AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 21, Number 12, 2005, pp. 1016–1030

Covariability of Selected Amino Acid Positions for HIV Type 1 Subtypes C and B

PETER B. GILBERT,[1] VLADIMIR NOVITSKY,[2] and MAX ESSEX[2]

ABSTRACT

We studied covariability of selected amino acid positions in globally dominant HIV-1 subtype C viruses. The analyzed sequences spanned the V3 loop, Gag p17, Gag p24, and five CTL epitope-rich regions in Gag, Nef, and Tat. The corresponding regions in HIV-1 subtype B were also evaluated. The analyses identified a great number of covarying pairs and triples of sites in the HIV-1B V3 loop (173 site pairs, 242 site triples). Several of these interactions were found in the earlier studies [e.g., the V3 loop covariability analyses by Korber *et al.* (Proc Natl Acad Sci USA 1993;90:7176–7180) and Bickel *et al.* (AIDS Res Hum Retroviruses 1996;12:1401–1411)] and have known biological significance. However, generally these key covarying sites did not covary in the HIV-1C V3 loop (total 17 covarying site pairs), suggesting that the V3 loop may have subtype differences in functional or structural operating characteristics. Covariability of positions 309 and 312 was observed in the immunodominant region HIV-1C Gag 291–320 but no covariability was found in the corresponding region of HIV-1B, and vice versa for Nef 122–141; these findings may reflect subtype-specific covariability within immunologically relevant regions. Gag p17 exhibited greater covariability and less diversity for HIV-1B than HIV-1C, raising the hypothesis that Gag p17 is highly immunodominant in HIV-1B and is especially important for HIV-1B vaccines. Information on covariability should be better exploited in assessments of HIV-1 diversity and how to surmount it with vaccine design.

INTRODUCTION

STATISTICAL COVARIABILITY OF AMINO ACID MUTATIONS AMONG SITES IN HIV-1 amino acid sequences may indicate interesting biological interactions between the sites. The interactions may reflect functional constraints of protein structure, motivating analyses that search for covarying sites. Korber *et al.*[1] analyzed for covariability 308 subtype B V3 loop amino acid sequences and identified seven highly significantly covarying pairs of sites. Bickel *et al.*[2] followed up this study by analyzing the same set of V3 loop sequences plus a second set of 248 subtype B V3 loop sequences and 192 non-subtype B sequences. This analysis identified 8–20 highly significantly covarying pairs of sites, including 5 of those found earlier.[1] Many of the covarying sites identified by these studies have known important biological functions, illustrating the potential value of covariability analyses for predicting critical structural features of HIV-1 proteins as targets for vaccines and therapies. In addition, Brown *et al.,*[3] Hoffman *et al.*,[4] and Wu *et al.*[5] studied covariability of HIV-1 protease amino acids and related the results to HIV-1 function or structure.

Due to the dearth of subtype C sequences in databases at the time of their analyses, the earlier work[1,2] could not study the covariability of subtype C sequences. Ample data on HIV-1C sequences have now accrued.[6–10] Given the recent expansion of the subtype C epidemic to its present global predominance,[11–14] and the increasing emphasis of HIV-1 vaccine research toward subtype C viruses, it is important to analyze for covariability subtype C sequences. Furthermore, up to the time of the analysis of Bickel *et al.*,[2] much of HIV vaccine research had focused on envelope-based vaccines for which studying the V3 loop was of paramount interest. Since then, the field has increasingly invested in vaccine design of nonenvelope genes such as *gag*, *nef*, and *tat*.[15,16] In this article, we analyze for covariability HIV-1 subtype C sequences predominantly from southern Africa, in the V3 loop as well as in Gag p17, Gag p24, and regions in Gag, Nef, and Tat that have been found to be rich with cytotoxic T lymphocyte (CTL) epitopes in HIV-1C.[17,18] The latter regions are important to study be-

[1] Department of Biostatistics, University of Washington, and Fred Hutchinson Cancer Research Center, Seattle, Washington 98109.
[2] Harvard AIDS Institute and Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, Massachusetts 02115.

cause several HIV-1 vaccine candidates under development are based on eliciting CTL responses to these regions.[15,16,19] In the selected regions, we also compare the covariability of subtype C sequences to that in subtype B sequences. This comparison provides a way to identify subtype-specific covariability.

Five immunodominant regions within HIV-1C were identified in our previous study;[18] two in Gag p24, spanning positions 171–190 and 291–320; two in Nef, in positions 67–96 and 122–141; and one in Tat, in positions 36–50. Throughout this article the amino acid positions are relative to the HXB2 numbering system.[20] In addition to the CTL epitopes within HIV-1B Gag and Nef that were described previously,[17] Elispot-based CTL responses to HIV-1B Gag p24 (positions 171–190) and HIV-1B Nef (positions 71–90 and 131–150) have recently been reported.[21,22]

In addition to reporting on covariability results for HIV-1 subtypes and proteins previously not evaluated, this article applies new quantitative methods for studying covariability that build upon the existing methods.[1,2] The new techniques include two statistics for quantitating the extent of covariation between pairs of sites with confidence intervals, including a normalized mutual information statistic that helps correct the problem of relatively low power to detect covariability among conserved positions. A normalized version of the mutual information statistic is also used for evaluating covariation of triples of sites.

This study identified a large number of covarying site pairs and triples in HIV-1C and HIV-1B, and located covariable-rich regions, some common to the subtypes and some subtype specific. To attempt to determine the biological significance of the covariability findings, $dN - dS$ rate differences and sequence diversity were examined in the regions evaluated for covariability. The analyses support biologically important covariation of certain amino acids sites, although data are limited for confirming the significance of the results. Such statistical findings are important because they generate hypotheses about biological function that can be tested in future experiments.

MATERIALS AND METHODS

The purpose of the methods is to detect pairs of sites i, j (or triples i, j, k) in HIV-1 amino acid sequences that covary in a statistically significant way. We briefly summarize the methods applied by Bickel *et al.*,[2] and then describe the new methods. For a pair of sites i and j, Bickel *et al.*[2] considered three statistical criteria for measuring the evidence for covariation. The first statistic is the mutual information M_{ij}, also used by Korber *et al.,*[1] which is the likelihood ratio statistic for testing independence of two sites versus arbitrary covariation. This statistic is equivalent to the maximum likelihood estimator of linkage disequilibrium developed by Hill.[23] For large samples M_{ij} is approximately equivalent to Pearson's χ^2 statistic for testing independence. The second statistic, G_{ij}, has a heuristic interpretation described in Goodman and Kruskal[24] as the average increase in the chance of guessing correctly the residue at i based on knowing both the residues at sites i and j compared to knowing only the residues at i. To describe the third statistic P_{ij}, consider Weir and Cockerham's[25] statistic $P_{ij}(a,b)$, which is the likelihood ratio statistic for testing independence of sites i and j versus the alternative that the pair of residues (a, b) is favored relative to what is expected by chance. Bickel *et al.*[2] defined P_{ij} as the maximum of the $P_{ij}(a, b)$ statistics over the 400 pairs of residues a and b. Like M_{ij}, P_{ij} is similar to the χ^2 statistic, except that P_{ij} is designed to detect covariation driven by a single pair of residues.

The statistic M_{ij} has relatively low power to detect covariability of conserved positions, which may be disadvantageous given the functional significance often associated with conserved regions. To remedy this problem, we also used a normalized version of M_{ij}, M^*_{ij}, that weights positions equally regardless of diversity, and ranges between 0 and 1. All four statistics M_{ij}, M^*_{ij}, G_{ij}, and P_{ij} are always nonnegative, with value zero reflecting no covariation and large values indicating covariation. To test whether HIV-1C and HIV-1B sequences had different degrees of covariability at two positions i and j, we used the difference in M^*_{ij} values as a test statistic, and applied the nonparametric bootstrap to obtain an unadjusted p-value. The statistics are defined mathematically in the Appendix.

The covariability statistics may potentially be improved by upweighting covarying mutations present in phylogenetically distantly related sequences, as they are more informative about covariation than mutations in phylogenetically similar sequences.[1] This was done for the M_{ij} and M^*_{ij} statistics by incorporating a weight function into the statistics as detailed in the Appendix. The results were similar to those obtained without the weighting.

For sequences of length q, there are $q \times (q - 1)/2$ pairs of sites that could potentially covary. Korber *et al.*[1] and Bickel *et al.*[2] studied globally gap-stripped V3 loop sequences of length $q = 31$, so that $31 \times 30/2 = 465$ pairs of sites were studied. Bickel *et al.*[2] computed the three statistics for each pair, and used a permutation procedure to obtain p-values for each pair of sites (i, j). The resulting p-values are not adjusted for the many tests that were conducted, and must be adjusted to avoid many false-positive discoveries. We used the same permutation procedure as Bickel *et al.*[2] for computing unadjusted p-values, but applied a more powerful procedure of p-value adjustment that detects covarying positions with greater probability.

Adjusting for the multiplicity of tests

Bickel *et al.*[2] used a Bonferroni correction to judge statistical significance of the hundreds of covariation tests, which amounts to using $p = 0.05/465 = 0.00011$ as a cut-off value for comparison with each unadjusted p-value $pval_{ij}$ to judge whether the site pair i, j "significantly covaries." The Bonferroni method provides stringent control of the false-positive rate, guaranteeing at most a 5% chance that there are any false rejections. However, if the test statistics are positively correlated across pairs of sites, which is likely the case, then the Bonferroni procedure can be extremely conservative. The cost of the conservative approach is a higher rate of false negatives (i.e., truly covarying pairs that are not identified), which is exacerbated most when q is large, because the Bonferroni correction becomes increasingly (and extremely) conservative with the number of tests. To improve power we used the now-popular false discovery rate (FDR) method for multiple testing adjustment.[26] Given K unadjusted p-values, the FDR procedure works as follows: order the K p-values from smallest to largest, and let k^* be the largest integer k such that the kth largest p-value is less than or equal to $(k \times \alpha)/K$. The tests corresponding to the k^* smallest p-values are then declared significant at level α. With this simple procedure, the expected number of significant test results that are false rejections is no greater than α. The FDR procedure has greater statistical power than the Bonferroni procedure, especially when there are a large number of tests (our case), since the FDR method does not become conservative as the number of tests increases. We applied both the Bonferroni and FDR methods.

Estimating the extent of covariability

We used a new measure, C_{ij}, of the degree of covariation for a pair (i, j) of sites, which is related to the test statistic P_{ij}. Let a and b be the consensus amino acids at positions i and j, respectively. The measure C_{ij} is the geometric mean of the proportion of sequences in which a non-a amino acid at site i is accompanied by a non-b amino acid at site j, and the proportion of sequences in which a non-b amino acid at site j is accompanied by a non-a amino acid at site i (see the Appendix for a formula for C_{ij}). C_{ij} measures general covariation of residues away from the consensus pair, and ranges between 0 (no covariation) and 1 (perfect covariation). A confidence interval (CI) for the degree of covariation was constructed using the nonparametric bootstrap. C_{ij} will be most useful when the modal amino acids have high prevalence, which is usually the case: 83.4%, 94.8%, and 83.8% of the positions in HIV-1C Gag p17, Gag p24, and the V3 loop that we studied had consensus amino acids in 80% or more of the sequences. Because M^*_{ij} and G_{ij} range between 0 and 1, they are also useful as estimates of the extent of covariation.

