ HIV-1 ■ *pol* gene sequences ■ protease ■ reverse transcriptase ■ subtyping

Future Medicine part of fsg

vaccine development, rather than the *pol* gene, which is central to drug-resistance interpretation. Sequence clustering is also used to infer historical links between epidemics in different geographic regions, such as Brazil, South Africa, South America and the UK [26–29]. More recently, phylogeographic methods have been applied to model the spread of such local epidemics within subtypes B, C, F and CRF02_AG populations [30–33]. In some cases, within-subtype clustering has led to sub-subtype definitions, although these designations are limited by the nomenclature standard, which requires full-length genome sequences [5]. Sub-subtypes are currently defined for subtypes A (A1, A2, A3 and A4) [34–36] and F (F1 and F2) [37,38]. A global characterization of within-subtype *pol* heterogeneity according to geographic region has not been reported.

HIV-1 subtypes and recombinants may be associated with various phenotypes, such as drug-resistance evolution [1], disease progression [39], transmission patterns [4] and neuropsychological outcomes [40]. Large-scale analyses to derive associations between genotypic diversity and such phenotypes across subtypes require data from multiple cohorts. Such analyses are facilitated by GenBank® [41] and curated HIV sequence databases [2,101], as well as investigator networks [1,42,43], which include tens of thousands of sequences linked to demographic, clinical and/or laboratory information. Genotypic associations with phenotype and experimental validation of such associations are based on mutations at individual sequence positions [1,6,9,44]. In the context of genotypic–phenotypic association studies, it is necessary to distinguish between mutational differences at individual positions and phylogenetic clustering. Phylogeny is effective for characterizing sequence-level clustering by aggregating variation across the entire sequence [45], but not at specific codons. These analyses depend on within-subtype mutation frequencies and geographic clustering. For example, mutations may increase in frequency in a population owing to migration, transmission bottlenecks or host selection in geographic regions. By contrast, some mutations that arise from a low genetic barrier may be similar across geographic regions. Sequence availability in databases across geographic regions is unequal, and it is important to distinguish true phenotypic mutational associations from artifactual correlations related to geographic clustering and unequal sampling within subtypes. Lack of awareness of such within-subtype diversity may lead to erroneous conclusions regarding genotype–phenotype associations.

In this study, we characterize sequence differences within HIV-1 subtypes using a large global *pol* gene sequence dataset from treatment-naive patients, by identifying within-subtype signature and nonsignature mutations at codons with geography-dependent variation; characterizing those positions, with a specific focus on India and South Africa, for which large datasets of *pol* sequences of the globally predominant subtype C are available; and examining whether within-subtype differences are consistent with evolutionary selection.

Materials & methods

Sequence alignment & preprocessing

Protease (PR; positions 1–99) and reverse transcriptase (RT; positions 1–240) nucleic acid sequences from a large, curated dataset of untreated HIV-infected patients [1] were translated to amino acids (AA), aligned with MUSCLE [46] and assigned a subtype using a previously described bootscanning subtyping algorithm [1]. To increase the number of sequences available per geographic region, we grouped countries of sequence origin into geographic regions as defined in the WHO survey on the regional distribution of HIV-1 (Table 1) [47].

Identification of signature sequence differences between geographic regions within subtypes

We first identified positions in which the most common AA differs across geographic regions within subtype populations. In order to accomplish this we developed a C++ program to automate multiple VESPA (viral epidemiology signature pattern analysis) [48] comparisons for all pairs of available geographic regions within each subtype. VESPA identifies signature sequence differences, defined as positions for which the most common AA differs between a query and a reference alignment. We use the term 'signature' to distinguish such positions from positions that only exhibit differences in AA distributions. We refer to the most common AAs at sequence positions as signature AA mutations, and to minority AAs as nonsignature AA mutations.

Identification of differences in AA distributions between geographic regions within subtypes, including nonsignature mutations

Signature sequence differences capture discordance in the most common AA between the sequences of viruses of the same subtype from different geographic regions. However, sequences

can differ in the relative frequencies of minority AAs at both signature and nonsignature positions. We extended the VESPA analysis and investigated such differences across geographic subpopulations within subtypes. For each HIV-1 subtype, *pol* gene (PR or RT) and sequence position, we tested whether the distribution of observed AAs (as opposed to only the most common AA, as per VESPA) differed by geographic region. We then identified geography-dependent positions, defined as sequence positions that have significantly distinct proportions of AA among sequences of the same subtype from different geographic regions. An omnibus test to determine whether the rate of occurrence of any AA differs between any two geographic regions for which sequences were available was performed. We compared a saturated log-linear regression model, including terms for AA, geographic region and their interaction to a reduced model, including only AA and geographic region. Log-linear models are a common approach to characterizing association between categorical variables (in this case, geographic region and the AA frequencies) [49,50]. A log-likelihood ratio test was used to determine whether the geography-independent model had a significantly worse fit than the fully saturated geography-dependent model. Since the test was repeated for each of 99 PR and 240 RT positions, we accounted for multiple hypothesis testing by adjusting p-values for each gene using a Benjamini–Hochberg false discovery rate [51] of 0.05.

Using the same statistical analysis, we performed a focused comparison of the globally predominant subtype C sequences from India and South Africa [47], for which large datasets are available. We compared sequences from these two regions to determine which specific AAs occur differently at the geography-dependent positions.

Characterization of geography-dependent positions: sequence conservation & drug resistance

We characterized geography-dependent positions by examining first, their degree of diversity by computing the Shannon entropy [52], a measure of diversity of the AA distribution at each sequence position; and second, their occurrence at positions unambiguously associated with drug resistance. The following positions that appear on all five major HIV drug-resistance mutation lists (ANRS/IAS-USA/REGA/Stanford/SDRM) are considered to be resistance associated: PR positions 24, 30, 32, 46, 47, 48, 50, 53, 54, 73, 76, 82, 84, 88 and 90; and RT positions

Table 1. Geographic distribution of sequences according to subtype and protein.

