Two-Sample Tests for Comparing Intra-Individual Genetic Sequence Diversity Between Populations

Peter B. Gilbert,[1,*] A.J. Rossini,[1] and Raj Shankarappa[2]

[1]*Statistical Center for HIV/AIDS Research and Prevention, Fred Hutchinson Cancer Research Center, and Department of Biostatistics, University of Washington, Seattle, Washington 98105, U.S.A.*

[2]*Thomas E. Starzl Transplantation Institute, Department of Surgery and Department of Infectious Diseases and Microbiology, University of Pittsburgh, Pittsburgh, Pennsylvania 15212, U.S.A.*

[*]*email: pgilbert@scharp.org*

SUMMARY. Consider a study of two groups of individuals infected with a population of genetically related heterogeneous mixture of viruses, and multiple viral sequences are sampled from each person. Based on estimates of genetic distances between pairs of aligned viral sequences within individuals, we develop four new tests to compare intra-individual genetic sequence diversity between the two groups. This problem is complicated by two levels of dependency in the data structure: (i) Within an individual, any pairwise distances which share a common sequence are positively correlated; and (ii) For any two pairings of individuals which share a person, the two differences in intra-individual distances between the paired individuals are positively correlated. The first proposed test is based on the difference in mean intra-individual pairwise distances pooled over all individuals in each group, standardized by a variance estimate that corrects for the correlation structure using U-statistic theory. The second procedure is a nonparametric rank-based analog of the first test, and the third test contrasts the set of subject-specific average intra-individual pairwise distances between the groups. These tests are very easy to use and solve correlation problem (i). The fourth procedure is based on a linear combination of all possible U-statistics calculated on independent,

identically distributed sequence sub-datasets, over the two levels (i) and (ii) of dependencies in the data, and is more complicated than the other tests but is generally more powerful. Although the proposed methods are empirical and do not fully utilize knowledge from population genetics, the tests reflect biology through the evolutionary models used to derive the pairwise sequence distances. The new tests are evaluated theoretically and in a simulation study, and are applied to a dataset of 200 HIV sequences sampled from 21 children.

KEY WORDS: Correlated data; CTL epitope; HIV genetic diversity; Hypothesis testing; Median test; Nonparametric statistics; Two-sample test; U-statistic; Wilcoxon test.

1. Introduction

Consider a recent study of 21 children who were infected with HIV at birth. Each child was dichotomized as slow/non-progressor (group 1; 9 children) or progressor (group 2; 12 children) based on the development of AIDS or death, immunologic parameters, and the presence of clinical symptoms (Shankarappa et al., 2002). Using independent PCR amplifications to eliminate resampling of virus templates, between 3 and 11 HIV gag p17 sequences were sampled per child, with mean 9.5 sequences. Sequences sampled from each individual formed monophyletic clusters consistent with their evolution from a unique common ancestral sequence and an absence of multiple infections. Sequences were evaluated for differences in regions predicted to encode 8-11 amino acid long epitopes recognized by cytotoxic T lymphocytes (CTLs). Predicted epitopes were identified using Epimatrix (De Groot et al., 1997), a computer algorithm based on a database of peptides that have been previously characterized for their binding to various human leukocyte antigen (HLA) molecules. For the purpose of this study, increased predicted binding to HLA is assumed to correlate with enhanced probability of the peptide being recognized by CTL. Nucleic acid sequence regions were portioned into those encoding potential CTL epitopes and those that did not, concatenated, and used to derive pairwise sequence distances.