A test statistic for higher-order covariability

To study covariability of triples of sites, we used the mutual information statistic M_{ijk} generalized to measure covariability of three sites (defined in the Appendix). This statistic is equivalent to the estimator of three-way disequilibrium developed by Weir.[28] We also used a normalized version of M_{ijk}, M^*_{ijk} that had been proposed[29] (defined in the Appendix). Studying covariability of site pairs and triples with the mutual information statistic has the advantage of providing a unified analysis of covariability, and its application is simpler and more straightforward than the second-order log-linear categorical models used by Bickel *et al.*[2]

$dN - dS$ estimates

For all positions in the evaluated regions, nonsynonomous (dN) minus synonomous (dS) rate differences $dN - dS$, scaled to represent the expected number of nucleotide substitutions per codon site, were estimated using the single most likely ancestral reconstruction maximum likelihood method.[30] A $dN - dS$ rate difference of 0 means neutral mutations, a difference <0 indicates purifying selection, and a difference >0 represents diversifying positive selection.[31] Spearman rank correlation coefficients and tests were used to assess associations between covariability statistics and $dN - dS$ values at pairs of positions.

Amino acid diversity

To analyze amino acid diversity of regions, pairwise amino acid distances were computed using the PROTDIST program with the PAM model[32] as described previously.[9] For each region, a two-sided Z-test was used to compare the mean diversity between subtypes. The test statistic equals the difference in mean pairwise diversity divided by the standard error of the difference, which was computed with appropriate account for the correlations in pairwise distances.[27]

Selection of amino acid sites for analysis

Many of the positions in the regions we studied were highly conserved across the studied sequences, and hence contained little information about covariability. Positions at which fewer than two sequences within a subtype had a nonconsensus amino acid were not evaluated.

First, we analyzed 264 HIV-1C V3 loop sequences in the Los Alamos database sampled from the southern African nations Botswana ($n = 51$), Malawi ($n = 80$), Mozambique ($n = 7$), South Africa ($n = 91$), Zambia ($n = 4$), and Zimbabwe ($n = 31$).[33] HIV-1C V3 loop sequences from other geographic regions were excluded to help avoid bias and to aid interpretability. The aligned sequences spanned 37 positions, 27 of which had enough diversity to be investigated for covariability. We evaluated all $27 \times 26/2 = 351$ viable pairs for covariability. Second, we analyzed the p17 and p24 regions of 73 nonrecombinant HIV-1C Gag sequences described in Novitsky *et al.*,[9] sampled from Botswana ($n = 51$), South Africa ($n = 5$), Tanzania ($n = 2$), Zambia ($n = 2$), Brazil ($n = 2$), Ethiopia ($n = 1$), India ($n = 9$), and Israel ($n = 1$). The analyzed p17 and p24 regions comprised 145 and 231 positions in alignment, respectively. The sequence lengths imply that there are 10,440 and 26,565 pairs of positions that potentially may covary. Analysis of this many pairs is highly demanding computationally; to make it viable, we reduced the number of sites to those at which the consensus amino acid had frequency less than 90%. This restriction reduced the number of positions to 38 and 22 for p17 and p24, respectively. Third, using the same set of 73 HIV-1C sequences, we analyzed separately the five immunodominant regions of Gag, Nef, and Tat described in the Introduction: Gag 171–190, Gag 291–320, Nef 67–96, Nef 122–141, and Tat 36–50. For the Tat region, one sequence was excluded because an approximately 500 base pair deletion eliminated its first exon. For these immunodominant regions, 1 of 20, 6 of 30, 10 of 30, 6 of 20, and 4 of 15 positions, respectively, had sufficient variability to be evaluated for covariability, so that 0, 15, 45, 15, and 6 pairs of positions were evaluated. After identifying the sites that significantly covaried within individual epitope-rich regions, we searched for cross-epitope region covariability by analyzing these positions together.

For each pair of sites (i, j) identified to significantly covary by the FDR method, we used the statistical criteria M_{ijk} and M^*_{ijk} to search for covariability of the triples (i, j, k), for k ranging over all sites other than i and j with at least one significant connection. These analyses were carried out separately within the eight regions studied.

The analyses described above were repeated in 264 HIV-1B V3 loop sequences and 73 HIV-1B sequences selected randomly from the Los Alamos sequence database[33] in the same regions as those studied in HIV-1C sequences. Thirty-two of the 37 positions in the HIV-1B V3 loop alignment were variable enough to assess; hence $32 \times 31/2 = 496$ pairs were evaluated for covariability. For the HIV-1B regions corresponding to the CTL epitope-rich regions in HIV-1C, 2 of 20, 3 of 30, 14 of 30, 8 of

20, and 3 of 15 positions were evaluable for regions Gag 171–190, Gag 291–320, Nef 67–96, Nef 122–141, and Tat 36–50, respectively. For Gag p17 and Gag p24, the HIV-1B positions corresponding to those evaluated in HIV-1C were analyzed. However, whereas all of the studied positions in Gag p17 and Gag p24 for HIV-1C had consensus amino acid frequency less than 90%, many of the corresponding positions for HIV-1B did not. To allow subtype comparisons of the extent of covariability, supplementary analyses were done at all HIV-1B Gag p17 and Gag p24 positions with more than 10% variability, which amounted to 35 and 15 positions, respectively.

The translated amino acid sequence sets were aligned using ClustalX v.1.81[34] and manually edited using BioEdit.[35] For the covariability analyses gaps were treated as a twenty-first residue. *p*-values in covariability analyses were calculated using 10,000 randomly permuted datasets. Bootstrap samples (5000) were used for computing CIs about the covariation parameters estimated by C_{ij} or M^*_{ij}.

RESULTS

In HIV-1C viral sequences, we found several significantly covarying pairs in the V3 loop, in Gag p17, and in epitope-rich region Nef 67–96, a small number of significantly covarying pairs in Gag p24, Gag 291–320, and Tat 36–50, and no covarying pairs in Gag 171–190 or Nef 122–141. In HIV-1B viruses, a very large number of site pairs significantly covaried in the V3 loop, a large number covaried in Gag p17, Gag p24, and Nef 122–141, a small number covaried in Nef 67–96 and Tat 36–50, and none covaried in Gag 171–190 and Gag 291–320. As judged by significance of at least one of the four statistical criteria, the FDR procedure was much more powerful in detecting covariation than the Bonferroni procedure, yielding a total of 321 discoveries of significantly covarying pairs, compared to 131 by the Bonferroni procedure (Table 1). Of the discoveries by FDR and by Bonferroni, respectively, 52 (16.2%) and 23 (17.6%) covarying pairs were in HIV-1C; the decreased covariability in HIV-1C was mostly due to the V3 loop and Gag p17. In contrast, only Nef 67–96 and Gag 291–320 showed more covariability in HIV-1C. We describe important details of the results by region.

V3 loop

Within HIV-1C, 17 pairs of V3 loop amino acids significantly covaried (Table 1, Fig. 1). Figure 2 displays the connections significant by two or more statistics. Figures 2–4

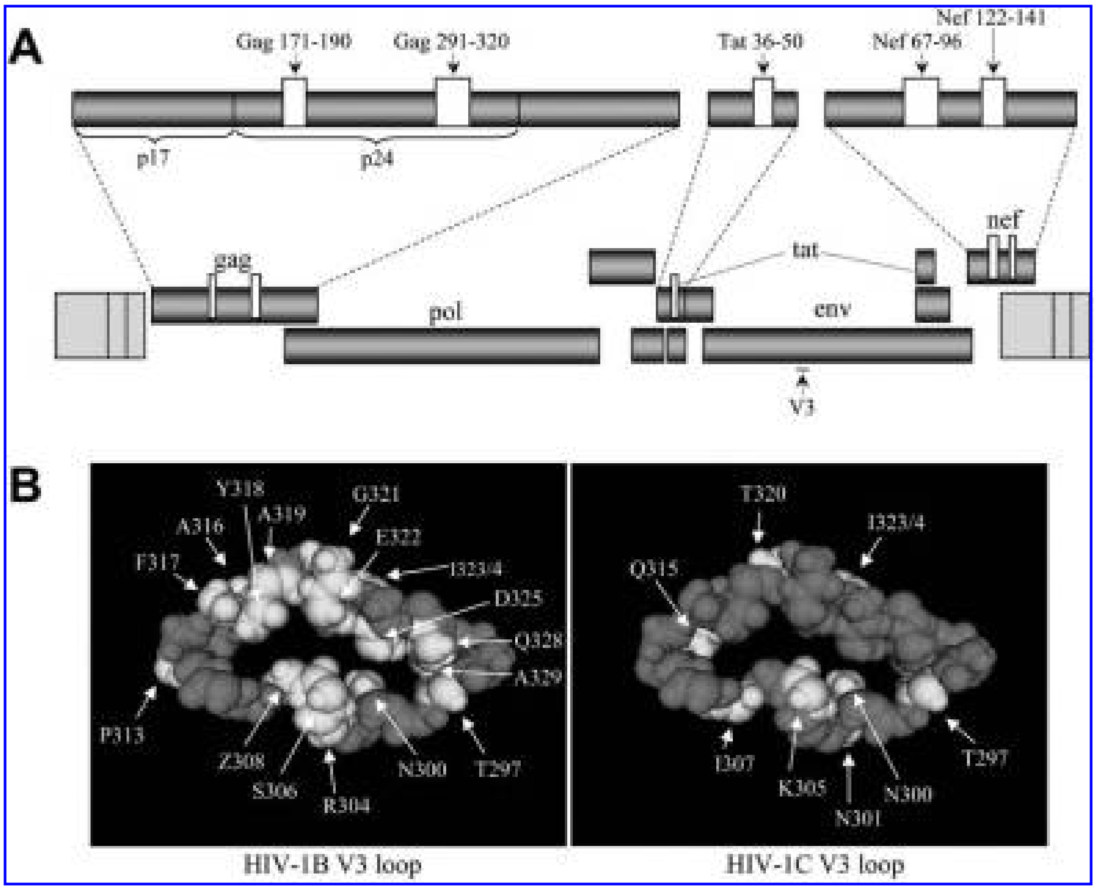

FIG. 1. (A) Whole proteins, location of cleavage products, and topology of the regions analyzed for covariability. (B) Structure of the V3 loop,[50] with strongly covarying sites indicated (listed in Table 2). T297 indicates that the consensus residue is T at position 297.

