PLOS ONE

Molecular Characterization of Ambiguous Mutations in HIV-1 Polymerase Gene: Implications for Monitoring HIV Infection Status and Drug Resistance

Du-Ping Zheng[1], Margarida Rodrigues[2], Ebi Bile[3], Duc B. Nguyen[4], Karidia Diallo[1], Joshua R. DeVos[1], John N. Nkengasong[1], Chunfu Yang[1*]

1 Division of Global HIV/AIDS, Center for Global Health, Centers for Disease Control and Prevention (CDC), Atlanta, Georgia, United States of America, **2** CDC-GAP Angola, Luanda, Angola, **3** CDC-GAP Botswana, Gaborone, Botswana, **4** Department of Health and Human Services/US CDC, Hanoi, Vietnam

Abstract

Detection of recent HIV infections is a prerequisite for reliable estimations of transmitted HIV drug resistance (t-HIVDR) and incidence. However, accurately identifying recent HIV infection is challenging due partially to the limitations of current serological tests. Ambiguous nucleotides are newly emerged mutations in quasispecies, and accumulate by time of viral infection. We utilized ambiguous mutations to establish a measurement for detecting recent HIV infection and monitoring early HIVDR development. Ambiguous nucleotides were extracted from HIV-1 *pol*-gene sequences in the datasets of recent (HIVDR threshold surveys [HIVDR-TS] in 7 countries; n=416) and established infections (1 HIVDR monitoring survey at baseline; n=271). An ambiguous mutation index of 2.04×10^{-3} nts/site was detected in HIV-1 recent infections which is equivalent to the HIV-1 substitution rate (2×10^{-3} nts/site/year) reported before. However, significantly higher index (14.41×10^{-3} nts/site) was revealed with established infections. Using this substitution rate, 75.2% subjects in HIVDR-TS with the exception of the Vietnam dataset and 3.3% those in HIVDR-baseline were classified as recent infection within one year. We also calculated mutation scores at amino acid level at HIVDR sites based on ambiguous or fitted mutations. The overall mutation scores caused by ambiguous mutations increased (0.54×10^{-2}3.48×10^{-2}/DR-site) whereas those caused by fitted mutations remained stable (7.50-7.89×10^{-2}/DR-site) in both recent and established infections, indicating that t-HIVDR exists in drug-naïve populations regardless of infection status in which new HIVDR continues to emerge. Our findings suggest that characterization of ambiguous mutations in HIV may serve as an additional tool to differentiate recent from established infections and to monitor HIVDR emergence.

Citation: Zheng D-P, Rodrigues M, Bile E, Nguyen DB, Diallo K, et al. (2013) Molecular Characterization of Ambiguous Mutations in HIV-1 Polymerase Gene: Implications for Monitoring HIV Infection Status and Drug Resistance. PLoS ONE 8(10): e77649. doi:10.1371/journal.pone.0077649

Received April 23, 2013; **Accepted** September 12, 2013; **Published** October 17, 2013

Funding: This research has been supported by the President's Emergency Plan for AIDS Relief (PEPFAR) through the Centers for Disease Control and Prevention. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

* E-mail: CYang1@cdc.gov

Introduction

With over 8.2 million HIV-infected patients on antiretroviral therapy (ART) in low- and middle-income countries at the end of 2011, emergence and transmission of HIV drug resistance (HIVDR) are ongoing public health challenges in the battle against HIV/AIDS [1-6]. Because of the incomplete suppression of viral replication by ART [2-4,7], even with the optimal adherence, HIVDR may still emerge in ART-patients; whereas under suboptimal ART situations, such as incomplete ART compliance and/or improper practice of regimen prescriptions, development of HIVDR could be enhanced [8]. Currently, mutations associated with HIVDR at 85 sites are identified, including 32 in reverse transcriptase (RT, 16 for nucleoside RT inhibitors [NRTIs] and 16 for non-nucleosides RTI [NNRTIs]), 36 in protease, seven in envelope, and 10 in integrase genes [9].

Estimates of transmitted HIVDR (t-HIVDR) and HIV incidence have been two important surrogates in measuring ART efficacy and prevention program effectiveness [10]. To assess the t-HIVDR in resource-limited countries, the World Health Organization (WHO) recommends conducting HIVDR threshold survey (HIVDR-TS) in recently HIV-infected populations enrolled by using WHO criteria [11,12]. Likewise, estimate of HIV incidence also requires recent infections that are determined by using serological assays with cross-sectional HIV-positive samples at population level [13-15]. Nevertheless, studies have indicated that these assays overestimated HIV incidences due to false classification of recent infections in certain populations and lack of validated

Table 1. Characteristics of sequence datasets and subtyping in partial HIV-1 *pol* gene.

Country[a]	Year	Data source	Infection route	Sequences	Subtype[b]						
					A	B	C	G	CRF01_AE	Others	Untypeable
Angola	2009	ANC (HIVDR-TS)	heterosexual	39			9	5		5	20
Botswana	2007	ANC (HIVDR-TS)	heterosexual	134			132*				2
China	2006	VCT (HIVDR-TS)	Multi-routes	45		8	3		20*	9	5
Kenya	2005-2006	ANC (HIVDR-TS)	heterosexual	33	19*		2			1	11
Malawi	2006	ANC (HIVDR-TS)	heterosexual	52			51*				1
Tanzania	2005-2006	ANC (HIVDR-TS)	heterosexual	45	14*		17			4	10
Vietnam	2006-2008	VCT (HIVDR-TS)	Multi-routes	68					68		
Nigeria	2008	ET (T1 baseline)	unknown	271	5		1	135			
Canada	2002-2008	Early infection	unknown	63		63*					

[a] Country data used in this study were collected from HIV drug resistance threshold survey (HIVDR-TS), T1 baseline survey, and published data (24) respectively; Antenatal care (ANC), Voluntary counseling and testing (VCT), Eligible for treatment (ET); Numbers labeled with asterisk (*) were sequences selected to represent subtypes for statistical analysis, see Table 3; [b] HIV-1 subtypes were primarily determined using REGA HIV subtyping tool (http://www.bioafrica.net/rega-genotype/html/subtypinghiv.html).

doi: 10.1371/journal.pone.0077649.t001

standards in accurately distinguishing recent (within 1 year) from established infections [13,14,16].

HIV evolves rapidly attributed partially to the high error rate of its RT [17,18], and accumulates mutations at a certain rate over time [19]. This results in generating a large number of variants (quasispecies) in a host and increasing genetic diversity in a viral population [20]. By analyzing the genetic relationships among quasispecies, the dynamic evolutional pathway of a variant can be tracked on a time scale within and among hosts [21,22]. In the early stage of viral replication, HIV variants with new point mutations account for only a small proportion of the total wild-type populations, thus a particular point mutation at an allele is detected as a mixture along with the wild-type by conventional population-based (Sanger) sequencing, which is termed as ambiguous mutation/ nucleotide. By applying the defined nucleotide substitution rate of HIV-1 [19,20], the quantified ambiguous mutations within a sequence could be used to estimate the duration of an HIV infection. Recent studies have demonstrated a constant increase of ambiguous nucleotides during the first 8 years of HIV infections at a rate of 0.2% per year in HIV subtype B *pol*-gene [23], and 0.45-0.5% of ambiguous nucleotides as a cutoff for distinguishing recent from established infections [23-25]. In this study, we characterized ambiguous nucleotides in non-B subtype sequences from eight population-based HIVDR-TS and HIVDR monitoring surveys and determined a predictive value for estimating HIV infection status using a molecular evolutional approach and compared ambiguous mutation evolving rate between sequences generated from non-B (A, C, and CRF01_AE) and B subtype viral strains, and evaluated the accuracy of WHO epidemiological criteria for the recent infection determination. We also used this approach to monitor HIVDR development at the early stage of HIV infections.

