JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 8, Number 2, 2001
Mary Ann Liebert, Inc.
Pp. 123–139

An Efficient Test for Comparing Sequence Diversity between Two Populations

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

ABSTRACT

We address the problem of comparing interindividual genomic sequence diversity between two populations. Although the methods are general, for concreteness we focus on comparing two human immunodeficiency virus (HIV) infected populations. From a viral isolate(s) taken from each individual in a sample of persons from each population, suppose one or multiple measurements are made on the genetic sequence of a coding region of HIV. Given a definition of genetic distance between sequences, the goal is to test if the distribution of interindividual distances differs between populations. If distances between all pairs of sequences within each group are used, then data-dependencies arising from the use of multiple sequences from individuals invalidates the use of a standard two-sample test such as the t-test. Where this problem has been recognized, a typical solution has been to apply a standard test to a reduced dataset comprised of one sequence or a consensus sequence from each patient. Disadvantages of this procedure are that the conclusion of the test depends on the choice of utilized sequences, often an arbitrary decision, and exclusion of replicate sequences from the analysis may needlessly sacrifice statistical power. We present a new test free of these drawbacks, which is based on a statistic that linearly combines all possible standard test statistics calculated from independent sequence subsamples. We describe statistical power advantages of the test and illustrate its use by application to nucleotide sequence distances measured from HIV-1 infected populations in southern Africa (GenBank accession numbers AF110959–AF110981) and North America/Europe. The test makes minimal assumptions, is maximally efficient and objective, and is broadly applicable.

Key words: HIV genetic diversity, HIV vaccine design, statistical hypothesis testing, two-sample test, U-statistic.

[1]Center for Biostatistics in AIDS Research and Department of Biostatistics, Harvard School of Public Health, Boston, 02115.
[2]Harvard AIDS Institute and Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, 02115.

1. INTRODUCTION

SINCE THE EXTENSIVE GENETIC DIVERSITY exhibited by the human immunodeficiency virus (HIV) poses one of the most difficult challenges to the development of AIDS vaccines and treatments, it is important to understand how sequence diversity differs in various HIV infected populations (Goudsmit *et al.*, 1991; McCutchan *et al.*, 1996; Essex, 1998; Fomsgaard, 1999; Zolla-Pazner *et al.*, 1999). Populations of interest to compare include ones defined by demographic characteristics, host genetic factors such as HLA type, HIV phylogenetic clusters, biological marker phenotypes, and clinical disease phenotypes. From two samples of HIV infected persons, suppose the data are measurements on the genetic sequence of a coding region of a virus or viruses isolated from each person and that multiple clones are sequenced from some or all persons. Given that genetic distances are calculated between all pairs of HIV sequences within each population sample, we address how to test for a difference in the distribution of interpatient HIV sequence distances between the two populations. The test can be used generally for any kind of sequence distance definition (e.g., from one based on nucleotides, amino acids, or a phenotype) and for arbitrary organisms.

The dependency in each set of pairwise sequence distances resulting from the tendency for intrapatient sequence diversity to be less than interpatient sequence diversity complicates the development of a valid statistical test. It implies that a standard two-sample test that compares all pairwise sequence distances between groups is not valid, as the clustering of intrapatient sequences will cause the Type I error rate to be too high. In the literature, a common way to deal with this problem has been to conduct a standard two-sample test on a reduced dataset formed from considering only one sequence or consensus sequence from each patient. Also, commonly, analyses of sequence diversity have not included application of a formal statistical test. Among many published analyses that have used these approaches, we cite three.

First, McCutchan *et al.* (1996) compared interpatient envelope sequence diversity between seven Thai patients in early HIV-1 disease and six Thai patients in late HIV-1 disease. The DNA sequencing included at least two clones per isolate from the V2–V5 region. Although a goal of the research was to infer differences in V2–V5 diversity between patients with early and late-stage disease, the analysis did not include use of a statistical test. Second, Murphy *et al.* (1993) compared HIV-1 V3 nucleotide diversity of 19 AIDS patients in Bagui, Central African Republic (CAR), infected with subtype A or E (10 A, 9 E) to V3 nucleotide diversity of 22 patients in Kampala, Uganda, infected with subtype A (Albert *et al.*, 1992), and to V3 nucleotide diversity of 16 patients in Thailand infected with subtype E (McCutchan *et al.*, 1992). On average, three sequences (range, 1–5) were available from CAR patients and four sequences (range, 3–5) were available from Ugandan patients, while one clone was sequenced from each Thai isolate. Using the consensus sequence from each patient, Murphy *et al.* (1993) concluded that subtype A CAR sequences were more diverse between patients than subtype A Ugandan sequences and that subtype E CAR sequences were more diverse between patients than subtype E Thai sequences. Although statistical significance of these differences was claimed, use of a statistical test was not reported. Third, Novitsky *et al.* (1999) compared interpatient HIV-1 nucleotide diversity between 23 subtype C sequences from 8 HIV-1 infected patients in Botswana and 27 subtype B sequences from 23 North American/European HIV-1 infected patients. Two-sample t-tests were used on reduced sequence sets consisting of one sequence per patient.

Not applying a formal statistical test is clearly not ideal, and the approach that uses only one sequence per patient has two disadvantages: 1) tests using different selected sequences from patients may give appreciably different results, with no objective way to identify the "right" test result, and 2) ignoring replicate sequences may needlessly sacrifice statistical power of the test. To solve both deficiencies, we propose a new valid global test that efficiently combines all possible test statistics based on one sequence per patient. The test builds on ideas of Wei and Johnson (1985), who developed a test procedure for combining statistically dependent test statistics with incomplete repeated measurements. They considered repeated measurements of the same characteristic taken under various conditions from each patient. Our problem is more complicated because the observations are distances between patients rather than measurements on individual patients.

Extending Wei and Johnson's (1985) procedure, our global test statistic is defined as the weighted sum of all possible unique valid individual test statistics. The weights are selected so that the test has locally optimal power or for other purposes, such as down-weighting patient isolates with many replicate sequences or down-weighting sequences obtained from less rigorous laboratory procedures. The global test allows for

different numbers of sequences from patients, and it makes no assumptions about the dependency structure of the replicate sequences measured from patients.

A procedure for constructing a global test by linearly combining t-statistics, Wilcoxon rank sum statistics, or general U-statistics is described in Section 2. Statistical details about this procedure are given in the appendix. A simple way to construct confidence intervals about group location differences in interpatient distances by inverting the global test statistic is given in Section 3. In Section 4 we investigate the extent of statistical power gained from using multiple sequences rather than only one sequence from patients. Two real examples are given in Section 5, and applications and extensions of the test procedure are discussed in Section 6.

2. AN EFFICIENT VALID GLOBAL TEST FOR COMBINING TWO-SAMPLE TEST STATISTICS

Let K and L be the maximum number of clones sequenced from patients in groups 1 and 2, respectively. Consider the k'th and k''th sequences from patients in group 1 ($k \le k' \in \{1, \cdots, K\}$) and the l'th and l''th sequences from patients in group 2 ($l \le l' \in \{1, \cdots, L\}$). For each fixed set $\{k, k', l, l'\}$, a two-sample t-statistic $U_{kk'll'}$ can be used to test if interpatient distances calculated between k'th and k''th sequences in group 1 differ in distribution from interpatient distances calculated between l'th and l''th sequences in group 2. More generally, any statistic within the family of two-sample U-statistics can be used, which includes the Wilcoxon rank sum statistic as well as the t-statistic. Our procedure linearly combines the $(K(K+1)/2) * (L(L+1)/2)$ possible unique U-statistics formed from considering all permutations of $\{k, k', l, l'\}$ with $k \le k'$ and $l \le l'$. To do this, define $V = \sum_{k \le k'}^{K} \sum_{l \le l'}^{L} \hat{w}_{kk'll'} U_{kk'll'}$, where the weight $\hat{w}_{kk'll'}$ may be data-dependent. Let Λ be the covariance matrix of the complete vector of U-statistics composed of the elements $\{U_{kk'll'} : k \le k' = 1, \cdots K, l \le l' = 1, \cdots, L\}$. Under the null hypothesis of no group difference in interpatient distance distributions, the global statistic $Z = V(\hat{w}'\hat{\Lambda}\hat{w})^{-\frac{1}{2}}$ is approximately normally distributed, where $\hat{\Lambda}$ is an estimate of Λ and $\hat{w}$ is the vector with elements $\{\hat{w}_{kk'll'} : k \le k' = 1, \cdots K, l \le l' = 1, \cdots, L\}$. The elements of $\hat{\Lambda}$ are given in the appendix by formulas (6.3) and (6.4). Significance levels of the test can be found by comparing Z to a standard normal distribution. Alternatively, when the assumption of normally distributed pairwise distances is in question, significance levels can be calculated by a Monte Carlo permutation procedure. In this procedure, the test statistic is repeatedly calculated using sequence datasets formed by randomly permuting group membership indices of individuals, and the p-value is computed as the fraction of these statistics greater than the observed test statistic. Fewer than eight individuals in either group may place the normality assumption in doubt.

The simplest and most easily interpreted global combined statistic weights all U-statistics equally, i.e., with all $\hat{w}_{kk'll'} = 1$. Alternatively, the weights $\hat{w}_{kk'll'}$ may be chosen to optimize the statistical power of the test for detecting a Pitman shift alternative hypothesis (Pitman, 1939; Lehmann, 1975). A major advantage of the global test is that it does not assume any model of dependence among the distances computed between various replicate sequences from patients.