Table 1. Number of Significantly Covarying Amino Acid Site Pairs

Region (number sequences)	*Multiple test technique*	M	M*	G	P	MM[a]	*≥1*[b]	*≥2*	*≥3*	*4*
C V3 loop (264)	Bonferroni	3	1	3	4	1	8	2	1	0
C V3 loop (264)	FDR	7	10	12	8	7	17	11	8	1
C Gag p17 (73)	Bonferroni	2	2	4	4	2	7	2	2	1
C Gag p17 (73)	FDR	13	12	13	11	12	18	17	9	5
C Gag p24 (73)	Bonferroni	0	0	0	0	0	0	0	0	0
C Gag p24 (73)	FDR	0	0	3	0	0	3	0	0	0
C Gag 171–190 (73)	Bonferroni	0	0	0	0	0	0	0	0	0
C Gag 171–190 (73)	FDR	0	0	0	0	0	0	0	0	0
C Gag 291–320 (73)	Bonferroni	0	0	1	1	0	1	1	0	0
C Gag 291–320 (73)	FDR	0	0	1	3	0	3	1	0	0
C Nef 67–96 (73)	Bonferroni	4	1	2	0	1	5	1	1	0
C Nef 67–96 (73)	FDR	7	7	9	4	7	9	8	7	3
C Nef 122–141 (73)	Bonferroni	0	0	0	0	0	0	0	0	0
C Nef 122–141 (73)	FDR	0	0	0	0	0	0	0	0	0
C Tat 36–50 (72)	Bonferroni	2	2	1	1	2	2	2	2	1
C Tat 36–50 (72)	FDR	2	2	1	2	2	2	2	2	1
B V3 loop (264)	Bonferroni	44	39	42	62	37	76	51	34	25
B V3 loop (264)	FDR	144	140	126	157	144	173	148	138	122
B Gag p17 (73)	Bonferroni	6	8	6	12	6	13	9	7	2
B Gag p17 (73)	FDR	29	27	22	52	27	59	36	23	12
B Gag p24 (73)	Bonferroni	8	6	5	6	6	9	7	5	4
B Gag p24 (73)	FDR	9	10	15	16	9	16	16	9	7
B Gag 171–190 (73)	Bonferroni	0	0	0	0	0	0	0	0	0
B Gag 171–190 (73)	FDR	0	0	0	0	0	0	0	0	0
B Gag 291–320 (73)	Bonferroni	0	0	0	0	0	0	0	0	0
B Gag 291–320 (73)	FDR	0	0	0	0	0	0	0	0	0
B Nef 67–96 (73)	Bonferroni	1	1	1	1	1	1	1	1	1
B Nef 67–96 (73)	FDR	1	1	1	1	1	1	1	1	1
B Nef 122–141 (73)	Bonferroni	8	6	6	7	6	8	7	7	5
B Nef 122–141 (73)	FDR	19	15	12	11	15	19	15	13	10
B Tat 36–50 (72)	Bonferroni	1	1	1	1	1	1	1	1	1
B Tat 36–50 (72)	FDR	1	1	1	1	1	1	1	1	1

[a] Significant by *M* and *M**.

[b] Significant by at least one of the statistics. ≥2 (≥3) indicates significance by at least 2 (3) of the statistics, and 4 indicates significance by all 4 statistics. Note that the total number of shaded boxes in Figs. 2–4 for a given HIV-1 region equals the FDR entry in column ≥2, and the total number of light/medium shaded boxes equals the FDR entry in column ≥3.

show only connections with relatively robust evidence of covariability (i.e., connections significant by two or more statistics), while Table 1 shows all connections significant by one or more statistics. One pair of positions, 316 and 321, stood out as covarying the strongest, where position 316 is adjacent to the crown of the V3 loop (positions 312 through 315) and position 321 is downstream of the crown. The consensus residues T, G covaried to A, N in 21 sequences. In addition, at positions 321, 322 the consensus pair G, D covaried to N, G in 11 sequences, and at positions 320, 321, T, G covaried to Z, N in 7 sequences and to N, N in 4 sequences. ("Z" denotes a gap in the subtype consensus sequence.)

In comparison, 10-fold more pairs significantly covaried in the HIV-1B V3 loop. To highlight the strongest covarying positions, in Table 2 site pairs significant by all four statistics by the Bonferroni method and that have M^*_{ij} in the highest 5% of $M*_{ij}$ across all site pairs in a given region are bolded, and 17 site pairs satisfied this condition in the HIV-1B V3 loop. At positions 304, 306, 322 and an insertion between HXB2 positions 323 and 324 (referred to as indel[323,324]), 25 HIV-1B sequences covaried from the residues R, S, D, I to I, G, E, V. At positions 319 and 322, 46 sequences covaried from A, D to T, E. Also note that adjacent sites often covaried with the same third position, reflecting the role of physical proximity. After FDR adjustment, three site pairs had significantly greater covariability in HIV-1C than HIV-1B (301–303, 303–315, 303–330), and 50 site pairs had significantly greater covariability in HIV-1B.

To place our results in the context of the earlier work, in Fig. 2 we marked the highly significant site pairs found by Korber *et al.*[1] ("K"), Bickel *et al.*[2] ("B"), and both studies ("KB"). Of the 7 HIV-1B site pairs identified as strongly covarying by Korber *et al.*[1] (see their Table 1), we found 6 of them to covary in HIV-1B, each by four statistics. For HIV-

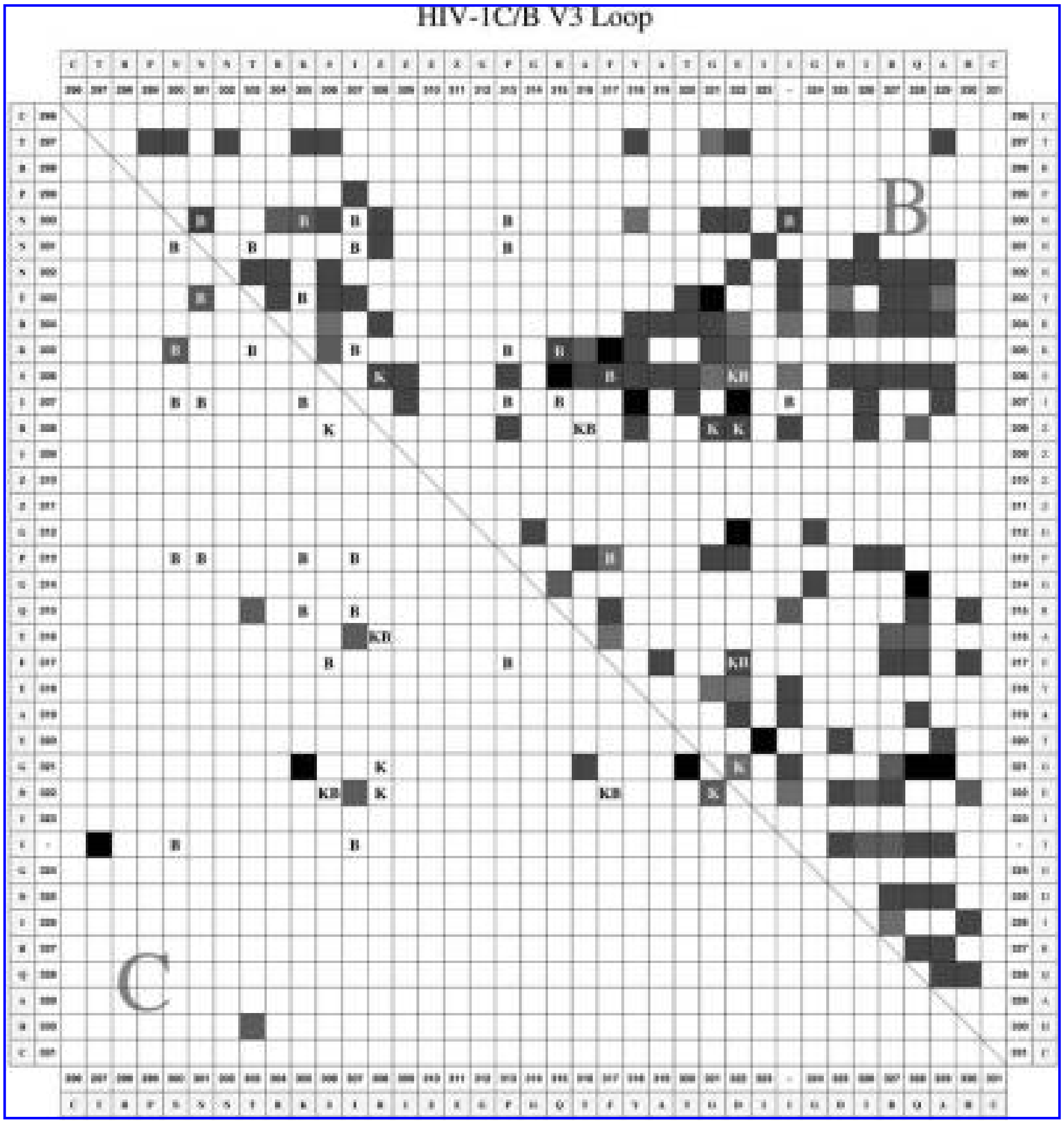

FIG. 2. Covarying pairs of HIV-1C and HIV-1B V3 loop amino acid sites, significant by at least three (light/medium shading), or two (darkest shading) of the statistical criteria M_{ij}, M^*_{ij}, G_{ij}, P_{ij}. Only pairs of sites with relatively robust evidence of covariability are shown, i.e., connections significant by only one statistic are not shown. The positions are marked at the margins of the display matrix, which are relative to the HXB2 numbering system.[20] The letters indicate the consensus amino acids (left and bottom for HIV-1C; right and top for HIV-1B), where "–" denotes a gap in alignment. The lower-left and upper-right parts of the matrix are for HIV-1C and HIV-1B, respectively. Pairs colored red have the greatest evidence of covariation: all four statistics were significant by the Bonferroni method and the M^*_{ij} statistic was in the upper fifth percentile of the M^*_{ij} values.