Genetic diversity of the viruses within a child is often described using the average or median of the estimates of pairwise evolutionary distances between sequences. The goal of our study was to assess if the level of HIV genetic diversity differed between the slow/non-progressor and progressor groups, to help identify the role of viral evolution in HIV pathogenesis (Shankarappa et al., 1999). This assessment can be based on a study of group-contrasts $D^1_{kij} - D^2_{k'i'j'}$ or on the corresponding rank-based contrasts, where D^g_{kij} is the distance between sequences i and j from child k in group $g, g = 1, 2$. Two layers of dependency in the data structure complicate this problem: (i) Within child k, any two pairwise

distances that share a sequence are positively correlated, i.e., cor$(D^g_{kij}, D^g_{ki'j'}) > 0$ whenever i or j equal one of i' or j'; and (ii) Any two contrasts which involve a common individual are positively correlated, i.e., cor$(D^1_{kij} - D^2_{li'j'}, D^1_{k'ij} - D^2_{l'i'j'}) > 0$ whenever $k = k'$ or $l = l'$. One approach to avoid the problem of correlated data is to use a standard two-sample test with the studied contrasts restricted to independent sub-samples of distances from each group. However, such a test would be inefficient because it only uses information from a subset of the available contrasts.

Another commonly employed approach to avoid the correlation problem is to use one genetic distance value per sequence by comparing each sequence within a subject to a consensus sequence derived for the subject's pool. Alternatively, estimating branch lengths of each sequence to the most recent common ancestor (MRCA) sequence derived from a phylogenetic analysis produces independent observations. While these methods are valid, they suffer from other procedural and population genetic constraints. The consensus sequence may not adequately represent the population because of its bias toward most frequently sampled sequences, and is difficult to derive when the sequences exhibit high heterogeneity. Estimating branch lengths to MRCA is disproportionately influenced by outlier sequences. To overcome these deficiencies, Nickle et al. (2003) outlined a Center of Tree (COT) approach that minimizes the distance between sequences while being less influenced by outliers.

By considering codon as the unit of reference, genetic changes can be portioned into synonymous and nonsynonymous substitutions. The ratio of relative rates of synonymous and nonsynonymous substitutions has been used extensively in inferring the presence of selection. Simple corrections for heterogeneity in rates of these substitutions, as provided by Kimura 2-parameter and Jukes-Cantor models of evolution, have been shown to be unrealistic and potentially inaccurate (Muse, 1996; Zanotto, 1999). However, using programs like Modeltest (Posada and Crandall, 1998), Paup* (Swofford, 2002), and PAML (Yang, 1997),

it is possible to derive more accurate models of evolution and identify selection operating at the level of individual amino acid sites (Nielsen and Yang, 1998). Such methods can be used to construct biologically meaningful pairwise distance estimates; then, empirical tests can be applied to the distances to provide biologically relevant comparisons of diversity between populations.

In this article, we develop four new valid testing procedures for comparing pairwise distances between groups. The first three tests accommodate the correlations (i) but not (ii) and are very simple to use, and the fourth test accommodates both correlations (i) and (ii) but is more complicated to use. The first procedure is referred to as a test of "pooled mean diversities", and is based on comparing the average of all within-individual pairwise distances in group 1 to the average of all within-individual distances in group 2, standardized by a variance estimate computed using U-statistic theory. The second procedure is referred to as a test of "pooled median diversities", and compares the ranks of the distances in group 2 to the pooled-group median distance, with variance estimate computed by the same technique used for the first statistic. The third test of "mean subject-specific diversities" compares within-subect average diversities between the groups. The fourth procedure is based on a linear combination of correlated test statistics, over the two levels of dependencies (i) and (ii) in the data described above, and is referred to as the linear combination of U-statistics (LCU) test. The elementary contributing test statistic for this procedure can be taken to be any statistic within the family of two-sample U-statistics, which includes both the t-statistic and the Wilcoxon rank sum statistic.

The new test procedures are described in Section 2. Through simulations, in Section 3 the power of the tests are compared to one another and to a standard t-test based on to-consensus sequence distances. The procedures are applied to the HIV genetic distances dataset in Section 4. The details about the covariance structure of the U-statistics for the

LCU test are provided in an Appendix.

2. Valid Tests for Combining Correlated Test Statistics

Let M_1 and M_2 be the number of subjects in groups $g = 1$ and $g = 2$, respectively, and let n_k^g be the number of sequences for individual k in group $g, k = 1, \cdots, M_g$. We develop four tests, which for validity require the assumption of non-informative cluster sizes, i.e., the intra-individual sequence diversities do not depend on the number of sequences sampled from individuals (Hoffman et al., 2001). Under informative cluster sizes, the tests can give biased results if one group has systematically more sequences per person than the other group.