3. CONFIDENCE INTERVAL

Suppose the true location difference in interpatient distances between groups, Δ_0, is the same for all $k \le k' = 1, \cdots, K, l \le l' = 1, \cdots, L$. This assumes exchangeability of the marginal distributions of interpatient pairwise distances for the various replicate sequence positions, for each group, which is usually a tenable hypothesis. Consider the statistic Z as a function of the location difference Δ between the two groups,

$$\begin{aligned} Z(\Delta) &= V(\Delta)(\hat{w}'(\Delta)\hat{\Lambda}(\Delta)\hat{w}(\Delta))^{-\frac{1}{2}} \\ &= \left[\sum_{k \le k'}^{K} \sum_{l \le l'}^{L} \hat{w}_{kk'll'}(\Delta) U_{kk'll'}(\Delta)\right] (\hat{w}'(\Delta)\hat{\Lambda}(\Delta)\hat{w}(\Delta))^{-\frac{1}{2}}, \end{aligned}$$

with $U_{kk'll'}(\Delta)$ equal to $U_{kk'll'}$ in (6.1) and $\hat{\Lambda}(\Delta)$ equal to $\hat{\Lambda}$ in (6.3)–(6.4) with $\theta_{kk'll'} = \theta_{rr'ss'} = \Delta$ (see the appendix). The estimated weights $\hat{w}(\Delta)$ may depend on Δ.

To construct a confidence interval, for a range of fixed Δ's the statistic $Z(\Delta)$ is calculated, which is a monotone function of Δ. A two-sided $1-\alpha$ level confidence interval about Δ_0 is then given by

$$(\sup\{\Delta \in [-\infty, \infty] : Z(\Delta) \leq -z_{1-\alpha/2}\}, \inf\{\Delta \in [-\infty, \infty] : Z(\Delta) \geq z_{1-\alpha/2}\}), \tag{3.1}$$

where z_α satisfies $\alpha = \Pr(Z \leq z_\alpha)$. A one-sided $1-\alpha$ level confidence interval for Δ_0 has confidence limit given by the appropriate element of (3.1) with $z_{1-\alpha/2}$ replaced by $z_{1-\alpha}$. This procedure gives an asymptotically correct confidence interval since $Z(\Delta)$ converges to a standard normal random variable for any fixed Δ. Plotting $Z(\Delta)$ versus Δ provides a useful visual picture of a confidence interval about Δ_0, as illustrated in the first example of Section 5.

4. STATISTICAL POWER ADVANTAGES OF THE GLOBAL TEST

We investigate potential efficiency advantages gained from using all sequences compared to one sequence in the comparison of interpatient sequence diversity. First we briefly consider the problem studied by Wei and Johnson (1985), as it provides a clear, simple characterization of conditions under which a combination global test has power advantages. We suppose K repeated measurements are made on some random variable on each of m patients in group 1 and on each of n patients in group 2. Consider the K U-statistics $U_1, \cdots, U_K$ calculated for each of the K measurement times (for a definition of these U-statistics, see page 360 of Wei and Johnson, 1985). Let $V = \sum_k w_k U_k$ be Wei and Johnson's (1985) test statistic with equal weights $w_k = 1, k = 1, \cdots, K$. For simplicity, suppose that $\mathrm{Var}(U_k) = \sigma_1$ for each k and $\mathrm{Cov}(U_k, U_l) = \sigma_{12}$ for each $k \neq l$. Then, an easy calculation shows that the asymptotic relative efficiency (*ARE*) of the global test based on V compared to a test based on a single U-statistic U is

$$ARE(V, U) = \frac{1}{K} + \frac{(K-1)}{K}\frac{\sigma_{12}}{\sigma_1}. \tag{4.1}$$

Thus, the combined statistic V offers no efficiency gains if the U_k's are perfectly correlated ($\mathrm{corr}(U_k, U_l) = \sigma_{12}/\sigma_1 = 1$) and has asymptotic relative efficiency increasing with the degree of independence between the U_k's to a maximum of K-fold superior efficiency for independent U_k's ($\sigma_{12} = 0$). The *ARE* translates into sample size requirements by implying that approximately $ARE * (m+n)$ patients are needed to achieve the same power to reject the null hypothesis using V compared to using U with $m+n$ patients (Gail, 1985).

Now consider our pairwise distance problem, with $K = L$. Since the asymptotic relative efficiency formula for the statistic $V = \sum_{k \leq k'}^{K} \sum_{l \leq l'}^{K} \hat{w}_{kk'll'} U_{kk'll'}$ is complicated, we consider the equivalent test statistic $V^* = \sum_{k \neq k'}^{K} \sum_{l \neq l'}^{K} \hat{w}_{kk'll'} U_{kk'll'}$. We suppose equal weights $\hat{w}_{kk'll'} = 1, k \neq k', l \neq l' = 1, \cdots, K$, and for simplicity assume $\mathrm{Var}(U_{kk'll'}) = \sigma_{1111}$ for all k, k', l, l'. Consider the permutations of $k, k', l, l', r, r', s, s'$ such that $\{k, k'\}$ and $\{r, r'\}$ compose the same set and $\{l, l'\}$ and $\{s, s'\}$ compose the same set. Suppose $\mathrm{Cov}(U_{kk'll'}, U_{rr'ss'}) = \sigma_{11111112}$ for indices with exactly one departure from equality of the two sets above, e.g., $k = r, k' = r', l = s'$, but l' different from s and s'. Also suppose $\mathrm{Cov}(U_{kk'll'}, Y_{rr'ss'}) = \sigma_{11111122}$ for indices with exactly two departures from equality of the two sets, and $\mathrm{Cov}(U_{kk'll'}, U_{rr'ss'}) = \sigma_{11111222}\,(\sigma_{11112222})$ for indices with exactly three (four) departures from equality of the two sets. Direct calculation then shows that the *ARE* of the global test based on V^* relative to a test based on a single U-statistic U is

$$ARE(V^*, U) = \frac{1}{K^4}\left(1 + 4(K-1)\frac{\sigma_{11111112}}{\sigma_{1111}} + 6(K-1)^2\frac{\sigma_{11111122}}{\sigma_{1111}} + 4(k-1)^3\frac{\sigma_{11111222}}{\sigma_{1111}} + (K-1)^4\frac{\sigma_{11112222}}{\sigma_{1111}}\right). \tag{4.2}$$

This expression is one if all of the covariance/variance ratios (correlations) are one and is K^{-4} if all of the covariance terms are zero. Thus, the asymptotic relative efficiency ranges from 1 to K^{-4} depending on the strength of dependency between the U-statistics.

This calculation compares the efficiency of the statistic V^* based on balanced, complete data (K sequences from all patients) to the statistic based on one sequence per patient. To accommodate imbalances in the numbers of sequences from patients in the calculation, K can be replaced by the square-root of a weighted average of numbers of between-sequence comparisons within a patient, $\overline{K} = \{\sum_{i=1}^{K} f_i * i^2\}^{1/2}$, where f_i is the fraction of the $m+n$ patients with exactly i sequences. Formula (4.2) holds approximately with $\overline{K}$ in place of K.

For complete sequence datasets with $K = L = 2$, 3, or 5 sequences per patient, and for an incomplete sequence dataset with $K = L = 3$ and an average of 1.25 sequences per patient (with $f_1 = 0.8$, $f_2 = 0.15$, $f_3 = 0.05$), Fig. 1 illustrates the *ARE* of the global test relative to a test based on the single statistic U_{1111} as a function of the covariance/variance ratios. We set $\sigma_{11111112} = \alpha\sigma_{11111122} = \alpha^2\sigma_{11112222} = \alpha^3\sigma_{11112222}$ and vary α and the ratio $\sigma_{11111112}/\sigma_{1111}$ between zero and one. When there are three or more sequences per patient, the efficiency gain is always substantial (at least 50%) unless all of the correlations are greater than 0.90. The efficiency gain is still large when two sequences per patient are used, with a minimum 30% gain when $\sigma_{11111112}/\sigma_{1111} < 0.90$ and $\alpha < 0.8$. With an average of 1.25 sequences per patient, the efficiency gain is considerably less, but is still appreciable (greater than 20%) as long as $\sigma_{11111112}/\sigma_{1111} < 0.60$. Notice that the influence of α on the *ARE* increases with the number of replicates per patient. For a specific example of how the *ARE* varies with the number of sequences per patient, suppose $\sigma_{11111112}/\sigma_{1111} = 0.80$ and $\alpha = 0.6$. The *ARE* is 0.78, 0.60, 0.48, and 0.40 when the number of sequences per person is 1.25, 2, 3, and 5, respectively.

For a particular sequence dataset, the observed efficiency gain of the equal-weight global statistic V relative to a particular U-statistic U can be estimated by

$$\widehat{ARE}(V, U) = \left\{\frac{U}{\overline{U}}\right\}^2 \frac{M^{-2} \sum_{k\le k'}^{K} \sum_{l\le l'}^{L} \sum_{r\le r'}^{K} \sum_{s\le s'}^{L} \hat{\sigma}_{kk'll'rr'ss'}}{M^{-1} \sum_{k\le k'}^{K} \sum_{l\le l'}^{L} \hat{\sigma}_{kk'll'}}, \tag{4.3}$$

where the covariance estimates $\hat{\sigma}_{kk'll'rr'ss'}$ are given by (6.3)–(6.4), $M = (K(K+1)/2) * (L(L+1)/2)$, and $\overline{U} = M^{-1} \sum_{k\le k'}^{K} \sum_{l\le l'}^{L} U_{kk'll'}$ is the average of the U-statistics. The estimated *ARE* informs about how much efficiency was gained by using multiple sequences from patients for the actual test conducted.

These kinds of analyses of efficiency trade-offs can assist in designing a sampling plan for a sequence diversity study so that the number of patients and the number of clones sequenced per patient can be chosen to best address the objectives of the study.