Geographic region	Subtype A		Subtype B		Subtype C		Subtype D		Subtype F		Subtype G		CRF01_AE		CRF02_AG		Total	
	PR	*RT*	*PR*	*RT*	*PR*	*RT*	*PR*	*(RT)*	*(PR)*	*(RT)*	*PR*	*(RT)*	*PR*	*(RT)*	*PR*	*RT*	*PR*	*RT*
Caribbean			99	79													99	79
E Africa[†]	254	109			50	46	168	79									472	234
E Asia			49	57													49	57
India					122	110											122	110
Latin America			442	464	74	85			87	38							603	587
N America			775	1033	45												820	1033
Other											24						24	
S Africa					201	228											201	228
S/SE Asia[‡]													109	85			109	85
W Africa							23				25		31		349	160	428	160
W Europe	69	35	612	881	162	102	20				65	48	23		105	71	1056	1137
Total	323	144	1977	2514	654	571	211	79	87	38	134	48	163	85	454	231	3983	3710

Empty cells represent subtype/protein/region combinations for which sequences were unavailable. Subtype categories without multiple WHO regions that were not available for analysis are indicated in parentheses.
[†]*Excluding Ethiopia.*
[‡]*Excluding India.*
E: East; N: North; PR: Protease; RT: Reverse transcriptase; S: South; SE: Southeast; W: West.

41, 65, 67, 69, 70, 74, 100, 101, 103, 106, 115, 151, 181, 184, 188, 190, 210, 215, 219, 225 and 230 [53]. These positions account for 15 and 21% of the total positions in the PR and RT sequences, respectively.

Evolutionary selection

We examined whether there was evidence of positive selection at geography-dependent positions by performing an exploratory nonsynonymous–synonymous substitution (dn–ds) analysis of subtype C sequences using synonymous nonsynonymous analysis program (SNAP)[48]. Positions for which the rate of nonsynonymous substitutions exceeds the rate of synonymous substitutions (dn–ds >0, or equivalently dn/ds >1) are suggestive of positive selection. dn–ds analysis was performed within each geographic region of subtype C. Owing to the fact that SNAP is unable to analyze large datasets, we conducted repeated (100 bootstrap replicates) analyses of 100 sequences, sampled with replacement, to calculate a 95% CI for the dn–ds estimates.

Finalized sequence dataset

Out of a total of 7693 PR and RT sequences, 3896 out of 3983 PR sequences of subtypes B, C, D, G, CRF01_AE and CRF02_AG; and 3460 out of 3710 RT sequences of subtypes A, B, C and CRF02_AG spanned more than one geographic region per subtype and were analyzed. Subtypes with data from only one region were not analyzed (Table 1). This study was approved by the Lifespan Institutional Review Board.

Results

Signature differences within subtypes

A total of 56 geographic population comparisons were performed within the five most common pure subtypes and two circulating recombinant forms. Signature differences were found at 11 PR positions (12, 13, 15, 19, 35, 37, 41, 63, 64, 67 and 93) and 11 RT positions (39, 60, 121, 122, 123, 135, 162, 177, 200, 207 and 211) within subtypes B, C, D, G and CRF01_AE (Tables 2 & 3). The number of signature differences for any particular comparison ranged from zero to five positions and the majority of signature differences were observed within subtypes B and C. Almost all positions noted in Tables 2 & 3 exhibit signature differences between multiple geographic region contrasts, with the exception of PR 67, which was unique to the western Africa/western Europe comparison in subtype G.

The signature differences identified in the 'Latin America' subtype C analyses are consistent with, and extend, previous reports from Brazil [22,54,55]. We verify PR signature mutations 12T, 19L, 37K and 41N, and identify additional signature mutations, including PR 63L and RT mutations 39D, 60V, 121D, 123D, 162S and 207Q. Some of these signature differences (PR 63L and RT 60V, 121D, 123D and 162S) result only from contrasts with subtype C sequences of Indian origin, which are distinct from all other regions at these positions.

Differences in AA distributions among geographic regions within subtypes, including nonsignature mutations

As extension to the VESPA analysis, we identified positions demonstrating differences in AA frequencies (including nonsignature mutations) within-subtypes across geographic regions (Figure 1). Thirty-six distinct PR positions and 80 distinct RT positions exhibited geographic variation across subtypes. These differences included all signature positions from Tables 2 & 3 and an additional 25 PR and 69 RT positions.

Figure 1A shows sequence positions in PR for which AA distributions differed by geographic region. Most (35 out of 36) geography-dependent positions in PR occur at loci not associated with drug resistance. The sequences of subtypes B, C, D, G, CRF01_AE and CRF02_AG had 23, 18, six, three, five and one positions that exhibited geography-dependent variation, respectively, of which, only position 82 in subtype C was at a drug-resistance position. Fourteen out of 36 geography-dependent positions appear in multiple subtypes; for example, position 93 is geography-dependent in subtypes B, C, D, G and CRF02_AG. Twenty two out of 36 positions only appear in one subtype. These geography-dependent positions occur at a wide range of both conserved and variable loci (Figure 1A). For example, PR position 25, which is region-dependent in subtype B, is close to the enzyme active site and is a relatively conserved sequence position (Shannon entropy [S] = 0.09 nats). By contrast, position 63, a known polymorphic site in PR that is region-dependent in subtypes B and C, is highly variable (S = 1.53 nats).

Figure 1B shows sequence positions in RT at which AA distributions exhibited a statistically significant dependence on geographic region. Subtype B had 65 positions that exhibited geography-dependent variation, nine of which were resistance-associated, subtype C had 32, one of which was resistance-associated; subtypes B and C shared 17 geography-dependent positions,

Table 2. Protease signature positions according to subtype and geographic region.

