HIV-1 Subtype B Protease and Reverse Transcriptase Amino Acid Covariation

Soo-Yon Rhee[1], Tommy F. Liu[1], Susan P. Holmes[2], Robert W. Shafer[1]*

1 Division of Infectious Diseases, Department of Medicine, Stanford University, Stanford, California, United States of America, **2** Department of Statistics, Stanford University, Stanford, California, United States of America

Despite the high degree of HIV-1 protease and reverse transcriptase (RT) mutation in the setting of antiretroviral therapy, the spectrum of possible virus variants appears to be limited by patterns of amino acid covariation. We analyzed patterns of amino acid covariation in protease and RT sequences from more than 7,000 persons infected with HIV-1 subtype B viruses obtained from the Stanford HIV Drug Resistance Database (http://hivdb.stanford.edu). In addition, we examined the relationship between conditional probabilities associated with a pair of mutations and the order in which those mutations developed in viruses for which longitudinal sequence data were available. Patterns of RT covariation were dominated by the distinct clustering of Type I and Type II thymidine analog mutations and the Q151M-associated mutations. Patterns of protease covariation were dominated by the clustering of nelfinavir-associated mutations (D30N and N88D), two main groups of protease inhibitor (PI)–resistance mutations associated either with V82A or L90M, and a tight cluster of mutations associated with decreased susceptibility to amprenavir and the most recently approved PI darunavir. Different patterns of covariation were frequently observed for different mutations at the same position including the RT mutations T69D versus T69N, L74V versus L74I, V75I versus V75M, T215F versus T215Y, and K219Q/E versus K219N/R, and the protease mutations M46I versus M46L, I54V versus I54M/L, and N88D versus N88S. Sequence data from persons with correlated mutations in whom earlier sequences were available confirmed that the conditional probabilities associated with correlated mutation pairs could be used to predict the order in which the mutations were likely to have developed. Whereas accessory nucleoside RT inhibitor–resistance mutations nearly always follow primary nucleoside RT inhibitor–resistance mutations, accessory PI-resistance mutations often preceded primary PI-resistance mutations.

Citation: Rhee SY, Liu TF, Holmes SP, Shafer RW (2007) HIV-1 subtype B protease and reverse transcriptase amino acid covariation. PLoS Comput Biol 3(5): e87. doi:10.1371/journal.pcbi.0030087

Introduction

HIV-1 is a highly mutable pathogen. In the decades since it entered human populations, it has accumulated extensive sequence variation leading to the development of different subtypes and recombinant forms [1]. Although the enzymatic targets of therapy are among the most conserved parts of the HIV-1 genome, these too can develop marked variation, particularly in the setting of selective antiretroviral drug pressure. Indeed, it is not uncommon for drug therapy to select for protease and reverse transcriptase (RT) variants containing substitutions at more than 10% of their amino acids [2]. However, despite this high degree of mutation, the spectrum of possible virus variants appears to be limited by patterns of amino acid covariation.

In 2003, we published two studies that examined the extent of covariation among RT and protease residues in the presence and absence of antiretroviral therapy [3,4]. Despite the relatively large size of the datasets in these studies—2,244 protease sequences and 1,210 RT sequences—there were insufficient data to examine patterns of covariation of different mutations at the same position. As more sequence data have become available, we are now analyzing covariation among mutations (rather than positions) in protease and RT. This expanded analysis uses a highly specific measure of covariation, the Jaccard similarity coefficient, and a multi-dimensional scaling based on this coefficient. In addition, we examine the relationship between conditional probabilities associated with a mutation pair and the order in which those mutations develop in viruses for which longitudinal sequence data are available.

Results

Protease

Protease sequences from 3,982 protease inhibitor (PI)–naive individuals and from 3,475 PI-experienced individuals were available for analysis. The PI-experienced individuals had received a median of 1 PI (interquartile range, 1–3).

Jaccard similarity coefficients and their standardized Z scores were calculated for all pairs of mutations at different positions present three or more times among the sequences from PI-naive and PI-experienced individuals. Among 19,203 pairs of mutations from the PI-experienced individuals, 161 pairs were significantly associated after adjusting for multiple comparisons by controlling the family-wise error rate at <0.01. Of these 161 pairs, 92 (57%) were positively associated ($Z > 5.1$, unadjusted $p < 4.4 \times 10^{-7}$) and 69 (43%) were

Editor: Greg Tucker-Kellogg, Lilly Systems Biology, Singapore

Received November 24, 2006; **Accepted** April 2, 2007; **Published** May 11, 2007

Abbreviations: NNRTI, nonnucleoside RT inhibitor; NRTI, nucleoside RT inhibitor; PI, protease inhibitor; RT, reverse transcriptase; TAM, thymidine analog mutation

* To whom correspondence should be addressed. E-mail: rshafer@stanford.edu

Author Summary

The identification of which mutations in a protein covary has played a major role in both structural and evolutionary biology. Covariation analysis has been used to help predict unsolved protein structures and to better understand the functions of proteins with known structures. The large number of published genetic sequences of the targets of HIV-1 therapy has provided an unprecedented opportunity to identify dependencies among mutations in these proteins that can be exploited to design inhibitors that have high genetic barriers to resistance. In our analysis, we identified many pairs of covarying drug-resistance mutations in HIV-1 protease and reverse transcriptase and organized them into clusters of mutations that often develop in a predictable order. Inhibitors that are active against early drug-resistant mutants are likely to be less prone to the development of resistance, whereas inhibitors that are active against fully evolved clusters of mutations may be useful drugs for salvage therapy.

negatively associated ($Z < -5.0$, unadjusted $p < 4.8 \times 10^{-7}$). Table 1 shows the Jaccard similarity coefficients and conditional probabilities of the 40 strongest positively associated protease mutation pairs and the ten strongest negatively associated protease mutation pairs. Table S1 shows the complete list of 161 statistically significant mutation pairs.

For the positively associated mutation pairs, Table 1 also contains two columns with data on the temporal order in which correlated mutations occurred in sequences with both mutations from persons in which an earlier sequence was available that contained only one of the two mutations. For example, the first row shows that among persons with both I54V and V82A in whom an earlier sequence contained only one of these two mutations was available, I54V occurred first in nine (26%) of 34 people, and V82A occurred first in 25 (74%) of 34 people ($p < 0.01$). In contrast, the fourth row shows that among persons with both A71V and L90M, each of the mutations was as likely to occur first (26 of 51 versus 25 of 51; $p = \text{NS}$). Figure S1 plots the relationship between the log of the ratio of the conditional probability of two mutations versus the log of the ratio in which two mutations develop, indicating that the conditional dependence between mutations is highly correlated with the order in which the mutations develop when they occur together ($r^2 = 0.56$, $p < 0.001$).