5. EXAMPLES

The global HIV-1 pandemic is most severe in southern African countries, including Zimbabwe, Zambia, Namibia, South Africa, and Botswana (UNAIDS and WHO, 1998), where subtype C predominates (Becker *et al.*, 1995; Janssens *et al.*, 1997; Bredell *et al.*, 1998; Essex, 1998). This region represents the center of the epidemic in terms of geographic position and HIV-1 prevalence, and therefore it is urgent to understand the diversity of local strains (Becker *et al.*, 1995; Williamson *et al.*, 1995; Janssens *et al.*, 1997; Salminen *et al.*, 1997; van Harmelen *et al.*, 1997, 1998; Bredell *et al.*, 1998). Recently, Novitsky *et al.* (1999) analyzed 23 nearly full-length genome HIV-1 clones derived from eight seropositive patients in Gaborone, Botswana.

Three clones from each of five patients, five clones from one patient, and two and one clones from one patient each were sequenced. A main objective of the study was to compare the interpatient diversity (as measured by percentage of nucleotide divergence) between these Botswana viruses and 27 full-length subtype B viruses available from the Los Alamos National Laboratory database. These B sequences were measured from 23 patients, so the majority had only one sequence, with two clones from two patients and three clones from one.

Novitsky *et al.* (1999) compared interpatient distances between virus subtype groups with two-sample t-tests using one sequence per patient, selected as the first sequenced clone. Separate tests were done

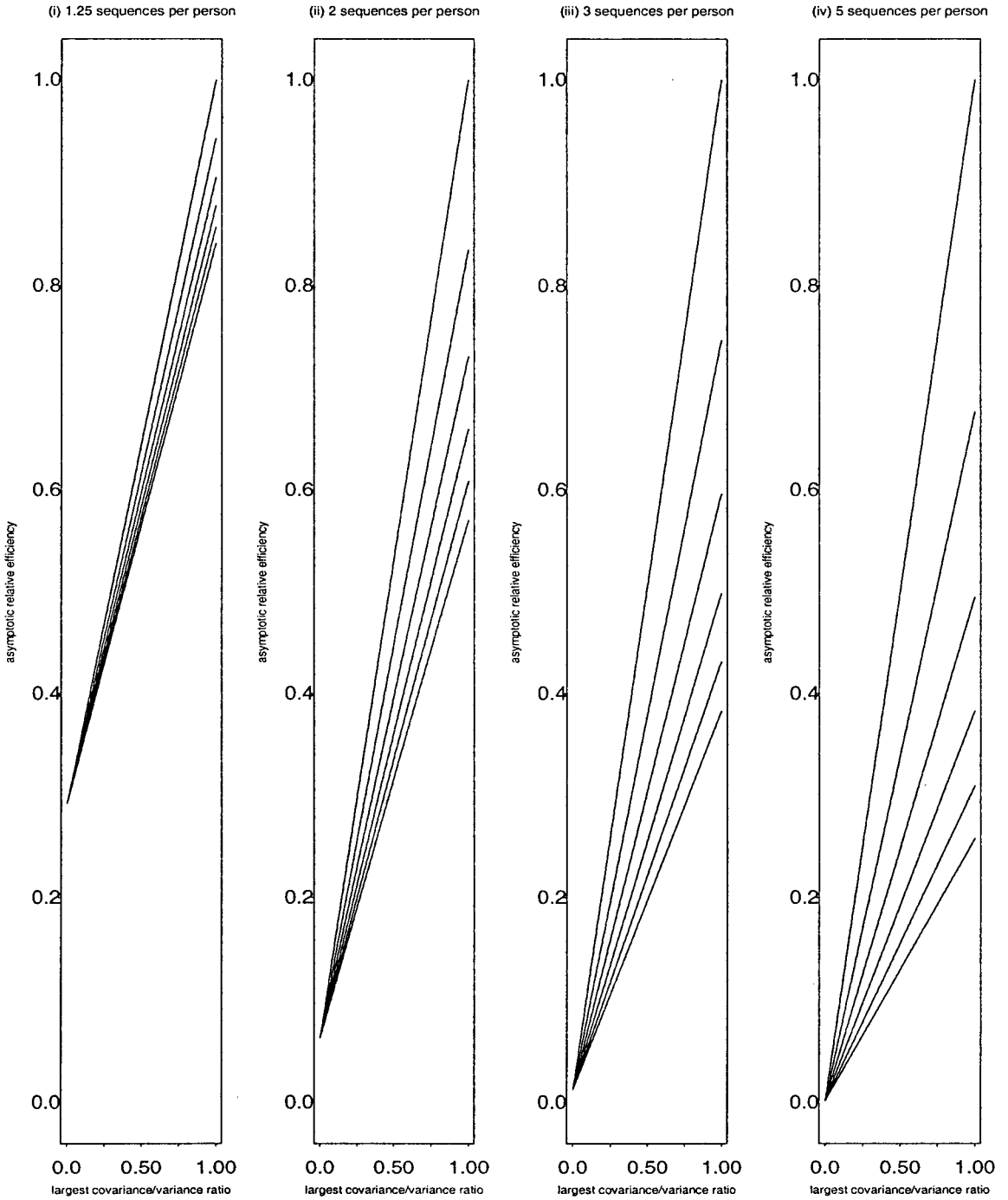

FIG. 1. Asymptotic relative efficiency of the equal-weight global test compared to a single U-statistic test as a function of the covariance ratio $\sigma_{11111112}/\sigma_{1111}$ when on average (**i**) 1.25 sequences, (**ii**) 2 sequences, (**iii**) 3 sequences, or (**iv**) 5 sequences are available from each patient. The rate α of covariance decrease equals 0, 0.2, 0.4, 0.6, 0.8, and 1.0 as the lines move from bottom to top.

for nucleotide distances based on 12 genes or regions across the HIV-1 genome. Note that 15 of the 23 Botswana sequences were not used in the analysis, and therefore it is possible that appreciably different results would be obtained if different sequenced clones had been used. We apply the global test to this problem, using the statistic that linearly combines all unique individual t-statistics with equal weight.

We briefly describe the methodology we used to obtain the 23 Botswana HIV-1 sequences. Genomic DNA was obtained directly from patients' peripheral blood mononuclear cells. Polymerase chain reactions were performed with the Expand Long Template PCR System (Boehringer Mannheim, Indianapolis, Ind.), using a previously described set of primers LA in a final concentration 300 nM (Fujii *et al.*, 1997). Cycling included denaturation (94°C, 2 min), 10 additional cycles of denaturation (94°C, 10 sec each), annealing (65°C, 30 sec), and elongation (68°C, 8 min) followed by 20 cycles for the first round (or 17–20 cycles for the second round) of denaturation (94°C, 10 sec), annealing (65°C, 30 sec), and extension (68°C, 8 min, with additional 20 sec per cycle). All PCRs were run on PTC-200 (MJ Research, Inc., Watertown, MA). Both round amplicons were patient to gel-purification by the gel extraction kit QIAEX II (QIAGEN Inc., Valencia, CA). Cloning was performed using pCR2.1 TOPO (Invitrogen, Carlsbad, CA) and cells JM109 (Promega Corporation, Madison, WI). Plasmid DNA was purified by QIAGEN Maxi Plasmid kit (QIAGEN Inc., Valencia, CA). A strategy of overlapping primers on both strands was used for the full-length genome sequencing. The reactions were run on an MJR Thermocycler utilizing Taq Dye Terminator FS Mix (ABI, Foster City, CA) and recommended cycling parameters (25 cycles of 96°C, 10 sec; 50°C, 5 sec; 60°C, 4 min). The sequencing product was then purified utilizing ethanol precipitation and run on an automated 373XL DNA sequencer (ABI, Foster City, CA).

For the entire set of subtype B and C sequences, alignments for the various genomic regions were derived from the complete genome alignment that was built using the hidden Markov model (Eddy, 1996; Korber *et al.*, 1997). All alignments were globally gap-stripped. Distances were computed by the DNADIST program from the PHYLIP package, v.3.572, under a Kimura 2-parameter model. The transition/transversion parameter was set to 3.0 for the *gag* gene, 1.5 for the envelope gene, 1.42 for the V1–V2 and V3 fragments, and 2.0 for all other genes and loci.

Since different numbers of clones were sequenced from patients, the number of patients from whom data is used in the calculation of each U-statistic, and therefore the size of the variance of each U-statistic, depends on the order in which the multiple sequences from a patient are arranged. For example, among the Botswana sequences, data from all eight patients are used for pairwise comparisons of first-sequenced clones, data from seven patients are used for comparisons of second-sequenced clones, and data from six patients are used for comparisons of third-sequenced clones. In general, the number of individuals from whom data are used for comparisons between the kth and k'th-sequenced clones equals the number with $\max\{k, k'\}$ clones, and is zero if this tally of individuals is one.

Since more than three sequences are available from only one Botswana patient (with five sequences), we simplify the example by only using the first sequences from this patient. Thus $K = L = 3$, and the number of individual test statistics linearly combined by the global statistic is $(K(K+1)/2) * (L(L+1)/2) = 36$. Thus, for this example, 21 of the 23 Botswana C sequences and all 27 of the B sequences are used.

The results of the global test that weights the 36 t-statistics equally are presented in Table 1. The results of the t-test calculated using only the first-sequenced clone from each patient are included for comparative purposes. The global test based on an equally weighted sum of Wilcoxon rank sum statistics gave highly similar answers (results not shown).