1B our analysis also found covariability of 9 of the 20 strongly covarying site pairs identified by Bickel *et al.*[2] (see the 20 bolded site pairs in their Fig. 1). In total we corroborated covariation of 13 site pairs in HIV-1B previously identified, and we found 27 covarying site-triples that involved one of these site pairs (further results on covarying site triples are given below). Both previous papers concluded that positions 306, 308, and 318–322 appeared to be most important for covariability, and interestingly these positions stood out in our analysis as participating in extensive covariability; notably positions 306, 308, 321, and 322 had 21, 11, 10, and 14 significant connections by all four statistics, respectively. Furthermore, Bickel *et al.*[2] noted covariabiality among the four sites 306–308–317–322, and we found 10 covarying site triples involved with a pair within this set: 300–306–322, 301–306–317, 302–317–322, 304–306–322, 306–314–322, 306–321–322, 306–322–324, 306–314–317, 306–317–324, and 306–317–327. Some of the aforementioned positions important for HIV-1B also covaried in HIV-1C, 320, 321, and 322, however positions 306, 308, 318, and 319 had no con-

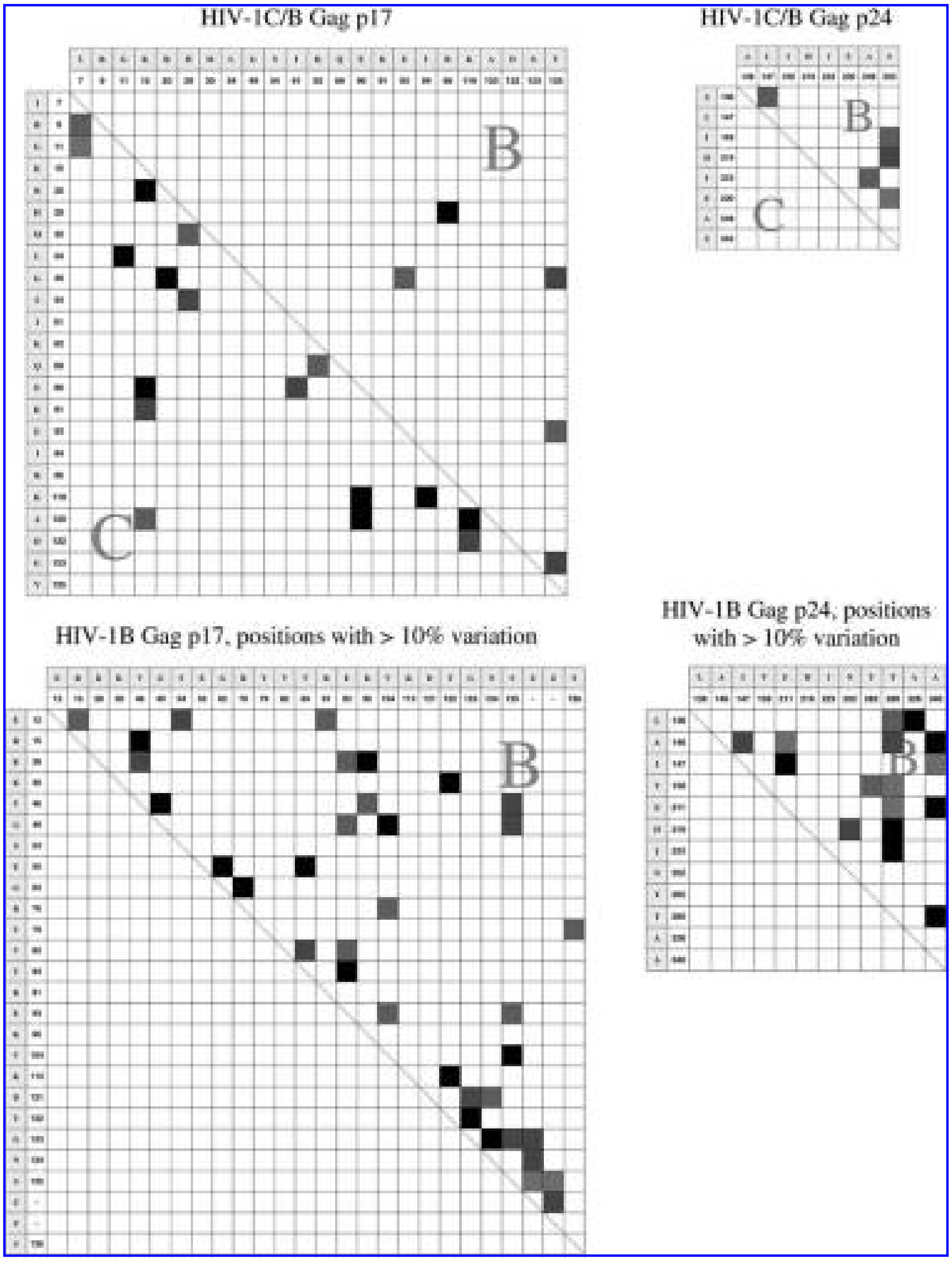

FIG. 3. Covarying pairs of HIV-1C and HIV-1B Gag p17 and Gag p24 amino acid sites, significant by at least three (light/medium shading), or two (darkest shading) statistical criteria M_{ij}, M^{*}_{ij}, G_{ij}, P_{ij}. Only pairs of sites with relatively robust evidence of covariability are shown, i.e., connections significant by only one statistic are not shown. Notation and meaning of the red color is as in Fig. 2. For both subtypes (lower-diagonal HIV-1C; upper diagonal HIV-1B) the upper panels show results for the 37 positions that were variable in the HIV-1C Gag p17 sequences (defined by <90% frequency of the consensus/majority amino acid) and for the 22 positions that were variable in HIV-1C Gag p24. For subtype B the bottom panels show results for the 35 HIV-1B Gag p17 variable positions and the 15 HIV-1B Gag p24 variable positions. For brevity, only positions at which there is some significant covariability are shown.

nections. Only connections 300–305, 301–303, and 321–322 found previously were identified in our analysis of HIV-1C. The fact that almost all of the site pairs that were found to strongly covary in HIV-1B by the earlier and current work did not covary in HIV-C sequences raises the hypothesis that the HIV-1C V3 loop may operate quite differently in functional or structural characteristics than the HIV-1B V3 loop.

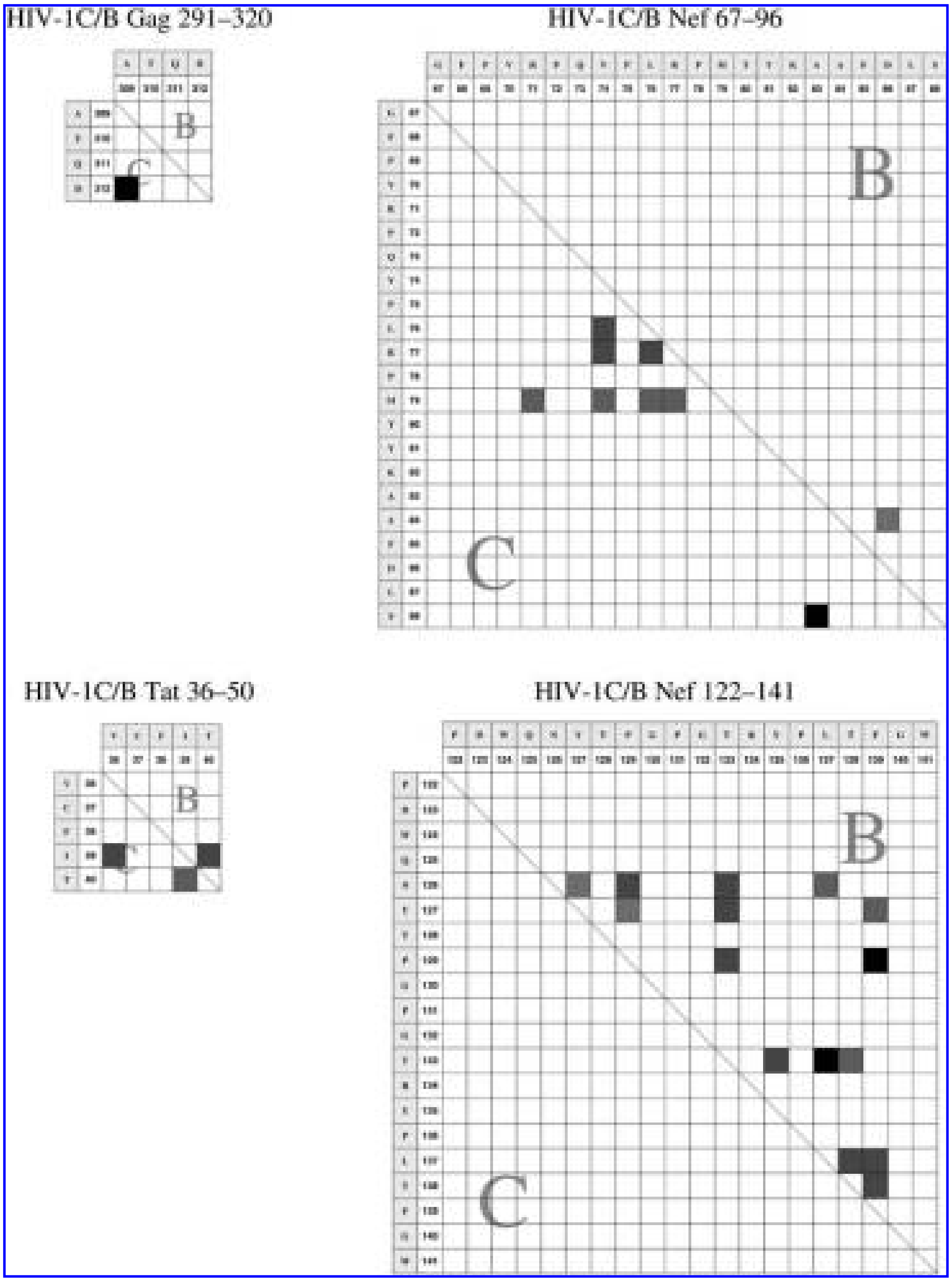

FIG. 4. Covarying pairs of Gag 291–320, Nef 67–96, Nef 122–141, and Tat 36–50 amino acid sites, for HIV-1C and HIV-1B, significant by at least three (light/medium shading), or two (darkest shading) statistical criteria M_{ij}, M^{*}_{ij}, G_{ij}, P_{ij}. Only pairs of sites with relatively robust evidence of covariability are shown, i.e., connections significant by only one statistic are not shown. Notation is as in Fig. 2.

Gag p17

In HIV-1C, 18 pairs significantly covaried in Gag p17 (Tables 1 and 2; Fig. 3, upper-left matrix). For the corresponding Gag p17 positions in HIV-1B, 7 pairs significantly covaried. Two pairs of sites had significantly greater covariation in HIV-1C than HIV-1B (15–90, 28–30), and one site pair had significantly greater covariation in HIV-1B than HIV-1C (93–125).

The analysis of HIV-1B Gag p17 positions with more than 10% variability showed a large number of covarying pairs, with 59 total connections (Tables 1 and 2; Fig. 3, lower-left matrix). The consensus pair V, T at positions 82, 84 covaried to I, V in 14 sequences.

