AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 19, Number 10, 2003, pp. 909–915
© Mary Ann Liebert, Inc.

# Sequence Note

# Protease and Reverse Transcriptase Mutation Patterns in HIV Type 1 Isolates from Heavily Treated Persons: Comparison of Isolates from Northern California with Isolates from Other Regions

MATTHEW J. GONZALES,[1] ILANA BELITSKAYA,[2] KATHRYN M. DUPNIK,[1]
SOO-YON RHEE,[1] and ROBERT W. SHAFER[1]

## ABSTRACT

**We compared HIV-1 subtype B reverse transcriptase (RT) and protease mutation patterns in isolates from heavily treated persons in Northern California with those from persons described in the published literature predominantly from other parts of the United States and Europe. There were few differences in the prevalence of single, double, and triple mutations between the two sets of sequences. More complex patterns of mutations could be characterized by clustering the sequences into eight groups of RT sequences and nine groups of protease sequences according to the presence of known drug-resistance mutations. This clustering accounted for 63% of the variation at RT inhibitor-resistance positions and 68% of the variation at protease inhibitor-resistance positions. The majority of clusters contained Northern California and literature sequences in similar proportions.**

## INTRODUCTION

HIV-1 DRUG RESISTANCE is caused by many different mutations that emerge in complex patterns. However, the patterns of drug-resistance mutations are not random and isolates with high levels of resistance to drugs of the same class often contain similar patterns of drug-resistance mutations.[1–3] Testing the activity of new compounds against a collection of such isolates would help determine the relative activity of new compounds against current drug-resistant isolates. Compounds retaining activity against a collection of virus isolates with such high-level class resistance would also be expected to be active against the vast majority of drug-resistant HIV-1 clinical isolates.

To create a collection of truly representative drug-resistant isolates, it is necessary to know whether the prevalence of mutation patterns in isolates from heavily treated persons varies among locations. In this study, we compare the prevalence of individual mutations and mutation patterns in HIV-1 isolates from heavily treated individuals in Northern California with those from other regions described in the published literature.

## MATERIALS AND METHODS

Sequences were obtained from Northern California isolates sequenced at the Stanford University Hospital (SUH) Diagnostic Virology Laboratory between July 1997 and December 2001 and from isolates described in published papers before October 2002 and catalogued in the Stanford HIV RT and Protease Sequence Database (*http://hivdb. stanford. edu*).[4] We analyzed HIV-1 protease and RT sequences of isolates from heavily treated persons with detectable plasma HIV-1 RNA who were receiving antiretroviral therapy. Analysis was restricted to protease sequences (positions 10–90) from persons treated with three or more protease inhibitors and RT sequences (positions 40–240) from persons treated with four or more nucleoside RT inhibitors including zidovudine, stavudine, didanosine, and

[1]Division of Infectious Diseases, Department of Medicine and [2]Department of Statistics, Stanford University, Stanford, California 94305.

lamivudine. For persons with multiple sequences meeting study criteria, only the latest isolate was analyzed.

### *Mutation patterns and statistical analysis*

Mutations were defined as differences from the consensus B reference sequence (*http://hiv-web.lanl.gov/*). Fisher's exact tests were used to compare the prevalence of mutation patterns involving one, two, or three positions between sequences from Northern California and the published literature. The method of Benjamini and Hochberg was used to identify differences in prevalence that were statistically significant in the presence of multiple-hypothesis testing.[5] In contrast to the Bonferroni correction, which divides the significance cutoff by the number of hypotheses tested ($n$), the Benjamini–Hochberg method ranks the hypotheses by their $p$ values. Each hypothesis of rank $r$ is compared with a significance cutoff—now called a false discovery rate (FDR)—divided by $(n - r)$. In this study, an FDR of 0.05 was used to determine statistical significance.

To identify mutation patterns involving four or more drug resistance positions, we used an implementation of $k$-medoids clustering known as partitioning around medoids (PAM).[6] The parameters of the PAM algorithm included a dissimilarity matrix and a prespecified number of clusters, $k$. The algorithm searched for $k$ representative sequences, or medoids, among the sequences to be clustered. After finding the set of $k$-medoids, clusters were constructed by assigning each sequence to the nearest medoid. PAM minimized the sum of dissimilarities between each sequence and the medoid of the cluster to which the sequence belongs.