2.1. Test of pooled mean diversities

To describe the first test, we first consider a general one-sample situation in which N sequences are randomly sampled from a population, and all pairwise distances D_{ij} are measured among the $N(N-1)/2$ sequence pairs. Let μ be the true mean pairwise distance (i.e., pooled mean diversity). We derive estimates of μ and its variance.

The mean μ can be estimated by the empirical mean $\widehat{\mu} = \{N(N-1)/2\}^{-1} \sum_{i<j} D_{ij}$. The standard variance estimate of $\widehat{\mu}$, $\widehat{\sigma}^2 = \{N(N-1)/2 - 1\}^{-1} \sum_{i<j} (D_{ij} - \widehat{\mu})^2$, assumes $N(N-1)/2$ independent observations. This estimate is too small because it ignores the positive correlation of distances D_{ij} and D_{ik} which share sequence i. On the other hand, the variance estimate $\widehat{\sigma}^2 = (N-1)^{-1} \sum_{i<j} (D_{ij} - \widehat{\mu})^2$ is too large unless the distances are perfectly correlated.

U-statistic theory provides a way to derive the correct variance estimate of $\widehat{\mu}$. The general form of a U-statistic of order k with kernel h is

$$U_{N,k} = \binom{N}{k}^{-1} \sum_{i_1 < \cdots < i_k} h(X_{i_1}, \cdots, X_{i_k}), \tag{1}$$

where $h(X_{i_1}, \cdots, X_{i_k})$ is a k-variate function which is symmetric in its arguments. For our

application, $k = 2, (X_{i_1}, X_{i_2}) = (X_i, X_j)$ represents a pair of sequences within a subject, and $\widehat{\mu}$ is a U-statistic with $h(X_i, X_j) = D_{ij}$. The variance of $U_{N,k}$ (Lee, 1990) is given by

$$Var(U_{N,k}) = \binom{N}{k}^{-1} \sum_{c=1}^{k} \binom{k}{c} \binom{N-k}{k-c} \sigma_c^2, \tag{2}$$

with $\sigma_c^2 = Cov(h(X_{i_1}, \cdots, X_{i_k}), h(X_{i'_1}, \cdots, X_{i'_k}))$ where c of the indices $\{i_1, \cdots, i_k\}$ and $\{i'_1, \cdots, i'_k\}$ coincide. For $k = 2$, (2) simplifies to $Var(\widehat{\mu}) = \{N(N-1)/2\}^{-1}\{2(N-2)\sigma_1^2 + \sigma_2^2\}$, where $\sigma_1^2 = Cov(D_{ij}, D_{ik})$ is the covariance of two pairwise distances that share one sequence i, and $\sigma_2^2 = Var(D_{ij})$. Empirical estimates of σ_1^2 and σ_2^2 are

$$\widehat{\sigma}_1^2 = \frac{1}{N(N-1)(N-2)/3 - 1} \sum_{i<j<l} \{(D_{ij} - \widehat{\mu})(D_{il} - \widehat{\mu}) + (D_{ij} - \widehat{\mu})(D_{jl} - \widehat{\mu})\}, \tag{3}$$

$$\widehat{\sigma}_2^2 = \frac{1}{N(N-1)/2 - 1} \sum_{i<j} (D_{ij} - \widehat{\mu})^2. \tag{4}$$

For the two-sample problem with groups $g = 1, 2$ with mean diversities μ_1, μ_2 and N_1, N_2 sequences per group, an asymptotically standard normal test statistic for evaluating $H_0 : \mu_1 = \mu_2$ is given by

$$T_{poolmn} = \frac{\widehat{\mu}_1 - \widehat{\mu}_2}{\left[\sum_{g=1}^{2} \{N_g(N_g-1)/2\}^{-1}\{2(N_g-2)\widehat{\sigma}_{g1}^2 + \widehat{\sigma}_{g2}^2\}\right]^{1/2}}. \tag{5}$$

With ρ_g the correlation of two pairwise distances sharing a sequence in group g, if $N^* = \sum_{g=1}^{2}[\{N_g(N_g-1)/2\}/\{2(N_g-2)\rho_g^2 + 1\}]$ is large, then a standard normal critical value can be used. Otherwise a t critical value can be used with degrees of freedom approximated by a technique such as Satterthwaite's method.