TABLE 2. C_{ij} COVARIATION MEASURES WITH CONFIDENCE INTERVALS FOR COVARYING PAIRS SIGNIFICANT BY ALL FOUR STATISTICAL CRITERIA, USING THE FDR METHOD

Region (no. sequences) *[average normalized* dN–dS*]*	*Site pair*	*Consensus pair*	*Modal substitution pair*	C_{ij}	*95% CI*	*Normalized* dN–dS[a]
C V3 loop (264) [−0.32]	316 321	T, G	A, N	0.555	0.366–0.686	−0.68, 0.23
C Gag p17 (73) [−0.47]	**7 11**[b]	I, G	V, E	0.592	0.333–0.877	−0.39, 0.89
	15 91	K, K	T, G	0.733	0.393–0.778	1.34, 0.64
	28 54	H, S	Q, A	0.560	0.347–0.734	−1.50, 0.81
	61 90	I, E	M, A	0.305	0.192–0.785	1.72, 1.91
	119 122	K, D	E, A	0.555	0.077–0.775	0.66, −0.67
C Nef 67–96 (73) [−0.84]	74 76	V, L	X, Z	0.459	0.000–0.816	−1.87, −0.29
	74 77	V, R	M, K	0.873	0.000–1.000	−1.87, −0.82
	76 77	L, R	Z, T	0.410	0.000–0.791	−0.29, −0.82
C Tat 36–50 (72) [−0.23]	**36 39**	V, I	A, L	0.577	0.316–0.775	1.23, 0.52
B V3 loop[c] (264) [−0.08]	297 318	T, Y	S, V	0.291	0.192–0.447	1.00, 0.33
	297 321	T, G	I, E	0.296	0.199–0.525	1.00, 0.01
	300 318	N, Y	G, F	0.547	0.289–0.602	0.05, 0.33
	303 304	T, R	V, Q	0.283	0.000–0.438	−0.67, −1.85
	303 325	T, D	V, K	0.289	0.160–0.430	−0.67, −1.00
	303 329	T, A	V, P	0.866	0.577–1.000	−0.67, −0.67
	304 306	R, S	I, G	0.565	0.473–0.646	−1.85, 0.29
	304 322	R, E	I, Q	0.466	0.397–0.530	−1.85, 1.63
	304 324	R, I	I, V	0.668	0.557–0.765	−1.85, 0.18
	304 —[d]	R, I	I, V	0.668	0.695–0.968	−1.85, 0.26
	304 325	R, D	Q, K	0.192	0.070–0.316	−1.85, −1.00
	304 329	R, A	Q, P	0.247	0.000–0.403	−1.85, −0.67
	305 318	K, Y	R, V	0.439	0.265–0.548	−0.57, 0.33
	305 321	K, G	R, K	0.351	0.181–0.467	−0.57, 0.01
	306 318	S, Y	G, F	0.239	0.140–0.330	0.29, 0.33
	306 321	S, G	G, D	0.447	0.393–0.634	0.29, 0.01
(Korber *et al.*,[1]	**306 322**	S, E	G, Q	0.589	0.520–0.676	0.29, 1.63
Bickel *et al.*[2])	**306 324**	S, I	G, V	0.651	0.563–0.727	0.29, 0.18
Bickel *et al.*[2])	**313 317**	P, F	S, I	0.405	0.242–0.546	−0.50, 0.69
	316 317	A, F	V, Y	0.422	0.243–0.550	0.33, 0.69
B V3 loop (264)	317 328	F, Q	L, K	0.479	0.333–0.605	0.69, 0.45
(Korber *et al.*[1])	**318 319**	Y, A	V, K	0.313	0.143–0.444	0.33, 1.34
	318 322	Y, E	V, R	0.475	0.266–0.456	0.33, 1.63
	319 322	A, E	T, D	0.424	0.357–0.730	1.34, 1.63
	321 322	G, E	K, R	0.420	0.375–0.549	0.01, 1.63
	322 324	E, I	Q, V	0.514	0.441–0.581	1.63, 0.18
B Gag p17[e] (73) [−0.52]	12 54	E, S	Q, A	0.314	0.119–0.755	1.53, 1.44
	28 46	K, V	M, L	0.395	0.153–0.497	0.55, −1.89
	46 125	V, S	I, G	0.605	0.480–0.893	−1.89, 1.41
	49 125	G, S	S, K	0.456	0.186–0.497	0.93, 1.41
	82 84	V, T	I, V	0.676	0.431–0.720	1.30, 2.34
	121 123	D, G	A, D	0.397	0.231–0.821	1.17, 1.26
	123 125	G, S	D, G	0.260	0.095–0.625	1.26, 1.41
	123 —[f]	G, Z	D, N	0.316	0.000–0.751	1.26, 0.46
	124 —[f]	N, Z	T, N	0.397	0.231–0.821	−1.03, 0.46
	125 —[f]	S, Z	G, S	0.510	0.355–0.781	1.41, 0.46
	125 —[g]	S, Z	G, T	0.516	0.250–0.707	1.41, 1.34
	—[f] —[g]	Z, Z	N, S	0.751	0.739–1.000	0.46, 1.34
B Gag p24[e] (73) [−1.11]	146 147	A, I	P, L	0.510	0.269–0.710	0.38, −0.44
	146 211	A, E	P, D	0.802	0.516–0.810	0.38, 0.35
	146 280	A, T	P, A	0.414	0.222–0.575	0.38, 2.21
	147 340	I, A	L, G	0.662	0.516–0.925	−0.44, 1.92
	159 280	V, T	I, I	0.548	0.513–0.910	−1.01, 2.21
	211 280	E, T	D, A	0.452	0.184–0.734	0.35, 2.21
	219 252	H, N	Q, S	0.240	0.086–0.541	2.15, −0.38
B Nef 67–96 (73) [−0.57]	**84 86**	A, D	G, V	0.570	0.131–0.785	−1.46, 0.34
B Nef 122–141 (73) [−0.74]	**126 127**	N, Y	T, T	0.548	0.302–0.791	−2.06, −0.60
	126 129	N, P	T, Q	0.316	0.000–0.630	−2.06, −0.19
	126 133	N, T	T, S	0.386	0.187–0.524	−2.06, −0.76

TABLE 2. C_{IJ} COVARIATION MEASURES WITH CONFIDENCE INTERVALS FOR COVARYING PAIRS SIGNIFICANT BY ALL FOUR STATISTICAL CRITERIA, USING THE FDR METHOD (*CONTINUED*)

Region (no. sequences) *[average normalized* dN–dS*]*	*Site pair*	*Consensus pair*	*Modal substitution pair*	C_{ij}	*95% CI*	*Normalized* dN–dS[a]
	127 129	Y, P	T, Q	0.577	0.000–1.000	−0.60, −0.19
	127 133	Y, T	T, S	0.264	0.146–0.397	−0.60, −0.76
	129 133	P, T	Q, S	0.305	0.073–0.433	−0.19, −0.76
	133 135	T, Y	P, F	0.227	0.096–0.771	−0.76, −0.95
	137 138	L, T	V, C	0.107	0.000–0.816	−1.17, −0.19
	137 139	L, F	T, L	0.176	0.000–0.671	−1.17, −0.49
	138 139	T, F	D, L	0.186	0.000–0.516	−0.19, −0.49
B Tat 36–50 (72) [0.34]	39 40	I, T	T, K	0.539	0.359–0.685	2.62, 5.70

[a] Normalized $dN - dS$ is the scaled $dN - dS$ estimate obtained by the single most likely ancestral reconstruction maximum likelihood method[30] for each of the two positions.

[b] Pairs significant by all four statistical criteria by the Bonferroni method and that have M^*_{ij} in the upper fifth percentile of M^*_{ij} across all pairs for the given region are bolded.

[c] Connections significant by all four statistical criteria by the Bonferroni method are included (due to space limitations connections significant only by the FDR method are omitted).

[d] —is an Indel between positions 323 and 324 of the HXB2 alignment.

[e] For HIV-1B Gag p17 and HIV-1B Gag p24, results are reported for the analyses of the HIV-1B positions at which the frequency of the consensus amino acid is less than 90%.

[f] Two Indels—were in the HIV-1B Gag p17 alignment, between positions 125 and 126 in the HXB2 alignment. — is adjacent to position 126.

[g] — is adjacent to position 125 in HXB2.

Gag p24

In HIV-1C Gag p24, three site pairs significantly covaried. There were twice as many (6) significant connections in the corresponding positions of HIV-1B (Fig. 3, upper-right matrix). Although none of the covarying pairs was common among the subtypes, notably position 252 was important for both subtypes. Four site pairs covaried significantly more in HIV-1C (159–312, 159–252, 219–228, 228–252) and two site pairs covaried significantly more in HIV-1B (91–116, 98–120).

The analysis of HIV-1B Gag p24 positions with more than 10% variability showed more extensive covariability than for HIV-1C Gag p24 (Fig. 3, lower-right matrix). Sixteen pairs significantly covaried, with pairs 159–280 and 211–280 showing strong covariation.

HIV-1C CTL epitope-rich regions

In HIV-1C Gag 171–190, all positions but one were too conserved to assess covariability; therefore no pairs were evaluated. Only one site pair was evaluable in HIV-1B Gag 171–190, without significant covariation. Three site pairs significantly covaried in HIV-1C Gag 291–320. The consensus pair A, D at positions 309 and 312 covaried to S, E in all 8 sequences for which positions 309 and 312 both contained nonconsensus residues. No pairs significantly covaried in HIV-1B Gag 291–320. Nine pairs significantly covaried in HIV-1C Nef 67–96 (Tables 1 and 2; Fig. 4), whereas only one site pair covaried in HIV-1B Nef 67–96, by all four statistics (84–86); 5 sequences covaried from A, D to G, V and 5 sequences from A, D to G, F. In HIV-1C, no positions significantly covaried in Nef 122–141, whereas 19 site pairs covaried in HIV-1B Nef 122–141. In HIV-1C Tat 36–50, 2 pairs significantly covaried (36–39 and 39–40), and in HIV-1B Tat 36–50, positions 39 and 40 significantly covaried, with 7 sequences covarying from I, T to T, K. For the five immunodominant regions no site pairs had significantly different covariation by subtype.

In the cross-epitope region analysis for HIV-1C, 13 significant connections were found. Of these, 12 pairs were within one of the epitope-rich regions (i.e., they were discovered in the region-specific analysis). The one (weakly) significant covariable pair with sites in different epitope-rich regions involved positions 88 in Nef 67–96 and 315 in Gag 291–320, with C_{ij} = 0.387 (95% CI 0.000–0.564). The consensus pair S, N covaried to G, A; G, G; and D, G in one sequence each.

Analysis for covariability of triples of sites

In the HIV-1C V3 loop, we evaluated 59 site triples for covariability, including positions 297, 300, 301, 303, 305, 307, 315, 316, 321, and 322 together with the significant pairwise connections 301–303, 303–315, 303–330, 305–321, 307–316, 307–322, 316–321, and 320–321. Two triplets of sites significantly covaried (297–316–321 and 297–307–316). In comparison, the analysis of 2498 site triples in the HIV-1B V3 loop involving 30 positions and 93 significant site pairs by all four statistics and Bonferroni revealed 242 significant connections, which did not include the significant connections in HIV-1C. Site pairs 322–325, 322–324, and 307–326 had the most connections with another position, 26, 14, and 12, respectively. Of the 17 site pairs with greatest evidence of covariation (bolded in Table 2), all but 3 covaried with at least one other position, and site pairs 303–325, 304–indel[323,324], 306–321, 321–322, and 322–324 had the most covarying site triples, each with between 4 and 6 other positions.