The dissimilarity matrix used by the PAM algorithm to generate RT clusters was based on the presence or absence of mutations at the 16 nucleoside RT inhibitor-resistance positions. (41, 62, 65, 67, 69, 70, 74, 75, 77, 115, 116, 151, 184, 210, 215, and 219). The dissimilarity matrix used by the PAM algorithm to generate protease clusters was based on the presence or absence of mutations at the 14 nonpolymorphic protease inhibitor resistance positions (24, 30, 32, 46, 47, 48, 50, 53, 54, 73, 88, 82, 84, and 90). These 14 protease positions were selected because they are associated with protease inhibitor resistance and are not polymorphic in the absence of therapy. Positions associated with nonnucleoside RT inhibitor resistance were not used for clustering. The gap statistic, a measure of naturally occurring clustering within a data set, was used to help determine the number of sequence clusters.[7]

A single set of clusters was created using sequences from both Northern California and the published literature. Fisher's exact tests were then used to compare the proportions of each type of sequence in each cluster.

## RESULTS AND DISCUSSION

### *RT sequences*

RT sequences from 487 persons who had received four or more nucleoside RT inhibitors met study criteria. Of these, 288 (59%) persons were from Northern California and 199 (41%) were from the published literature. The median duration of nucleoside RT inhibitor therapy was 263 weeks for persons from Northern California and 183 weeks for persons in the published literature. Eighty (28%) persons from Northern California and 53 (27%) from the literature also received abacavir, one person from Northern California and one from the literature also received tenofovir, and 204 (71%) persons from Northern California and 109 (55%) from the literature also received one or more nonnucleoside RT inhibitors. The persons described in the literature were predominantly from parts of the United States other than Northern California (45%), Europe (44%), and South America (11%). The median isolation date for the Northern California sequences was October 1999 (range: June 1997–December 2001) and for the literature sequences was January 1999 (range: June 1995–May 2001).

There were no statistically significant differences in the prevalence of single, double, or triple mutations between the Northern California sequences and the literature sequences. Figure 1 shows the plots of empirical cumulative distribution function for the $p$ values obtained using Fisher's exact test to compare the rates of single, double, and triple mutations between the two sets of sequences. The figures show that although there were differences between the two sets of sequences, these differences did not reach statistical significance after adjustment for multiple comparisons.

### *Clusters of RT inhibitor-resistance mutations*

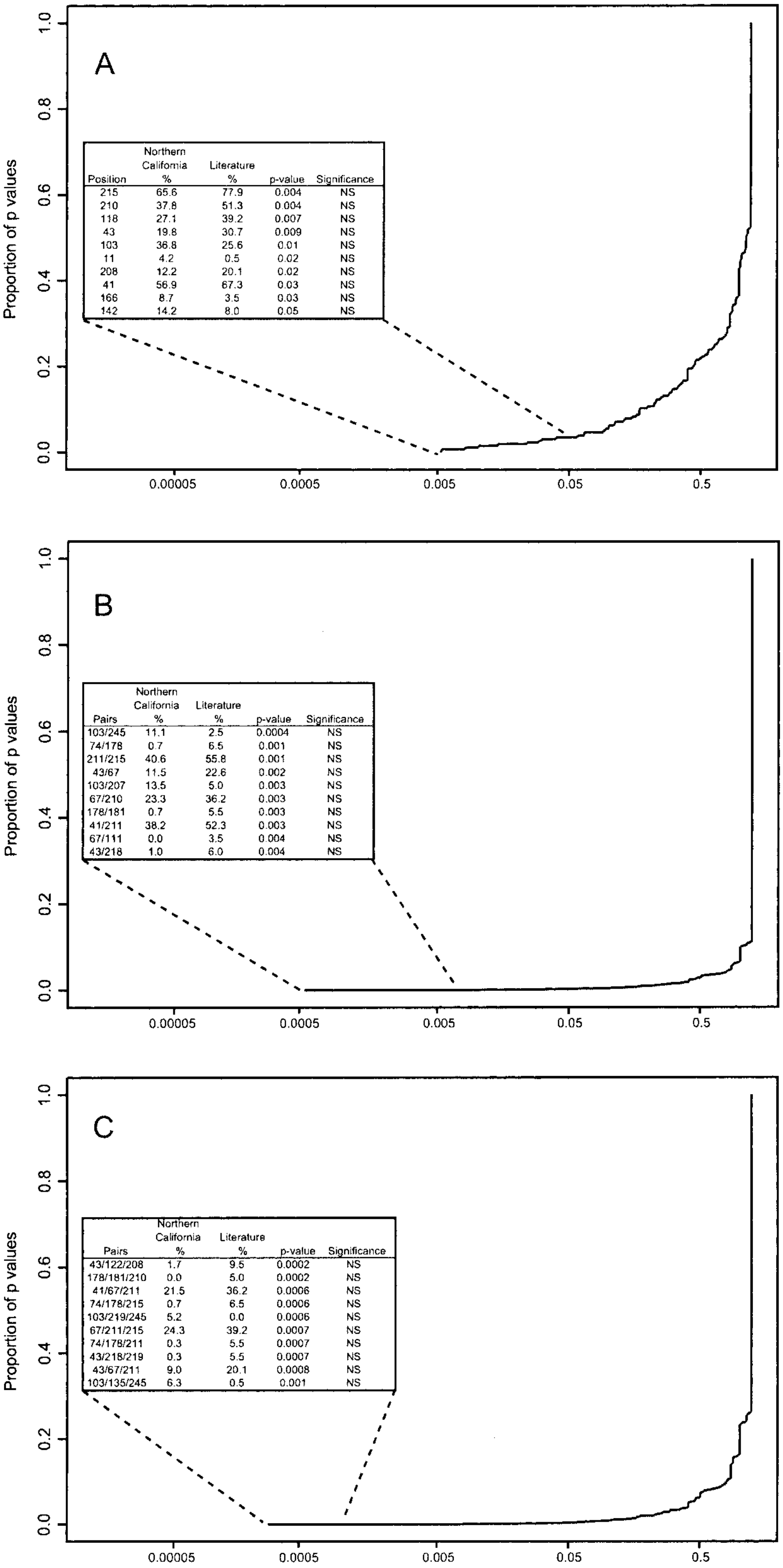
$k$-Medoids clustering was used to identify common patterns of mutations at drug-resistance positions. The gap statistic suggested that the 487 RT sequences optimally grouped into seven clusters ($k = 7$), but we increased $k$ to eight because this led to the addition of a cluster of sequences containing mutations associated with the multidrug-resistance mutation, Q151M. Figure 2 shows a profile of each cluster. Within a cluster, the mean number of different drug-resistance mutations between each sequence and the medoid was 1.33 (total sum of differences: 646). In contrast, the mean number of different mutations between each of the 487 sequences and the medoid of the complete data set was 3.53 ($k = 1$; total sum of differences: 1725). Thus, the eight clusters explain approximately 63% ($1 - 646/1725$) of the variability at the 16 NRTI-resistance positions.