Materials and Methods

Data sources and types

Sequence data were from HIVDR-TS conducted in seven countries (Angola, Botswana, China, Kenya, Malawi, Tanzania, and Vietnam) (n=416) and from a baseline survey (n=271) in monitoring HIVDR development in patients commencing ART in Nigeria during 2005-2009 (Table 1). The detail demographic and clinical data of participants were previously described [26-32]. In brief, all patients in HIVDR-TS were enrolled as recent infections according to WHO criteria [11,12]. For those from Angola, Botswana, Kenya, Malawi and Tanzania, they were pregnant women who were <25 years old, attending antenatal clinics (ANC), diagnosed with HIV infections for the first time, and ART-naïve; and for those from China and Vietnam, they were individuals attending voluntary counseling and testing (VCT) sites and were partially intravenous drug users (IDU). The WHO criteria were designed to increase the likelihood of identifying recently infected individuals for the HIVDR-TS. The participants in baseline survey were patients eligible for ART according to the Nigeria national guidelines for HIV/AIDS care and treatment in adolescents and adults (CD4 ≤200 cells/μl, WHO stage III or IV or AIDS defined illness) [33], they were most likely established or chronic HIV-infected individuals. This dataset was used to compare to those of recent infections in the HIVDR-TS for ambiguous mutation calibration.

We also included a dataset of subtype B sequences (n=63) from published resources [24] for the comparison of ambiguous mutation preference with our non-B subtype sequences. These sequences were generated from individuals who had been infected within 155 days confirmed by serological tests.

DNA sequencing, genotyping and subtyping

All partial *pol*-gene sequences were generated using a validated HIV-1 genotyping assay using a conventional population-based bi-directional sequencing procedure [34-36]. The lengths of sequences were 981 (HIVDR-TS data) and

1,002 (Baseline HIVDR monitoring survey data) nucleotides (nts) containing HIVDR mutation sites of protease and RT region [9].

Sequences were primarily subtyped using Stanford REGA HIV-1 Subtyping Tool version 2.0 (http://dbpartners.stanford.edu/RegaSubtyping/). The sequences used in this study included those published previously (JQ617150-JQ617250), and new submissions (JX083986-JX123826).

Detection and analysis of ambiguous mutations

Ambiguous mutations, which consist of mixed nucleotides at a sequence position and named using the standard IUPAC ambiguous nucleotide codes, were determined and automatically called using customized software, Recall [37] when the sequencing signal intensity of the minor base was ≥20% of the major base signal at a nucleotide position on bi-directional sequences after subtracting background noise. Ambiguous mutations were extracted from each of sequences and tallied at country level. The mean of the ambiguous mutations was then calculated using the formula: $M_{AM} = \sum N_{AM}/N$ (M_{AM}: mean of ambiguous mutations per sequence; N_{AM}: number of ambiguous mutations of a sequence; $\sum N_{AM}$: sum of ambiguous mutations in a dataset, N: total number of sequences in the dataset). The index (I) of ambiguous mutations was calculated using the formulas: $I_{AM} = N_{AM}/Ls$ for an index at sequence level, or $I_{AM} = M_{AM}/Ls$ for an index at a dataset level (I_{AM}: index of ambiguous mutations per site; Ls: length of a sequence by nucleotide), (note: Ls is 1/3 of full length when calculation was for 1st, 2nd, or 3rd codon position).

The composition of nucleotides or ambiguous nucleotides in a sequence dataset was obtained using BioEdit with the algorithm of base composition and mass export [38]. Ambiguous mutations were then stratified at 1st, 2nd, 3rd and all codon positions by dataset of threshold, baseline and Vietnam (VT), or by HIVDR and non-HIVDR sites based on the 2013 HIVDR List [9]. The index of ambiguous mutation was calculated using the same formulas as described previously.

At the AA level, we scored 1 for a pure mutated AA and 0.5 for an ambiguous mutated AA because of its ambiguity, and calculated the total DR mutation score at each of the HIVDR sites [9] with the formula: DR mutation % = $([N_{MAA}+N_{AMAA}/2]/N_{SEQ})\times 100\%$ (N_{MAA}: number of mutated AA; N_{AMAA}: number of ambiguous mutated AA).

Recent infection determination

Based on the estimated HIV nucleotide substitution rate of 2×10^{-3} nts per site per year [19,20], a sequence with ≤2 ambiguous mutations, or 2×10^{-3} ambiguous nucleotides per site, was considered to be derived from a subject who was infected within one year. In a dataset, the percentage of recent infections was assessed by calculating the proportion of sequences that had ≤2 ambiguous mutations.

Statistical analysis

Statistical analyses were performed using IBM SPSS Statistics 20 (IBM), or otherwise were indicted. Data of non-normal distribution were determined by one-Sample Kolmogorov-Smirnov Test. Plot of dataset median, interquartile range (IQR), and range with and without outliers was made using online resource (http://www.physics.csbsju.edu/stats/). The overall significant difference of values in all datasets was determined by Kruskal-Wallis test, and the difference of pairwise comparison was determined by Mann-Whitney test when Kruskal-Wallis P value was <0.05. For multiple comparisons, the P values were corrected by the Bonferroni method.

Ethics Statement

This is a data mining study based on the sequences generated from our previously published survey studies [26-32] in which all the surveys had been approved by the local Institutional Review Board (IRB) from Angola, Botswana, China, Kenya, Malawi, Tanzania, Vietnam, and Nigeria as well as the Associate Director for Science at the Center for Global Health of CDC, USA who determined that the anonymous specimen testing performed at CDC was a non-human subject research.

Results

Range and index of ambiguous mutations for country-based datasets

Ambiguous mutations were extracted from 8 datasets of a total of 687 sequences (Table 1) and were used for statistical descriptive and significant analyses (Figure 1). We observed that two HIVDR-TS datasets (Angola and China) each contained one sequence with much higher number of ambiguous mutations, 25 and 19, respectively than those in the remaining HIVDR-TS sequence datasets (Figure 1, indicated by dash box). For study purpose, we analyzed the two datasets with and without the outlier sequences.

For datasets from Angola, Botswana, China, Kenya, Malawi and Tanzania, the maximal range of ambiguous mutations without the two outliers was 0-15 nts with mean ranging from 1.13-2.68 nts per sequence. However, for those from Vietnam and Nigeria, the ranges were 0-38, and 0-82 nts with means of 17.96 and 14.42 nts per sequence, respectively. Likewise, the ranges of ambiguous mutation index were $1.16\text{-}2.74\times10^{-3}$ nts per site among the datasets of Angola, Botswana, China, Kenya, Malawi and Tanzania; whereas the ranges of ambiguous mutation index for datasets from Vietnam and Nigeria were 18.32 and 14.40×10^{-3} nts per site, respectively (Figure 1).

When the two outliers were included in the analysis, the mean and index of ambiguous mutations were 3.26 nts per sequence and 3.31×10^{-3} nts per site for Angola, and 2.91 nts per sequence and 2.96×10^{-3} nts per site for China, respectively, which were slightly higher than the values generated without outlier sequences but they were still much lower than those from Vietnam and Nigeria (3.26 and 2.91 vs 17.96 and 14.42 for the mean of ambiguous mutations and 3.31 and 2.96 vs 18.32 and 14.40×10^{-3} nts per site for the index of ambiguous mutations, Figure 1).