For all 12 genomic regions, the average interpatient genetic distance is larger for subtype C viruses than for subtype B viruses. The mean group difference Δ_0 in interpatient distance is statistically significantly different from zero by the global combined test and by the individual t-test in all cases, with two-sided p-values ranging between 0.0000015 and 0.026. The two procedures produce estimates of Δ_0, test statistics, and p-values that tend to have similar values. Confidence intervals about Δ_0 derived from the equal-weight global statistic tend to be only slightly narrower than confidence intervals derived from the individual t-statistic, as can be seen for the analyses of envelope, *gag*, and LTR sequences (Fig. 2). This suggests that substantial gains in efficiency have not been realized for this example. To evaluate the reasons for this, we calculated the observed *ARE* $\widehat{ARE}(V, U_{1111})$ using formula (4.3) for several of the genes. For envelope, *gag*, and LTR, for example, it equals 0.92, 0.89, and 0.65, respectively. The efficiency gains are modest because U_{1111} compares distances between first-sequenced clones, and very few (only 3) of the 23 persons in the subtype B group had more than one clone sequenced, so that U_{1111} is based on a large portion of the

TABLE 1. COMPARISON OF INTERPATIENT NUCLEOTIDE SEQUENCE DISTANCE DISTRIBUTIONS, BOTSWANA HIV-1 SUBTYPE C VIRUSES VERSUS LOS ALAMOS DATABASE HIV-1 SUBTYPE B VIRUSES[a]

	t-test using first sequences only				*Global t-test using all sequences*			
Genomic region	*Mean difference* $\Delta(\%)$	*Confidence interval for* Δ_0^b	*t-statistic*	*p-value*	*Mean difference* $\Delta(\%)$	*Confidence interval for* Δ_0^b	*Z-statistic*	*p-value*
gag	3.0	(2.6,3.5)	4.81	$1.5e^{-6}$	3.3	(2.8,3.6)	4.65	$3.4e^{-6}$
pol	2.1	(1.7,2.6)	4.56	$5.0e^{-6}$	2.4	(1.9,2.9)	4.38	$1.2e^{-5}$
vif	1.2	(0.2,2.2)	2.23	$2.6e^{-2}$	1.7	(0.1,2.6)	2.64	$8.4e^{-3}$
vpr	4.1	(3.0,5.3)	4.21	$2.6e^{-5}$	4.9	(3.5,5.8)	4.09	$4.3e^{-6}$
tat	3.1	(2.3,4.0)	4.25	$2.1e^{-5}$	3.0	(2.1,3.9)	4.04	$5.4e^{-6}$
rev	2.4	(1.4,3.3)	3.74	$1.8e^{-4}$	2.6	(1.7,3.4)	3.77	$1.6e^{-4}$
vpu	4.5	(3.6,5.4)	4.56	$5.0e^{-6}$	4.7	(3.8,5.4)	4.52	$6.3e^{-6}$
env	3.1	(2.6,3.5)	4.78	$1.7e^{-6}$	3.4	(2.8,3.7)	4.52	$6.1e^{-6}$
V1-V2	7.2	(5.0,9.3)	4.12	$3.8e^{-5}$	7.5	(5.3,9.2)	4.07	$4.8e^{-5}$
V3	2.7	(1.8,3.5)	4.06	$5.0e^{-5}$	3.1	(2.3,3.8)	4.13	$3.7e^{-5}$
nef	2.3	(1.5,3.2)	3.72	$2.0e^{-4}$	2.6	(1.8,3.5)	3.99	$6.7e^{-5}$
3′LTR	2.4	(1.3,3.2)	3.55	$3.8e^{-4}$	2.4	(1.1,2.9)	3.42	$6.3e^{-4}$

[a]The subtype B sequences used in the analysis are: AUMBCC54, C18MBC, DH123, 89.6, RF, WEAU, HAN, MN, BCSG3, OYI, CAM1, NY5, LAI, pNL43, HXB2, JRFL, AUMBC925, AUMBC200, YU2, YU10, ACH320A, ACH320B, SF2, AD8, D31, MANC, and WR27. Sequences AUMBCC54, C18MBC, and NY5 were excluded from the *nef* and 3′LTR analyses because of deletions or the absence of sequences for these regions.

[b]Ninety-five percent confidence intervals are calculated by inverting the t-statistic or the equal-weight global combined statistic, as described in Section 3.

total sequence dataset. When the observed *ARE* is calculated for V relative to the U-statistic that compares distances between first- and second-sequenced clones, or between second- and second-sequenced clones, or between second- and third-sequenced clones, the efficiency advantage is substantial. For the envelope, *gag*, and LTR regions, respectively, we calculate $\widehat{ARE}(V, U_{1212}) = 0.48, 0.46, 0.38$, $\widehat{ARE}(V, U_{2222}) = 0.61, 0.58, 0.23$, and $\widehat{ARE}(V, U_{2323}) = 0.11, 0.10, 0.13$. Thus, when another U-statistic besides U_{1111} is used for the one-sequence-per-patient analysis, the global test is typically 2 to 10 times more efficient.

Figure 2 also displays a weighted global statistic as a function of Δ, where the weight $\hat{w}_{kk'll'}$ for $U_{kk'll'}$ is defined by the fraction of pairwise comparisons between the k'th and k''th sequence replicates among the maximum possible $m(m-1)/2$ in group 1 multiplied by this fraction for the l'th and l''th sequence replicates in group 2, divided by the estimated variance $\hat{\sigma}_{kk'll'}$ of $U_{kk'll'}$, that is, normalized by the average variance estimate of the set of U-statistics that use the same number of pairwise comparisons. The weighted global statistic consistently produces larger lower and upper confidence limits than the other two test statistics, illustrating that the choice of weights for the global statistic can appreciably affect the results.

Graphical comparison of the frequency distributions of interpatient pairwise distances across groups is useful for evaluating the nature of group differences. For example, in addition to showing a greater level of envelope sequence pairwise distance in HIV-1 C viruses than in HIV-1 B viruses, Fig. 3 shows less variability in HIV-1 C envelope pairwise distances.

The illustrative power of this example is limited by the unavailability of multiple sequences from most patients in group 2. Therefore, we briefly consider another subtype B comparison population from which there are many sequences per patient. Wolinsky *et al.* (1996) studied longitudinally a large number of partial envelope sequences obtained from six participants in the Chicago MACS cohort. We consider a database comprised of one sequence from each patient for each sampled time point, with 5 time points for 2 patients, 6 time points for 3 patients, and 8 time points for 1 patient, spanning an average 43.2 months of follow-up after seroconversion. Thus, an average of 6 (range, 5–8) sequences are available from each patient.

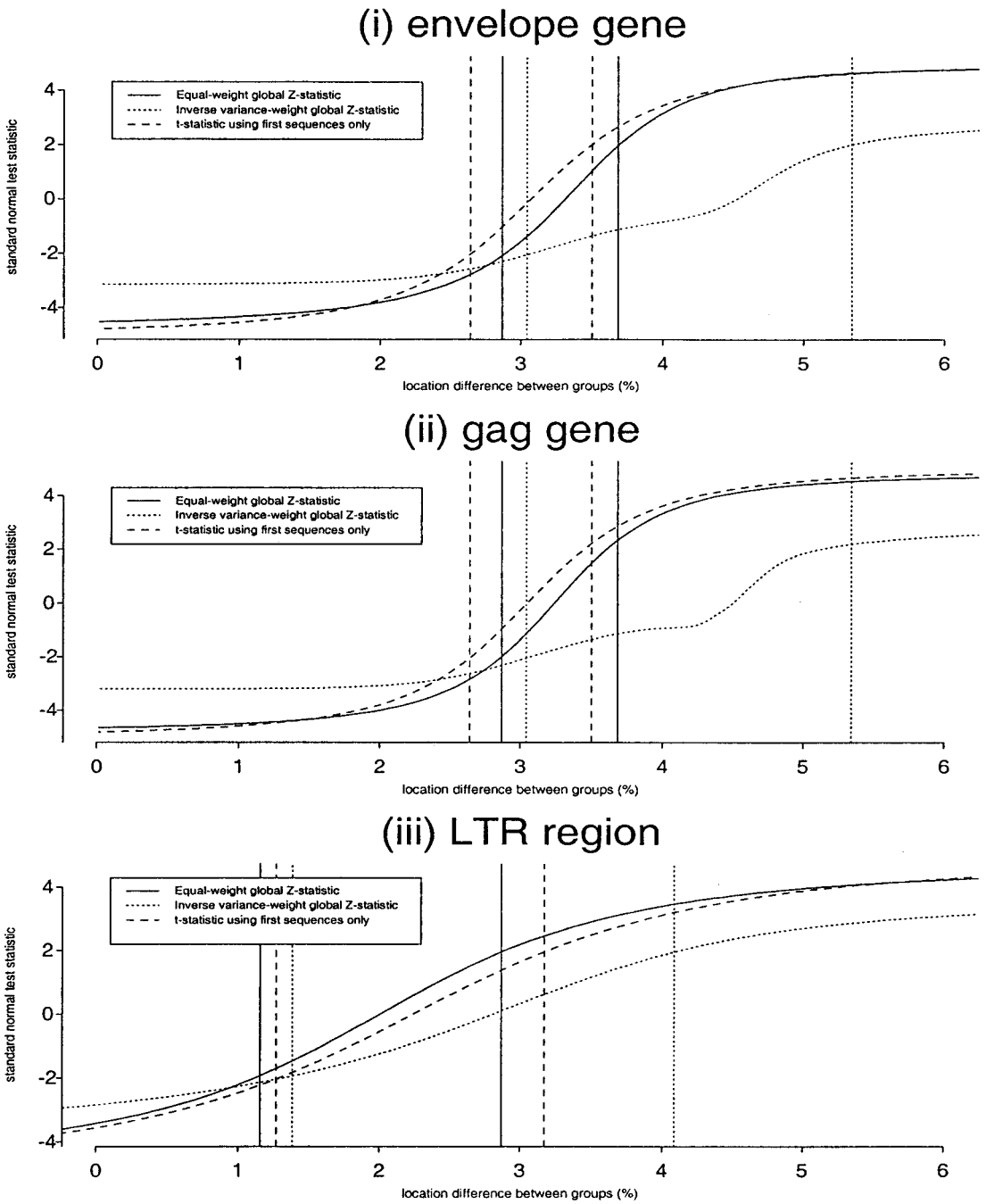

FIG. 2. Test statistics for comparing distributions of Botswana HIV-1 C and Los Alamos database HIV-1 B inter-patient nucleotide sequence distances, plotted as a function of the mean location difference Δ. This gives confidence intervals for Δ_0 based on the global test that weights all sequence comparisons equally, the global test that weights sequence comparisons by the proportions of pairwise comparisons in each group with the sequence number in question divided by the normalized estimated variance of the t-statistic, and the t-test performed on first sequences only.