The procedure described above must be adapted slightly to fit our example because there are multiple children within a group, and only pairwise distances between sequences within the same individual are used. The number of intra-individual pairwise distances in group g equals $N_g^{pair} = \sum_{k=1}^{M_g} n_k^g(n_k^g - 1)/2$, and $\widehat{\mu}_g$ is obtained as the average of these distances.

There are $\sum_{k=1}^{M_g} 2(n_k^g - 2)$ covariance terms for intra-individual distances sharing one sequence, and therefore $T_{poolmn} = \{\widehat{\mu}_1 - \widehat{\mu}_2\}/\{\widehat{Var}(\widehat{\mu}_1) + \widehat{Var}(\widehat{\mu}_2)\}^{1/2}$ is modified to

$$T_{poolmn} = \frac{\widehat{\mu}_1 - \widehat{\mu}_2}{\left[\sum_{g=1}^{2} \left(N_g^{pair}\right)^{-1} \left\{\sum_{k=1}^{M_g} 2(n_k^g - 2)\widehat{\sigma}_{g1}^2 + \widehat{\sigma}_{g2}^2\right\}\right]^{1/2}}, \tag{6}$$

with $\widehat{\sigma}_{g1}^2 = \{\sum_{k=1}^{M_g} n_k^g(n_k^g - 1)(n_k^g - 2)/3 - 1\}^{-1} \sum_{k=1}^{M_g} \sum_{i<j<l} \{(D_{kij}^g - \widehat{\mu}_g)(D_{kil}^g - \widehat{\mu}_g) + (D_{kij}^g - \widehat{\mu}_g)(D_{kjl}^g - \widehat{\mu}_g)\}$ and $\widehat{\sigma}_{g2}^2 = \{N_g^{pair} - 1\}^{-1} \sum_{k=1}^{M_g} \sum_{i<j} (D_{kij}^g - \widehat{\mu}_g)^2$.

2.2. Test of pooled median diversities

Let ω denote the median of all within-individual pairwise distances pooled over both groups. Consider the kernel $h(X_i, X_j) = I(D_{ij}^2 \leq \omega)$, which indicates whether the pairwise distance between sequences X_i and X_j in group 2 is less than or equal to the pooled median. Then, $\widehat{P}_{med2} = \{N_2(N_2-1)/2\}^{-1} \sum_{i<j} I(D_{ij}^2 \leq \omega)$ is a rank-based U-statistic that estimates $P_{med2} = \Pr(D_{ij}^2 \leq \omega)$. The same variance calculations used in Section 2.1 apply, yielding $Var(\widehat{P}_{med2}) = \{N_2(N_2 - 1)/2\}^{-1}\{2(N_2 - 2)\sigma_{med21}^2 + \sigma_{med22}^2\}$.The variance of $I(D_{ij}^2 \leq \omega)$ is estimated by $\widehat{\sigma}_{med22}^2 = \{N_2^{pair}/(N_2^{pair} - 1)\}\widehat{P}_{med2}(1 - \widehat{P}_{med2})$. Based on (3) with D_{ij} replaced with $I(D_{ij}^2 \leq \omega)$, σ_{med21}^2 is estimated by $\widehat{\sigma}_{med21}^2 =$

$$\frac{1}{N_2(N_2 - 1)(N_2 - 2)/3 - 1} \sum_{i<j<l} \left\{I(D_{ij}^2 \leq \omega, D_{il}^2 \leq \omega) + I(D_{ij}^2 \leq \omega, D_{jl}^2 \leq \omega)\right\} - (\widehat{P}_{med2})^2.$$