Among the 18 positively associated pairs in Table 1 containing a major and an accessory PI-resistance mutation (as defined in Methods), the accessory mutation appeared first more often in 12 of the 18 pairs. There were several striking patterns of temporal association among these 18 pairs of correlated major and accessory mutations. The major mutation L90M preceded the accessory mutation G73S in 31 of 34 persons for whom temporal data were available. In contrast, the accessory mutation L63P preceded L90M in 160 of 172 persons, and the accessory mutations L10I and A71V preceded the major mutation I84V in 51 of 59 and 35 of 38 persons, respectively.

The Jaccard dissimilarity coefficients associated with 595 pairs of 35 mutations were used for a multidimensional scaling. The mutations included in this analysis were the 22 positively associated mutations in Table 1 and 13 additional clinically relevant PI-resistance mutations (L10F, V32I, L33F, I47V, I50V/L, F53L, I54L/M, Q58E, L76V, V82T, and N88S). Figure 1 plots the mutations along axes representing the first two principal components. The first principal component accounted for 10% of the total inertia and separates the nelfinavir-resistance mutations D30N and N88D from the main group of PI-resistance mutations. The second principal component accounted for 7% of the total inertia and separates V82A-associated mutations (I54V, L24I, and M46L) from L90M-associated mutations (M46I, G73S, and I84V). Finally, the lower-left part of the figure contains a cluster with seven of the 11 mutations recently reported to be associated with phenotypic and clinical resistance to the newest PI, darunavir (V32I, L33F, I47V, I50V, I54L/M, and L76V).

At several positions, there was sufficient data to contrast covariation patterns for different mutations. For example, M46I/L were each significantly associated with L10I, L24I, V32I, L33F, I54V, V82A, and L90M. However, M46I was uniquely associated with F53L, G73S/T, V82F/T, I84V, and N88S. I54V was significantly associated with L10F, L24I, L33F, M46I/L, G48V, F53L, V82A/F/T, I84V, and L90M. In contrast, I54L/M were significantly associated only with L33F, M46I, I47V, I84V, and L90M. N88D was positively associated with D30N and negatively associated with M46I, whereas N88S was negatively associated with D30N and positively associated with M46I. Of note, the divergent associations of different mutations at positions 46 and 88 have previously been reported by Hoffman and coworkers [5].

Among 7,131 pairs of mutations in sequences from PI-naive persons, 65 pairs were significantly associated (family-wise error rate < 0.01; Table S2). All but three of the positive associations among PI-naive persons were weaker (i.e., had a lower Z score) than the positive associations among treated persons in Table 1.

Reverse Transcriptase

RT sequences from 2,601 RT inhibitor–naive and from 5,188 RT inhibitor–experienced individuals were available for analysis. The RT inhibitor experienced individuals had received a median of three nucleoside RT inhibitors (NRTIs; interquartile range, 2–4) and zero nonnucleoside RT inhibitors (NNRTIs; interquartile range, 0–1).

Jaccard similarity coefficients and their standardized Z scores were calculated for all pairs of RT mutations at different positions present three or more times among the sequences from RT inhibitor–experienced and –naive persons. Among 65,624 pairs of mutations from the RT inhibitor–experienced persons, 327 pairs were significantly associated after adjusting for multiple comparisons by controlling the family-wise error rate at <0.01. Of these 327 pairs, 213 (65%) were positively associated ($Z > 5.2$, unadjusted $p < 2 \times 10^{-7}$) and 114 (35%) were negatively associated ($Z < -5.0$, unadjusted $p < 5 \times 10^{-7}$). Table 2 shows the Jaccard similarity coefficients and conditional probabilities of the 40 strongest positively associated RT mutation pairs and the ten strongest negatively associated RT mutation pairs. Table S3 shows the complete list of 327 statistically significant RT mutation pairs.

Positively associated mutation pairs consisted primarily of Type I or II thymidine analog mutations (TAMs; as defined in Methods); accessory NRTI mutations that occurred in combination with Type I or II TAMs (K43E, E44D, V118I, H208Y, D218E); and Q151M-associated mutations (V75I,

Table 1. Forty Highest Positively Correlated Protease Mutation Pairs and Ten Highest Negatively Correlated Protease Mutation Pairs from PI-Experienced Persons