Subtype	Region 1	Region 2	12	13	15	19	35	37	41	63	64	67	93
A	E Africa	W Europe											
B	E Asia	Caribbean								T/P			L/I
B	Latin America	Caribbean											
B	Latin America	E Asia								P/T			I/L
B	N America	Caribbean											
B	N America	E Asia								P/T			I/L
B	N America	Latin America											
B	W Europe	Caribbean											
B	W Europe	E Asia								P/T			I/L
B	W Europe	Latin America											
B	W Europe	N America											
C	E Africa	India								L/P			
C	E Africa	N America											
C	India	N America								P/L			
C	Latin America	E Africa	T/S			L/I		K/N	N/K				
C	Latin America	India	T/S			L/I		K/N	N/K	L/P			
C	Latin America	N America	T/S			L/I		K/N	N/K				
C	Latin America	S Africa	T/S			L/I		K/N	N/K				
C	Latin America	W Europe	T/S			L/I		K/N	N/K				
C	S Africa	E Africa											
C	S Africa	India								L/P			
C	S Africa	N America											
C	W Europe	E Africa											
C	W Europe	India								L/P			
C	W Europe	N America											
C	W Europe	S Africa											
D	E Africa	W Africa		V/I	I/V		E/D			L/P	V/I		
D	E Africa	W Europe											
D	W Africa	W Europe		I/V	V/I		D/E			P/L	I/V		
G	W Africa	W Europe										E/C	
CRF01_AE	S/SE Asia	W Africa						N/D					
CRF01_AE	S/SE Asia	W Europe											
CRF01_AE	W Africa	W Europe						D/N					
CRF02_AG	W Africa	W Europe											

Region 1 and Region 2 columns indicate the reference and comparison populations in the viral epidemiology signature pattern analysis contrast. Numbered columns indicate sequence positions where signature differences were detected by viral epidemiology signature pattern analysis. Signature mutations at these positions are described as most frequent amino acid in region 1/most frequent amino acid in region 2. If a position was not a signature position for a particular contrast, the table cell is empty.
E: East; N: North; S: South; SE: Southeast; W: West.

while the remaining positions were exclusively geography-dependent in either subtype B or C. Overall, ten out of 80 geography-dependent positions occurred at resistance-associated loci (positions 41, 67, 69, 70, 100, 101, 184, 210, 215 and 225). Fewer sequences were available for subtype A and CRF02_AG, and, after adjusting for multiple comparisons, none of the sequence positions were above the threshold of significance. Similar to PR, these geography-dependent positions occur at a wide range of both conserved and variable positions.

Table 3. Reverse transcriptase signature positions according to subtype and geographic region.

Subtype	Region 1	Region 2	39	60	121	122	123	135	162	177	200	207	211
A	E Africa	W Europe											
B	E Asia	Caribbean				E/K				D/E	I/T	Q/E	R/K
B	Latin America	Caribbean						T/I		D/E		Q/E	R/K
B	Latin America	E Asia				K/E		T/I			T/I		
B	N America	Caribbean								D/E		Q/E	R/K
B	N America	E Asia				K/E					T/I		
B	N America	Latin America						I/T					
B	W Europe	Caribbean								D/E		Q/E	R/K
B	W Europe	E Asia				K/E					T/I		
B	W Europe	Latin America						I/T					
B	W Europe	N America											
C	E Africa	W Europe											
C	India	E Africa	D/E	I/V	Y/D		D/S		A/S				
C	India	Latin America		I/V	Y/D				A/S			E/Q	
C	India	S Africa	D/E	I/V	Y/D		D/G		A/S				
C	India	W Europe	D/E	I/V	Y/D		D/S		A/S				
C	Latin America	E Africa	D/E				D/S					Q/E	
C	Latin America	W Europe	D/E				D/S					Q/E	
C	S Africa	E Africa					G/S						
C	S Africa	Latin America	E/D				G/D					E/Q	
C	S Africa	W Europe					G/S						
CRF02_AG	W Africa	W Europe											R/K

Region 1 and Region 2 columns indicate the reference and comparison populations in the viral epidemiology signature pattern analysis contrast. Numbered columns indicate sequence positions where signature differences were detected by viral epidemiology signature pattern analysis (VESPA). Signature mutations at these positions are described as most frequent amino acid in region 1/most frequent amino acid in region 2. If a position was not a signature position for a particular contrast, the table cell is empty.
E: East; N: North; S: South; SE: Southeast; W: West.