Six of the eight clusters were composed primarily of the thymidine analog mutations (TAMs) at positions 41, 67, 70, 210, 215, and 219; one cluster was composed of Q151M and its associated mutations at positions 75, 77, and 116;and one cluster had a medoid with no mutations. M184V was in the medoid of five of the clusters, including four of the clusters with TAMs and the Q151M cluster. M184V also occurred in about one-third of the sequences in the cluster whose medoid had no mutations.

The proportion of isolates belonging to cluster 1, characterized by the absence of mutations, was higher in isolates from Northern California than isolates from the literature (24% vs. 14%; $p = 0.01$). The proportion of isolates belonging to cluster 3, characterized by mutations at positions 41, 67, 69, 210, and 215, was higher in isolates from the literature than isolates from Northern California (20% vs. 11%; $p = 0.009$).

### *Protease sequences*

Protease sequences from 485 persons who had received three or more protease inhibitors met study criteria. Of these 299 (62%) were from Northern California and 186 (38%) were from



| Position | Northern Califomia % | Literature % | p-value | Significance |
|---|---|---|---|---|
| 215 | 65.6 | 77.9 | 0.004 | NS |
| 210 | 37.8 | 51.3 | 0.004 | NS |
| 118 | 27.1 | 39.2 | 0.007 | NS |
| 43 | 19.8 | 30.7 | 0.009 | NS |
| 103 | 36.8 | 25.6 | 0.01 | NS |
| 11 | 4.2 | 0.5 | 0.02 | NS |
| 208 | 12.2 | 20.1 | 0.02 | NS |
| 41 | 56.9 | 67.3 | 0.03 | NS |
| 166 | 8.7 | 3.5 | 0.03 | NS |
| 142 | 14.2 | 8.0 | 0.05 | NS |

| Pairs | Northern California % | Literature % | p-value | Significance |
|---|---|---|---|---|
| 103/245 | 11.1 | 2.5 | 0.0004 | NS |
| 74/178 | 0.7 | 6.5 | 0.001 | NS |
| 211/215 | 40.6 | 55.8 | 0.001 | NS |
| 43/67 | 11.5 | 22.6 | 0.002 | NS |
| 103/207 | 13.5 | 5.0 | 0.003 | NS |
| 67/210 | 23.3 | 36.2 | 0.003 | NS |
| 178/181 | 0.7 | 5.5 | 0.003 | NS |
| 41/211 | 38.2 | 52.3 | 0.003 | NS |
| 67/111 | 0.0 | 3.5 | 0.004 | NS |
| 43/218 | 1.0 | 6.0 | 0.004 | NS |

| Pairs | Northern California % | Literature % | p-value | Significance |
|---|---|---|---|---|
| 43/122/208 | 1.7 | 9.5 | 0.0002 | NS |
| 178/181/210 | 0.0 | 5.0 | 0.0002 | NS |
| 41/67/211 | 21.5 | 36.2 | 0.0006 | NS |
| 74/178/215 | 0.7 | 6.5 | 0.0006 | NS |
| 103/219/245 | 5.2 | 0.0 | 0.0006 | NS |
| 67/211/215 | 24.3 | 39.2 | 0.0007 | NS |
| 74/178/211 | 0.3 | 5.5 | 0.0007 | NS |
| 43/218/219 | 0.3 | 5.5 | 0.0007 | NS |
| 43/67/211 | 9.0 | 20.1 | 0.0008 | NS |
| 103/135/245 | 6.3 | 0.5 | 0.001 | NS |