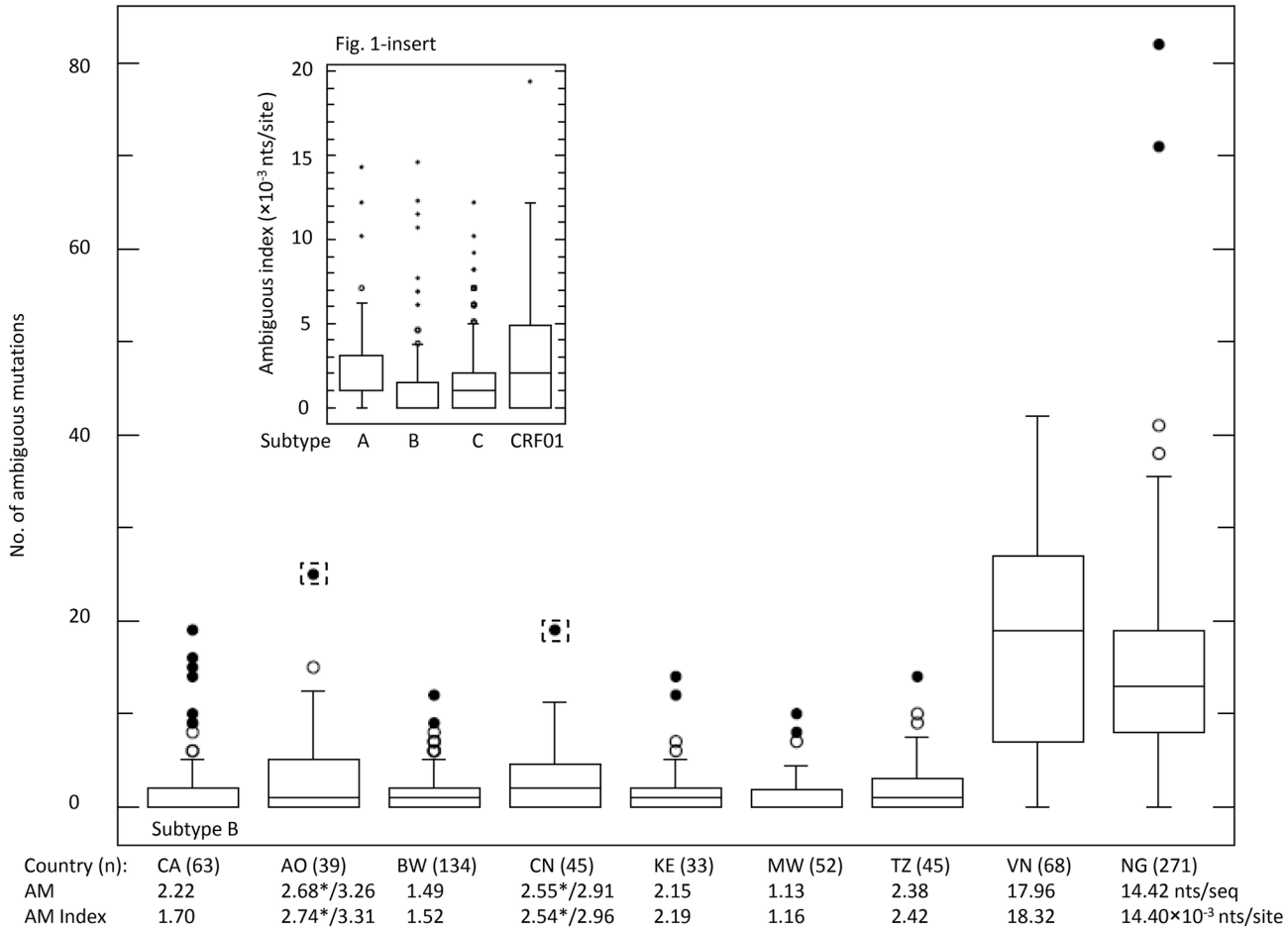


Figure 1. Descriptive statistics of ambiguous mutation in various sequence datasets. Plot of ambiguous mutations with descriptive statistics was performed using online statistical tool (http://www.physics.csbsju.edu/stats/). Individual country dataset was described for minimal and maximal ranges (short horizontal line at the bottom and top of the box), interquartile range (IQR, at 1st to 3rd quartile, box), median (line inside box), suspected outlier (open dot), and outlier (solid dot). Number in the bracket is the number of sequences from the country, Angola (AO), Botswana (BW), China (CN), Kenya (KE), Malawi (MW), Tanzania (TZ), Vietnam (VN), Nigeria (NG), and Canada (CA) [24]. Numbers with asterisk were calculated without the outlier in dash square box. Figure 1-insert shows the descriptive statistics of ambiguous mutation index in the dataset based on subtype (Table 3).

doi: 10.1371/journal.pone.0077649.g001

Statistical analyses indicated that both mean and index of ambiguous mutation in the HIVDR-TS datasets (likely from recent infections) were significantly lower than those in the HIVDR monitoring baseline survey (established infection) (p<0.001) with an exception of those from Vietnam which is described in the later sections.

Ambiguous mutation in recent and established HIV infected cohorts

To calibrate the difference of ambiguous mutations between recent and established HIV infections on a larger scale, we reorganized the data into three subsets: 1) threshold surveys for which samples were collected mainly from heterosexually transmitted individuals (N= 346), 2) baseline (N=271), and 3) Vietnam (N=68), and characterized the ambiguous mutations at each setting (Table 2). For the subset of threshold surveys, the mean of ambiguous mutations was 2.00 nts with a range of 0-15 nts per sequence, and the index was 2.04×10^{-3} (95% confidence intervals, [CI]: $1.13\text{-}2.71\times10^{-3}$) ambiguous mutations per site. However, the subset from Vietnam, which was also from threshold surveys but might consist of subjects with different routes of transmission [30,39], and the one from Nigeria, which contained established and chronic infections, yielded means of 17.96 and 14.42 ambiguous mutations per sequence, and index of 18.30 (95% CI: 16.59-20.02) and 14.40 (95% CI: 13.50-15.29) $\times10^{-3}$ ambiguous mutations per site (p=0.538), respectively, which were significantly higher than the values of the threshold data (p<0.001) (Table 2, Figure 2D). These combined data further confirmed that ambiguous

Table 2. Characteristics of ambiguous mutations (AMs) and rates between threshold and baseline sequence datasets.

	Threshold	Baseline	Vietnam-TS	*p*-value
Number of Sequences	346	271	117	
Sequence length (nts)	981	1002	867	
AMs and occurrence rate				
AM Range	0-15	0-82	0-38	
Mean (AMs/sequence)	2.00	14.42	12.78	
Rate (AMs/site, $\times 10^{-3}$) (95% C.I.)	2.04 (1.13-2.71)	14.40 (13.50-15.29)	14.74 (13.38-16.09)	<0.001 (TS vs BL, VN); 0.538 (VN vs BL)
AM occurrence rate (AMs/site, $\times 10^{-3}$)				
All codon positions				DR vs nDR: 0.889 (TS), 0.590 (BL), 0.441 (VN)
Non-DR sites (nDR)	2.02	14.97	14.35	
DR sites	2.07	12.09	12.26	
1st codon position				
Non-DR sites	0.90	6.76	5.86	
DR sites	1.41	7.21	8.97	
2nd codon position				
Non-DR sites	1.01	5.06	5.34	
DR sites	0.55	3.34	3.39	
3rd codon position				
Non-DR sites	4.16	33.05	31.86	
DR sites	4.24	25.71	24.42	

[a] The AMs of 2 outlier sequences were excluded for calculation.

doi: 10.1371/journal.pone.0077649.t002

mutation index was significantly lower in the individuals in the HIVDR-TS than those in the baseline and Vietnam surveys.

We further characterized the sequence distribution by ambiguous mutations. Results exhibited completed different patterns between the 3 datasets (Figure 2A-C). In the threshold survey subset, the highest distribution of sequences peaked at 0 ambiguous mutation and 80% of the sequences had <3 ambiguous mutations; however, the peak of distribution curve shifted to the range of 3-28 ambiguous mutations in the baseline subset, and diversified with various cluster ranges of ambiguous mutation in the Vietnam subset (Figure 2A-C). These patterns of distribution curve indicated the uniformity of infection status among the subjects in the dataset: the narrower the distribution range is, the more uniformity the subjects are and a wide and diversified distribution pattern suggests a wide range of mixed infections, e.g. those shown in VN dataset. By using the estimated HIV nucleotide substitution of 2×10^{-3} nts per site per year [19,20], sequences of 75.2% in threshold, 3.3% in baseline, and 10.3% in Vietnam subsets could be classified as coming from people infected with HIV within one year. This result implies that around 75% of patients enrolled in the HIVDR-TS using WHO epidemiological criteria could be recently infected individuals.