An asymptotically normal Z-statistic for testing $H_0 : P_{med2} = 1/2$ is given by $T_{poolmed} = (\widehat{P}_{med2} - 1/2)/(\widehat{Var}(\widehat{P}_{med2}))^{1/2}$. Adapted to our application with multiple individuals, this test statistic equals

$$T_{poolmed} = \frac{\widehat{P}_{med2} - 1/2}{\left[\left(N_2^{pair}\right)^{-1} \sum_{k=1}^{M_2} \{n_k^2(n_k^2 - 1)/2\}^{-1}\{2(n_k^2 - 2)\widehat{\sigma}_{med21}^2 + \widehat{\sigma}_{med22}^2\}\right]^{1/2}}. \tag{7}$$

2.3. Test of mean subject-specific diversities

The statistics T_{poolmn} and $T_{poolmed}$ weight each intra-individual pairwise distance equally, so that subjects with more sequences contribute more information. Alternative related test

statistics that weight each subject equally treat the averages or medians of the pairwise distances within each subject as the observations. For example, the average diversity for subject k in group g is $\widehat{\mu}_{gk} = \{n_k^g(n_k^g-1)/2\}^{-1}\sum_{i<j} D_{kij}^g$, and a test can be designed to compare the two i.i.d. samples $\widehat{\mu}_{11},\cdots,\widehat{\mu}_{1M_1}$ and $\widehat{\mu}_{21},\cdots,\widehat{\mu}_{2M_2}$.

The U-statistic variance calculations of Section 2.1 imply that the variance of the "mean of the averages", $M_g^{-1}\sum_{k=1}^{M_g}\widehat{\mu}_{gk}$, is $M_g^{-2}\sum_{k=1}^{M_g}\{n_k^g(n_k^g-1)/2\}^{-1}\{2(n_k^g-2)\sigma_{gk1}^2+\sigma_{gk2}^2\}$. Then, a standardized test statistic for comparing mean subject-specific diversities is given by

$$T_{subjmn} = \frac{M_1^{-1}\sum_{k=1}^{M_1}\widehat{\mu}_{1k} - M_2^{-1}\sum_{k=1}^{M_2}\widehat{\mu}_{2k}}{\left[\sum_{g=1}^{2} M_g^{-2}\sum_{k=1}^{M_g}\{n_k^g(n_k^g-1)/2\}^{-1}\{2(n_k^g-2)\widehat{\sigma}_{gk1}^2+\widehat{\sigma}_{gk2}^2\}\right]^{1/2}}, \tag{8}$$

where $\widehat{\sigma}_{gk1}^2$ and $\widehat{\sigma}_{gk2}^2$ are equal to $\widehat{\sigma}_{g1}^2$ and $\widehat{\sigma}_{g2}^2$ as in (3) and (4), with sums restricted to pairwise distances on subject k's sequences. For balanced data (equal number of sequences per subject), $T_{subjmn} = T_{poolmn}$; for unbalanced data the testing procedures differ.

2.4. Linear combination of U-statistics (LCU) test

The derivation of the LCU test statistic is outlined in four steps:

Step 1: Consider the comparison of intra-individual distances between person i in group 1 and person j in group 2. For each person, consider a maximal sub-collection of pairwise distances calculated from his or her set of sequences that is genuinely an i.i.d. sample (illustrated in Figure 1A; a sample is i.i.d. if no two distances share a sequence). From the two samples of distances, calculate a two-sample U-statistic.

Step 2: Given the two individuals i and j considered in Step 1, linearly combine all possible U-statistics formed from the different ways of taking maximal i.i.d. samples of pairwise distances from each individual. The variance of the resulting statistic V_{ij} can be estimated using the limiting multivariate normal distribution of the vector of U-statistics.

Step 3: Consider a "correspondence pairing" between the M_1 individuals in group 1 and M_2 individuals in group 2, with $M_1 \leq M_2$. The correspondence pairing is a mapping that

connects each individual in group 1 with a unique individual in group 2 (illustrated in Figure 1B). For each pair of individuals (with one in each group), the V_{ij}-statistic described in Step 2 is calculated. Then, a new statistic is formed as the average of the V'_{ij}s over the distinct pairs of individuals in the correspondence pairing, weighted by the inverse variance estimates.