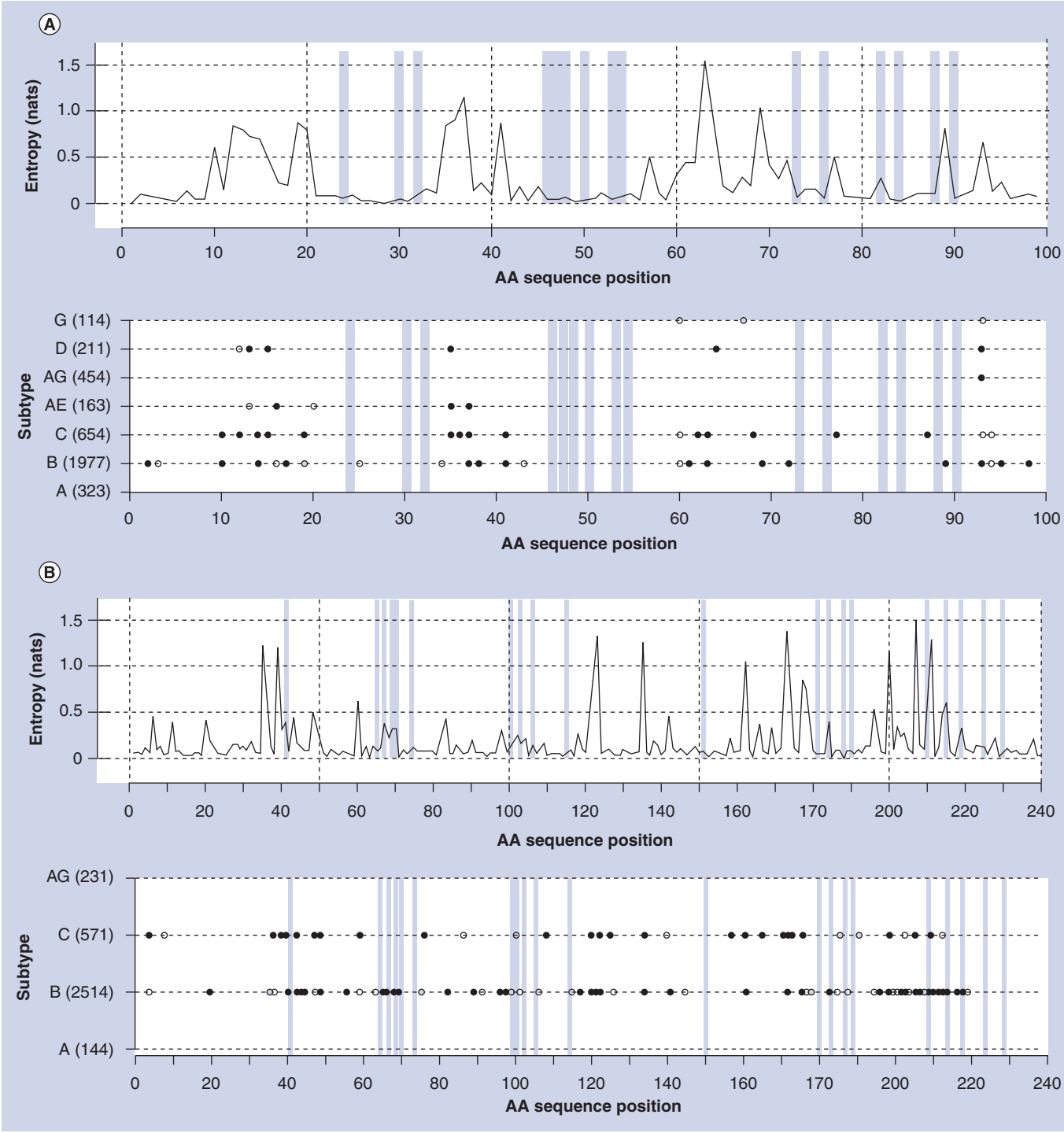

Figure 1. Geography-dependent positions according to subtype. Graphs show linked Shannon entropy (top) and geography-dependent positions (bottom) in protease **(A)** and reverse transcriptase **(B)**. Numbers in parentheses on the vertical axis indicate number of available sequences. Sequence positions that are statistically significant for geography-dependent amino acid distributions are shown as circles. The color scale ranges continuously from white (false-discovery rate = 0.05) to black (false-discovery rate <0.005). Only subtypes and sequence positions for which significant positions were identified are shown.
AA: Amino acid.

Characterization of geography-dependent positions in subtype C from India & South Africa

The globally predominant HIV-1 subtype (subtype C), is the major subtype in India and South Africa. We compared sequences from these two regions and examined changes in AA frequency at positions with different AA distributions (Table 4). Changes in AA distribution between Indian and South African subtype C

Table 4. Protease and reverse transcriptase positions with different amino acid distributions between subtype C sequences from India and South Africa.

Position	Amino acid with prevalence increase (%)	Amino acid with prevalence decrease (%)	Signature position	Most common amino acid(s)
Subtype C protease: India versus South Africa				
12	S (+10), A (+3), P (+2)	T (-15)	No	S
14	R (+29), I (+2)	K (-29)	No	K
15	I (+21)	V (-21)	No	V
19	T (+10), M (+3)	V (-11), I (-2), L (-2)	No	I
35	E (+15)	D (-15)	No	E
36	V (+21), L (+6), M (+3)	I (-28), T (-2)	No	I
63	P (+47), S (+5), T (+4)	L (-45), V (-7), A (-2)	Yes	P/L
77	V (+6)	I (-6)	No	V
82	I (+13)	V (-13)	No	V
Subtype C reverse transcriptase: India versus South Africa				
39	D (+66), N (+4)	E (-65), K (-3)	Yes	D/E
60	I (+60)	V (-60)	Yes	I/V
121	Y (+72), H (+2)	D (-72)	Yes	Y/D
123	D (+56), E (+4)	G (-34), N (-16), S (-10)	Yes	D/G
135	T (+20), R (+14), K (+5), M (+2)	I (-36), V (-4)	No	I
162	A (+39), H (+5), Y (+3)	S (-44), C (-4)	Yes	A/S
174	R (+18), Q (+13)	K (-27), N (-3), E (-2)	No	Q
177	E (+18)	D (-14), G (-2)	No	E
214	F (+14)	L (-14)	No	F

'Position' indicates subtype C sequence positions at which amino acid distributions are significantly different between India and South Africa. 'Amino acid with prevalence increase' indicates amino acids for which the prevalence is higher in India than South Africa. 'Amino acid with prevalence decrease' indicates amino acids for which the prevalence is lower in India than South Africa. The relative percentage differences for each amino acid are shown in parentheses, and the direction of differences is shown by a plus or minus sign. The differences represent absolute percentage point differences, not multiplicative percentage changes. Differences ≤1% are not shown. The 'signature position' column indicates whether or not the position is also a signature position based on the viral epidemiology signature pattern analysis (VESPA) of **Table 1**. *'Most common amino acid(s)' indicates the most common amino acid in both regions (if the position is not a signature difference) or the most common amino acid in India and South Africa, separated by a '/' (if the position is a signature difference).*

sequences were seen in nine PR and nine RT positions. For example, although valine is not a signature AA at PR position 36, its relative prevalence is higher in India by 21 percentage points. Isoleucine at this position is the signature AA in both India and South Africa, and is lower by 28 percentage points in India compared with South Africa. Additionally, in one PR (63) and five RT (39, 60, 121, 123 and 162) positions, the signature AA differs between India and South Africa. These six signature positions are consistent with recent local epidemiological studies that characterized naturally occurring polymorphisms in India [24,25].

Evolutionary selection

An exploratory analysis of evolutionary selection effects in subtype C sequences with SNAP suggests that some geography-dependent positions demonstrate evidence for positive selection (**Figure 2**).

Overall, the highest dn–ds differences were found at PR positions 36 (most prominent in India) and 63 (most prominent in East Africa, Latin America, North America, South Africa and western Europe); and RT positions 123 (most prominent in east Africa, South Africa, and western Europe), 162 (India) and 173 (Latin America). Other positions also exhibit comparably high estimates of dn–ds, including PR positions 12, 14, 15, 19 and 37 and RT positions 48, 135, 174 and 207.

Within subtype C, all positions that were determined to be signature differences between India and South Africa (PR 63 and RT 39, 60, 121, 123 and 162) exhibited positive 95% CIs in either one of the two regions, consistent with the possibility of positive selection as an explanatory factor for these signature differences.