Ambiguous mutation at HIV drug resistant and non-resistant sites

To explore if there was any site preference for ambiguous mutation to occur, we measure the ambiguous mutation at DR or non-DR sites and all codon positions [9] . In general, no significant difference of ambiguous mutation was observed between DR and non-DR sites of all codon positions within each of the subsets (p=0.889 [HIVDR-TS], 0.590 [baseline], and 0.110 [VN-TS]) (Table 2), indicating a random mutation mechanism. However, at the 1st codon position the ambiguous mutation at DR sites was always higher than the non-DR sites across the datasets which was in reverse at the 2nd codon position. For example, in the threshold subset, the ambiguous mutation at the DR sites was 56.7% higher at the 1st codon position but 45.5% lower at the 2nd codon position than those at the non-DR sites. At the 3rd codon position, the ambiguous mutation between the DR and non-DR sites varied. They were similar in threshold (4.24 and 4.16), but were higher at the non-DR sites than the DR sites for the baseline and Vietnam subsets (33.05 vs 25.71; 42.50 vs 32.21, respectively), implying the evolutional pressure applied to DR and non-DR codon positions is different during the course of viral replication and/or infection.

Amino acid fitness of ambiguous and mutated nucleotides at HIV DR sites

To obtain the DR-score attributed to the ambiguous or non-ambiguous mutated nucleotides, we stratified the non-synonymous DR-associated mutations at amino acid (AA) level based on the 2013 HIVDR list [9] (Figure 3). Results showed that 12 of the 65 DR sites had patterns on the mutation score. It was constantly high at 3 DR sites (M36, 76.36-99.07%; H69, 90.10-98.53%; L89, 71.61-98.34%), moderate at 1 DR site (L63, 40.17-56.62%), and low at 1 DR site (V179, 10.52-20.59%) across all 3 subsets. The mutation score increased substantially at 2 DR sites (V82, 0.432.2148.52%; Q151, 0.2998.5398.52%), and decreased at 5 DR sites (K20, 20.3919.854.06%; D60, 15.375.882.58%; T74,

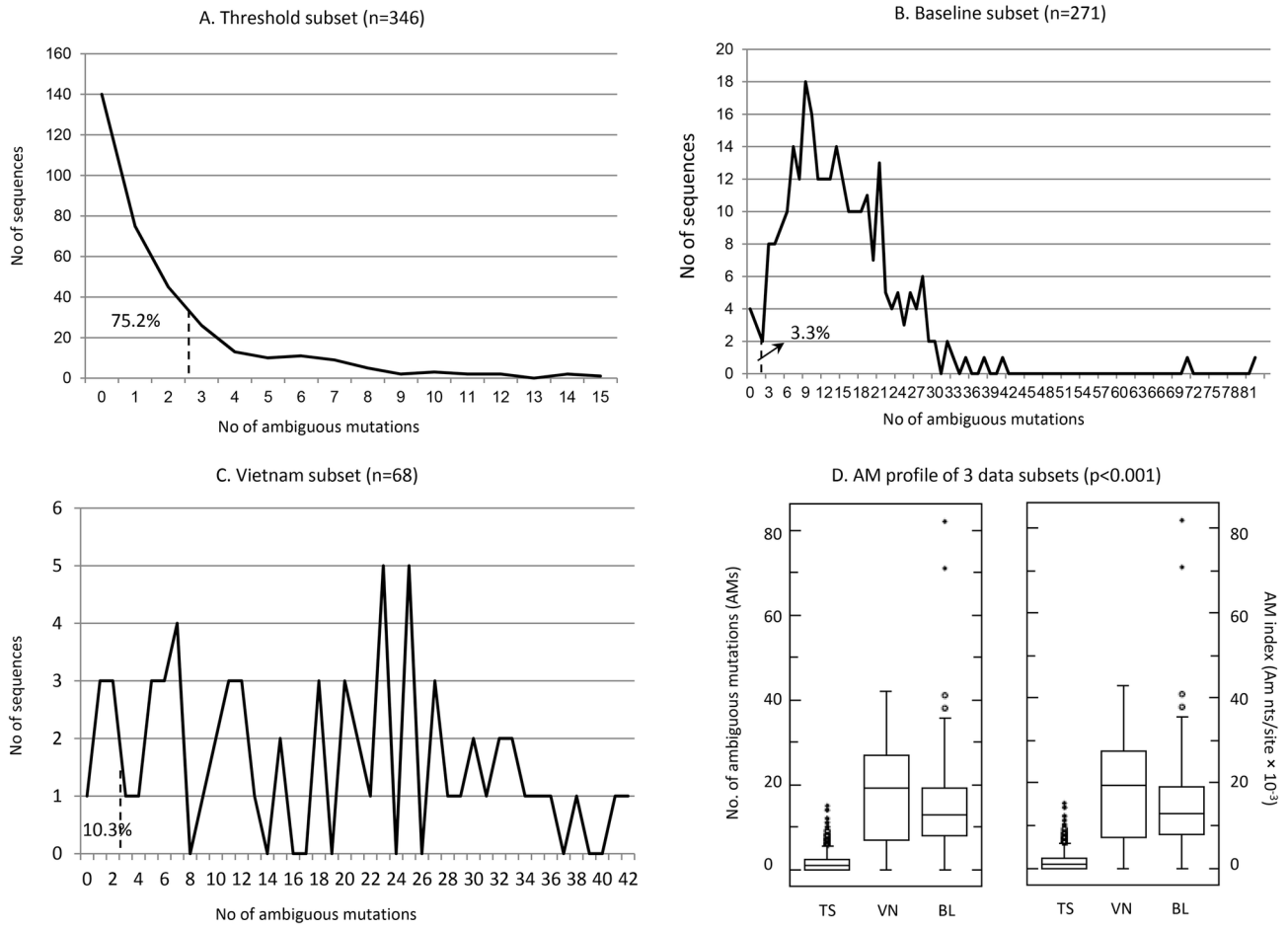


Figure 2. Distribution of ambiguous mutations and data statistical description of three data subsets. Sequence frequency distribution with number of ambiguous mutations (AMs) was plotted by subset: (A). Threshold (n=346), (B). Baseline (n=271), and (C). Vietnam (VN) (n=68); and the statistical description of the 3 data subsets was plot (D) by number and index of ambiguous mutations using the same method as described in Figure 1. The percentage in A-C indicated recent infections in a dataset classified by having ≤2 AMs per sequence (indicated by dash line).

doi: 10.1371/journal.pone.0077649.g002

12.790.743.32%; V77, 15.370.742.40%; I93, 62.1522.791.85%) in the order of threshold, Vietnam and baseline subsets. Interestingly, of these 12 DR sites, 10 were located in the protease gene and only two (Q151 and V179) were in the RT gene. Some might associate with polymorphism or have combination effects on DR and viral replication depending on subtypes [40,41]. It was also evident that the DR score of ambiguous mutated AAs were gradually accumulating in the Vietnam and baseline sequences comparing to the threshold ones.

By calculating the overall DR-associated mutations derived from pure mutated AAs or ambiguous mutated AAs (Figure 4), we found that the index of pure mutated AAs was constant across the 3 subsets (7.50-7.89×10^{-2} per DR site) (p=0.681); in contrast, the index of ambiguous mutated AAs increased significantly from 0.54×10^{-2} per DR site for threshold to 4.30-3.48×10^{-2} per DR site for Vietnam and baseline datasets (p<0.001), indicating that background of t-HIVDR existed in ART-naïve populations of these cohorts. Under such background, new HIVDR continued to develop in the studied populations.