Step 4: Linearly combine all possible weighted average statistics formed from different correspondence pairings. Form the LCU Z-statistic by dividing this linear combination by an estimate of its standard deviation, which is calculated using the limiting multivariate normal distribution of the vector of weighted average statistics.

Note that in Steps 2 and 4, a linear combination of dependent statistics is taken, and U-statistic theory is needed to characterize the limiting distribution of the combination. In contrast, in Step 3 a linear combination of independent statistics is taken. Further note that in Steps 2 and 4, the linear combinations sum over $(n_i^1)^{pair} \times (n_j^2)^{pair}$ and $M_2!/(M_2 - M_1)!$ terms, respectively, where $(n_k^g)^{pair}$ is the number of different ways of taking a maximal i.i.d. sample of pairwise distances from individual k in group g. The number $(n_k^g)^{pair}$ is equal to the product of odd integers $\leq n_k^g$. For example, Patient 1 in Figure 1A has 9 sequences, and there are $9 * 7 * 5 * 3 = 945$ ways to sample 4 pairwise distances that do not share any sequences. If the average number of sequences per person is $\bar{n}$, then the LCU Z-statistic sums over approximately $M = \{(\bar{n})^{pair}\}^2 M_2!/(M_2 - M_1)!$ elementary statistics. Because M can be huge, the computation time can be prohibitive; in this case the LCU test statistic can be based on a random subset of the possible correspondence pairings for Step 4 and on a random subset of the possible maximal i.i.d. samples of pairwise distances between each pair of individuals within a correspondence pairing for Step 2. This approach is analogous to a Monte Carlo permutation test in which a random sample of the possible permutations are used. For both procedures, taking enough random samples assures that the result is insensitive to the particular samples taken. In Sections 3 and 4 we consider what constitutes

sufficient samples for the LCU test.

FIGURE 1 HERE

We now develop the LCU test in more detail. Let SP_i^1 denote the collection of the $(n_i^1)^{pair}$ i.i.d. pairwise distance samples for individual i in group 1, and SP_j^2 denote the collection of the $\left(n_j^2\right)^{pair}$ i.i.d. pairwise distance samples for individual j in group 2. In Step 1, for each pair of sets $(SP_{ik}, SP_{jl}) \in (SP_i^1, SP_j^2)$, a two-sample U-statistic U_{ijkl} can be used to test if intra-individual distances calculated between the pairs of sequences in SP_{ik} for individual i in group 1 differ in distribution from intra-individual distances calculated between the pairs of sequences in SP_{jl} for individual j in group 2. In Step 2, we linearly combine a random sample (or the full sample) of the $(n_i^1)^{pair} \times (n_j^2)^{pair}$ possible unique U-statistics formed from the pairs of sets $\{(SP_{ik}, SP_{jl}) : k = 1, \cdots, (n_i^1)^{pair}, l = 1, \cdots, (n_j^2)^{pair}\}$. Set $V_{ij} = \sum_{(SP_{ik}, SP_{jl}) \in (SP_i^1, SP_j^2)} \widehat{w}_{ijkl} U_{ijkl}$, where the weight $\widehat{w}_{ijkl}$ may be data-dependent. Let Λ_{ij} be the covariance matrix of the complete vector of U-statistics composed of the elements $\{U_{ijkl} : (SP_{ik}, SP_{jl}) \in (SP_i^1, SP_j^2)\}$. Under the null hypothesis of no group difference in intra-individual distance distributions, the LCU statistic $Z_{ij} = V_{ij}(\widehat{w}_{ij}' \widehat{\Lambda}_{ij} \widehat{w}_{ij})^{-\frac{1}{2}}$ is approximately standard normally distributed, where $\widehat{\Lambda}_{ij}$ is an estimate of Λ_{ij} and $\widehat{w}_{ij}$ is the vector with elements $\{\widehat{w}_{ijkl} : (SP_{ik}, SP_{jl}) \in (SP_i^1, SP_j^2), k = 1, \cdots, (n_i^1)^{pair}, l = 1, \cdots, (n_j^2)^{pair}\}$. The elements of $\widehat{\Lambda}_{ij}$ are given in the Appendix by formulas (13) and (14). The test statistic Z_{ij} is equivalent to the test statistic proposed by Wei and Johnson (1985) for combining dependent U-statistics across repeat measurement times; the difference is that our statistic combines over sets of sequence pairs instead of measurement times.