Correlations	Mut X	Mut Y	Association[a]					P(Y|X)	P(X|Y)	Temporal Order[b]		Mutation Classifications[c]
			J	J_{RAND}	J_{SE}	Z	p			$N_{X0\rightarrow XY}$	$N_{0Y\rightarrow XY}$	
Positive Correlations	54V	82A	0.51	0.11	0.02	23.8	0	0.73	0.63	9	25	Maj
	30N	88D	0.59	0.04	0.03	21.0	0	0.62	0.91	13	2	Maj
	35D	36I	0.40	0.15	0.01	17.8	0	0.55	0.60	57	19	Maj – Misc
	71V	90M	0.38	0.17	0.01	15.8	0	0.59	0.51	26	25	Maj – Acc
	10I	54V	0.34	0.14	0.01	14.8	0	0.39	0.69	35	14	Maj – Acc
	10I	90M	0.38	0.19	0.01	14.8	0	0.54	0.56	31	34	Maj – Acc
	10I	82A	0.34	0.15	0.01	14.5	0	0.42	0.65	28	25	Maj – Acc
	71V	82A	0.34	0.14	0.01	14.4	0	0.45	0.57	15	27	Maj – Acc
	54V	71V	0.32	0.13	0.01	14.2	0	0.60	0.41	16	29	Maj – Acc
	10I	71V	0.36	0.18	0.01	14.1	0	0.48	0.58	40	31	Acc
	20R	36I	0.29	0.07	0.02	14.1	0	0.89	0.30	3	47	Acc
	77I	93L	0.34	0.20	0.01	11.8	0	0.54	0.47	26	36	Acc
	46L	82A	0.25	0.07	0.02	11.7	0	0.68	0.28	1	17	Maj
	73S	90M	0.21	0.07	0.01	11.5	0	0.87	0.22	3	31	Maj – Acc
	36I	62V	0.30	0.16	0.01	11.3	0	0.50	0.43	39	29	Acc – Misc
	84V	90M	0.23	0.09	0.01	11.2	0	0.70	0.26	3	44	Maj
	46I	90M	0.26	0.13	0.01	10.6	0	0.57	0.33	18	44	Maj
	10I	84V	0.22	0.09	0.01	10.5	0	0.25	0.69	51	8	Maj – Acc
	62V	90M	0.30	0.18	0.01	10.1	0	0.48	0.45	47	28	Maj – Acc
	71V	73S	0.19	0.06	0.01	10.0	0	0.21	0.73	24	4	Maj – Acc
	20R	54V	0.20	0.06	0.01	9.5	0	0.54	0.24	7	16	Maj – Acc
	63P	90M	0.38	0.29	0.01	9.4	0	0.40	0.89	160	12	Maj – Acc
	54V	90M	0.24	0.14	0.01	9.3	0	0.53	0.31	14	41	Maj
	46L	54V	0.20	0.06	0.01	9.2	0	0.53	0.24	11	10	Maj
	90M	93L	0.31	0.20	0.01	9.2	0	0.49	0.45	19	73	Maj – Acc
	24I	54V	0.17	0.04	0.01	9.0	0	0.71	0.18	3	2	Maj
	36I	54V	0.23	0.12	0.01	9.0	0	0.32	0.45	35	22	Maj – Acc
	24I	82A	0.16	0.04	0.01	8.9	0	0.75	0.17	0	7	Maj
	10I	46I	0.24	0.13	0.01	8.8	0	0.30	0.53	57	22	Maj – Acc
	20R	35D	0.17	0.07	0.01	8.8	0	0.64	0.19	1	53	Acc – Misc
	46I	84V	0.20	0.08	0.01	8.8	0	0.27	0.43	21	13	Maj
	71V	84V	0.19	0.09	0.01	8.4	0	0.22	0.54	35	3	Maj – Acc
	20R	82A	0.17	0.06	0.01	8.3	0	0.53	0.20	10	22	Maj – Acc
	10I	93L	0.30	0.20	0.01	8.2	2E-16	0.47	0.44	21	67	Acc
	10I	62V	0.28	0.18	0.01	8.0	1E-15	0.41	0.46	38	46	Acc – Misc
	35D	37D	0.19	0.10	0.01	7.9	2E-15	0.24	0.49	25	7	Misc
	48V	82A	0.13	0.03	0.01	7.8	6E-15	0.77	0.13	4	11	Maj
	20I	90M	0.13	0.05	0.01	7.7	1E-14	0.77	0.13	7	16	Maj – Acc
	24I	46L	0.18	0.03	0.02	7.5	6E-14	0.43	0.24	1	3	Maj
	62V	93L	0.28	0.19	0.01	7.5	7E-14	0.47	0.40	24	47	Acc – Misc
Negative correlations	30N	82A	0.00	0.08	0.00	−82.5	0	0.00	0.00	—	—	Maj
	82A	88D	0.00	0.06	0.00	−53.5	0	0.00	0.00	—	—	Maj
	24I	90M	0.00	0.04	0.00	−37.7	0	0.01	0.00	—	—	Maj
	36I	77I	0.03	0.17	0.00	−31.9	0	0.07	0.06	—	—	Acc
	30N	73S	0.00	0.05	0.00	−27.4	0	0.00	0.00	—	—	Maj – Acc
	30N	54V	0.01	0.07	0.00	−22.9	0	0.02	0.01	—	—	Maj
	30N	90M	0.02	0.09	0.00	−18.4	0	0.08	0.02	—	—	Maj
	73S	88D	0.00	0.04	0.00	−17.9	0	0.00	0.00	—	—	Maj – Acc
	16E	63P	0.00	0.02	0.00	−17.3	0	0.11	0.00	—	—	Acc – Misc
	64V	73S	0.01	0.06	0.00	−15.4	0	0.01	0.03	—	—	Acc – Misc

[a]J: Jaccard similarity coefficients; J_{RAND}: expected value of the Jaccard similarity coefficient assuming Mut X and Mut Y occur independently based on a permutation procedure described in Methods; J_{SE}: the standard error of J; Z: $(J - J_{RAND}) / J_{SE}$. Mutation pairs are ordered by their Z statistic.

[b]Based on the number of persons with both mutations (Mut X and Mut Y) in whom an earlier sequence containing only Mut X or Mut Y: $N_{X0\rightarrow XY}$ is the number in whom Mut X developed first; $N_{0Y\rightarrow XY}$ is the number in whom Mut Y developed first.

[c]Major (Maj) indicates mutations in the protease substrate cleft or that directly reduce drug susceptibility. Accessory (Acc) indicates mutations that are commonly believed to reduce susceptibility only in the presence of a major mutation or to compensate for the decreased replication associated with enzymes containing major mutations. Miscellaneous (Misc) includes the remaining mutations, including some that are treatment-associated (see Methods) and others that are not. When both mutations belong to the same category, only one abbreviation is shown.

doi:10.1371/journal.pcbi.0030087.t001

F77L, F116Y). Among the top 40 associated mutation pairs, there were only three positive associations between Type I and II TAMs (M41L, L210W, and T215Y with D67N). The strongest significant association between an NRTI and an NNRTI mutation was between L74V and Y181C ($J = 0.17$, $Z = 8.9$, unadjusted $p < 1 \times 10^{-11}$). Of note, the associations between the five accessory mutations listed above and Type I and II TAMs have also previously recently been described by

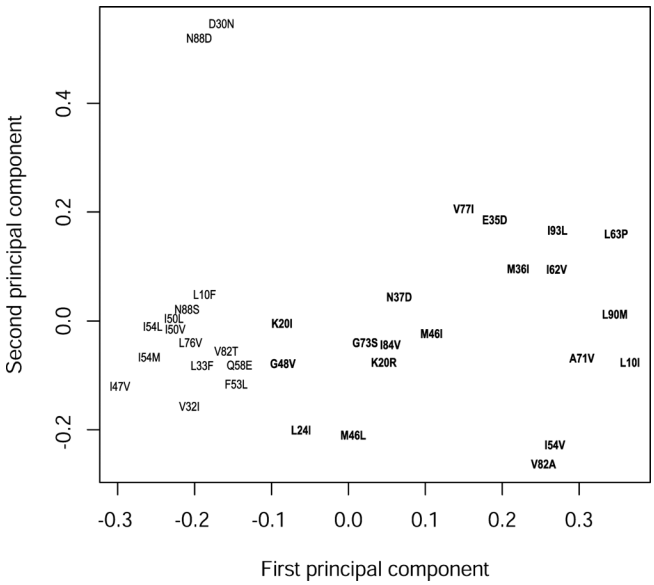

Figure 1. Multidimensional Scaling of 35 HIV-1 Protease Mutations

Includes the 22 mutations obtained from the mutation pairs with the highest positive association (Table 1) in bold, and 13 additional clinically relevant protease inhibitor resistance mutations (L10F, V32I, L33F, I47V, I50V/L, F53L, I54L/M, Q58E, L76V, V82T, and N88S). The graph is a 2-D projection of the distances among the 35 mutations, in which the distance between any two mutations is measured by their Jaccard dissimilarity coefficient among persons who have received at least one protease inhibitor.