Subtyping and ambiguous mutation preference between subtypes

Subtype of HIV sequences in each dataset was primarily determined using the online REGA HIV subtyping tool (Table 1). A single dominant subtype was found in Botswana and Malawi (subtype C), Kenya (subtype A), Nigeria (subtype G), China and Vietnam (CRF01-AE); whereas multiple subtypes and recombinants were identified in Tanzania and Angola. Among those HIVDR-TS sequences, 46% were subtype C, 29% CRF01_AE, 7% subtype A and 10.5% untypeable.

To explore the difference of ambiguous mutation between subtypes, we collected a dataset of 63 subtype B sequences

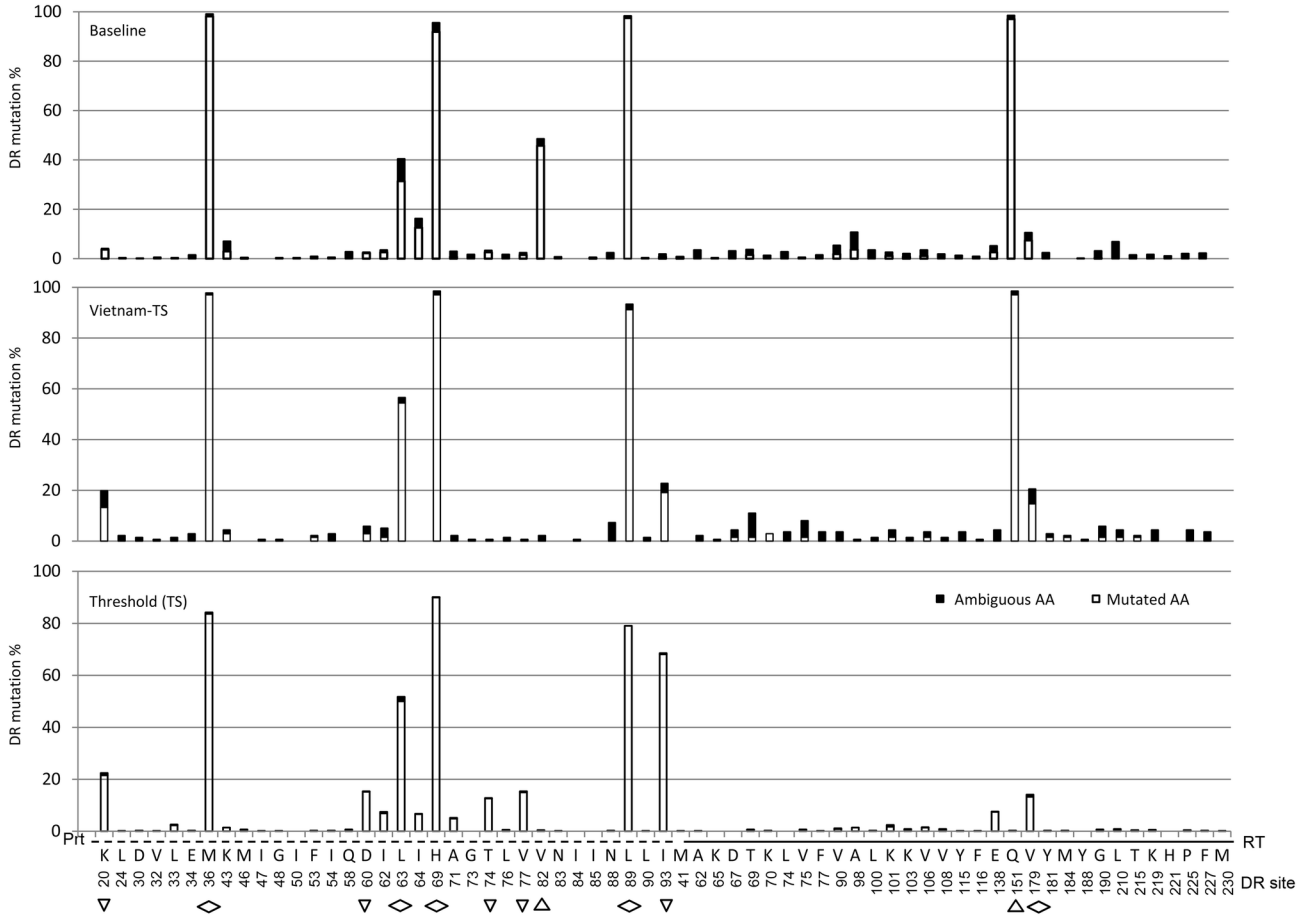


Figure 3. Proportional distribution of mutated and ambiguous mutated amino acids at HIVDR sites. The mutation score at each of the drug resistance sites [9] was proportionally calculated with the mutated and ambiguous mutated amino acids for all the sequences in the datasets. A mutated or ambiguous mutated amino acid was defined as an amino acid had mutated from a wild type to a pure non-synonymous mutation or an ambiguous mutation in the mixture allele. The scores were summed by 1 for a pure amino acid mutation and 0.5 for an ambiguous amino acid mutation, and then converted to percentages against the total number of wild-type amino acids at the site. The distribution of drug resistance mutation scores was plot by the dataset of Threshold (bottom panel), Vietnam (VN, central panel) and Baseline (top panel). The *x*-axis is the wild-type amino acids at drug resistance sites; the *y*-axis is the drug resistance mutation score (%). The sites with obvious score changes across the 3 datasets from bottom to top panel were labeled by up-triangle (increased), rhombus (remained), and down- triangle (decreased). Amino acids of protease gene (Prt) were top-dash lined, and of reverse transcriptase gene (RT) were top-solid lined.