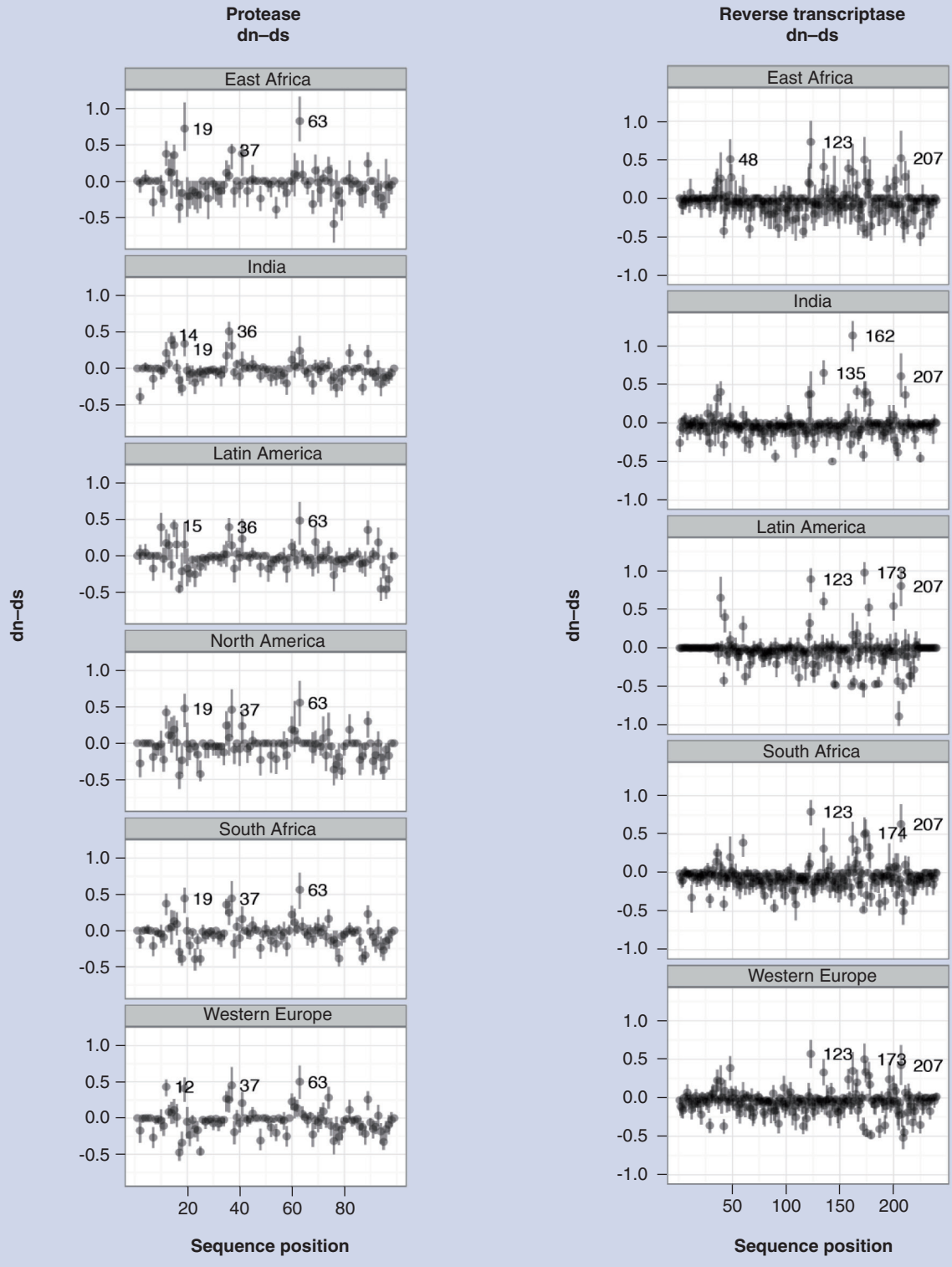

Figure 2. Nonsynonymous–synonymous substitution analysis of selection in subtype C *pol*. 95% CI of dn–ds estimates are shown for all sequence positions of protease (amino acid 1–99; left) and reverse transcriptase (amino acid 1–240; right) in all geographic regions from which subtype C sequences were available. The three positions with the highest ranked estimated difference between nonsynonymous and synonymous mutation rates are labeled.
dn–ds: Nonsynonymous–synonymous substitution.

Discussion

We characterized systematic differences in AA frequencies of PR and RT within HIV-1 group M subtypes from different geographic regions. We found that 11 PR and 11 RT positions exhibited differences in sequence signatures across geographic regions within subtypes. Positions with different AA frequency distributions across geographic regions within subtypes accounted for ~35% of the *pol* gene (37% PR and 33% RT positions). Geography-dependent positions occur throughout PR and RT, in both conserved and diverse loci, primarily in nonresistance positions. An exploratory analysis suggested that positive selection driven by host immune responses contributed to systematic population differences within subtypes.

These results emphasize that homogeneous genetic variation within subtypes should not be assumed. Our findings further highlight the limitations of the subtyping nomenclature in capturing genetic diversity [56], which may not be as discrete as earlier suggested. Systematic diversity within subtypes may contribute to genetic background influencing drug-resistance evolution or other phenotypic characteristics, although further data and research are needed to establish these associations. We showed that there were multiple AA differences within subtypes from different geographic regions. Ignoring such information can result in erroneous conclusions regarding genotype–phenotype associations. The geographical dependence reported here is distinct from previously reported HIV-1 sub-subtypes, which are currently only defined for subtypes A and F [34–38]. Epidemiological reports of regional sequence characteristics have mostly focused on *env* gene characterizations [11–21] and sequence-level clustering based on phylogeny. We extend previous characterization of mutational *pol* geographic signatures, reported among subtype C sequences from Brazil [22], and describe unique *pol* signature and non-signature differences in multiple subtypes and geographic regions. There were no signature differences at drug-resistance positions, although 11 resistance positions (one PR and ten RT) exhibited differences in AA distributions across regions. The implications of these findings on the evolution of drug resistance mandate further study.