doi:10.1371/journal.pcbi.0030087.g001

Svicher and coworkers [6] and Cozzi-Lepri and coworkers in independent datasets [7]. The conditional probabilities and the temporal data columns show that each of the accessory NRTI mutations consistently follows the Type I or II TAMs. Among 12 pairs with a TAM and an accessory mutation, the TAM occurred first more often in all 12 pairs and was preceded by the accessory mutation in only 6% of pairs. In addition to the five accessory mutations in Table 2 (K43E, E44D, V118I, H208Y, and D218E), other NRTI mutations that consistently followed TAMs included the known treatment-selected mutations T69D and T69N. Figure S2 plots the relationship between the log of the ratio of the conditional probability of two mutations versus the log of the ratio in which two mutations develop, indicating that the conditional dependence between mutations is highly correlated with the order in which the mutations develop when they occur together ($r^2 = 0.81$, $p < 0.001$).

The Jaccard dissimilarity coefficients associated with the 561 pairs of 34 mutations were used for a multidimensional scaling. The mutations included in this analysis were the 23 positively associated mutations in Table 2 and 11 additional clinically relevant NRTI-resistance mutations (K65R, A62V, T69ins, L74I/V, V75M, Y115F, M184V, and K219R/E/N). Figure 2 plots the mutations along axes representing the first two principal components. The first principal component accounts for 13% of the total inertia and separates the TAMs from the Q151M-associated mutations, whereas the second principal component accounts for 9% of the total inertia and separates the Type I and Type II TAMs. A62V, K65R, and Y115F are mutations that cluster with Q151M but may also occur with Type II (but not Type I) TAMs. D67N is a Type II TAM that can also occur with Type I TAMs, and it therefore occurs between Type I TAMs and Type II TAMs in terms of the second principal component. The non-TAM mutations, M184V and L74V, demonstrated no clustering with other NRTI-associated mutations.

At several positions, there was sufficient data to contrast covariation patterns for different mutations (Table 2, Figure 2, and Table S3). The Type I TAM, T215Y, clustered with other Type I TAMs, whereas the Type II TAM, T215F, clustered with other Type II TAMs. K219Q/E were Type II TAMs that cluster with other Type II TAMs. In contrast, two less common mutations at this position (K219N/R) were positively associated with Type I TAMs. T69D was associated with both Type I and Type II TAMs, whereas T69N was associated only with Type II TAMs. L74V was associated with the NNRTI-resistance mutations L100I, K103N, and Y181C, whereas L74I was associated with M41L. V75I was associated with Q151M-associated mutations, whereas V75M was positively associated with the Type I TAMs.

Among 19,431 pairs of mutations in sequences from RT inhibitor–naive persons, 41 pairs were significantly associated (family-wise error rate <0.01; Table S4). However, all of the positive associations among RT inhibitor–naive persons were weaker (i.e., had a lower Z score) than the positive associations among treated persons in Table 2.

Discussion

In this analysis of amino acid covariation in protease and RT sequences from more than 7,000 persons infected with HIV-1 subtype B viruses, we confirmed several previously reported patterns of amino acid covariation and identified many new patterns of covariation. Multidimensional scaling further organized many of the correlations into clusters of co-occurring mutations. RT covariation was dominated by the distinct clustering of the TAMs and Q151M-associated mutations, and by the separation of the Type I and Type II TAMs. Protease covariation was dominated by the clustering of nelfinavir-associated mutations (D30N and N88D), two main groups of PI-resistance mutations associated either with V82A or L90M, and a newly identified cluster of the mutations V32I, L33F, I47V, I50V, I54L/M, and L76V. This new cluster of mutations is associated with decreased susceptibility to all PIs, including the salvage therapy PIs amprenavir and lopinavir and the recently approved PI darunavir. Although none of the sequences in this study were from patients who received darunavir, this drug is highly similar to amprenavir and is affected by the same PI-resistance mutations.

Previous studies of HIV-1 covariation have used either the Pearson correlation for binomial random variables or mutual information [3–6,8–10]. The correlation coefficient is overly sensitive to rare pairs of mutations because its statistical significance is based on a departure from equality between the diagonal and off-diagonal products of a 2×2 contingency table. In contrast, mutual information is insensitive to rare pairs of mutations, approaching a high level only for commonly occurring pairs of mutations. We therefore used the Jaccard similarity coefficient, which uses only those sequences in which at least one of a pair of mutations is present, and we assessed the significance of this coefficient using a distribution based on the underlying data.

We also used a conservative correction for multiple

Table 2. Forty Highest Positively Correlated RT Mutation Pairs and Ten Highest Negatively Correlated RT Mutation Pairs from RTI-Experienced Persons