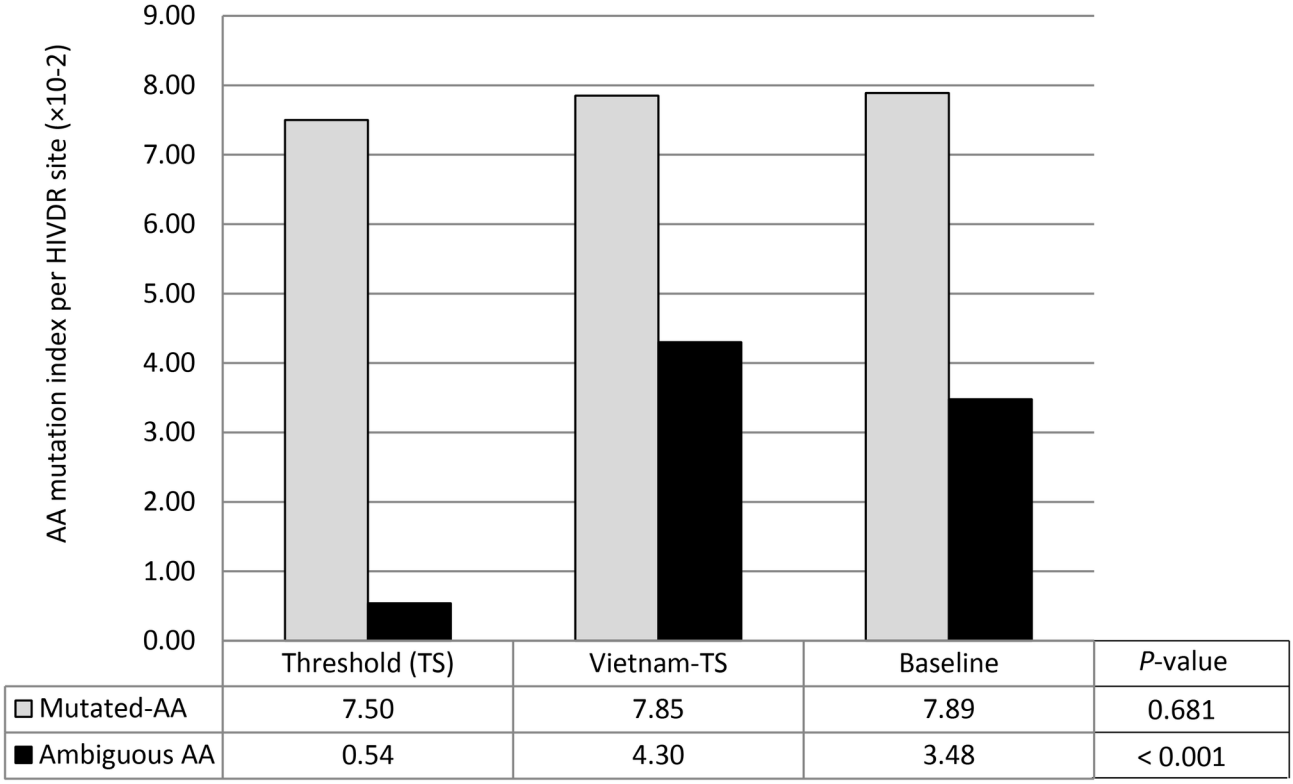
doi: 10.1371/journal.pone.0077649.g003

generated from specimens collected people infected with HIV-1 within 155 days [24], and compared them with all the non-B subtype or stratified non-subtype B sequences. These stratified sequences were selected from the dominant subtype(s) in the datasets, including subtype C from Botswana and Malawi, subtype A from Kenya and Tanzania, and CRF01_AE from China. Statistical analysis indicated that no significant difference of ambiguous mutation was found between subtype B and non-B subtypes (p=0.16) or subtype C (p=0.107); however, significant differences were noticed between subtypes B and A (p=0.001) or between subtype B and CRF01_AE (p=0.011, Table 3, Figure 1-insert).

We couldn't perform analysis of ambiguous mutation preference on the basis of infection route due to the incomplete infection route data from the China and Vietnam HIVDR-TS and the Nigeria baseline monitoring survey.

Discussion

Detection of HIV recent infections is challenging and crucial for accurate HIV incidence and t-HIVDR estimations. We pursued an investigational molecular approach using ambiguous mutation for determining HIV infection status and monitoring early development of HIVDR. We characterized ambiguous mutations in recent (HIVDR-TS) and established



	Threshold (TS)	Vietnam-TS	Baseline	*P*-value
Mutated-AA	7.50	7.85	7.89	0.681
Ambiguous AA	0.54	4.30	3.48	< 0.001

Figure 4. Index of mutated and ambiguous mutated amino acids at HIVDR sites by data subset. The total score of drug resistance mutations caused by pure mutated amino acids or by ambiguous mutated amino acids was calculated separately for each of the data subsets, and divided by the number of total drug resistance sites [9] to obtain the index of mutated or ambiguous mutated amino acids by subset. The definition and score calculation of pure mutated and ambiguous mutated amino acids were described in Figure 3.

doi: 10.1371/journal.pone.0077649.g004

infections (HIVDR baseline), and demonstrated the lower ambiguous mutation index was associated significantly with recent infections. We defined 2.04×10^{-3} ambiguous mutations/site as a measure for infections within one year, referred as recent infection in the current study. With the substitution rate defined, the proportion of subjects: 75.2% in the threshold, 3.3% in the baseline, and 10.3% in the Vietnam dataset, was classified being recent infections. These results provided data on the accuracy of defining HIV-1 recent infections using the WHO epidemiological criteria.

The dataset from Nigerian HIVDR monitoring baseline survey representing established infections exhibited a significantly higher mutation index (14.40×10^{-3} nts/site), which clearly differentiated them from the recent infections in the threshold subset, and served as a great calibrator for determining the ambiguous mutation index for recent infections. Based on our analyses, HIV infection status could range from 1 to 10 years for the majority of subjects in the Nigeria dataset, which reflects the reality because the subjects enrolled for ART included established and chronic infections according to the Nigeria HIV treatment guidelines [26,33]. The distribution curve of ambiguous mutations reflects the distribution of infection status of subjects in the dataset. For instance, a narrow sharp curve indicates a uniformity of subjects who had been infected around the same period of time whereas a wide curve indicates a wider range of infection status from recent to chronic infections. These could serve as a tool to evaluate uniformity of infection status in subjects from a dataset or cohort.

Table 3. Comparison of ambiguous mutation (AM) in the recent infections of subtype B and other subtypes.

Subtype	Number of sequences	AM index ($\times10^{-3}$ nts/site)	95% CI	*P*-value
Non-B subtypes	346	2.03	1.62-2.21	0.16
A	33	2.55	1.59-3.52	0.001
B	63	1.71	1.01-2.40	
C	183	1.40	0.99-1.81	0.107
CRF01_AE	20	3.26	2.02-4.50	0.011

doi: 10.1371/journal.pone.0077649.t003

Although the Vietnam dataset was also from HIVDR-TS, only 10.3% of the subjects could be identified for being recently infected. The overall high ambiguous mutation index

(18.32×10^{-3} nts/site) similar to the one found in the Nigeria baseline dataset was somewhat not surprising because in Vietnam the two highest HIV risk groups were IDUs and commercial sex workers [39,42] and the subjects in the Vietnam HIVDR-TS were enrolled at VCT [28,30]. In contrast, those from other HIVDR-TS surveys except for China were enrolled at ANC clinics. They were women attending their first pregnant visits and most likely infected through heterosexual transmission [27,29,31,32]. Because the mechanism of HIV evolution is transmission route dependent [2,21], IDUs have only about 40% of the chance being infected with a single virion and transmission of multi-viral strains through injection would amplify the founder effect in folds, leading to higher genetic diversities [20,43,44]. These might explain the Vietnam dataset and the two subjects with outlier sequences from China (6 IDUs were identified in the dataset, personal communication) and Angola. Thus, those with higher ambiguous mutations specifically in Vietnam dataset might not be established infections, but IDUs.

Emergence and transmission of HIVDR is an on-going concern in the scale-up ART programs in resource-limited settings. To increase the elements of HIVDR surveillance and monitoring, we utilized the ambiguous mutation approach to monitor early emerged DR mutations and distinguish them from fitted DR mutations. Our data revealed an increase trend in DR mutations caused by ambiguous mutations ($0.543.48\times10^{-2}$) but a constant level of fitted DR mutations ($7.50\text{-}7.89\times10^{-2}$) from recent to established infections, indicating that t-HIVDR background exists in ART-naïve populations in which new HIVDR mutations continue to emerge over time. We also identified multiple DR sites that had mutational scores increased, remained stable or deceased over the time of infections. This dynamic distribution profile may be valuable for predicting early development of possible DRs in HIV-infected populations which provide data to decision maker for regimen considerations and/or selections.

Utilization of ambiguous mutations for predicting time of HIV infections may represent a relatively new sensitive approach. There are limited data available which were mainly focused on subtype B sequences, including two publications on HIV *pol*-gene [23-25] and one on *env*-gene [45]. The analyses focused on *pol*-gene found that >0.45-0.5% of ambiguous nucleotides provided strong evidence against a recent infection within 1 year and that ambiguous mutation rate constantly increased by 0.2% per year for the first 8 years of HIV infections. Our finding in which recent infections were defined as having an ambiguous mutation rate of 2.03×10^{-3} nts/site/year is in agreement with these studies and this is also corroborated with the substitution rate of 2×10^{-3} nts/site/year as reported by others [19,20].