Positions exhibiting sequence variation within subtypes could potentially have phenotypic consequences, although direct evidence is needed to demonstrate this. RT I135X, a well-characterized escape mutation associated with the *HLA-B*51* allele [57,58], was found here to be a signature mutation differentiating geographic subpopulations within subtype B (Table 3). It also exhibits different AA frequency distributions across geographic regions within both subtypes B and C, and was estimated to have a high rate of nonsynonymous mutation in subtypes B (not shown, data available upon request) and C (Figure 2, India). Furthermore, in our exploratory dn–ds analysis, RT positions 123 and 162 were found to have high estimated rates of nonsynonymous mutations in subtype C sequences from multiple geographic regions. Both positions are within cytotoxic T lymphocyte epitope clusters that include a HLA-B35 epitope associated with rapid progression to AIDS in multiple studies [59–62]. Unfortunately, host genotypes were unavailable for the patients in this dataset, and thus further study is needed to clarify the interaction between sequence heterogeneity, HLA-mediated host selection [5,63] and geographic sequence clustering within HIV-1 subtypes.

A recent epidemiological analysis of subtype C Indian PR sequences described positions noted for positively selected sites using three different methods. Four positions (12, 19, 36 and 82) were found to be significant in all three methods [24]. In our analysis, these four positions were also estimated to have significantly positive rates of nonsynonymous mutation, and were some of the highest dn–ds estimates in India. Out of the 99 codons in PR, positions 36, 19, 82 and 12 were associated with the first-, third-, seventh- and eighth-highest dn–ds values in India, respectively. Given that these analyses were conducted using different datasets and applied different methodologies, this suggests that these findings are fairly robust. We extend these analyses of selection to other regional populations within subtype C, as well as RT sequences.

We and others have previously identified significant differences in resistance-associated and nonassociated mutations across subtypes [1,64]. We extend earlier findings and show that a number of previously identified positions exhibit geographic variation. For example, of 14 PR and 18 RT previously identified subtype C-specific positions, seven (50%) PR and 15 (83%) RT positions were also found here to vary depending on the geographic origin of subtype C sequences. This finding further stresses the importance of

accounting for within-subtype subpopulation heterogeneity, such as geography, in global sequence analyses.

Several limitations of this study should be noted. First, geographical information is only one of many informative indicators of population bottlenecks, and other information such as HIV transmission networks, temporal, ethnic and socioeconomic indicators may also be useful. Second, aggregating data to the WHO defined-regions does not capture systematic variation across smaller subpopulations. Local epidemiologic studies continue to contribute to the expanding global dataset, which will enable more detailed analyses. Third, as the global molecular epidemiology of HIV is constantly in flux, these analyses represent a snapshot in time; a limitation of all surveillance studies. Lastly, since both the VESPA and dn–ds analyses are derived from the same dataset, they should be considered as complementary and not as independent validation.

Conclusion

In summary, we performed a comprehensive analysis of HIV-1 subtypes across geographic regions utilizing a global *pol* sequence dataset and characterized within-subtype sequence heterogeneity. We describe AA differences between geographic regions that are distinct from sequence-level phylogenetic clustering and previously described sub-subtypes. The current subtype nomenclature does not capture the additional subpopulation heterogeneity described here. Taking these differences into account can improve the interpretation of HIV-1 sequence datasets from multiple cohorts and locations. The sequence differences identified here can help improve the interpretation of genotype–phenotype association studies of HIV-1 subtypes.

Future perspective

Future studies on HIV-1 genotypic–phenotypic associations, which utilize global, multicohort datasets, should account for systematic geographic differences within subtypes that may not be captured by the subtyping nomenclature. Geographic origin is only one of many indicators of subpopulations within subtypes, and tighter integration between sequence analysis and subpopulation metadata should be explored. Finally, alternative statistical and algorithmic approaches for optimally capturing genetic background information should be considered alongside the subtyping nomenclature.

Financial & competing interests disclosure

A DeLong, JW Hogan, A Huang, D Katzenstein and R Kantor are funded by R01AI66922; A Huang, A DeLong, JW Hogan and R Kantor are funded by CFAR-P30AI042853; and A Huang was funded by T32 grant T32DA13911-10, all from the NIH. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Ethical conduct of research

The authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations.

Executive summary

- We performed a comprehensive analysis of HIV-1 subtypes across geographic regions utilizing a global *pol* sequence dataset to identify subtype-specific geography-dependent positions.
- Sequence signature differences and amino acid distribution differences in treatment-naive *pol* gene sequences within subtypes are demonstrated in approximately 35% of the *pol* gene, including 22 signature sequence differences. These results confirm and extend previous local epidemiological studies. An exploratory evolutionary selection analysis suggests that these differences may arise from local host-selection effects.
- The current subtype nomenclature does not capture the additional subpopulation heterogeneity described here. Homogeneous genetic variation within subtype should not be assumed.
- Sequence differences identified here provide a basis for improving the interpretation of genotype–phenotype association studies.

References

1. Kantor R, Katzenstein DA, Efron B *et al.* Impact of HIV-1 subtype and antiretroviral therapy on protease and reverse transcriptase genotype: results of a global collaboration. *PLoS Med.* 2(4), e112 (2005).
2. Rhee SY, Gonzales MJ, Kantor R, Betts BJ, Ravela J, Shafer RW. Human immunodeficiency virus reverse transcriptaseand protease sequence database. *Nucleic Acids Res.* 31(1), 298–303 (2003).
3. Johnson VA, Calvez V, Günthard HF *et al.* 2011 Update of the drug resistance mutations in HIV-1. *HIV Med.* 18, 156–163 (2010).