The simplest and most easily interpreted combined statistic Z_{ij} weights all U-statistics equally, i.e. with all $\widehat{w}_{ijkl} = 1$. Alternatively, the weights $\widehat{w}_{ijkl}$ may be chosen as inverse variance estimates of the U_{ijkl}, or to optimize the statistical power of the test for detecting a

Pitman shift alternative hypothesis (Pitman, 1939; Lehmann, 1975). Wei and Johnson (1985) described the general form of the optimal weight functions, and the specific forms for the cases of combining Wilcoxon statistics and of combining t-statistics.

To combine the test statistics Z_{ij} calculated from different individuals (Step 3), let CP_z denote a correspondence pairing between the M_1 individuals in group 1 and M_2 individuals in group 2 (illustrated in Figure 1B). Let CP denote the collection of all unique sets CP_z. Define a weighted average statistic $\bar{V}_z$ by

$$\bar{V}_z = \frac{\sum_{(i,j)\in CP_z} \widehat{Var}^{-1}(V_{ij}) V_{ij}}{\sum_{(i,j)\in CP_z} \widehat{Var}^{-1}(V_{ij})}, \tag{9}$$

and a standardized version $Z_z = \bar{V}_z / \sqrt{\widehat{Var}(\bar{V}_z)}$, which equals

$$Z_z = \frac{\sum_{(i,j)\in CP_z} \widehat{Var}^{-1}(V_{ij}) V_{ij}}{\left(\sum_{(i,j)\in CP_z} \widehat{Var}^{-1}(V_{ij}) \left[1 + \sum_{(i',j')\in CP_z} \widehat{Var}^{-1}(V_{i'j'}) \widehat{Cov}(V_{ij}, V_{i'j'}) \right] \right)^{1/2}}, \tag{10}$$

where expressions for $Cov(V_{ij}, V_{i'j'})$ and its estimate are given in the Appendix. Each statistic Z_z is asymptotically standard normal under H_0. Now, for Step 4 define an overall test statistic V_{LCU} by

$$V_{LCU} = \sum_{CP_z \in CP} \widehat{w}^{\star}_z Z_z, \tag{11}$$

where $\widehat{w}^{\star} = (\widehat{w}^{\star}_z : CP_z \in CP)'$. Since each statistic Z_z has unit variance, a reasonable choice of weight functions is $\widehat{w}^{*}_z = 1$ for all CP_z. The covariance matrix $\Lambda^{\star}$ for the vector $Z = (Z_z : CP_z \in CP)'$ can be estimated by substituting estimates into the covariance formulas (15)-(19) in the Appendix. Then under H_0, the statistic $Z_{LCU} = V_{LCU}(\widehat{w}^{\star\prime} \widehat{\Lambda}^{\star} \widehat{w}^{\star})^{-1/2}$ is asymptotically standard normal and provides an overall test of H_0 using all of the available data. If the normality assumption is in question, significance levels of the test could be approximated using a Monte Carlo permutation procedure; however, currently this approach

would be excessively burdensome computationally. Therefore, currently the most useful application of the LCU test is to datasets with moderate-to-large sample sizes.