Our observation on subtype preference in viral ambiguity showed no significant difference between subtype B and overall non-B subtypes (p=0.16), which is consistent with a recent study [25]. However, our subtype-stratified analyses appear to show somewhat different pictures since subtype A (p=0.001) and CRF01-AE (p=0.011) did show higher ambiguity when they were compared to subtype B. We believe that the discrepant results may be attributed to the smaller sample size that we have for these subtypes (Table 3). In this study, we classified 75.2% of subjects in the threshold surveys were infected within one year and this is in agreement with a recent study in which 73% of the pregnant women were identified as recent infections within one year from the datasets generated using the same WHO epidemiological criteria for identifying recent infections in resource-limited countries [25].

However, there are limitations in our study. We characterized ambiguous mutations with datasets based on recent and established infections. However, due to the limited epidemiological data, we couldn't conduct analysis and interpret the data based on transmission modes. We could only confirm the heterosexual transmission route for the women recruited at ANC clinics because they were all at their first pregnancy and diagnosed the first time with HIV infections. Therefore, the founder effect would likely be the viral replication mechanism, and the ambiguous mutation detected in the HIVDR-TS datasets by Sanger sequencing would reflect the genetic diversity in the viral population. However, for the dataset with suspected IDUs, Sanger sequencing might not be able to resolve the genetic diversity due to the multi-virion infection nature. With the progress on new generation sequencing technologies, e.g. deep or single genome sequencing, the multi-virion infection could be resolved at individual variant level to reflect the true genetic diversity [44-46]. We identified significant higher ambiguous mutations occurred in subtype A and CRF01_AE based on relatively small sample size. Studies on a larger number of these subtypes using epidemiologically defined cohorts of HIV infections would be useful to confirm our findings and further our understanding of ambiguous mutation preferences. Lastly, we detected an increase of ambiguous mutations at HIVDR sites. Due to the lack of clinical data, we couldn't determine a threshold of early developed minor HIVDR mutations that would have clinical significance for treatment and regimen decision-making [3,47].

In summary, we characterized ambiguous mutations in HIV-1 protease and reverse transcriptase gene regions with likely recent and established infections. We defined an ambiguous mutation index for detecting HIV recent infections and characterized the distribution of ambiguous mutations for monitoring the early development of HIVDR. Our data suggest that molecular characterization of ambiguous mutations in HIV-1 may serve as an additional tool along with serologic assays to differentiate recent from established infections, evaluate infection status, and monitor the early development of HIVDR.

Acknowledgements

We thank our colleagues in the countries of Angola, Botswana, China, Kenya, Malawi, Tanzania and Vietnam for their support and contributions to the HIVDR surveys, and members in the Drug Resistance and Molecular Bioinformatics Team for generating the sequence data and technical support.
Dr. Duc B. Nguyen received training support from Emory AIDS International Training and Research Program (NIH/FIC D43 TW01042).

Disclaimer: Use of trade names is for identification only and does not constitute endorsement by the U.S. Department of Health and Human Services, the Public Health Service, or the Centers for Disease Control and Prevention. The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

Author Contributions

Conceived and designed the experiments: DPZ CY. Performed the experiments: MR EB DBN KD JRD. Analyzed the data: DPZ. Contributed reagents/materials/analysis tools: MR EB DBN CY JNN. Wrote the manuscript: DPZ CY JNN.

References

1. Bennett DE, Camacho RJ, Otelea D, Kuritzkes DR, Fleury H et al. (2009) Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. PLOS ONE 4: e4724. doi:10.1371/journal.pone.0004724. PubMed: 19266092.
2. Wensing AM, Boucher CA (2003) Worldwide transmission of drug-resistant HIV. AIDS Res 5: 140-155. PubMed: 14598563.
3. Johnson JA, Li JF, Wei X, Lipscomb J, Irlbeck D et al. (2008) Minority HIV-1 drug resistance mutations are present in antiretroviral treatment-naive populations and associate with reduced treatment efficacy. PLOS Med 5: e158. doi:10.1371/journal.pmed.0050158. PubMed: 18666824.
4. Sungkanuparph S, Oyomopito R, Sirivichayakul S, Sirisanthana T, Li PC et al. (2011) HIV-1 drug resistance mutations among antiretroviral-naive HIV-1-infected patients in Asia: results from the TREAT Asia Studies to Evaluate Resistance-Monitoring Study. Clin Infect Dis 52: 1053-1057. doi:10.1093/cid/cir107. PubMed: 21460324.
5. WHO (2012). WHO HIV Drug Resistance Report 2012. World Health Organization.
6. Who U, UNICEF (2011) Global HIV/AIDS Response: Epidemic Update and Health Sector Progress Towards Universal Access. WHO Publication
7. Oette M, Schülter E, Rosen-Zvi M, Peres Y, Zazzi M et al. (2012) Efficacy of antiretroviral therapy switch in HIV-infected patients: a 10-year analysis of the EuResist Cohort. Intervirology 55: 160-166. doi: 10.1159/000332018. PubMed: 22286887.
8. Ekstrand ML, Shet A, Chandy S, Singh G, Shamsundar R et al. (2011) Suboptimal adherence associated with virological failure and resistance mutations to first-line highly active antiretroviral therapy (HAART) in Bangalore, India. Int Health 3: 27-34. doi:10.1016/j.inhe.2010.11.003. PubMed: 21516199.
9. Johnson VA, Calvez V, Gunthard HF, Paredes R, Pillay D et al. (2013) Update of the drug resistance mutations in HIV-1: March 2013. Top Antivir Med 21: 6-14. PubMed: 23596273.
10. Brookmeyer R (2010) Measuring the HIV/AIDS epidemic: approaches and challenges. Epidemiol Rev 32: 26-37. doi:10.1093/epirev/mxq002. PubMed: 20203104.
11. Bennett DE, Myatt M, Bertagnolio S, Sutherland D, Gilks CF (2008) Recommendations for surveillance of transmitted HIV drug resistance in countries scaling up antiretroviral treatment. Antivir Ther 13 Suppl 2: 25-36. PubMed: 18575189.
12. Myatt M, Bennett DE (2008) A novel sequential sampling technique for the surveillance of transmitted HIV drug resistance by cross-sectional survey for use in low resource settings. Antivir Ther 13 Suppl 2: 37-48. PubMed: 18575190.
13. Cohen MS, Gay CL, Busch MP, Hecht FM (2010) The detection of acute HIV infection. J Infect Dis 202 Suppl 2: S270-S277. doi: 10.1086/655651. PubMed: 20846033.
14. Guy R, Gold J, Calleja JM, Kim AA, Parekh B et al. (2009) Accuracy of serological assays for detection of recent infection with HIV and estimation of population incidence: a systematic review. Lancet Infect Dis 9: 747-759. doi:10.1016/S1473-3099(09)70300-7. PubMed: 19926035.
15. Parekh BS, McDougal JS (2005) Application of laboratory methods for estimation of HIV-1 incidence. Indian J Med Res 121: 510-518. PubMed: 15817960.
16. Laeyendecker O, Brookmeyer R, Oliver AE, Mullis CE, Eaton KP et al. (2011) Factors Associated with Incorrect Identification of Recent HIV Infection Using the BED Capture Immunoassay. AIDS Res Hum Retrovir, 28: 816–22. PubMed: 22014036.
17. Preston BD, Poiesz BJ, Loeb LA (1988) Fidelity of HIV-1 reverse transcriptase. Science 242: 1168-1171. doi:10.1126/science.2460924. PubMed: 2460924.
18. Roberts JD, Bebenek K, Kunkel TA (1988) The accuracy of reverse transcriptase from HIV-1. Science 242: 1171-1173. doi:10.1126/science.2460925. PubMed: 2460925.
19. Suzuki Y, Yamaguchi-Kabata Y, Gojobori T (2000) Nucleotide substitution rates of HIV-1. AIDS Res 2: 39-47.
20. Rambaut A, Posada D, Crandall KA, Holmes EC (2004) The causes and consequences of HIV evolution. Nat Rev Genet 5: 52-61. doi: 10.1038/nrg1246. PubMed: 14708016.
21. Lemey P, Rambaut A, Pybus OG (2006) HIV evolutionary dynamics within and among hosts. AIDS Res 8: 125-140. PubMed: 17078483.
22. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT et al. (2008) Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci U S A 105: 7552-7557. doi:10.1073/pnas.0802203105. PubMed: 18490657.
23. Kouyos RD, von Wyl V, Yerly S, Böni J, Rieder P et al. (2011) Ambiguous nucleotide calls from population-based sequencing of HIV-1 are a marker for viral diversity and the age of infection. Clin Infect Dis 52: 532-539. doi:10.1093/cid/ciq164. PubMed: 21220770.
24. Ragonnet-Cronin M, Aris-Brosou S, Joanisse I, Merks H, Vallée D et al. (2012) Genetic diversity as a marker for timing infection in HIV-infected patients: evaluation of a 6-month window and comparison with BED. J Infect Dis 206: 756-764. doi:10.1093/infdis/jis411. PubMed: 22826337.
25. Andersson E, Shao W, Bontell I, Cham F, Cuong DD et al. (2013) Evaluation of sequence ambiguities of the HIV-1 pol gene as a method to identify recent HIV-1 infection in transmitted drug resistance surveys. Infect Genet Evol 18C: 125-131. PubMed: 23583545.
26. Ugbena R, Aberle-Grasse J, Diallo K, Bassey O, Jelpe T et al. (2012) Virological Response and HIV Drug Resistance 12 Months After Antiretroviral Therapy Initiation at 2 Clinics in Nigeria. Clin Infect Dis 54 Suppl 4: S375-S380. doi:10.1093/cid/cir1064. PubMed: 22544206.
27. Bussmann H, de la Hoz Gomez F, Roels TH, Wester CW, Bodika SM et al. (2011) Prevalence of transmitted HIV drug resistance in Botswana: lessons learned from the HIVDR-Threshold Survey conducted among women presenting for routine antenatal care as part of the 2007 national sentinel survey. AIDS Res Hum Retrovir 27: 365-372. doi:10.1089/aid.2009.0299. PubMed: 21034246.
28. Duc NB, Hien BT, Wagar N, Tram TH, Giang le T et al. (2012) Surveillance of Transmitted HIV Drug Resistance Using Matched Plasma and Dried Blood Spot Specimens From Voluntary Counseling and Testing Sites in Ho Chi Minh City. Vietnam: 2007-2008. Clin Infect Dis 54 Suppl 4: S343-347.
29. Kamoto K, Aberle-Grasse J (2008) Surveillance of transmitted HIV drug resistance with the World Health Organization threshold survey method in Lilongwe, Malawi. Antivir Ther 13 Suppl 2: 83-87. PubMed: 18575195.
30. Nguyen HT, Duc NB, Shrivastava R, Tran TH, Nguyen TA et al. (2008) HIV drug resistance threshold survey using specimens from voluntary counselling and testing sites in Hanoi, Vietnam. Antivir Ther 13 Suppl 2: 115-121. PubMed: 18575200.
31. Somi GR, Kibuka T, Diallo K, Tuhuma T, Bennett DE et al. (2008) Surveillance of transmitted HIV drug resistance among women attending antenatal clinics in Dar es Salaam, Tanzania. Antivir Ther 13 Suppl 2: 77-82. PubMed: 18575194.
32. Zhang J, Kang D, Fu J, Sun X, Lin B et al. (2010) Surveillance of transmitted HIV type 1 drug resistance in newly diagnosed hiv type 1-infected patients in Shandong Province, China. AIDS Res Hum Retrovir 26: 99-103. doi:10.1089/aid.2009.0184. PubMed: 20121622.
33. National AIDS/STDs Control Program FMoH, Abuja, Nigeria (2007) National Guideline for HIV and AIDS Treatment and Care in Adolescents and Adults.
34. Yang C, McNulty A, Diallo K, Zhang J, Titanji B et al. (2010) Development and application of a broadly sensitive dried-blood-spot-based genotyping assay for global surveillance of HIV-1 drug resistance. J Clin Microbiol 48: 3158-3164. doi:10.1128/JCM.00564-10. PubMed: 20660209.
35. Zhou Z, Wagar N, DeVos JR, Rottinghaus E, Diallo K et al. (2011) Optimization of a low cost and broadly sensitive genotyping assay for HIV-1 drug resistance surveillance and monitoring in resource-limited settings. PLOS ONE 6: e28184. doi:10.1371/journal.pone.0028184. PubMed: 22132237.