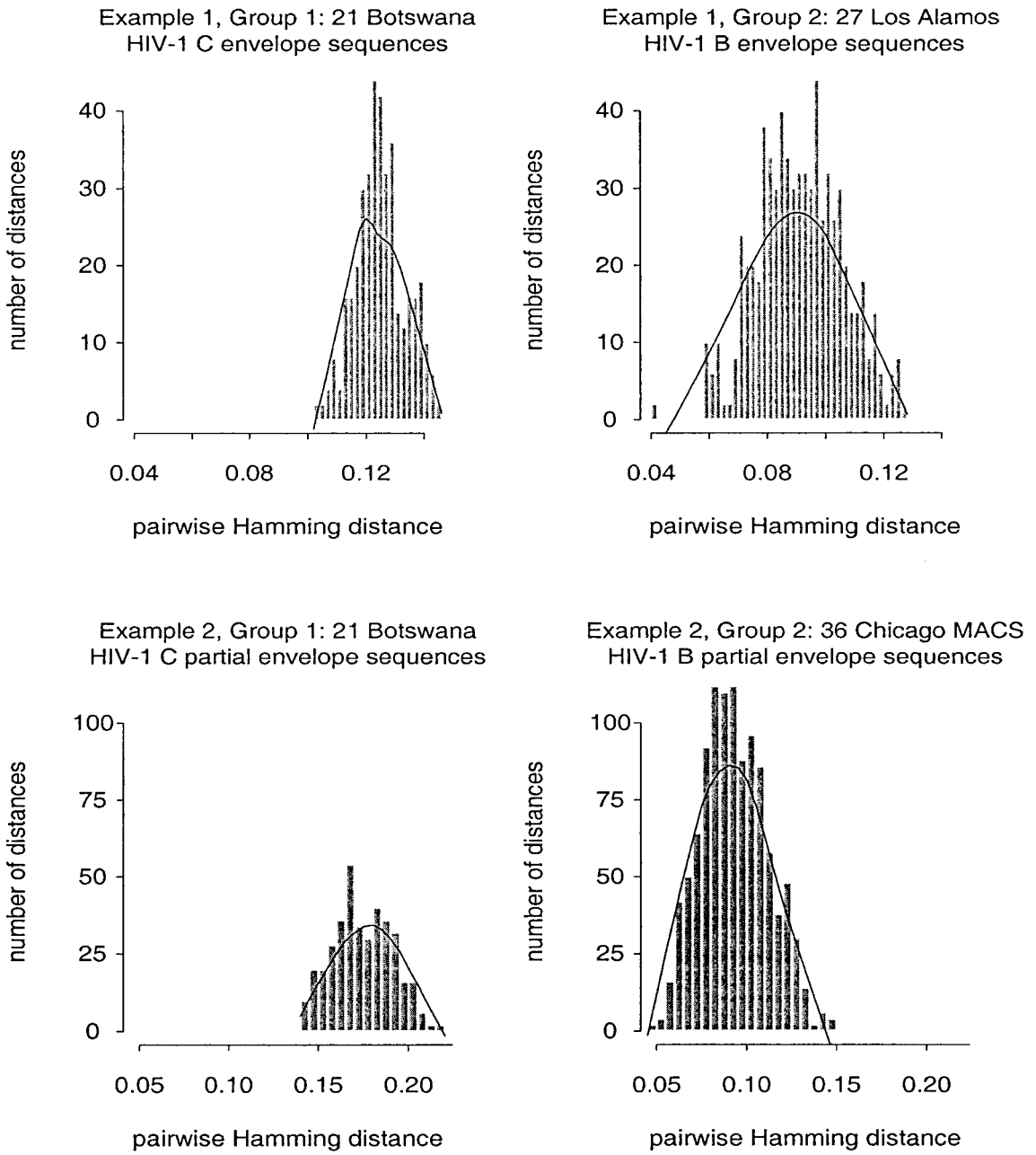

FIG. 3. Frequency distributions of pairwise nucleotide Hamming distances for Examples 1 and 2. The smooth line through each histogram is calculated by the robust, local smoother function lowess in Splus with two-thirds of the data smoothed at each histogram break point.

For the comparison of interpatient partial envelope sequence diversity between these sequences and the 23 Botswana sequences, $K = 3$ and $L = 8$, and there are 198 two-sample tests that can be performed based on unique sequence orderings. Alignments and nucleotide distance calculations were done as described for the first example. Figure 3 depicts the interpatient pairwise distance distributions, showing greater interpatient diversity in Botswana viruses. As depicted in Fig. 4, the individual t-statistics range from 1.63 to 3.08. Given the small sample size of individuals, the Monte Carlo permutation procedure (with 200 randomly sampled datasets) is used to calculate the corresponding significance levels. The two-sided p-values range from 0.001 to 0.102, illustrating that the inference about population differences in sequence

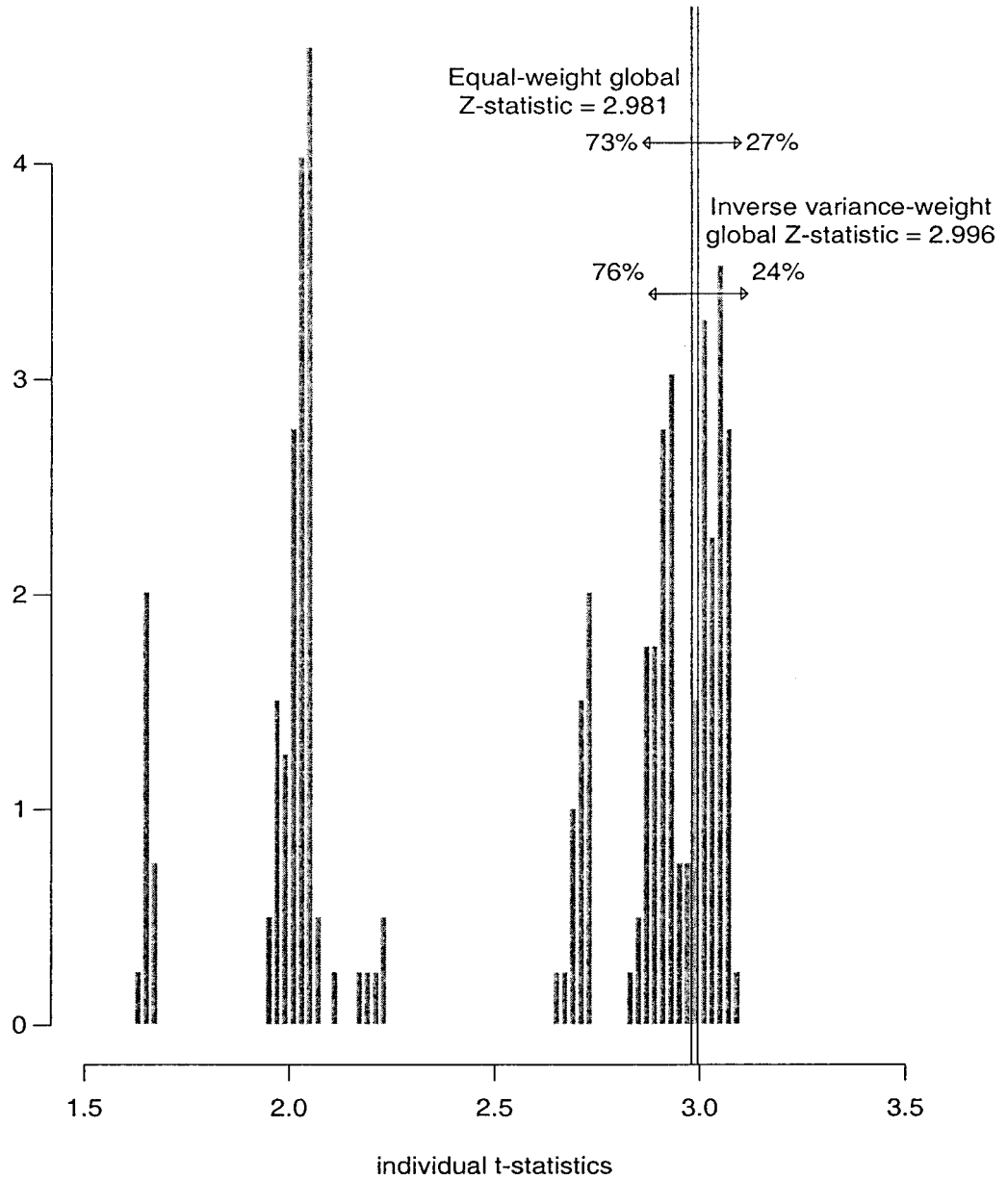

FIG. 4. Frequency distribution of all possible individual t-statistics (n = 198) calculated from unique replicate sequence positions. The comparison is between samples of interpatient Hamming distances calculated from 21 Botswana HIV-1 C partial envelope sequences (Novitsky *et al.*, 1999) and from 36 Chicago MACS HIV-1 B partial envelope sequences (Wolinsky *et al.*, 1996). The equal-weight and inverse-variance-weight (weighting system defined fully in the legend of Fig. 2) global Z-statistics are greater than 73% and 76% of the individual t-statistics, respectively.

diversity can vary considerably by the particular sequences selected from patients. The clustering of t-statistics into four groups can be explained by variations in numbers of sequences per patient, which causes some t-statistics to utilize more sequence distances. The equal-weight global combined statistics is 2.981, with permutation-based p-value 0.0027. The observed *ARE* of the global test is 0.739 compared to an average individual t-test, indicating its $([1/0.739] - 1) * 100\% = 35.3\%$ efficiency advantage relative to a t-test conducted on a randomly selected sequence subdataset. The high correlation of the individual t-statistics, with average correlation $\text{corr}(U_{kk'll'}, U_{rr'ss'}) = 0.738$, explains why the efficiency gain for this dataset with 3–6 sequences from most patients is not larger.