Of 324 site triples evaluated in HIV-1C Gag p17 involving 20 positions and 18 pairwise connections, we found 87 significantly covarying triplets. Seventeen of the connected site triples involved site pair 15–20, 17 connections involved 15–120, 17 involved 20–49, and 14 involved 90–120. Of the 826 site triples evaluated in HIV-1B Gag p17 over 25 positions and 36 pairs, 84 significant connections were found, 14 of which involved site pair 125–indel[125,126], 12 with 28–95, 7 with 62–76, 5 with strongly covarying pair 49–125 (49–62–125, 49–93–125, 49–123–125, 49–125–indel[125,126], 49–125–indel[125,126]), and 3 with strongly covarying pair 82–84 (46–82–84, 62–82–84, 82–84–105). For HIV-1C Gag p24, 9 site triples were evaluated for covariability, based on 5 positions and 3 site pairs, and no significant connections were found. For HIV-B Gag p24, 160 site triples were evaluated (12 positions and 16 site pairs), and there were 9 significant connections, all of which involved key position 280 (with site pairs 146–326, 146–340, 147–159, 147–223, 159–223, 159–340, 219–223, 219–252, 223–340). Within the immunodominant regions, only 3 significant connections were found, all in HIV-1B Nef 67–96: 72–74–78, 74–84–93, and 84–90–93.

Normalized $dN - dS$ differences associated with degree of covariability

For each region, correlations between the average normalized $dN - dS$ value for each pair (i, j) of positions and the C_{ij} covariability statistics were assessed. There were moderate or weak correlations in both subtypes for the V3 loop, Gag p17, and Gag p24. For the HIV-1C V3 loop, the Spearman rank correlation r was 0.41 ($p < 0.0001$) between average normalized $dN - dS$ and C_{ij}, and for the HIV-1B V3 loop $r = 0.30$ ($p < 0.0001$). For HIV-1C Gag p17, $r = 0.25$ ($p < 0.0001$), whereas for HIV-1B Gag p17, $r = 0.12$ ($p = 0.0054$). In addition, for HIV-1C Gag p24 $r = 0.14$ ($p = 0.04$) and for HIV-1B Gag p24 $r = 0.39$ ($p < 0.0001$). For the immunodominant regions a significant correlation was found only for HIV-1C Gag 291–320, with $r = 0.52$ ($p = 0.05$). Based on the hypothesis testing procedure,[30] for none of the site pairs with greatest evidence of covariability (those listed in Table 2) was there significant evidence ($p < 0.05$) that both sites were under selection pressure (i.e., $dN - dS > 0$).

Association of amino acid site covariability and diversity

Of the 8 regions evaluated for covariability, 2 showed comparable levels of covariability in subtypes C and B (Gag 171–190, and Tat 36–50), 4 showed less covariability in HIV-1C (V3 loop, Gag p17, Gag p24, Nef 122–141), and 2 showed more covariability in HIV-1C (Nef 67–96 and Gag 291–320). To help interpret these results, Table 3 summarizes the pairwise amino acid diversity of the regions. Six of the regions had significantly different mean diversity, including 5 of the 6 regions with subtype differences in the degree of covariability. For the V3 loop and Nef 122–141, the subtype with more covariability also had higher diversity. This correspondence may be explained partly by the fact that the statistical criteria have greater statistical power for detecting covariability in regions with greater diversity. This observation may help explain the extensive amount of covariability identified in the HIV-1B V3 loop, with its large mean diversity of 22.2%, compared to 14.5% for the HIV-1C V3 loop. The subtype difference in covariability for Nef 122–141 is too large to be explained fully by differential diversity, however. In addition, Gag p17 had considerably less covariation for HIV-1C but more diversity. This finding may be explained by the fact that Gag p17 is highly immunodominant in HIV-1B, containing many known CTL epitopes[17]; the greater covariability may reflect the functional importance of HIV-1B Gag p17. This finding may also support

TABLE 3. AMINO ACID DIVERSITY BASED ON PAIRWISE DISTANCES

		Diversity (%)		*Difference in diversity (C − B)*			*Number (%) of unique sequences*
Region	*Subtype*	*Mean*	*Range*	*Mean*	*95% CI*[a]	p-*value*[b]	
V3 loop	C	14.5	0.8–70.8	−7.6	(−9.6, −5.7)	<0.0001*	196 (74.2%)
	B	22.2	0.2–98.7				144 (54.5%)
Gag p17	C	14.8	1.0–32.3	4.0	(2.4, 5.5)	<0.0001*	73 (100.0%)
	B	10.9	0.0–25.6				67 (91.8%)
Gag p24	C	5.1	0.4–10.4	1.8	(1.1, 2.4)	<0.0001*	73 (100.0%)
	B	3.3	0.0–8.6				60 (82.2%)
Gag 171–190	C	2.5	0.0–9.2	2.0	(0.5, 3.5)	0.008*	15 (20.5%)
	B	0.5	0.0–10.4				5 (6.8%)
Gag 291–320	C	4.4	0.0–18.6	2.3	(0.5, 3.9)	0.010*	16 (21.9%)
	B	2.1	0.0–21.6				11 (15.1%)
Nef 67–96	C	11.8	0.0–57.6	−4.0	(−7.6, −0.3)	0.035	31 (42.5%)
	B	15.7	0.0–100.0				46 (63.0%)
Nef 122–141	C	7.8	0.0–39.7	−10.1	(−14.6, −5.6)	<0.0001*	11 (15.1%)
	B	17.9	0.0–100.0				19 (26.0%)
Tat 36–50	C	11.9	0.0–54.2	−0.4	(−4.0, 3.1)	0.81	11 (15.3%)
	B	12.4	0.0–42.7				13 (18.1%)

[a] 95% confidence intervals (CIs) were computed as the mean difference in diversity plus or minus 1.96 times the standard error of the difference computed as described in Novitsky *et al.*[9]

[b] Unadjusted *p*-value, computed using the two-sample Z-test summarized in Materials and Methods and detailed in Novitsky *et al.*[9] An asterisk implies the result is statistically significant at level 0.05 after FDR adjustment for eight hypothesis tests.

the higher fitness of HIV-1C, which is indirectly supported by the relatively high prevalence of HIV-1C.

Despite the observed relative conservancy of subtype C's V3 loop, the data suggest that HIV-1C may have a broader range of viable viruses than HIV-1B: of the 264 V3 loop sequences studied, 196 (74%) were distinct for HIV-1C, but only 144 (54%) were distinct for HIV-1B.

DISCUSSION

This study evaluated for covariability eight regions in HIV-1C and HIV-1B, and identified hundreds of significantly covarying pairs and triples of amino acids. We summarize the results, connecting the covariability findings to biological knowledge where possible.

In HIV-1C, several covarying pairs were identified in the V3 loop, Gag p17, Gag p24, Gag 291–320, Nef 67–96, and Tat 36–50. The greatest covariability of amino acids to specific residues was found at positions 316–321 (from T, G to A, N) and 321–322 (from G, D to N, G) in the V3 loop, positions 309–312 (from A, D to S, E) in Gag 291–320, and positions 36–39 in Tat 36–50. For HIV-1B, the greatest covariability was found at positions 304, 306, 322, indel[323,324] in the V3 loop (from R, S, D, I to I, G, E, V), and at site pairs 82–84 in Gag p17 (from V, T to I, V) and 159–280, 211–280 in Gag p24. Table 2 lists the site pairs with robust evidence of covariability.

Notably, greater covariability was found in the V3 loop and Gag p17 for subtype B, perhaps suggesting greater functional constraints in these regions for HIV-1B. In addition, the finding of strong covariability of site pair 309–312 in Gag 291–320 but no covarying pairs in the corresponding region for HIV-1B may reflect the fact that this region contains T cell epitopes for subtype C but not for subtype B.[18,36] Similarly, in Nef 122–141, there were no significant covariable pairs in HIV-1C, but a plethora of connections in HIV-1B. This finding may reflect that this region, known to be part of the four-stranded antiparallel β-sheet of Nef,[37] is T cell epitope rich for subtype B but not for subtype C. In HIV-1C many T cell epitopes remain to be mapped, and the analysis reported here illustrates that covariability analyses complement direct mapping analyses, in these cases supporting that Gag 291–320 and Nef 122–141 merit special attention for HIV-1C and HIV-1B, respectively. Covariability within a type II polyproline helix that represents the main binding site for the Src family kinases[37] among positions 71–79 in Nef was found for subtype C, which may suggest the existence of compensatory mutations or linkage. This hypothesis is supported by the direct epitope mapping analyses, which showed that Nef 71–79 contains a T cell epitope for subtype C.[9,38] In addition, there was strong evidence that amino acid positions 39, 40 in the core region of Tat covaried in both subtypes.

A recent study of a rhesus macaque infected with SHIV-89.6P demonstrates the utility of covariability analysis of CTL epitope regions; Peyerl *et al.*[39] used covariability analysis to help establish that a CTL epitope was structurally constrained from mutating to escape from CTL recognition. Amino acid mutation T47I at position 2 in an immunodominant Gag Mamu-A*01-restricted epitope coincident with mutation I71V in a flanking position led to SHIV-89.6P escape from the dominant epitope-specific CTL response. As measured by Gag protein expression, viral infectivity, and replication kinetics, the fitness of the virus with T47I mutation was greatly reduced compared to the wild-type virus, but when the T47I and I71V mutations were both present, the fitness was restored to wild-type levels. Through the analysis of SIV and HIV-2 sequences from the Los Alamos sequence database,[33] Peyerl *et al.*[39] showed strong covariability of positions 47 and 71 (Fisher's exact test $p < 0.0005$), and this result can be confirmed using the methods used here, which are more appropriate than Fisher's exact test. This example illustrates that covariability analyses are useful for understanding how viruses escape from CTL recognition.

For the V3 loop, Gag p17, and Gag p24 in both subtypes, we found significant correlations between the C_{ij} statistics quantifying evidence for covariability and the normalized $dN - dS$ differences. However, for most of the strongly covarying amino acid sites (Table 2), there was no significant evidence that $dN - dS > 0$, and therefore the extent to which the covarying amino acids might be under selection pressure is unclear. We also identified several triples of sites that covaried in the HIV-1C V3 loop and a great number in the HIV-1B V3 loop (242 connections), a large number of covarying site triples in Gag p17 for both subtypes (87 for HIV-1C and 84 for HIV-1B), and several covarying site triples in HIV-1B Gag p24 (9, all involving key position 280).