36. Inzaule S, Yang C, Kasembeli A, Nafisa L, Okonji J et al. (2013) Field evaluation of a broadly sensitive HIV-1 in-house genotyping assay for use with both plasma and dried blood spot specimens in a resource-limited country. J Clin Microbiol 51: 529-539. doi:10.1128/JCM.02347-12. PubMed: 23224100.
37. Woods CK, Brumme CJ, Liu TF, Chui CK, Chu AL et al. (2012) Automating HIV drug resistance genotyping with RECall, a freely accessible sequence analysis tool. J Clin Microbiol 50: 1936-1942. doi:10.1128/JCM.06689-11. PubMed: 22403431.
38. Hall AG (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41: 95-98.
39. Thanh DC, Hien NT, Tuan NA, Thang BD, Long NT et al. (2009) HIV risk behaviours and determinants among people living with HIV/AIDS in Vietnam. AIDS Behav 13: 1151-1159. doi:10.1007/s10461-008-9451-8. PubMed: 18787940.
40. Kearney M, Palmer S, Maldarelli F, Shao W, Polis MA et al. (2008) Frequent polymorphism at drug resistance sites in HIV-1 protease and reverse transcriptase. AIDS 22: 497-501. doi:10.1097/QAD.0b013e3282f29478. PubMed: 18301062.
41. Lisovsky I, Schader SM, Martinez-Cajas JL, Oliveira M, Moisi D et al. (2010) HIV-1 protease codon 36 polymorphisms and differential development of resistance to nelfinavir, lopinavir, and atazanavir in different HIV-1 subtypes. Antimicrob Agents Chemother 54: 2878-2885. doi:10.1128/AAC.01828-09. PubMed: 20404123.
42. Dean J, Ta Thi TH, Dunford L, Carr MJ, Nguyen LT et al. (2011) Prevalence of HIV type 1 antiretroviral drug resistance mutations in Vietnam: a multicenter study. AIDS Res Hum Retrovir 27: 797-801. doi:10.1089/aid.2011.0013. PubMed: 21366425.
43. Cohen MS, Shaw GM, McMichael AJ, Haynes BF (2011) Acute HIV-1 Infection. N Engl J Med 364: 1943-1954. doi:10.1056/NEJMra1011874. PubMed: 21591946.
44. Bar KJ, Li H, Chamberland A, Tremblay C, Routy JP et al. (2010) Wide variation in the multiplicity of HIV-1 infection among injection drug users. J Virol 84: 6241-6247. doi:10.1128/JVI.00077-10. PubMed: 20375173.
45. Park SY, Love TM, Nelson J, Thurston SW, Perelson AS et al. (2011) Designing a genome-based HIV incidence assay with high sensitivity and specificity. AIDS 25: F13-F19. doi:10.1097/QAD.0b013e328349f089. PubMed: 21716075.
46. Kearney M, Maldarelli F, Shao W, Margolick JB, Daar ES et al. (2009) Human immunodeficiency virus type 1 population genetics and adaptation in newly infected individuals. J Virol 83: 2715-2727. doi:10.1128/JVI.01960-08. PubMed: 19116249.
47. Gianella S, Richman DD (2010) Minority variants of drug-resistant HIV. J Infect Dis 202: 657-666. doi:10.1086/655397. PubMed: 20649427.