**FIG. 1. (A–C)** Plots of empirical cumulative distribution functions of the *p* values obtained comparing the prevalence of single, double, and triple reverse transcriptase (RT) mutations in HIV-1 isolates from heavily treated persons in Northern California and the remaining published literature. This analysis includes mutations or differences from the consensus B sequence at all RT positions between codons 1 and 240. *p* values were determined using Fisher's exact test. The small height of the graph at low *p* values indicates the overall similarity between the two data sets.

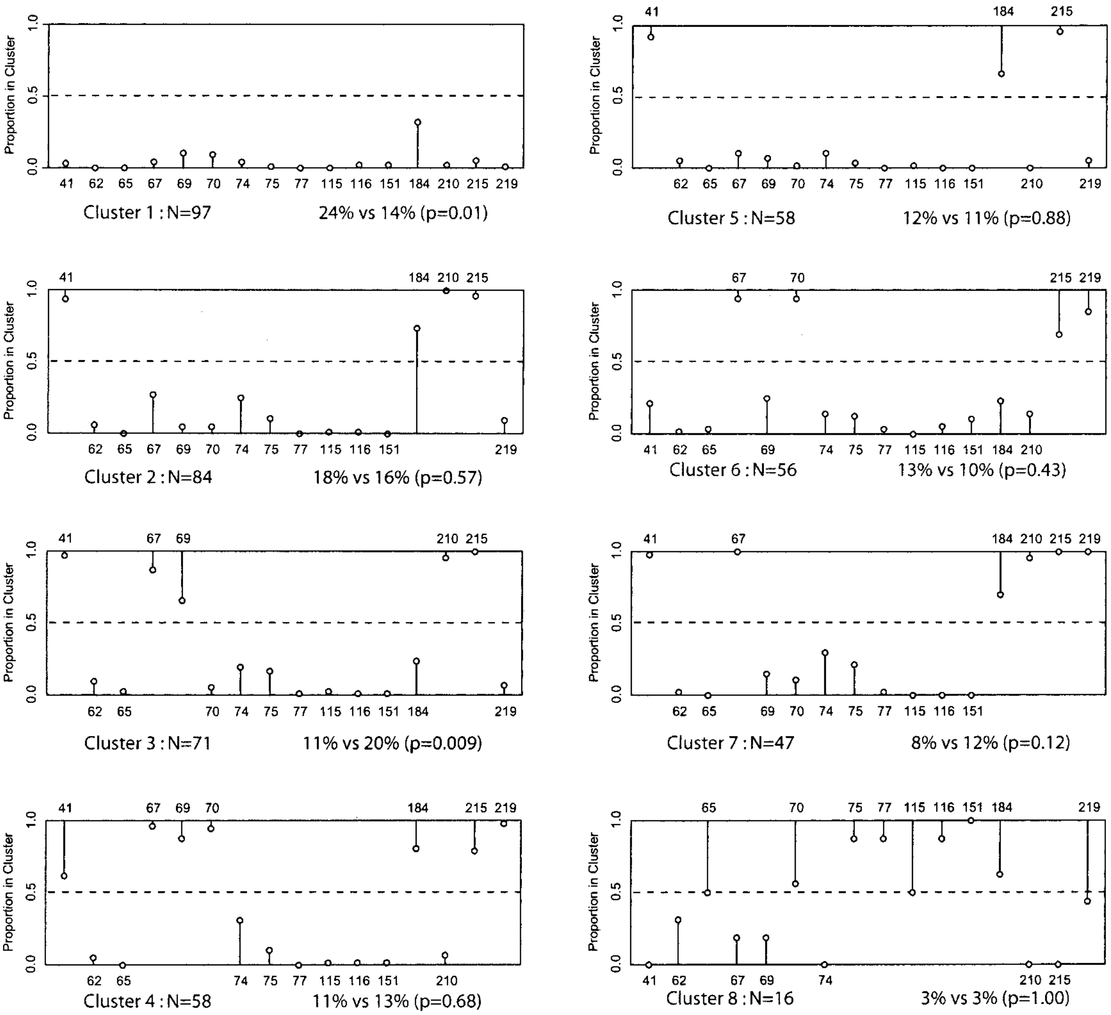
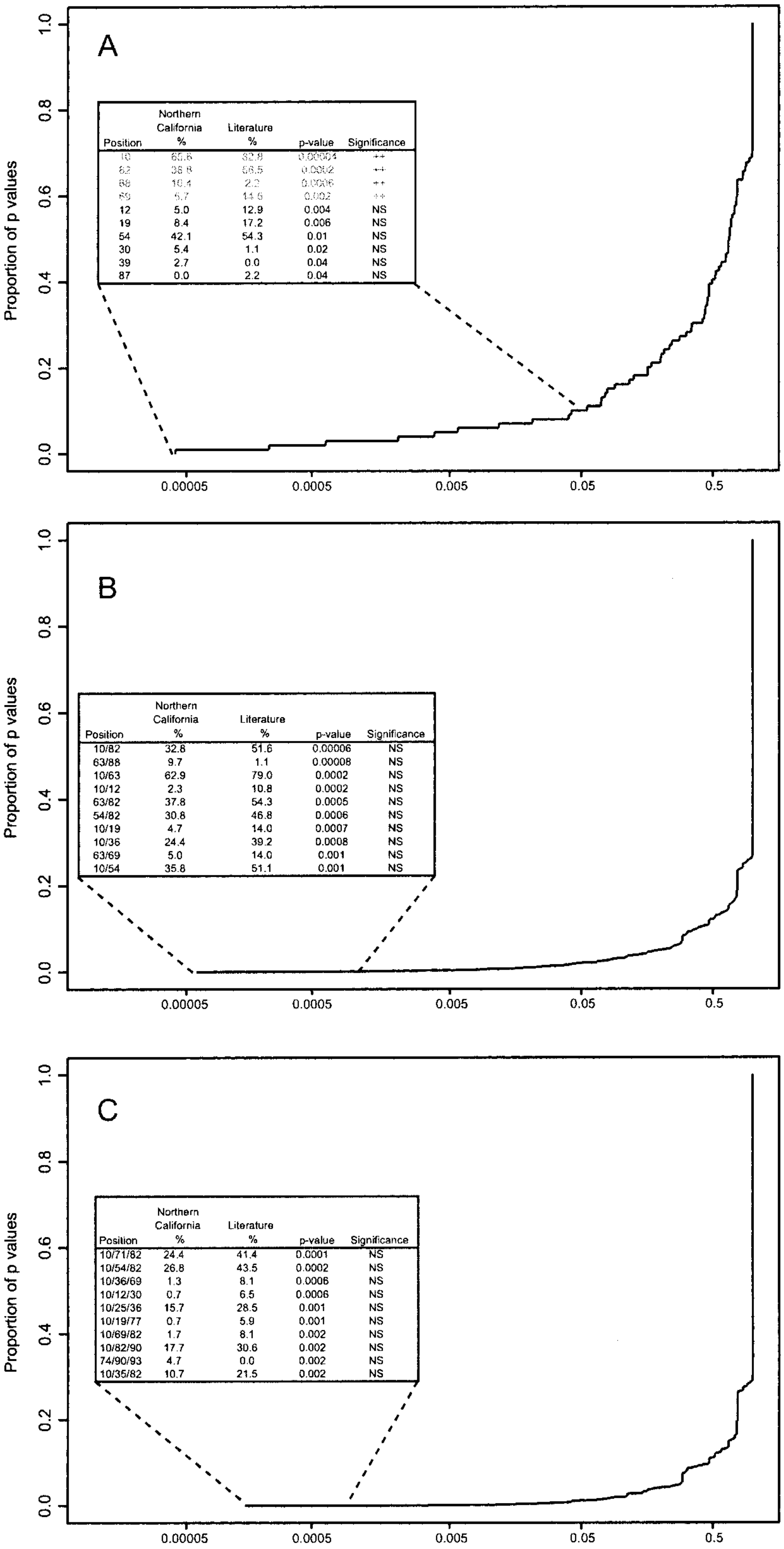


**FIG. 2.** Eight clusters of reverse transcriptase (RT) drug-resistance mutations in isolates from persons receiving four or more nucleoside RT inhibitors determined by $k$-medoids clustering. The $x$-axis indicates the 16 mutations used for clustering. The $y$-axis indicates the proportion of positions in a cluster that was mutant. The medoid consists of isolates containing mutations at each of the positions listed above the top bold line in each panel. Open circles indicate the proportion of isolates within a cluster that is mutant at each drug-resistance position. For positions that are mutant in the medoid, the open circles are attached to the top line. For positions that are wild type in the medoid, open circles are attached to the bottom line. The lower left part of each panel shows the number of sequences in each cluster. The lower right part of each panel shows the percent of Northern California (first number) and literature sequences (second number) belonging to each cluster.

the published literature. The median duration of protease inhibitor therapy was 148 weeks for persons from Northern California and 84 weeks for persons described in the literature. Sixty-four (21%) persons from Northern California and 28 (15%)from the literature received amprenavir and 11 (4%) persons from Northern California and 40 (22%) from the literature received lopinavir. Of the persons described in the literature 46% were from parts of the United States other than Northern California, 50% were from Europe, 3% were from South America, and 1% were from Asia. The median isolation date for the Northern California sequences was August 1999 (range: June 1997–November 2001) and for the literature sequences was May 1998 (range: January 1996–February 2001).