Our analysis of the HIV-1B V3 loop corroborated several of biological observations.[1,2] de Jong et al.[40] found that mutations at positions 306 and positions within 319–322 are necessary for the HIV-1 phenotype to completely convert from NSI to SI, and we found strong covariability of position 306 with all four sites 319–322. Chesebro *et al.*[41] found that mutations at site 308 and sites within 318–322 are needed for a phenotype switch from T cell to macrophage tropic, and we found strong covariability of site 308 with sites 318, 321, 322. Site pairs 306–322[42] and 308–322[43] were found to be important for viral tropism. In addition to the observations made by Korber *et al.*[1] and Bickel *et al.*[2] on functional significance, positions 317 and 330 strongly covaried in HIV-1B (but not HIV-1C), and in mutagenesis experiments these sites were found to be important for CCR5 binding, soluble CD4 binding, and monoclonal antibody binding.[44] Position 297 showed extensive covariability in HIV-1B, and this position is in a neutralization epitope of monoclonal antibody 2G12.[45] Furthermore, position 301 is a potential N-linked glycosylation site for both subtypes (with NNT motif at positions 301–303 for most sequences), and site 301 covaried with site 303 for HIV-1C and with sites 300, 308, 323, and 326 for HIV-1B. Positions 301 and 303 have been implicated in immune escape.[46] These results may help the design of experiments that investigate the glycan shield as a mechanism of immune evasion.[47]

The HIV-1B V3 loop exhibited more diversity and covariability than the HIV-1C V3 loop, yet had a smaller range of unique viruses. This observation suggests that mutations in the HIV-1B V3 loop occur "as a package" (several amino acid replacements tend to appear together), and supports interpreting the wealth of covariability as reflecting greater functional constraints within the HIV-1B V3 loop compared to HIV-1C. The decreased functional constraints in the V3 loop of HIV-1C relative to HIV-1B may reflect that HIV-1C is highly fit and is in some sense ideal for the subtype C environment. This hypothesis of greater fitness of the C versus B V3 loop could be tested experimentally by generating a number of mutated plasmids and analyzing their viral fitness. If changes in the HIV-1C V3 loop result in weaker viruses, then the experiment could explain why HIV-1C accumulates less mutations within the V3 loop, and could partially explain why HIV-1C rarely switches to the X4 phenotype during the course of HIV-1 disease.[48,49]

There are several limitations to our analysis that could cause observed covariability to be a sampling artifact. As also discussed by Korber *et al.*[1] and Bickel *et al.*,[2] biases could result from (1) founder-virus effects, whereby statistical covariation may be observed between two sites because a group of viruses descended from a single ancestor virus; (2) unknown epidemiological clustering of some of the analyzed viruses; (3) the fact that data are sparse, with consensus amino acids presenting with high frequency with typically only a small number of other amino acids appearing; and (4) the sequences were not sampled randomly from a target population of interest (such as recent seroconverters in a geographic region where an HIV vaccine efficacy trial is being planned).

Unfortunately, because limited data are available on longitudinal sequence samples, especially for HIV-1C, it is not possible at this time to ascertain the potential impact of limitation (1); this is a general challenge faced in covariability analyses. The $dN - dS$ analyses are limited by the lack of longitudinal sampling of sequences, which will be required to definitively test the hypothesis that covariable pairs are under evolutionary selection pressure. Moreover, the pattern for covarying sites to have higher $dN - dS$ differences may have occurred in part because the test statistics have relatively high power for detecting covariability of sites with large $dN - dS$ differences.

Problem (2) was partially alleviated by sampling the HIV-1C viruses across several countries in southern Africa, and by sampling the HIV-1B viruses randomly from the Los Alamos database. Limitation (3) complicates the study of higher order covariability of sites, and can make p-values sensitive to a small number of sequences. Focusing attention away from p-values and toward the newly proposed covariability summary measures C_{ij} and M^*_{ij} eases the difficulty. The lower 95% confidence limits may be superior to p-values as measures of the reliability of the covariability findings.

To minimize limitation (4), to the extent possible, the analyzed HIV-1C viruses were selected from individuals with common characteristics. The V3 loop sequence dataset was geographically restricted to viruses known to come from southern African countries, and most of the Gag, Nef, and Tat sequences also came from southern Africa. In addition, most of the HIV-1C viruses were likely transmitted heterosexually, and most of the 51 HIV-1C viruses sampled from Botswana came from asymptomatic blood donors.[9,18] Other than these selection criteria, the analyses of the sequence data are susceptible to limitation (4). In particular, common selection factors were not available for determining the set of HIV-1B sequences for the analysis.

In the face of the potential biases, a case for biologically important covariability of two sites can be built using information on the functional covariation of the sites *in vitro* and on the biological function of the sites. Verification of cross-clade covariability also lends support to biological covariability, as bias from a founder-virus effect is less likely.

Due to the limited knowledge about the functional significance of positions in HIV-1C, it is not possible at this time to verify that the identified covarying positions in HIV-1C have biological importance. The analyses generate hypotheses about functionally important sites, which merit investigation in future studies of viral function. If the covariability could be linked directly to functional or structural significance, then knowledge of the covarying sites may be useful for selecting the amino acid sequence(s) of the antigen(s) used in an HIV-1 vaccine, and for selecting the peptides to use in the evaluation of immunogenicity of HIV-1 vaccines. Even in the absence of biological links, information on covariability helps delineate the distribution of HIV-1 quasispecies in a population, which guides the selection of antigen sequences to give maximal vaccine coverage. Typically, assessments of HIV-1 amino acid diversity and variability have ignored information on covariability. We propose that covariability data should be used to improve such assessments.

Given the difficulty in interpreting results of analyses of haphazardly sampled sequence sets, we think that analyses of sequence sets with carefully constructed sampling plans will be of greatest value. The sample size for such studies may be modest, and therefore it is important to use powerful statistical procedures for maximizing signal detection, such as the FDR procedure used here. Color figures and online software for graphical display and testing of covariability are available from the first author (pgilbert@scharp.org).

APPENDIX: STATISTICS FOR MEASURING COVARIABILITY

In the notation of Bickel *et al.*,[2] the statistics are defined as follows. For a set of N equal-length aligned sequences, let

$$\begin{aligned}
\hat{p}_i(a) &= \text{(number of sequences with residue } a \text{ at site } i)/N \\
\hat{p}_{ij}(a,b) &= \text{(number of sequences with residue } a \text{ at site } i \text{ and residue } b \text{ at site } j)/N \\
\hat{p}_{ij}(a,\max) &= \max_b \hat{p}_{ij}(a,b),\ \hat{p}_{ij}(\max,b) = \max_a \hat{p}_{ij}(a,b), \\
&\quad \hat{p}_i(\max) = \max_a \hat{p}_i(a)
\end{aligned}$$

Then

$$M_{ij} = \sum_{a,b} \hat{w}_{ij}(a,b)\hat{p}_{ij}(a,b)\log\{\hat{p}_{ij}(a,b)/[\hat{p}_i(a)\hat{p}_j(b)]\}$$

$$G_{ij} = \frac{1}{2}\,\frac{\Sigma_a \hat{p}_{ij}(a,\max) + \Sigma_b \hat{p}_{ij}(\max,b) - \hat{p}_i(\max) - \hat{p}_j(\max)}{1 - (1/2)[\hat{p}_i(\max) + \hat{p}_j(\max)]}$$

$$P_{ij} = \max_{a,b} P_{ij}(a,b)$$

where $P_{ij}(a,b)$ is the sum of the four terms $\hat{p}_{ij}(a',b')\log\{\hat{p}_{ij}(a',b')/[\hat{p}_i(a')\hat{p}_j(b')]\}$ with (a',b') set to (a,b), $(\bar{a},b)$, $(a,\bar{b})$, and $(\bar{a},\bar{b})$, where $\bar{a}$ represents all residues other than a and $\bar{b}$ represents all residues other than b. We took the estimated weight $\hat{w}_{ij}(a,b)$ to be the normalized average pairwise amino acid sequence distance (computed using PROTDIST in the PHYLIP phylogeny inference package, ver. 3.572c) of all sequences with amino acid a at position i and amino acid b at position j. A more appropriate weight may be this quantity exponentiated, to reflect that linkage disequilibrium declines exponentially with the number of generations. The normalized mutual information statistic is given by $M^*_{ij} = 2M_{ij}/\{\Sigma_a\hat{p}_i(a)\log[\hat{p}_i(a)] + \hat{p}_j(a)\log[\hat{p}_j(a)]\}$.

To define C_{ij}, consider the two-by-two table formed by considering a, b, $\bar{a}$, $\bar{b}$, with cell probabilities $p_{ij}(a,b)$, $p_{ij}(\bar{a},b)$, $p_{ij}(a,\bar{b})$, $p_{ij}(\bar{a},\bar{b})$. Then $C_{ij} = \hat{w}^C_{ij}(\bar{a},\bar{b})\{[\hat{p}_{ij}(\bar{a},\bar{b})/\hat{p}_i(\bar{a})] \times [\hat{p}_{ij}(\bar{a},\bar{b})/\hat{p}_j(\bar{b})]\}^{1/2} = \hat{w}^C_{ij}(\bar{a},\bar{b})\hat{p}_{ij}(\bar{a},\bar{b})/[\hat{p}_i(\bar{a})\hat{p}_j(\bar{b})^{1/2}$, where the weight $\hat{w}^C_{ij}(\bar{a},b)$ is the normalized average pairwise amino acid distance (defined the same as for M_{ij} and M^*_{ij}) between all sequences with a non-a amino acid at site i and non-b amino acid at site j.

For three sites (i, j, k), the mutual information statistic M_{ijk} is given by

$$M_{ijk} = \sum_{a,b,c} \hat{p}_{ijk}(a,b,c)\log\left[\frac{\hat{p}_{ijk}(a,b,c)\hat{p}_i(a)p_j(b)\hat{p}_k(c)}{\hat{p}_{ij}(a,b)\hat{p}_{ik}(a,c)\hat{p}_{jk}(b,c)}\right]$$

where $\hat{p}_{ijk}(a,b,c)$ is the fraction of the N sequences with residues a, b, and c at sites i, j, and k, respectively. The normalized version of M_{ijk} is given by

$$M^*_{ijk} = 3M_{ijk}\bigg/\left[\sum_a \hat{p}_i(a)\log\hat{p}_i(a) + \hat{p}_j(a)\log\hat{p}_j(a) + \hat{p}_k(a)\log\hat{p}_k(a)\right]$$

ACKNOWLEDGMENTS

This research was supported by Grants 1 RO1 AI054165-01, 1 U01 A146703-01, and RO1 AI 47067 from the National Institutes of Health. The authors thank the referees for providing insightful comments and Paul Moon for helping construct the figures.