4. Chan PA, Kantor R. Transmitted drug resistance in nonsubtype B HIV-1 infection. *HIV Therapy* 3(5), 447–465 (2009).
5. Frahm N, Goulder PJR, Brander C. Total assessment of HIV-specific CTL responses: epitope clustering, processing preferences, and the impact of HIV sequence heterogeneity. *In: HIV Molecular Immunology 2002*. Korber BT, Brander C, Haynes BF *et al.* (Eds). Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM, USA 3–22 (2002).
6. Grossman Z, Paxinos EE, Averbuch D *et al.* Mutation D30N is not preferentially selected by human immunodeficiency virus type 1 subtype C in the development of resistance to nelfinavir. *Antimicrob. Agents Chemother.* 48(6), 2159–2165 (2004).
7. Myers RE, Gale CV, Harrison A, Takeuchi Y, Kellam P. A statistical model for HIV-1 sequence classification using the subtype analyser (STAR). *Bioinformatics* 21(17), 3535–3540 (2005).
8. Brenner BG, Oliveira M, Doualla-Bell F *et al.* HIV-1 subtype C viruses rapidly develop K65R resistance to tenofovir in cell culture. *AIDS* 20(9), F9–F13 (2006).
9. Calazans A, Brindeiro R, Brindeiro P *et al.* Low accumulation of L90M in protease from subtype F HIV-1 with resistance to protease inhibitors is caused by the L89M polymorphism. *J. Infect. Dis.* 191, 1961–1970 (2005).
10. Taylor BS, Sobieszczyk ME, McCutchan FE, Hammer SM. The challenge of HIV-1 subtype diversity. *N. Engl. J. Med.* 358(15), 1590–1602 (2008).
11. Ou CY, Weniger BG, Luo CC *et al.* Independent introduction of two major HIV-1 genotypes into distinct high-risk populations in Thailand. *Lancet* 341(8854), 1171–1174 (1993).
12. Abebe A, Lukashov VV, Rinke de Wit TF *et al.* Timing of the introduction into Ethiopia of subcluster C′ of HIV type 1 subtype C. *AIDS Res. Hum. Retroviruses* 17(7), 657–661 (2001).
13. Abebe A, Pollakis G, Fontanet AL *et al.* Identification of a genetic subcluster of HIV type 1 subtype C (C′) widespread in Ethiopia. *AIDS Res. Hum. Retroviruses* 16(17), 1909–1914 (2000).
14. Ou C, Takebe Y, Luo C *et al.* Wide distribution of two subtypes of HIV-1 in Thailand. *AIDS Res. Hum. Retroviruses* 8(8), 1471–1472 (1992).
15. Kuiken CL, Goudsmit J. Silent mutation pattern in V3 sequences distinguishes virus according to risk group in Europe. *AIDS Res. Hum. Retroviruses* 10(3), 319–320 (1994).
16. Adwan G, Papa A, Kouidou S *et al.* HIV type 1 sequences with GGC substitution in injecting drug users in Greece. *AIDS Res. Hum. Retroviruses* 15(7), 679–680 (1999).
17. Shankarappa R, Chatterjee R, Learn GH *et al.* Human immunodeficiency virus type 1 env sequences from Calcutta in eastern India: identification of features that distinguish subtype C sequences in India from other subtype C sequences. *J. Virol.* 75(21), 10479–10487 (2001).
18. Seth P. Evolution of HIV-1 in India. *Indian J. Virol.* 1–5 (2010).
19. Kalpana A, Srikanth T, Abhay J, Sushama J, Swarali K, Ramesh P. gp120 sequences from HIV type 1 subtype C early seroconverters in India. *AIDS Res. Hum. Retroviruses* 20(8), 889–894 (2004).
20. Kalish ML, Baldwin A, Raktham S *et al.* The evolving molecular epidemiology of HIV-1 envelope subtypes in injecting drug users in Bangkok, Thailand: implications for HIV vaccine trials. *AIDS* 9(8), 851 (1995).
21. Subbarao S, Limpakarnjanarat K, Mastro TD *et al.* HIV type 1 in Thailand, 1994–1995: persistence of two subtypes with low genetic diversity. *AIDS Res. Hum. Retroviruses* 14(4), 319–327 (1998).
22. Soares MA, de Oliveira T, Brindeiro RM *et al.* A specific subtype C of human immunodeficiency virus type 1 circulates in Brazil. *AIDS* 17(1), 11 (2003).
23. Ojesina AI, Kanki PJ. HIV-1 subtype and reverse transcriptase genotype: role for geographical location and founder effects. *PLoS Med.* 3(12) e540 (2006).
24. Neogi U, Sahoo PN, Kumar R, De Costa A, Shet A. Characterization of HIV type 1 subtype C protease gene: selection of L63P mutation in protease inhibitor-naive indian patients. *AIDS Res. Hum. Retroviruses* 27(11), 1249–1253 (2011).
25. Neogi U, Prarthana BS, Gupta S *et al.* Naturally occurring polymorphisms and primary drug resistance profile among antiretroviral-naive individuals in Bangalore, India. *AIDS Res. Hum. Retroviruses* 26(10), 1097–1101 (2010).
26. Bello G, Passaes CP., Guimarães ML *et al.* Origin and evolutionary history of HIV-1 subtype C in Brazil. *AIDS* 22(15), 1993 (2008).
27. De Oliveira T, Pillay D, Gifford RJ. The HIV-1 subtype C epidemic in South America is linked to the United Kingdom. *PloS One* 5(2), e9311 (2010).
28. Morgado MG, Sabino EC, Shpaer EG *et al.* V3 region polymorphisms in HIV-1 from Brazil: prevalence of subtype B strains divergent from North American/European prototype and detection of subtype F. *AIDS Res. Hum. Retroviruses* 10(5), 569–576 (1994).
29. Ana V, Koko O. Close phylogenetic relationship between Angolan and Romanian HIV-1 subtype F1 isolates. *Reterovirology* 6, 39 (2009).
30. Paraskevis D, Pybus O, Magiorkinis G *et al.* Tracing the HIV-1 subtype B mobility in Europe: a phylogeographic approach. *Retrovirology* 6(1), 49 (2009).
31. Mehta SR, Wertheim JO, Delport W *et al.* Using phylogeography to characterize the origins of the HIV-1 subtype F epidemic in Romania. *Infect. Genet. Evol.* 11(5), 975–979 (2011).
32. Véras NM, Gray RR, de Macedo Brígido LF, Rodrigues R, Salemi M. High-resolution phylogenetics and phylogeography of human immunodeficiency virus type 1 subtype C epidemic in South America. *J. Gen. Virol.* 92(7), 1698–1709 (2011).
33. Véras NM, Santoro MM, Gray RR *et al.* Molecular epidemiology of HIV type 1 CRF02_AG in Cameroon and African patients living in Italy. *AIDS Res. Hum. Retroviruses* 27(11), 1173–1182 (2011).
34. Gao F, Vidal N, Li Y *et al.* Evidence of two distinct subsubtypes within the HIV-1 subtype A radiation. *AIDS Res. Hum. Retroviruses* 17(8), 675–688 (2001).
35. Meloni ST, Kim B, Sankale JL *et al.* Distinct human immunodeficiency virus type 1 subtype A virus circulating in west Africa: sub-subtype A3. *J. Virol.* 78(22), 12438–12445 (2004).
36. Vidal N, Mulanga C, Bazepeo SE, Lepira F, Delaporte E, Peeters M. Identification and molecular characterization of subsubtype A4 in central Africa. *AIDS Res. Hum. Retroviruses* 22(2), 182–187 (2006).
37. Triques K, Bourgeois A, Saragosti S *et al.* High diversity of HIV-1 subtype F strains in central Africa. *Virology* 259(1), 99–109 (1999).
38. Triques K, Bourgeois A, Vidal N *et al.* Near-full-length genome sequencing of divergent African HIV type 1 subtype F viruses leads to the identification of a new HIV type 1 subtype designated K. *AIDS Res. Hum. Retroviruses* 16(2), 139–151 (2000).
39. Kanki PJ, Hamel DJ, Sankalé JL *et al.* Human immunodeficiency virus type 1 subtypes differ in disease progression. *J. Infect. Dis.* 179, 68–73 (1999).
40. Boivin MJ, Ruel TD, Boal HE *et al.* HIV-subtype A is associated with poorer neuropsychological performance compared with subtype D in antiretroviral therapy-naive Ugandan children. *AIDS* 24(8), 1163–1170 (2010).