6. DISCUSSION

The original Wei–Johnson test, upon which our test is based, has important applications in its own right to the analysis of HIV sequence diversity. This test uses a linear combination of U-statistics to compare two populations by a characteristic measured repeatedly from each individual. For example, if multiple clones per individual are sequenced, and the distance from each sequence to a reference sequence is calculated, then the Wei–Johnson test can be used to assess if one population of sequences is more divergent from the reference sequence than the other population of sequences. To illustrate how the test could have been usefully applied in a published paper, consider Albert *et al.* (1992). They measured 85 V3 sequences from 22 HIV-1 subtype A infected patients in Uganda and compared V3 nucleotide divergence from the United States/European consensus sequence between two patient populations defined by phylogenetic clusters. Whereas Albert *et al.* (1992) used the consensus sequence from each patient to make the comparison, the Wei-Johnson test would use all the individual sequences. Among many other applications of the Wei–Johnson test to sequence diversity research questions, we mention two. First, suppose multiple sequences are measured from isolates sampled from vaccinated and unvaccinated trial participants infected while enrolled in a preventive HIV vaccine trial. Then the test would be ideal for assessing if infecting viruses in vaccinated patients tend to be more divergent from the prototype virus represented in the vaccine construct than infecting viruses in unvaccinated patients, where "divergence" is measured by a genetic distance or a phenotypic distance (e.g., a "neutralization serotype" distance or a "CTL epitope" distance). Thus, the Wei–Johnson test can be used to test "sieve analysis" hypotheses of differential vaccine protection (Gilbert *et al.*, 1998, 1999) when multiple clones from infected vaccine trial participants are sequenced. Second, given two vaccine strain candidates and sequence data from persons in the population where a vaccine trial is planned, the test could be used to determine if one vaccine strain sequence is significantly "better matched" to the local population. Thus, the test can assist in the selection of viral antigens to include in a vaccine to test locally.

The global test we have developed appropriately accounts for data-dependencies arising from multiplicities of sequences from individuals. However, even when only one sequence per individual is used, the distances within each sample are not completely independent of one another, as each sequence is used in multiple pairwise distance calculations. The global statistic can be modified to combine individual test statistics that compare samples of distances whose members are completely independent. To do this, notice that for each individual in group 1, the distances between a sequence from that individual and a sequence from the other $m - 1$ individuals is an independent, identically distributed sample. Thus, for fixed replicate sequence indicators k and k', there are m independent identically distributed samples. Using the same procedure for group 2 distances, this suggests defining a global test-statistic V' by $V' = \sum_{k \le k', l \le l'} \sum_{i=1}^{m} \sum_{j=1}^{n} \hat{w}_{kk'll'ij} U_{kk'll'ij}$, where now each U-statistic $U_{kk'll'ij}$ is specific to index individuals i and j as well as to k, k', l, l' (see (6.7) in the appendix for a mathematical representation of $U_{kk'll'ij}$). The asymptotically standard normal statistic $Z' = V'/\{\widehat{\mathrm{Var}}(V')\}^{1/2}$ can then be used to test the null hypothesis. Confidence intervals, tests based on t and Wilcoxon statistics, and the characterization of optimal weight functions can be derived in a similar manner as shown for the global statistic V.

The principle of constructing global test statistics to use all available sequences and to appropriately account for data-dependencies can be applied to many other important sequence diversity questions. For example, suppose multiple sequences per person are available from a population and the goal is to assess if the mean interpatient sequence distance is different than some fixed value. In a manner directly analogous to the approach presented here, a valid global test statistic could be constructed as a linear combination of one-sample U-statistics (which include one-sample t and Wilcoxon statistics), and the global statistic could be inverted to generate a confidence interval.

A second example is signature analysis. For instance, suppose viruses are sequenced from two infected populations and the goal is to identify signature sites or motifs in one of the sequence populations. Examples of comparative populations of interest are vaccinated versus unvaccinated infected trial participants (e.g., Berman *et al.*, 1997; Connor *et al.*, 1998; Graham *et al.*, 1998), HIV infected mothers who transmit versus those who do not transmit the virus to their infant (e.g., Scarlatti *et al.*, 1993; Ahmad *et al.*, 1995; Contag *et al.*, 1997), and antiretroviral-treated patients whose virus is resistant versus nonresistant to a particular drug regimen. When one sequence per patient is used, popular signature programs such as VESPA (Korber and Meyers, 1992; Korber *et al.*, 1993) and MotifScan (Connor *et al.*, 1998) use Fisher's exact tests to guide in determining significance levels. If more than one sequence is available from some patients, then

it is natural to use a global statistic that linearly combines Fisher's exact test statistics. This global statistic and its properties can be constructed using the principles presented here.

A third example is correlation analysis. Korber *et al.* (1994) calculated the linear correlation coefficient of pairwise sequence distances between envelope regions C2-V3 and gp120. As Korber *et al.* (1994, p. 6736) recognized, the standard Pearson test for positive linear correlation is not completely valid because all points in the 1326 pairwise comparisons between the 52 patients are not independent. To address this problem, they conducted Pearson tests on unlinked sequence subsets to verify significance for valid test datasets. A more systematic approach would use a global statistic that linearly combines Pearson statistics across all unique unlinked datasets.

The magnitude of positive correlations between U-statistics determines how much efficiency is gained by using a global test with all sequences compared to a test with only one sequence per individual. If intraindividual sequence diversity is very small relative to interindividual sequence diversity, then these correlations are expected to be near one, in which case the global test will give an answer consistent with results from individual tests on subsequence sets. But if intra- versus interindividual sequence diversity is substantial, then the correlations will tend to be substantially less than one, in which case the global test can achieve large (2- to 10-fold) efficiency gains compared to individual tests, and the various individual tests may give different conclusions. Since HIV-1 exhibits significant relative levels of intrapatient sequence diversity in some regions (e.g., in the envelope gene [Wain-Hobson, 1995; Mullins, 1995]), the global test is expected to enjoy appreciable efficiency advantages for some HIV-1 applications.

In conclusion, for sequence diversity studies in which multiple sequences are measured from isolated samples from two populations, the two-sample global test developed here will often be the best available test. Its advantage over the common alternative, a two-sample t-test or Wilcoxon rank sum test conducted using one sequence or consensus sequence per patient, is interpretability and efficiency. It is maximally objective because it gives an overall result that does not depend on the selected sequence subset, and it is efficient because it uses all of the available sequences. Some important applications of these tests are viral diversity studies to evaluate local strains in potential vaccine trial sites, to estimate the extent of HLA profile matching with the predominant local viruses, to study the role of viral escape variants, and to compare populations by different disease stages, vaccine or drug treatment protocols, HLA types/profiles, or transmission modes and risk factors.

Beyond HIV sequence diversity applications, the global test applies generally to two-sample comparisons where repeated measurements of the same characteristic are taken between pairs of experimental units within each sample. The program for the global test and the associated confidence intervals is available by request from the first author (P.B.G.).

APPENDIX: GLOBAL TEST BASED ON LINEARLY COMBINED U-STATISTICS

For measurements from group 1, let $X_{kk'ii'}$ denote the distance between replicate sequence k from patient i and replicate sequence k' from patient $i'(i, i' = 1, \cdots, m, k, k' = 1, \cdots, K)$. For measurements from group 2, let $Y_{ll'jj'}$ denote the distance between replicate sequence l from patient j and replicate sequence l' from patient $j'(j, j' = 1, \cdots, n, l, l' = 1, \cdots, L)$. Let $X_{ii'} = (X_{11ii'}, \cdots, X_{KKii'})'$ and $Y_{jj'} = (X_{11jj'}, \cdots, X_{LLjj'})'$ denote independent random samples with distribution functions F and G whose marginals are denoted by $F_{kk'}$ and $G_{ll'}$, respectively $(k, k' = 1, \cdots, K, l, l' = 1, \cdots, L)$. Let M be the maximum of K and L. The null hypothesis to test is $H_0 : F(z_{11}, \cdots, z_{MM}) = G(z_{11}, \cdots, z_{MM})$ for all $z_{11}, \cdots, z_{MM} \in R^{M^2}$. The null hypothesis assumes exchangeability in that the X's and Y's have equal marginal distributions. The data of some components of $X_{ii'}$ and $Y_{jj'}$ may be missing; i.e., there may be fewer than K or L replicates from some patients. Set the indicator function $\delta_{kk'ii'}$ to 1 if $X_{kk'ii'}$ is observed, 0 otherwise, and define $\epsilon_{kk'jj'}$ similarly for $Y_{kk'jj'}$. The indicators $\delta_{ii'} = (\delta_{11'ii'}, \cdots, \delta_{KK'ii'})'(i, i' = 1, \cdots, m)$ and $\epsilon_{jj'} = (\epsilon_{11'jj'}, \cdots, \epsilon_{LL'jj'})'(j, j' = 1, \cdots, n)$ are assumed to be independent random samples from possibly different populations and to be independent of the underlying vectors $X_{ii'}$ and $Y_{jj'}$.