The prevalence of mutations at position 10 (83% vs. 66%; $p = 0.00004$) and 82 (57% vs. 39%; $p = 0.0002$) was higher in the sequences from the literature than in sequences from Northern California. The prevalence of mutations at position 88 (10% vs. 2%; $p = 0.0006$) was higher in sequences from Northern California than in sequences from the literature. There were no statistically significant differences between the two data sets for the remaining 20 protease mutations. There were no statistically significant differences in the prevalence of double or triple mutations between the two groups of sequences. Figure 3 shows the plots of empirical cumulative distribution function for the $p$ values obtained using Fisher's exact test to compare the rates of single, double, and triple mutations between the two sets of sequences.



| Position | Northern California % | Literature % | p-value | Significance |
|---|---|---|---|---|
| 10 | 65.6 | 32.8 | 0.00004 | ++ |
| 62 | 38.8 | 56.5 | 0.0002 | ++ |
| 88 | 10.4 | 2.2 | 0.0006 | ++ |
| 69 | 5.7 | 14.6 | 0.002 | ++ |
| 12 | 5.0 | 12.9 | 0.004 | NS |
| 19 | 8.4 | 17.2 | 0.006 | NS |
| 54 | 42.1 | 54.3 | 0.01 | NS |
| 30 | 5.4 | 1.1 | 0.02 | NS |
| 39 | 2.7 | 0.0 | 0.04 | NS |
| 87 | 0.0 | 2.2 | 0.04 | NS |

| Position | Northern California % | Literature % | p-value | Significance |
|---|---|---|---|---|
| 10/82 | 32.8 | 51.6 | 0.00006 | NS |
| 63/88 | 9.7 | 1.1 | 0.00008 | NS |
| 10/63 | 62.9 | 79.0 | 0.0002 | NS |
| 10/12 | 2.3 | 10.8 | 0.0002 | NS |
| 63/82 | 37.8 | 54.3 | 0.0005 | NS |
| 54/82 | 30.8 | 46.8 | 0.0006 | NS |
| 10/19 | 4.7 | 14.0 | 0.0007 | NS |
| 10/36 | 24.4 | 39.2 | 0.0008 | NS |
| 63/69 | 5.0 | 14.0 | 0.001 | NS |
| 10/54 | 35.8 | 51.1 | 0.001 | NS |

| Position | Northern California % | Literature % | p-value | Significance |
|---|---|---|---|---|
| 10/71/82 | 24.4 | 41.4 | 0.0001 | NS |
| 10/54/82 | 26.8 | 43.5 | 0.0002 | NS |
| 10/36/69 | 1.3 | 8.1 | 0.0006 | NS |
| 10/12/30 | 0.7 | 6.5 | 0.0006 | NS |
| 10/25/36 | 15.7 | 28.5 | 0.001 | NS |
| 10/19/77 | 0.7 | 5.9 | 0.001 | NS |
| 10/69/82 | 1.7 | 8.1 | 0.002 | NS |
| 10/82/90 | 17.7 | 30.6 | 0.002 | NS |
| 74/90/93 | 4.7 | 0.0 | 0.002 | NS |
| 10/35/82 | 10.7 | 21.5 | 0.002 | NS |

**FIG. 3.** (A–C) Plots of empirical cumulative distribution functions of the *p* values obtained comparing the prevalence of single, double, and triple protease mutations in HIV-1 isolates from heavily treated persons in Northern California and the remaining published literature. This analysis includes mutations or differences from the consensus B sequence at all protease positions between codons 1 and 99. *p* values were determined using Fisher's exact test. The small height of the graph at low *p* values indicates the overall similarity between the two data sets.