41. Benson DA, Boguski MS, Lipman DJ *et al.* GenBank. *Nucleic Acids Res.* 27(1), 12 (1999).
42. Land S, Cunningham P, Zhou J *et al.* TREAT Asia Quality Assessment Scheme (TAQAS) to standardize the outcome of HIV genotypic resistance testing in a group of Asian laboratories. *J. Virol. Methods* 159(2), 185–193 (2009)
43. de Oliveira T, Shafer RW, Seebregts C. Public database for HIV drug resistance in southern Africa. *Nature* 464(7289), 673 (2010).
44. Novitsky V, Wester CW, DeGruttola V *et al.* The reverse transcriptase 67N 70R 215Y genotype is the predominant TAM pathway associated with virologic failure among HIV type 1C-infected adults treated with ZDV/ddI-containing HAART in southern Africa. *AIDS Res. Hum. Retroviruses* 23(7), 868–878 (2007).
45. Kuiken C, Leitner T. HIV-1 subtyping. In: *Computational and Evolutionary Analysis of HIV Molecular Sequences*. Rodrigo AG, Learn GH Jr (Eds). Kluwer Academic Publishers, MA, USA, 27–53 (2002).
46. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32(5), 1792–1797 (2004).
47. Hemelaar J, Gouws E, Ghys PD, Osmanov S. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *AIDS* 20(16), W13–W23 (2006).
48. Korber B, Myers G. Signature pattern analysis: a method for assessing viral sequence relatedness. *AIDS Res. Hum. Retroviruses* 8(9), 1549–1560 (1992).
49. Agresti A. *Categorical Data Analysis (2nd Edition)*. Wiley-Interscience, NJ, USA (2002).
50. Quinn GP, Keough MJ. *Experimental Design and Data Analysis for Biologists.* Cambridge University Press, UK (2002).
51. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* 289–300 (1995).
52. Shannon CE, Weaver W. A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423 (1948).
53. Korber B. HIV signature and sequence variation analysis. In: *Computational Analysis of HIV Molecular Sequences*. Kluwer Academic Publishers, MA, USA 4, 55–72 (2000).
54. Rodrigues R, Scherer LC, Oliveira CM *et al.* Low prevalence of primary antiretroviral resistance mutations and predominance of HIV-1 clade C at polymerase gene in newly diagnosed individuals from south Brazil. *Virus Res.* 116(1–2), 201–207 (2006).
55. Brígido LFM, Ferreira JLP, Almeida VC *et al.* Southern Brazil HIV type 1 C expansion into the State of São Paulo, Brazil. *AIDS Res. Hum. Retroviruses* 27(3), 339–344 (2011).
56. Castro-Nallar E, Pérez-Losada M, Burton GF, Crandall KA. The evolution of HIV: inferences using phylogenetics. *Mol. Phylogenet. Evol.* 62(2), 777–792 (2011).
57. Tomiyama H, Sakaguchi T, Miwa K *et al.* Identification of multiple HIV-1 CTL epitopes presented by HLA-B*5101 molecules. *Hum. Immunol.* 60(3), 177–186 (1999).
58. Kawashima Y, Pfafferott K, Frater J *et al.* Adaptation of HIV-1 to human leukocyte antigen class I. *Nature* 458(7238), 641–645 (2009).
59. Kaslow RA, Carrington M, Apple R *et al.* Influence of combinations of human major histocompatibility complex genes on the course of HIV–1 infection. *Nature Med.* 2(4), 405–411 (1996).
60. Scorza Smeraldi R, Fabio G, Lazzarin A *et al.* HLA-associated susceptibility to AIDS: HLA B35 is a major risk factor for Italian HIV-infected intravenous drug addicts. *Hum. Immunol.* 22(2), 73–79 (1988).
61. Scorza Smeraldi R, Lazzarin A, Moroni M, Fabio G, Eisera NB, Zanussi C. HLA-associated susceptibility to acquired immunodeficiency syndrome in Italian patients with human-immunodeficiency-virus infection. *Lancet* 328(8517), 1187–1189 (1986).
62. Just JJ. Genetic predisposition to HIV-1 infection and acquired immune deficiency virus syndrome: a review of the literature examining associations with HLA. *Hum. Immunol.* 44(3), 156–169 (1995).
63. Yusim K, Kesmir C, Gaschen B *et al.* Clustering patterns of cytotoxic T-lymphocyte epitopes in human immunodeficiency virus type 1 (HIV-1) proteins reveal imprints of immune evasion on HIV-1 global variation. *J. Virol.* 76(17), 8757–8768 (2002).
64. Shafer RW, Rhee SY, Bennett DE. Consensus drug resistance mutations for epidemiological surveillance: basic principles and potential controversies. *Antivir. Ther.* 13, 59–68 (2008).

Website

101. Los Alamos HIV databases. www.hiv.lanl.gov