For each $k \le k' \in \{1, \cdots, K\}$ and $l \le l' \in \{1, \cdots, L\}$, consider a two-sample U-statistic with kernel ϕ:

$$U_{kk'll'} = \sqrt{N}\left\{\frac{m(m-1)}{2}\frac{n(n-1)}{2}\right\}^{-1}\sum_{i<i'}^{m}\sum_{j<j'}^{n}[\delta_{kk'ii'}\epsilon_{ll'jj'}\{\phi(X_{kk'ii'}, Y_{ll'jj'}) - \theta_{kk'll'}\}] \qquad (6.1)$$

with $\theta_{kk'll'} = E\{\phi(X_{kk'ii'}, Y_{ll'jj'})\}$ and $N = m(m-1)/2 + n(n-1)/2$. Under H_0, let $\theta_{kk'll'} = \theta_{kk'll'0}$, a known constant, and let $U_{kk'll'} = U_{kk'll'0}$. The interpatient distances calculated from sequence replicates k and k' in group 1 and replicates l and l' in group 2 differ between the two populations if $\theta_{kk'll'} \neq \theta_{kk'll'0}$.

For given distribution functions F and G, under the hypotheses

$$E\{\phi^2(X_{kk'ii'}, Y_{ll'jj'})\} < \infty \qquad (k \leq k' = 1, \cdots, K, l \leq l' = 1, \cdots, L)$$

and m/n converges to a constant $\rho \in (0, 1)$, the multivariate generalized U-statistic $(U_{1111}, \cdots, U_{KKLL}) \in R^{\frac{K(K+1)}{2}\frac{L(L+1)}{2}}$ converges to a mean-zero multivariate normal distribution. Let $\Lambda = ((\sigma_{kk'll'rr'ss'}))$, $k \leq k', r \leq r' = 1, \cdots, K, l \leq l', s \leq s' = 1, \cdots, L$ be the limiting covariance matrix under H_0. If in addition

$$E\{\phi^4(X_{kk'ii'}, Y_{ll'jj'})\} < \infty \qquad (k \leq k' = 1, \cdots, K, l \leq l' = 1, \cdots, L), \tag{6.2}$$

then $\sigma_{kk'll'rr'ss'}$ is consistently estimated by $\hat{\sigma}_{kk'll'rr'ss'} = (2N/m(m-1))\hat{\sigma}^2_{1kk'll'rr'ss'} + (2N/n(n-1))\hat{\sigma}^2_{2kk'll'rr'ss'}$. Here,

$$\hat{\sigma}^2_{1kk'll'rr'ss'} = \left\{\frac{m(m-1)}{2}\frac{n(n-1)}{2}\left(\frac{n(n-1)}{2} - 1\right)\right\}^{-1} * \sum_1 \delta_{kk'ii'}\delta_{rr'aa'}\epsilon_{ll'jj'}\epsilon_{ss'bb'}\{\phi(X_{kk'ii'}, Y_{ll'jj'}) - \theta_{kk'll'0}\}\{\phi(X_{rr'aa'}, Y_{ss'bb'}) - \theta_{rr'ss'0}\} \tag{6.3}$$

and

$$\hat{\sigma}^2_{2kk'll'rr'ss'} = \left\{\frac{n(n-1)}{2}\frac{m(m-1)}{2}\left(\frac{m(m-1)}{2} - 1\right)\right\}^{-1} * \sum_2 \delta_{kk'ii'}\delta_{rr'aa'}\epsilon_{ll'jj'}\epsilon_{ss'bb'}\{\phi(X_{kk'ii'}, Y_{ll'jj'}) - \theta_{kk'll'0}\}\{\phi(X_{rr'aa'}, Y_{ss'bb'}) - \theta_{rr'ss'0}\} \tag{6.4}$$

where $\sum_1$ denotes summation over $i < i'$, $a < a' = 1, \cdots, m$ with $i = a$ or $i' = a'$ and $j < j'$, $b < b' = 1, \cdots, n$ with $j \neq b$ and $j' \neq b'$; and $\sum_2$ denotes summation over $j < j'$, $b < b' = 1, \cdots, n$ with $j = b$ or $j' = b'$ and $i < i'$, $a < a' = 1, \cdots, m$ with $i \neq a$ and $i' \neq a'$.

For the statistic $V = \sum_{k \leq k'}^{K} \sum_{l \leq l'}^{L} \hat{w}_{kk'll'} U_{kk'll'0}$, the possibly data-dependent weight $\hat{w}_{kk'll'}$ is assumed to converge in probability, as $N \to \infty$, to a deterministic quality $w_{kk'll'}$ that is a function of the underlying distributions F and G under H_0. If Λ is positive definite, then under H_0 the statistic $Z = V(\hat{w}'\hat{\Lambda}\hat{w})^{-\frac{1}{2}}$ has a limiting standard normal distribution, as $N \to \infty$, where $\hat{w} = (\hat{w}_{1111}, \cdots, \hat{w}_{KKLL})' \in R^{\frac{K(K+1)}{2}\frac{L(L+1)}{2}}$ and $\hat{\Lambda} = ((\hat{\sigma}_{kk'll'rr'ss'}))$.

Adapting the theorem from Wei and Johnson (1985), it can be shown that weight functions defined in the following way are optimal for testing the sequence of local alternative hypotheses H_{1N} of the form

$$H_{1N} : \theta^N_{kk'll'} = \theta_{kk'll'0} + t\lambda_{kk'll'N}N^{-\frac{1}{2}} \qquad (k \leq k' = 1, \cdots, K, l \leq l' = 1, \cdots, L), \tag{6.5}$$

where t is any positive constant and $\{\lambda_{kk'll'N}\}$ is a sequence of positive numbers that converges to a deterministic function $\lambda_{kk'll'}$ of F_0, the distribution F under H_0. Let $\hat{\lambda}_{kk'll'}$ be a consistent estimate of $\lambda_{kk'll'}$ under H_0, and define $\hat{\tau}_{kk'} = \{\frac{m(m-1)}{2}\}^{-1}\sum_{i<i'}^{m}\delta_{kk'ii'}$ and $\hat{\eta}_{ll'} = \{\frac{n(n-1)}{2}\}^{-1}\sum_{j<j'}^{n}\delta_{ll'jj'}(k \leq k' = 1, \cdots, K, l \leq l' = 1, \cdots, L)$. Then it can be shown that, asymptotically, the upper-tailed test based on $\tilde{V} = \sum \tilde{w}_{kk'll'}U_{kk'll'0}$, with

$$\begin{aligned}\tilde{w} &= (\tilde{w}_{1111}, \cdots, \tilde{w}_{KKLL})' = \hat{\Lambda}^{-1}v \\ &= \hat{\Lambda}^{-1}(\hat{\tau}_{11}\hat{\eta}_{11}\hat{\lambda}_{1111}, \cdots, \hat{\tau}_{1K}\hat{\eta}_{11}\hat{\lambda}_{1K11}, \cdots, \hat{\tau}_{KK}\hat{\eta}_{L1}\hat{\lambda}_{KKL1}, \cdots, \hat{\tau}_{KK}\hat{\eta}_{LL}\hat{\lambda}_{KKLL})',\end{aligned} \tag{6.6}$$

maximizes the power against all alternatives H_{1N} for which $t > 0$ in (6.5). Thus, the statistic $\tilde{Z} = \tilde{V}(\tilde{w}'\hat{\Lambda}\tilde{w})^{-\frac{1}{2}}$ provides an asymptotically consistent test, optimal in the above sense, for equal distributions of interpatient sequence distances in the two groups.

We consider important special cases of the global test statistic. With kernel $\theta(x, y) = y - x$, each statistic $U_{kk'll'0}$ is a t-statistic. Under the Pitman location alternative H_{1N}, $\theta_{kk'll'0} = 0$, and the optimal weights are defined by (6.6) with $\hat{\lambda}_{kk'll'} = 1$. Inspection of (6.6) shows that the optimal weight $\tilde{w}_{kk'll'}$ equals the inverse variance $\hat{\Lambda}^{-1}$ multiplied by the k, k', l, l''th element of the vector v, equal to the fraction of pairwise comparisons between the kth and k''th sequence replicates among the maximum possible $m(m-1)/2$ in group 1 multiplied by this fraction for the lth and l''th sequence replicates in group 2. Thus, it weights each t-statistic by the inverse variance and by the amount of data used in the test.

Next consider the global test formed by combining Wilcoxon statistics. For this test to be well-defined, it is necessary to assume that the distribution functions F and G are continuous. With kernel $\phi(x, y) = 1$ if $y < x$ and 0 otherwise, $\theta_{kk'll'0} = 1/2$, and $U_{kk'll'0}$ is a Wilcoxon statistic. The weights for the locally most powerful test statistic are defined as follows. In (6.5) $t = \Delta$, and

$$\lambda_{kk'll'N} = \Delta^{-1} N^{\frac{1}{2}} \int_{-\infty}^{\infty} \{F_{kk'll'0}(x) - F_{kk'll'0}(x - \Delta N^{-\frac{1}{2}})\} dF_{kk'll'0}(x),$$

which converges to $\lambda_{kk'll'} = \int_{-\infty}^{\infty} f_{kk'll'0}^2(x)dx < \infty$ as $N \to \infty$, where $f_{kk'll'0}$ is the density function of $F_{kk'll'0}$. Extending the formula in Wei and Johnson (1985), optimal weights for testing H_{1N} are defined by

$$\hat{\lambda}_{kk'll'} = \frac{v_\gamma (m_{kk'} + n_{ll'})^{\frac{1}{2}}}{(3m_{kk'}n_{ll'})^{\frac{1}{2}}(D^{(m_{kk'}n_{ll'}+1-c_\gamma)} - D^{(c_\gamma)})},$$

where v_γ is the standard normal γ critical value, $m_{kk'} = \sum_{i<i'} \delta_{kk'ii'}$, $n_{ll'} = \sum_{j<j'} \epsilon_{ll'jj'}$, $D^{(1)} < \cdots < D^{(m_{kk'}n_{ll'})}$ are the ordered set of $m_{kk'}n_{ll'}$ differences $Y_{ll'jj'} - X_{kk'ii'}$, and

$$c_\gamma = \frac{1}{2} m_{kk'}n_{ll'} - \frac{1}{12} v_\gamma \{m_{kk'}n_{ll'}(m_{kk'} + n_{ll'})\}^{\frac{1}{2}}.$$

For the modified global statistic $V' = \sum_{k \le k', l \le l'} \sum_{i=1}^{m} \sum_{j=1}^{n} \hat{w}_{kk'll'ij} U_{kk'll'ij}$ introduced in the Discussion section, the U-statistics are defined by

$$U_{kk'll'ij} = \frac{\sqrt{m+n-2}}{(m-1)(n-1)} \sum_{i' \neq i}^{m} \sum_{j' \neq j}^{n} [\delta_{kk'ii'} \epsilon_{ll'jj'} \{\phi(X_{kk'ii'}, Y_{ll'jj'}) - \theta_{kk'll'ij}\}]. \tag{6.7}$$

The variance of V' can be calculated similarly as displayed in (6.3) and (6.4).