Correlations	Mut X	Mut Y	Association[a]					P(Y|X)	P(X|Y)	Temporal Order[b]		Mutation Classifications[c]
			J	J_{RAND}	J_{SE}	Z	p			$N_{X0\rightarrow XY}$	$N_{0Y\rightarrow XY}$	
Positive correlations	41L	215Y	0.73	0.24	0.01	52.0	0	0.84	0.86	22	29	I
	210W	215Y	0.60	0.17	0.01	37.2	0	0.94	0.62	16	64	I
	41L	210W	0.59	0.17	0.01	36.5	0	0.61	0.94	59	10	I
	70R	219Q	0.53	0.10	0.01	29.7	0	0.56	0.90	22	12	II
	67N	70R	0.45	0.15	0.01	25.2	0	0.55	0.70	21	18	II
	67N	219Q	0.41	0.11	0.01	23.6	0	0.43	0.88	9	6	II
	118I	210W	0.39	0.10	0.01	22.0	0	0.71	0.47	3	61	I – Acc
	116Y	151M	0.73	0.01	0.04	18.5	0	0.95	0.76	0	4	151
	44D	210W	0.30	0.06	0.01	17.7	0	0.92	0.31	4	44	I – Acc
	215F	219Q	0.36	0.06	0.02	17.7	0	0.65	0.44	0	21	II
	41L	118I	0.29	0.12	0.01	16.7	0	0.32	0.78	102	4	I – Acc
	70R	215F	0.29	0.07	0.01	16.6	0	0.31	0.78	27	5	II
	210W	211K	0.33	0.18	0.01	15.7	0	0.73	0.38	23	48	I – Misc
	44D	118I	0.30	0.05	0.02	15.3	0	0.68	0.34	15	13	Acc
	118I	215Y	0.27	0.12	0.01	15.1	0	0.73	0.30	5	100	I – Acc
	67N	118I	0.27	0.11	0.01	14.9	0	0.32	0.63	41	8	II – Acc
	67N	210W	0.30	0.15	0.01	14.4	0	0.41	0.52	21	35	I – II
	41L	44D	0.20	0.07	0.01	14.4	0	0.20	0.99	71	0	I – Acc
	41L	67N	0.34	0.20	0.01	14.4	0	0.45	0.57	50	28	I – II
	44D	215Y	0.21	0.07	0.01	14.3	0	0.98	0.21	6	74	I – Acc
	67N	69D	0.20	0.06	0.01	13.6	0	0.21	0.85	14	1	II – Misc
	67N	215F	0.22	0.08	0.01	13.5	0	0.24	0.75	26	1	II
	208Y	210W	0.20	0.05	0.01	12.8	0	0.81	0.21	4	41	I – Acc
	123E	177E	0.25	0.12	0.01	12.7	0	0.36	0.45	13	7	Misc
	75I	116Y	0.57	0.01	0.04	12.6	0	0.76	0.70	1	2	151
	77L	116Y	0.58	0.01	0.05	12.5	0	0.77	0.70	0	1	151
	43E	210W	0.19	0.04	0.01	12.4	0	0.86	0.19	2	30	I – Acc
	77L	151M	0.52	0.01	0.04	12.2	0	0.81	0.60	0	3	151
	69N	70R	0.18	0.04	0.01	12.0	0	0.83	0.19	3	19	II – Misc
	44D	208Y	0.26	0.03	0.02	11.7	0	0.37	0.48	6	5	Acc
	218E	219Q	0.21	0.04	0.01	11.7	0	0.68	0.23	0	28	II – Acc
	75I	151M	0.49	0.01	0.04	11.7	0	0.78	0.57	0	6	151
	75I	77L	0.54	0.01	0.05	11.5	0	0.68	0.72	0	1	151
	211K	215Y	0.37	0.26	0.01	11.5	0	0.49	0.59	69	43	I – Misc
	60I	67N	0.21	0.09	0.01	11.5	0	0.59	0.24	7	31	II – Misc
	44D	67N	0.17	0.06	0.01	11.4	0	0.72	0.19	5	30	II – Acc
	118I	208Y	0.20	0.04	0.01	11.3	0	0.23	0.58	14	6	Acc
	41L	208Y	0.15	0.05	0.01	11.3	0	0.15	0.94	55	2	I – Acc
	67N	215Y	0.30	0.20	0.01	11.2	0	0.53	0.41	17	57	I – II
	214L	219Q	0.21	0.07	0.01	11.2	0	0.37	0.33	12	9	II – Misc
Negative correlations	210W	214L	0.01	0.09	0.00	−50.2	0	0.01	0.02	—	—	I – Misc
	214L	215Y	0.01	0.10	0.00	−48.8	0	0.04	0.01	—	—	I – Misc
	44D	215F	0.00	0.04	0.00	−35.9	0	0.00	0.00	—	—	II – Acc
	44D	214L	0.00	0.05	0.00	−34.6	0	0.01	0.00	—	—	Acc – Misc
	65R	215Y	0.00	0.02	0.00	−33.7	0	0.01	0.00	—	—	I – II
	41L	214L	0.03	0.11	0.00	−25.8	0	0.03	0.10	—	—	I – Misc
	70R	210W	0.04	0.13	0.00	−24.5	0	0.07	0.07	—	—	I – II
	162C	166R	0.00	0.05	0.00	−24.1	0	0.00	0.01	—	—	Misc
	121H	123E	0.00	0.02	0.00	−23.4	0	0.01	0.00	—	—	Misc
	122E	123E	0.06	0.16	0.00	−22.6	0	0.09	0.13	—	—	Misc

[a]J: Jaccard similarity coefficients; J_{RAND}: expected value of the Jaccard similarity coefficient assuming Mut X and Mut Y occur independently based on a permutation procedure described in Methods; J_{SE} : the standard error of J; Z: $(J - J_{RAND}) / J_{SE}$. Mutation pairs are ordered by their Z statistic.

[b]Based on the number of persons with both mutations (Mut X and Mut Y) in whom an earlier sequence containing only Mut X or Mut Y: $N_{X0\rightarrow XY}$ is the number in whom Mut X developed first; $N_{0Y\rightarrow XY}$ is the number in whom Mut Y developed first.

[c]I: Type I TAMs; II: Type II TAMs; Acc: accessory NRTI-resistance mutations; 151: Q151M-associated mutations; Miscellaneous (Misc) includes the remaining mutations, including some that are treatment associated (see Methods) and others that are not. When both mutations belong to the same category, only one abbreviation is shown.

doi:10.1371/journal.pcbi.0030087.t002

comparisons (Holm's method) because our analysis was not designed to identify all covarying mutations but only those with the strongest association. Without a correction for multiple comparisons, 753 pairs of protease mutations from PI-experienced persons and 2,061 pairs of RTI mutations from RTI-experienced persons had a significant Jaccard similarity coefficient at a p-value of 0.01 but with the Holm's correction, only 161 pairs of protease mutations and 327

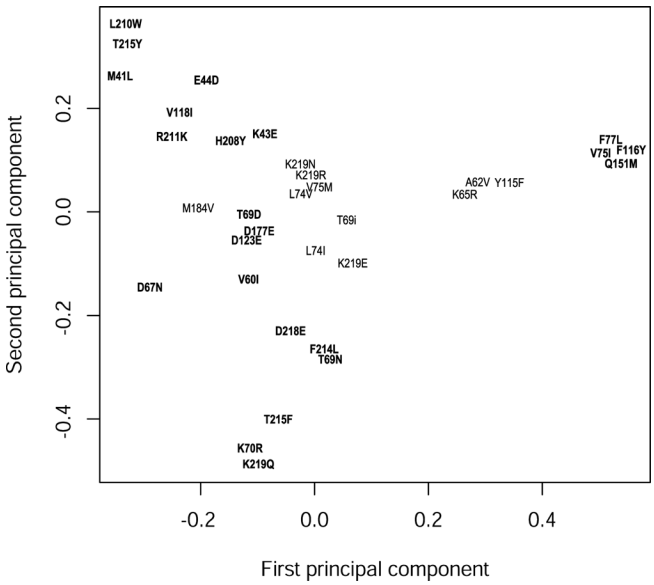

Figure 2. Multidimensional Scaling of 34 HIV-1 Reverse Transcriptase Mutations

Includes the 23 mutations obtained from the mutation pairs with highest positive association (Table 2) in bold, and 11 additional clinically relevant nucleoside RT inhibitor resistance mutations (K65R, A62V, T69ins, L74I/V, V75M, Y115F, M184V, and K219R/E/N). The graph is a 2-D projection of the distances among the 37 mutations, in which the distance between any two mutations is measured by their Jaccard dissimilarity coefficient among persons who have received at least one nucleoside RT inhibitor.
doi:10.1371/journal.pcbi.0030087.g002

pairs of RTI mutations were significantly associated using a family-wise error rate of 0.01.