### *Clusters of protease inhibitor-resistance mutations*

*k*-Medoids clustering identified common patterns of mutations at drug-resistance positions. The gap statistic suggested that the sequences grouped optimally into eight clusters ($k = 8$), b¬t we increased $k$ to 9 because this led to the addition of a cluster characterized by mutations at positions 30 and 88, which are common in persons treated with nelfinavir but which were not present in the first eight clusters. Figure 4 shows the profile of each cluster. Within a cluster, the mean number of different mutations between each sequence and the medoid was 0.94 (total sum of differences: 454). In contrast, the mean num-

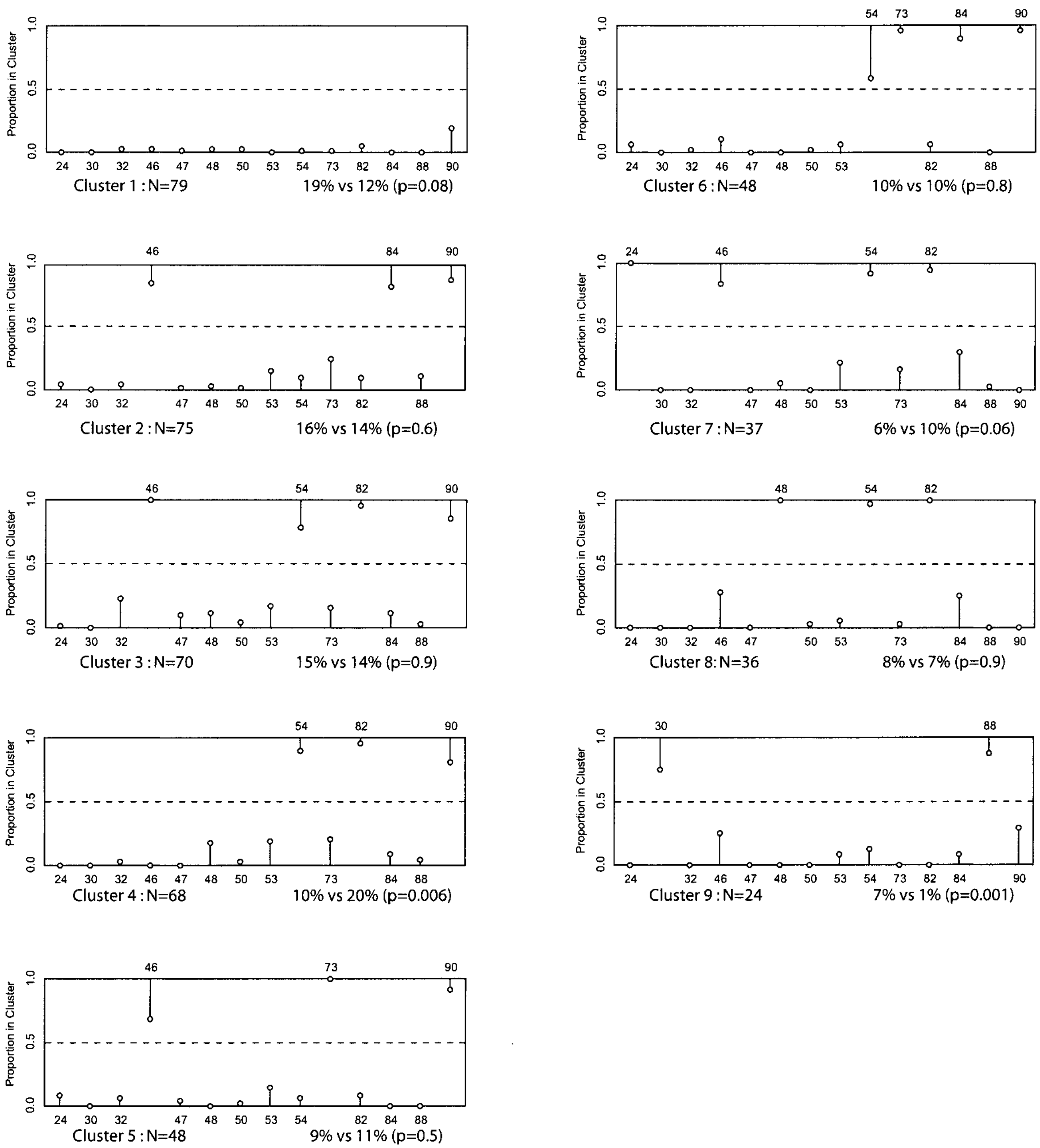


**FIG. 4.** Nine clusters of protease drug-resistance mutations in isolates from persons receiving four or more protease inhibitors determined by *k*-medoids clustering. The *x*-axis indicates the 16 mutations used for clustering. The *y*-axis indicates the proportion of positions in a cluster that was mutant. The medoid consists of isolates containing mutations at each of the positions listed above the top bold line in each panel. Open circles indicate the proportion of isolates within a cluster that is mutant at each drug-resistance position. For positions that are mutant in the medoid, the open circles are attached to the top line. For positions that are wild type in the medoid, open circles are attached to the bottom line. The lower left part of each panel shows the number of sequences in each cluster. The lower right part of each panel shows the percent of Northern California (first number) and literature sequences (second number) belonging to each cluster.

ber of different mutations between each of the 485 sequences and the medoid of the complete data set was 2.90 ($k = 1$; total sum of differences: 1417). Thus, the nine clusters explained approximately 68% (1 − 454/1417) of the variability at the 14 protease inhibitor-resistance positions.