ACKNOWLEDGMENTS

This research was supported by grants OIG R35 CA39805-13, AI46703-01, 5-U01-AI38855, 5-U01-AI28076, and 1 U01 A146703-01 from the National Institutes of Health and by grant D43 TW00004 from the Fogarty International Center, National Institutes of Health. We thank L.J. Wei from spurring this work with helpful discussions about U-statistics.

REFERENCES

Ahmad, N., Baroudy, B.M., Baker, R.C., and Chappey, C. 1995. Genetic analysis of human immunodeficiency virus type 1 envelope V3 region isolates from mother and infants after perinatal transmission. *J. Virol.* 69, 1001–1012.

Albert, J., Franzen, L., Jansson, M., Scarlatti, G., Kataaka, P.K., *et al.* 1992. Ugandan HIV-1 V3 loop sequences closely related to the U.S./European consensus. *Virology* 190, 674–681.

Becker, M.L.B., de Jager, G., and Becker, W.B. 1995. Analysis of partial *gag* and *env* gene sequences of HIV type 1 strains from Southern Africa. *AIDS Res. Hum. Retroviruses* 11, 1265–1267.

Berman, P.W., Gray, A.M., Wrin, T., Vennari, J.C., Eastman, D.J., *et al.* 1997. Genetic and immunologic characterization of viruses infecting MN-rgp120-vaccinated volunteers. *J. Infect. Dis.* 176, 384–397.

Bredell, H., Williamson, C., Sonnenberg, P., Martin, D.J., Morris, L. 1995. Genetic characterization of HIV type 1 from migrant workers in three South African gold mines. *AIDS Res. Hum. Retroviruses* 14, 677–684.

Connor, R.I., Korber, B.T.M., Graham, B.S., Hahn, B.H., Ho, D.D., *et al.* 1998. Immunological and virological analyses of persons infected by human immunodeficiency virus type 1 while participating in trials of recombinant gp120 subunit vaccines. *J. Virol.* 72, 1552–1576.

Contag, C., Ehrnst, A., Duda, J., Bohlin, A.-B., Lindgren, S., *et al.* 1997. Mother-to-infant transmission of human immunodeficiency virus type 1 involving five envelope sequence subtypes. *J. Virol.* 71, 1292–1300.

Eddy, S.R. 1996. Hidden Markov models. *Curr. Opin. Struct. Biol.* 6, 361–365.

Essex, M. 1998. State of the HIV pandemic. *J. Hum. Virol.* 1, 427–429.

Fomsgaard, A. 1999. HIV-1 DNA vaccines. *Immunol. Lett.* 65, 127–131.

Fujii, S., Obaru, K., Matsushita, S., Morikita, T., Higuchi, H., *et al.* 1997. Characterization of proviral DNA from an individual with long-term, nonprogressive infection with HIV-1 and nonrecoverable virus. *J. Acq. Immune Def. Syn. Hum. Retrovirol* 15, 247–256.

Gail, M.H. 1985. Applicability of sample size calculations based on a comparison of proportions for use with the logrank test. *Cont. Clin. Trials* 6, 112–119.

Gilbert, P.B., Self, S.G., and Ashby, M.A. 1998. Statistical methods for assessing differential vaccine protection against human immunodeficiency virus types. *Biometrics* 54, 799–814.

Gilbert, P.B., Lele, S.R., and Vardi, Y. 1999. Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika* 86, 27–43.

Goudsmit, J., Back, N.K., and Nara, P.L. 1991. Genomic diversity and antigenic variation of HIV-1: Links between pathogenesis, epidemiology and vaccine development. *FASEB J.* 5, 2427–2436.

Graham, B.S., McElrath, J.M., Connor, R.I., Schwartz, D.H., Gorse, G.J., *et al.* 1998. Analysis of intercurrent human immunodeficiency virus type 1 infections in Phase I and II trials of candidate AIDS vaccines. *J. Infect. Dis.* 177, 310–319.

Janssens, W., Buve, A., Nkengasong, J.N. 1997. The puzzle of HIV-1 subtypes in Africa. *AIDS* 11, 705–712.

Korber, B., and Myers, G. 1992. Signature pattern analysis: A method for assessing viral sequence relatedness. *AIDS Res. Hum. Retroviruses* 8, 1549–1558.

Korber, B., Farber, R.M., Wolpert, D.H., and Lapedes, A.S. 1993. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope glycoprotein: An information theoretical analysis. *Proc. Natl. Acad. Sci. USA* 90, 2176–2180.

Korber, B., MacInnes, K., Smith, R., and Myers, G. 1994. Mutational trends in V3 loop protein sequences observed in different genetic lineages of human immunodeficiency virus type-1. *J. Virol.* 68, 6730–6744.

Korber, B., Foley, B., Leitner, T., McCutchan, F., Hahn, B., *et al.* 1997. *Human retroviruses and AIDS.* Theoretical Biology and Biophysics, Group T-10, Los Alamos, NM.

Lehmann, E.L. 1975. *Nonparametrics: Statistical methods based on ranks.* Holden Day, San Francisco, CA.

McCutchan, F.E., Hegerich, P.A., Brennan, T.P., Phanuphak, P., Singharaj, P., *et al.* 1992. Genetic variants of HIV-1 in Thailand. *AIDS Res. Hum. Retroviruses* 8, 1887–1895.

McCutchan, F.E., Artenstein, A.W., Sanders-Buell, E., Salminen, M.O., Carr, J.K., *et al.* 1996. Diversity of the envelope glycoprotein among human immunodeficiency virus type 1 isolates of clade E from Asia and Africa. *J. Virol.* 70, 3331–3338.

Mullins, J.I. 1995. *Participant Principles in STD and HIV Research*, University of Washington, Seattle, WA.

Murphy, E., Korber, B., Georges-Courbot, M.-C., You, B., Pinter, A., *et al.* 1993. Diversity of V3 region sequences of human immunodeficiency virus type 1 from the Central African Republic. *AIDS Res. Hum. Retroviruses* 9, 997–1006.

Novitsky, V.A., Montano, M.A., McLane, M.F., Renjifo, B., Vannberg, F., *et al.* 1999. Molecular cloning and phylogenetic analysis of HIV-1 subtype C: A set of 23 full-length clones from Botswana. *J. Virol.* 73, 4427–4432.

Pitman, E.J.G. 1939. Tests of hypotheses concerning location and scale parameters. *Biometrika* 31, 200–215.

Salminen, M.O., Carr, J.K., Robertson, D.L., Hegerich, P., Gotte, D., *et al.* 1997. Evolution and probable transmission of intersubtype recombinant human immunodeficiency virus type 1 in a Zambian couple. *J. Virol.* 71, 2647–2655.

Scarlatti, G., Leitner, T., Halapi, E., Wahlberg, J., Marchisio, P., *et al.* 1993. Comparison of variable region 3 sequences of human immunodeficiency virus type 1 from infected children with the RNA and DNA sequences of the virus populations of their mothers. *Proc. Natl. Acad. Sci. USA* 90, 1721–1725.

UNAIDS and WHO. 1998. *Report on the Global HIV/AIDS Epidemic, June 1998,* Author, Geneva.

van Harmelen, J., Wood, R., Lambrick, M., Rybicki, E.P., and Williamson, A.L. 1997. An association between HIV-1 subtypes and mode of transmission in Cape Town, South Africa. *AIDS* 11, 81–87.

van Harmelen, J., Bredell, H., Morris, L., van der Ryst, E., Lyons, S., *et al.* 1998. Determination of prevalent HIV-1 subtypes in South Africa and the construction of a gp120 DNA vaccine. In *12th World AIDS Conference, Geneva*, Abstract 33218.

Wain-Hobson, S. 1995. Virological mayhem. *Nature* 373, 102.

Wei, L.J., and Johnson, W.E. 1985. Combining dependent tests with incomplete repeated measurements. *Biometrika* 72, 359–364.

Williamson, C., Engelbrecht, S., Lambrick, M., van Rensburg, E., Wood, R., *et al.* 1995. HIV-1 subtypes in different risk groups in South Africa. *Lancet* 346, 782.

Wolinsky, S.M., Korber, B.T.M., Neumann, A.U., Daniels, M., Kunstman, K.J., *et al.* 1996. Adaptive evolution of human immunodeficiency virus type 1 during the natural course of infection. *Science* 272, 537–542.

Zolla-Pazner, S., Gomy, M.K., and Nyambi, P.N. 1999. The implications of antigenic diversity for vaccine development. *Immunol. Lett.* 66, 159–164.

Address correspondence to:
Peter Gilbert
Department of Biostatistics
Harvard School of Public Health
655 Huntington Avenue
Boston, MA 02115

E-mail: pgilbert@hsph.harvard.edu