REFERENCES

1. Korber BT, Farber RM, Wolpert DH, and Lapedes AS: Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. Proc Natl Acad Sci USA 1993;90:7176–7180.
2. Bickel PJ, Cosman PC, Olshen RA, Spector PC, Rodrigo AG, and Mullins JI: Covariability of V3 loop amino acids. AIDS Res Hum Retroviruses 1996;12:1401–1411.
3. Brown AJ, Korber BT, and Condra JH: Associations between amino acids in the evolution of HIV type 1 protease sequences under indinavir therapy. AIDS Res Hum Retroviruses 1999;12:247–253.
4. Hoffman NG, Schiffer CA, and Swanstrom R: Covariation of amino acid positions in HIV-1 protease. Virology 2003;314:536–548.
5. Wu TD, Schiffer CA, Gonzales MJ, *et al.*: Mutation patterns and structural correlates in human immunodeficiency virus type 1 protease following different protease inhibitor treatments. J Virol 2003;77:4836–4847.
6. Gao F, Robertson DL, Carruthers CD, *et al.*: A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. J Virol 1998;72:5680–5698.
7. Lole KS, Bollinger RC, Paranjape RS, *et al.*: Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J Virol 1999;73:152–160.
8. Ndung'u T, Renjifo B, Novitsky VA, McLane MF, Gaolekwe S, and Essex M: Molecular cloning and biological characterization of full-length HIV-1 subtype C from Botswana. Virology 2000;278: 390–399.
9. Novitsky V, Smith UR, Gilbert P, *et al.*: HIV-1 subtype C molecular phylogeny: Consensus sequence for an AIDS vaccine design? J Virol 2002;76:5435–5451.
10. Rodenburg CM, Li Y, Trask SA, *et al.*: Near full-length clones and reference sequences for subtype C isolates of HIV type 1 from three different continents. AIDS Res Hum Retroviruses 2001;17:161–168.
11. Esparza J and Bhamarapravati N: Accelerating the development and future availability of HIV-1 vaccines: Why, when, where, and how? Lancet 2000;355:2061–2066.
12. Essex M: Human immunodeficiency viruses in the developing world. Adv Virus Res 1999;53:71–88.
13. Osmanov S, Pattou C, Walker N, Schwardlander B, and Esparza J: Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000. J Acquir Immune Defic Syndr Hum Retrovirol 2002;29:184–190.
14. UNAIDS and WHO: Global HIV/AIDS and STD surveillance. Epidemiological fact sheets by country. http:/www.who.int/emc-hiv/fact sheets/, 2000.
15. Graham BS: Clinical trials of HIV vaccines. Ann Rev Med 2002; 53:207–221.
16. Nabel GJ: Challenges and opportunities for development of an AIDS vaccine. Nature 2001;410:1002–1007.
17. Korber BT, Brander C, Haynes B, Koup R, Kuiken C, Moore JP, Walker BD, and Watkins DI (eds.): *HIV Molecular Immunology 2000*. Theoretical Biology and Biophysics Group T-10, Los Alamos National Laboratory, Los Alamos, NM, 2000.
18. Novitsky V, Rybak N, McLane MF, *et al.*: Identification of human immunodeficiency virus type 1 subtype C Gag-, Tat-, Rev-, and Nef-specific Elispot-based CTL responses for AIDS vaccine design. J Virol 2001;75:9210–9228.
19. Shiver JW, Fu TM, Chen L, *et al.*: Replication-incompetent adenoviral vaccine vector elicits effective anti-immunodeficiency-virus immunity. Nature 2002;415:331–335.
20. Korber BT, Foley BT, Kuiken CL, Pillai SK, and Sodroski JG: Numbering positions in HIV relative to HXB2CG. In: *Human Retroviruses and AIDS 1998* (Korber B, Kuiken C, Foley B, Hahn B, McCutchan F, Mellors J, and Sodroski J, eds.). Theoretical Biology and Biophysics Group T-10, Los Alamos National Laboratory, Los Alamos, NM, 1998, pp. III-102–III-111.
21. Bansal A, Sabbaj S, Edwards BH, Ritter D, Perkins C, Tang J, Szinger JJ, Weiss H, Goepfert PA, Korber B, Wilson CM, Kaslow RA, and Mulligan MJ: T cell responses in HIV type 1-infected adolescent minorities share similar epitope specificities with whites despite significant differences in HLA class I alleles. AIDS Res Hum Retroviruses 2003;19:1017–1026.
22. Edwards BH, Bansal A, Sabbaj S, Bakari J, Mulligan MJ, and Goepfert PA: Magnitude of functional CD8+ T-cell responses to the gag protein of human immunodeficiency virus type 1 correlates inversely with viral load in plasma. J Virol 2002;76:2298–2305.
23. Hill, WG: Tests for association of gene frequencies at several loci in random mating diploid populations. Biometrics 1975;31:881–888.
24. Goodman LA and Kruskal WH: *Measures of Association for Cross Classification*. Springer-Verlag, New York, 1979.
25. Weir BS and Cockerham CC: Testing hypotheses about linkage disequilibrium with multiple alleles. Genetics 1978;88:633–642.
26. Benjamini Y and Hochberg Y: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc B 1995;57:289–300.
27. Gilbert PB, Rossini AJ, Shankarappa R; Two-sample tests for comparing intra-individual genetic sequence diversity between populations. Biometrics 2005;61:107–118.
28. Weir BS: *Genetic Data Analysis*. Sinauer, Sunderland, MA, 1990.

29. Park PJ and Kohane IS: Identifying three-way interactions among gene expression and chemosensitivity profiles using ternary mutual information. Technical report, Harvard University, Boston, MA, 2001.
30. Kosakovsky Pond SL and Frost SDW: Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 2005;22:1208–1222.
31. Yang Z, Nielsen R, Goldman N, and Pedersen AM: Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 2000;1:431–439.
32. Felsenstein J: PHYLIP: Phylogeny Inference Package, 3.572c. University of Washington, Seattle, WA, 1996.
33. Kuiken C, Foley B, Hahn B, Marx P, McCutchan F, Mellors J, Mullins J, Wolinsky S, and Korber B (eds.): *HIV Sequence Compendium 2000*. Theoretical Biology and Biophysics Group T-10, Los Alamos National Laboratory, Los Alamos, NM, 2000.
34. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, and Higgins DG: The ClustalX windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997;25:4876–4882.
35. Hall TA: BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 1999;41:95–98.
36. Goulder PJ, Brander C, Annamalai K, *et al*.: Differential narrow focusing of immunodominant human immunodeficiency virus gag-specific cytotoxic T-lymphocyte responses in infected African and caucasoid adults and children. J Virol 2000;74:5679–5690.
37. Piguet V and Trono D: A structure-function analysis of the Nef protein of primate lentiviruses. In: *Human Retroviruses and AIDS 1999* (Kuiken C, Foley B, Hahn B, Marx P, McCutchan F, Mellors J, Mullins J,Wolinsky S, and Korber B, eds.). Theoretical Biology and Biophysics, Group T-10, Los Alamos National Laboratory, Los Alamos, NM, 1999, pp. 448–459.
38. Mashishi T, Loubser S, Hide W, *et al*.: Conserved domains of subtype C Nef from South African HIV type 1-infected individuals include cytotoxic T lymphocyte epitope-rich regions. AIDS Res Hum Retroviruses 2001;17:1681–1687.
39. Peyerl FW, Barouch DH, Yeh WW, *et al*.: Simian-human immunodeficiency virus escape from cytotoxic T-lymphocyte recognition at a structurally constrained epitope. J Virol 2003;77:12572–12578.
40. de Jong J-J, Goudsmit J, Keulen W, Klaver B, Krone W, Tersmette M, and de Ronde A: Human immunodeficiency virus type 1 clones chimeric for the envelope V3 domain differ in syncytium formation and replication capacity. J Virol 1992;66:757–765.
41. Chesebro B, Wehrly K, Nishio J, and Perryman S: Macrophage-tropic human immunodeficiency virus isolates from different patients exhibit unusual V3 envelope sequence homogeneity in comparison with T-cell tropic isolates: Definition of critical amino acids involved in cell tropism. J Virol 1992;66:6547–6554.
42. Fouchier RAM, Groenink M, Kootstra NA, Tersmette M, Huisman HG, Meidema F, and Schuitemaker H: Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. J Virol 1992;66:3183–3187.
43. Westervelt P, Trowbridge DB, Epstein LG, *et al*.: Macrophage tropism determinants of human immunodeficiency virus type 1 *in vivo*. J Virol 1992;66:2577–2582.
44. Rizzuto CD, Wyatt R, Hernandez-Ramos N, Sun Y, Kwong PD, Hendrickson WA, and Sodroski JG: A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. Science 1998;280:1949–1953.
45. Wyatt R, Kwong PD, Desjardins E, Sweet RW, Robinson J, Hendrickson WA, and Sodroski JG: The antigenic structure of the HIV gp120 envelope glycoprotein. Nature 1998;393:705–711.
46. Davis D, Stephens M, Willers C, and Lachmann PJ: Glycosylation governs the binding of antipeptide antibodies to regions of hypervariable amino acid sequence within recombinant gp120 of human immunodeficiency virus type 1. J Gen Virol 1990;71:2889–2898.
47. Wei X, Decker JM, Wang S, *et al*.: Antibody neutralization and escape by HIV-1. Nature 2003;422:307–312.
48. Ping LH, Nelson JA, Hoffman IF, *et al*.: Characterization of V3 sequence heterogeneity in subtype C human immunodeficiency virus type 1 isolates from Malawi: Underrepresentation of X4 variants. J Virol 1999;73:6271–6281.
49. Tscherning C, Alaeus A, Fredriksson R, *et al*.: Differences in chemokine coreceptor usage between genetic subtypes of HIV-1. Virol 1998;241:181–188.
50. Vranken WF, Budesinsky M, Fant F, Boulez K, and Borremans FA: The complete consensus V3 loop peptide of the envelope protein gp120 of HIV-1 shows pronounced helical character in solution. FEBS Lett 1995;374:117–121.

Address reprint requests to:
Peter B. Gilbert
Statistical Center for HIV/AIDS Research and Prevention
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue North
MW-500
Seattle, Washington 98109

E-mail: pgilbert@scharp.org

This article has been cited by:

1. Nobubelo G. Ngandu , Helba Bredell , Clive M. Gray , Carolyn Williamson , Cathal Seoighe , And the HIVNET028 Study Team . 2007. CTL Response to HIV Type 1 Subtype C Is Poorly Predicted by Known Epitope Motifs. *AIDS Research and Human Retroviruses* **23**:8, 1033-1041. [Abstract] [PDF] [PDF Plus]
2. Vlad Novitsky , C. William Wester , Victor DeGruttola , Hermann Bussmann , Simani Gaseitsiwe , Ann Thomas , Sikhulile Moyo , Rosemary Musonda , Erik Van Widenfelt , Richard G. Marlink , M. Essex . 2007. The Reverse Transcriptase 67N 70R 215Y Genotype Is the Predominant TAM Pathway Associated with Virologic Failure among HIV Type 1C-Infected Adults Treated with ZDV/ddI-Containing HAART in Southern Africa. *AIDS Research and Human Retroviruses* **23**:7, 868-878. [Abstract] [PDF] [PDF Plus]