Covariation between two mutations may result from the shared inheritance of the mutations from a founder virus, from a shared evolutionary pressure (e.g., an antiretroviral drug) that independently selects for each mutation, or from a functional dependency between the mutations. In our analysis, covariation was unlikely to result from shared inheritance because the most strongly covarying mutations occurred solely among treated HIV-1 isolates, consistent with the repeated selection of the correlated mutations in many different isolates as a result of selective drug pressure rather than the inheritance of the correlated mutations from a small number of ancestral viruses.

However, the possibility that many of the covarying residues resulted from similar selective pressures rather than from functional dependency cannot be excluded. For example, it is possible that some pairs of covarying protease amino acids result from the selective pressure of the same PI or possibly pair of PIs. Shared selective pressure is a possible explanation for why covarying mutations are not necessarily close to one another in tertiary structures (Figure S3) [4]. An analysis of covariation that controls for treatment history would be better able to distinguish functional dependency from shared selective pressure. However, for most PIs and NRTI combinations, insufficient data are available for such an analysis. Identifying similar patterns of covariation in one or more independent lineages (e.g., other non-B subtypes) would also provide additional independent evidence for functional dependency.

Our examination of conditional dependency between mutation pairs, the temporal order in which mutations occur, and the relationship between these two types of data provided new insights into the evolution of protease and RT in persons receiving antiretroviral therapy. A strong positive relationship between the conditional dependency ratio of two mutations and the order in which the mutations occur would represent the most parsimonious mechanism for HIV-1 to develop multiple mutations (i.e., the mutation that occurs more often in a pair of mutations would be on average more likely to occur first). Nonetheless, we found that the positive relationship between conditional dependency and the order of mutation occurrence was stronger for covarying RT ($r^2 = 0.81$) compared with protease ($r^2 = 0.56$) mutation pairs. This suggests that the number of mutational steps required to develop multiple PI-resistance mutations may be greater on average than that required for developing the same number of multiple NRTI-resistance mutations.

We also found that accessory NRTI-resistance mutations nearly always followed primary NRTI-resistance mutations (particularly the TAMs). In contrast, the commonly recognized accessory PI-resistance mutations were as likely to precede as to follow major PI-resistance mutations. This frequent precedence of accessory PI-resistance mutations results in part from the fact that many of the accessory PI-resistance mutations are polymorphic and therefore present prior to the start of therapy. However, this alone does not explain the marked dependency of some major mutations on polymorphic accessory PI-resistance mutations that occur only at low levels in untreated persons.

The strong positive relationship between conditional probabilities and temporal data that we describe support the validity of previous research, which used cross-sectional data to infer mutational pathways [11] and causality [12,13]. Our results also suggest that there is a complex process underlying the order in which major and accessory PI-resistance mutations develop during PI therapy, and that the designation of major PI-resistance mutations as primary and accessory PI-resistance mutations as secondary often refers only to their roles in causing resistance and not to the order in which they develop.

Materials and Methods

Virus sequence data. Sequences included HIV-1 subtype B RT and protease sequences from published studies in the Stanford HIV Drug Resistance Database (http://hivdb.stanford.edu) [14]. For patients with more than one sequence, only the latest sequence obtained while receiving treatment was analyzed. For each gene, separate analyses were done for the sequences from treatment-experienced and treatment-naive individuals.

RT positions 1–240 and protease positions 1–99 were analyzed. Mutations were defined as differences from the consensus wild-type subtype B amino acid reference sequence (http://hivdb.stanford.edu/pages/asi/releaseNotes/index.html). For each pair of mutations (*X*, *Y*), the numbers of sequences containing both mutations (*X* and *Y*), only one mutation (*X* or *Y*), or neither mutation (not *X*, not *Y*) were counted and used to populate a contingency table. Sequences containing mixtures at either of the two positions were excluded from analysis of that pair of positions.

Antiretroviral treatment–selected mutations were defined based on the results of a previous study, as mutations that were significantly more common in treated than untreated persons after adjusting for multiple comparisons [15]. PI-selected mutations included L10I/V/F/R, V11I, K20R/M/I/T, L23I, L24I, D30N, V32I, L33F/I, E34Q, E35G, M36I/V, K43T, M46I/L/V, G48V/M, I50V/L, F53L, I54V/M/L/T/A/S, K55R, Q58E, L63P, I66F, C67F, A71V/T/I, V72L, G73S/T/C/A, T74A/P/S, L76V, V77I, V82A/T/F/S/L/M, I84V/A/C, I85V, N88D/S/T/G, L89V,

L90M, T91S, Q92R/K, I93L, and C95F. Several PI-resistance mutations—particularly those that occur in the substrate cleft or that have a major impact on drug susceptibility—are considered major PI-resistance mutations [2,16]. For the purposes of this study, we defined mutations at positions 24, 30, 32, 46, 47, 48, 50, 53, 54, 76, 82, 84, 88, and 90 as being major PI-resistance mutations. Several PI-resistance mutations—including several that are polymorphic in untreated persons—are commonly considered accessory drug resistance mutations that either compensate for the decreased replication associated with many of the major mutations or that reduce drug susceptibility further when present with a major mutation. Mutations at positions 10, 20, 33, 36, 58, 63, 71, 73, 74, 77, and 93 are usually considered to be accessory mutations. Little attention has been given to the remaining PI-selected mutations, and for the purposes of this paper, we leave them unclassified with respect to the designations major and accessory.