The proportion of literature sequences belonging to the cluster with mutations at positions 54, 82, and 90 (cluster 4) was higher than the proportion of Northern California sequences. belonging to this cluster (20% vs. 10%; $p = 0.006$). The proportion of Northern California sequences belonging to the cluster with mutations at positions 30 and 88 (cluster 9) was higher than the proportion of literature sequences belonging to this cluster (7% vs. 1%; $p = 0.001$).

## CONCLUSIONS

In an analysis of 487 isolates from persons receiving four or more nucleoside RT inhibitors and 485 isolates from persons receiving three or more protease inhibitors, we found few significant differences in the prevalence of single mutations, pairs of mutations, and mutation triplets between sequences from Northern California and those described in the literature. To compare the complex combinations of mutations occurring in multidrug-resistant isolates, we restricted our analysis to clusters of 16 RT and 14 protease drug-resistance positions in the two sequence sets. Clusters were determined using a standard approach known as $k$-medoid clustering. Eight RT and nine protease clusters accounted for approximately two-thirds of the variation at these drug-resistance positions. Clustering is therefore a useful way to split sequences into representative groups based on shared mutations at drug-resistance positions.

Although the subtype of an isolate and the HLA type of the person from whom an isolate is obtained affect the amino acid composition of RT and protease,[8,9] variation at drug-resistance positions is primarily explained by selective drug pressure. All isolates analyzed in this study were from heavily treated persons, ensuring that although the treatment order and drug combinations may have differed between the Northern Californians and the persons described in the literature, the specific drugs received were similar.

However, our approach for choosing representative highly drug-resistant isolates has several unavoidable limitations. First, no sequences were available from heavily treated persons infected with non-B isolates. Had there been a sizable proportion of non-B isolates in our sample, there probably would have been differences at subtype-specific polymorphic positions. Second, several simplifications were required to identify mutation patterns. We confined our cluster analysis to a subset of positions associated with drug resistance and did not consider the effects of different mutations at the same position. Third, reporting bias may have played a role in the composition of the sequences described in the literature.

Our study has two main conclusions. First, our data suggest that mutation patterns in isolates from heavily treated persons are not random and that clustering RT and protease into eight and nine groups, respectively, accounts for about two-thirds of the variation at these positions. Second, our data suggest that subtype B RT and protease isolates from heavily treated persons in Northern California are similar to those described in the published literature from multiple geographic locations. The identification of mutational clusters should not preclude the testing of a new compound against a broad range of drug-susceptible and drug-resistant HIV-1 isolates. However, isolates with these clusters provide a reasonable target for drug-development efforts and will also make it possible to compare the relative activity of new compounds against a standard set of isolates.

## APPENDIX

The GenBank accession number of the sequences analyzed in this study can be found at: *http://hivdb.stanford.edu/pages/data/clusters.html.* One hundred new sequences were submitted together with this manuscript (AY305901–AY306000).

## REFERENCES

1. Condra JH, Schleif WA, Blahy OM, *et al.*: In vivo emergence of HIV-1 variants resistant to multiple protease inhibitors. Nature 1995;374:569–571.
2. Shafer RW, Winters MA, Palmer S, and Merigan TC: Multiple concurrent reverse transcriptase and protease mutations and multidrug resistance of HIV-1 isolates from heavily treated patients. Ann Intern Med 1998;128:906–911.
3. Hertogs K, Bloor S, Kemp SD, *et al.*: Phenotypic and genotypic analysis of clinical HIV-1 isolates reveals extensive protease inhibitor cross-resistance: A survey of over 6000 samples. AIDS 2000; 14:1203–1210.
4. Shafer RW, Stevenson D, and Chan B: Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res 1999;27:348–352.
5. Benjamini Y and Hochberg Y: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol 1995;57:289–300.
6. Kaufman L and Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Series in Probability and Mathematical Statistics.* John Wiley & Sons, Inc., New York, 1990.
7. Tibshirani R, Walther G, and Hastie T: Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc 2001;63:411–423.
8. Gonzales MJ, Machekano RN, and Shafer RW: Human immunodeficiency virus type 1 reverse-transcriptase and protease subtypes: Classification, amino acid mutation patterns, and prevalence in a northern California clinic-based population. J Infect Dis 2001;184: 998–1006.
9. Moore CB, John M, James IR, Christiansen FT, Witt CS, and Mallal SA: Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. Science 2002;296:1439–1443.

Address reprint requests to:
*Robert W. Shafer*
*Division of Infectious Diseases*
*Room S-156, Stanford University Medical Center*
*Stanford, California 94305*

*E-mail:* rshafer@stanford.edu