NRTI-selected mutations included T39A, M41L, K43E/Q/N, E44D/A, A62V, K65R, D67N/G/E, T69D/N/S/insertion, K70R, L74V/I, V75I/M/T/A, F77L, V90I, K104N, Y115F, F116Y, V118I, Q151M, M184V/I, E203K, H208Y, L210W, T215Y/F/D/C/E/S/I/V, D218E, K219Q/E/N/R, H221Y, K223Q, and L228H/R. These mutations included the Type I TAMs M41L, L210W, and T215Y, and the Type II TAMs D67N, K70R, T215F, and K219Q/E [7]. Recently described accessory NRTI mutations included T39A, K43E/Q/N, E44D/A, V118I, E203K, H208, D218E, H221Y, K223Q, and L228H/R [3,6,17]. Q151M-associated mutations included A62V, V75I, F77L, F116Y, and Q151M [18,19].

NNRTI-selected mutations included A98G, L100I, K101E/P/N/H, K103N/S, V106A/M, V108I, V179D/E, Y181C/I/V, Y188L/C/H, G190A/S/E/Q, P225H, F227L, M230L, P236L, and K238T.

Pairwise correlation. We used the Jaccard similarity coefficient (J) to assess covariation among protease and RT mutations. For a given pair of mutations X and Y, the Jaccard similarity coefficient is calculated as $J = N_{XY} / (N_{XY} + N_{X0} + N_{0Y})$ where N_{XY} represents the number of sequences containing X and Y, N_{X0} represents the number of sequences containing X but not Y, and N_{0Y} represents the number of sequences containing Y but not X. This coefficient represents the probability of both mutations occurring together when either mutation occurs and, therefore, does not inflate the correlation between two mutations that may appear correlated by other measures when both mutations are nearly always absent.

To test whether observed Jaccard similarity coefficients were statistically significant, the expected value of the Jaccard similarity coefficients (J_{RAND}) and its standard error (J_{SE}) assuming two mutations (X and Y) occur independently were calculated for each pair of mutations. J_{RAND} was calculated as the mean Jaccard similarity coefficient after 2,000 random rearrangements of the X or Y vector (containing 0 or 1 for presence or absence of a mutation, respectively). J_{SE} was calculated using a jackknifed procedure, which removed one sequence at a time, repeatedly for each sequence. The standardized score Z, $Z = (J - J_{RAND}) / J_{SE}$, indicates a significant positive association ($Z > 2.56$) or a significant negative association ($Z < -2.56$) at an unadjusted $p < 0.01$.

Holm's method was used to control the family-wise error rate for multiple hypothesis testing [20]. The p-values of observed Jaccard similarity coefficients for all pairs of mutations were ranked in descending order. Starting from the smallest p (rank $r = n$, where n is the number of pairs), we compared each p of rank r with a significance cutoff of $0.01 / r$ as long as $p_r \leq 0.01 / r$. All p-values from $p_r \ldots p_n$ were considered to be statistically significant.

To deal with contingency tables containing 0 for N_{XY} (potentially leading to Z scores of $-\infty$), we generated a conservative nonzero approximation of J_{SE} using the following procedure. Given a dataset of n sequences, x with mutation X and y with mutation Y, we computed the probability of both mutations (P_{XY}), mutation X but not Y (P_{X0}), mutation Y but not X (P_{0Y}), and neither mutation (P_{00}) under the null hypothesis of independence by $P_{XY} = (x / n) \times (y / n)$, $P_{X0} = (x / n) \times (y / n) / \mathrm{n}$, $P_{0Y} = (n - x) / n \times (y / n)$ and $P_{00} = 1 - P_{XY} - P_{X0} - P_{0Y}$. These probabilities were used to create 200 two-by-two contingency tables with cells containing randomly distributed numbers adding up to 20,000 based on the null hypothesis probabilities of independence.

Multidimensional scaling. Given the matrix of dissimilarity coefficients (1 − Jaccard similarity coefficient) for a list of mutations (X_1, X_2, ..., X_n), multidimensional scaling was used to construct points in 2-D space such that the Euclidean distances between these points approximate the entries in the dissimilarity matrix [21]. For a given k, it computes points $X_1, X_2, \ldots, X_n$ in 2-D space such that $S = \sum_{i,j} (dist(X_i, X_j) - d_{ij})^2$ is minimized where $dist(X_i, X_j)$ is the Euclidean distance between X_i and X_j and d_{ij} is the dissimilarity between Xi and Xj in the matrix D. This was performed using the R function cmdscale (classical multidimensional scaling).

Multidimensional scaling captures the inertia in a dataset in terms of a set of variables (or principal components) that define a projection that encapsulates the maximum amount of inertia in a dataset and is orthogonal (and therefore uncorrelated) to the previous principal component. Using the first and second principal components, we summarized the relationship among mutations in a graphical model, placing pairs of mutations with low Jaccard dissimilarity coefficients close together and mutations with high Jaccard dissimilarity coefficients far apart.

Supporting Information

Figure S1. Relationship between Conditional Probability and Order of Occurrence within Pairs of Covarying Protease Mutations

The relationship between the log of the ratio of the conditional probability of two protease mutations and the log of the ratio in which mutation develops first. A total of 38 protease mutation pairs from Table 1, of which the sum of $(X,0 \rightarrow X,Y)$ and $(0,Y \rightarrow X,Y) \geq 5$, the count of $(X,0 \rightarrow X,Y)$ or $(0,Y \rightarrow X,Y)$ is not zero, were plotted.

Found at doi:10.1371/journal.pcbi.0030087.sg001 (22 KB PDF).

Figure S2. Relationship between Conditional Probability and Order of Occurrence within Pairs of Covarying RT Mutations

The relationship between the log of the ratio of the conditional probability of two RT mutations and the log of the ratio in which mutation develops first. A total of 31 RT mutation pairs from Table 2, of which the sum of $(X,0 \rightarrow X,Y)$ and $(0,Y \rightarrow X,Y) \geq 5$, the count of $(X,0 \rightarrow X,Y)$ or $(0,Y \rightarrow X,Y)$ is not zero, were plotted.

Found at doi:10.1371/journal.pcbi.0030087.sg002 (22 KB PDF).

Figure S3. Structural Locations and Distances For Three Sets of Covarying Protease Mutations

Locations of residues present in the three most relevant clusters of PI-resistance mutations superimposed on the crystallographic structure of wild-type HIV-1 (1HPV.pdb): L24, M46, I54, and V82 (A); M46, G73, I84, and L90 (B); and V32, L33, I47, I50, I54, and L76 (C). Each panel shows the wild-type residues superimposed on the substrate cleft surface of the protease monomer. The shortest interatomic distances between selected residues are shown. Cluster A usually contains the mutations L24I, M46L > M46I, I54V, and V82A. Cluster B usually contains the mutations M46I, G73S, I84V, and L90M. Cluster C usually contains the mutations V32I, L33F, I47V, I50V, I54L > I54M, and L76V. The cluster consisting of mutations at positions 30 and 88 is not shown, as it is associated with resistance to a single PI (nelfinavir) rather than to multiple PIs.

Found at doi:10.1371/journal.pcbi.0030087.sg003 (2.0 MB TIF).

Table S1. 161 Highly Correlated Protease Mutation Pairs from PI-Experienced Persons

Found at doi:10.1371/journal.pcbi.0030087.st001

Table S2. Highly Correlated HIV-1 Subtype B Protease Mutation Pairs from PI-Naive Persons

Found at doi:10.1371/journal.pcbi.0030087.st002

Table S3. 327 Highly Correlated Protease Mutation Pairs from RTI-Experienced Persons

Found at doi:10.1371/journal.pcbi.0030087.st003

Table S4. Highly Correlated HIV-1 Subtype B RT Mutation Pairs from RTI-Naive Persons

Found at doi:10.1371/journal.pcbi.0030087.st004

Text S1. Accession Numbers

Found at doi:10.1371/journal.pcbi.0030087.sd001

Accession Numbers

The 11,355 GenBank (http://www.ncbi.nlm.nih.gov/Genbank) accession numbers of the sequences used in this study are provided in Text S1.

Acknowledgments

Author contributions. RWS, SYR, and SPH conceived and designed the experiments. SYR performed the experiments. All authors contributed to analyzing the data. TFL and SPH contributed reagents/materials/analysis tools. SYR and RWS wrote the paper.

Funding. SYR, TFL, SPH, and RWS were supported by US National Institute of Allergy and Infectious Diseases grant AI068581.

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

References

1. Korber B, Muldoon M, Theiler J, Gao F, Gupta R, et al. (2000) Timing the ancestor of the HIV-1 pandemic strains. Science 288: 1789–1796.
2. Baxter JD, Schapiro JM, Boucher CA, Kohlbrenner VM, Hall DB, et al. (2006) Genotypic changes in human immunodeficiency virus type 1 protease associated with reduced susceptibility and virologic response to the protease inhibitor tipranavir. J Virol 80: 10794–10801.
3. Gonzales MJ, Wu TD, Taylor J, Belitskaya I, Kantor R, et al. (2003) Extended spectrum of HIV-1 reverse transcriptase mutations in patients receiving multiple nucleoside analog inhibitors. AIDS 17: 791–799.
4. Wu TD, Schiffer CA, Gonzales MJ, Taylor J, Kantor R, et al. (2003) Mutation patterns and structural correlates in human immunodeficiency virus type 1 protease following different protease inhibitor treatments. J Virol 77: 4836–4847.
5. Hoffman NG, Schiffer CA, Swanstrom R (2003) Covariation of amino acid positions in HIV-1 protease. Virology 314: 536–548.
6. Svicher V, Sing T, Santoro MM, Forbici F, Rodriguez-Barrios F, et al. (2006) Involvement of novel human immunodeficiency virus type 1 reverse transcriptase mutations in the regulation of resistance to nucleoside inhibitors. J Virol 80: 7186–7198.
7. Cozzi-Lepri A, Ruiz L, Loveday C, Phillips AN, Clotet B, et al. (2005) Thymidine analogue mutation profiles: Factors associated with acquiring specific profiles and their impact on the virological response to therapy. Antivir Ther 10: 791–802.
8. Korber BT, Farber RM, Wolpert DH, Lapedes AS (1993) Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. Proc Natl Acad Sci U S A 90: 7176–7180.
9. Svicher V, Ceccherini-Silberstein F, Erba F, Santoro M, Gori C, et al. (2005) Novel human immunodeficiency virus type 1 protease mutations potentially involved in resistance to protease inhibitors. Antimicrob Agents Chemother 49: 2015–2025.
10. Kagan RM, Cheung PK, Huard TK, Lewinski MA (2006) Increasing prevalence of HIV-1 protease inhibitor–associated mutations correlates with long-term non-suppressive protease inhibitor treatment. Antiviral Res 71: 42–52.
11. Beerenwinkel N, Rahnenfuhrer J, Daumer M, Hoffmann D, Kaiser R, et al. (2005) Learning multiple evolutionary pathways from cross-sectional data. J Comput Biol 12: 584–598.
12. Deforche K, Silander T, Camacho R, Grossman Z, Soares MA, et al. (2006) Analysis of HIV-1 pol sequences using Bayesian networks: Implications for drug resistance. Bioinformatics 22: 2975–2979.
13. Chen L, Lee C (2006) Distinguishing HIV-1 drug resistance, accessory, and viral fitness mutations using conditional selection pressure analysis of treated versus untreated patient samples. Biol Direct 1: 14.
14. Rhee SY, Gonzales MJ, Kantor R, Betts BJ, Ravela J, et al. (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res 31: 298–303.
15. Rhee SY, Fessel WJ, Zolopa AR, Hurley L, Liu T, et al. (2005) HIV-1 protease and reverse-transcriptase mutations: Correlations with antiretroviral therapy in subtype B isolates and implications for drug-resistance surveillance. J Infect Dis 192: 456–465.
16. Rhee SY, Taylor J, Wadhera G, Ben-Hur A, Brutlag DL, et al. (2006) Genotypic predictors of human immunodeficiency virus type 1 drug resistance. Proc Natl Acad Sci U S A 103: 17355–17360.
17. Hertogs K, Bloor S, De Vroey V, van Den Eynde C, Dehertogh P, et al. (2000) A novel human immunodeficiency virus type 1 reverse transcriptase mutational pattern confers phenotypic lamivudine resistance in the absence of mutation 184V. Antimicrob Agents Chemother 44: 568–573.
18. Iversen AK, Shafer RW, Wehrly K, Winters MA, Mullins JI, et al. (1996) Multidrug-resistant human immunodeficiency virus type 1 strains resulting from combination antiretroviral therapy. J Virol 70: 1086–1090.
19. Shirasaka T, Kavlick MF, Ueno T, Gao WY, Kojima E, et al. (1995) Emergence of human immunodeficiency virus type 1 variants with resistance to multiple dideoxynucleosides in patients receiving therapy with dideoxynucleosides. Proc Natl Acad Sci U S A 92: 2398–2402.
20. Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6: 65–70.
21. Schoenberg IJ (1935) Remarks to Maurice Frechet's article "Sur la definition axiomatique d'une classe d'espace distancies vectoriellement applicable sur l'espace de Hilbert." Ann Mathematics 36: 724–732.