For large datasets, the LCU test is computationally burdensome even if an asymptotic critical value is used, in which case it may be necessary to use limited random samples of U-statistics in the test statistic. Use of a random sample of comparisons may make the test performance sensitive to the particular samples used if some pairwise comparisons are especially informative or are outliers. These considerations support using large random samples in the LCU test procedure; in fact the procedure should not be used if the result is sensitive to the particular random samples used. In practice the procedure can be repeated for several sets of random samples to verify reliability of the outcome. Based on the Example dataset, with N_{step2} the number of U-statistics linearly combined in Step 2 and N_{step4} the number of statistics Z_z linearly combined in Step 4, we have found that selecting N_{step2} and N_{step4} at least 100 produces stable answers that do not depend appreciably on the particular random samples used.

R and Fortran implementations of the testing procedures are available upon request.

3. Evaluation of power of the test statistics

Using limited random samples for the LCU test may reduce its power. To evaluate this possibility, we examined ratios of the variance of Z_{LCU} for the maximal numbers N_{step2} and N_{step4} versus the variance of Z_{LCU} for smaller numbers N_{step2} and N_{step4}. With $M_1 = M_2$, K (L) the number of sequences per person in group 1 (2), variance $\sigma^2_{12}(\sigma^2_{22})$ for an intra-subject pairwise distance in group 1 (2), covariance $\sigma^2_{11}(\sigma^2_{21})$ for two intra-subject pairwise distances that share a sequence in group 1 (2) (with $\rho^2_g \equiv \sigma^2_{g1}/\sigma^2_{g2}$), covariance σ^2_{V1} for $\bar{V}_z$ and $\bar{V}_{z'}$ (with $\rho^2_V = \sigma^2_{V1}/Var(\bar{V}_z)$), the variance expression is given by

$$Var(Z_{LCU}) = \frac{1+\rho^2_V(N_{step4}-1)}{M_1 N_{step2} N_{step4}} \left[\frac{(N_{step2}-1)\rho^2_1 + 1}{[K/2]}\sigma^2_{12} + \frac{(N_{step2}-1)\rho^2_2 + 1}{[L/2]}\sigma^2_{22} \right], \quad (12)$$

where $[x]$ is the integer part of x. We suppose equal variance parameters for the two groups, with values estimated based on the DNA distance data of the slow/non-progressor group studied in the Example. This yields $\sigma_{12}^2 = \sigma_{22}^2 = 0.0003316$. Since V_z and $\bar{V}_{z'}$ share many pairwise distances, the correlation ρ_V may be expected to be high; we choose $\rho_V = 0.80$.

For $M_1 = M_2 = 10$, $K = L = 8$, and $N_{step2} = N_{step4} = 50$, the variance ratio ranged monotonically from 0.47 for zero correlation $\rho_g = 0$ to 0.99 for perfect correlation $\rho_g = 1$, with ratio 0.64, 0.87, 0.96, 0.98, 0.99 for ρ_g = 0.10, 0.25, 0.50, 0.75, 0.90, respectively. Thus, the variance is 1-53% greater for 50 versus complete samples, and the inflation depends on the strength of correlation. When $N_{step2} = N_{step4}$ are increased to 100, the variance ratio increases to 0.95, 0.97, $>$ 0.99 for correlation 0.10, 0.25, $>$ 0.50, respectively. Therefore using at least 100 samples makes the variance inflation quite small.

Next, a simulation study was carried out to compare the power of the newly proposed test statistics T_{poolmn} and $T_{poolmed}$. In addition, we consider a procedure commonly used in the literature, a two-sample t-statistic T_{cons} used for comparing the $n_c^1 = \sum_{k=1}^{M_1} n_k^1$ distances to consensus sequences in group 1 to the n_c^2 distances to consensus sequences in group 2. Since the tests evaluate different hypotheses that reflect different scientific/biological questions, the procedures should not be viewed as competitors on an equal footing; nevertheless, information on relative power is helpful for guiding use of the tests. In particular, to-consensus distances and pairwise distances are biologically different ways to measure diversity, so that T_{cons} should be viewed as distinct in purpose relative to the other statistics. The statistic T_{subj} was not included in the simulations because it is equivalent to T_{poolmn} for balanced data, and Z_{LCU} was not included because of the relatively large computational burden.

Data were simulated under parameters estimated using the DNA data for the slow/non-progressor group in the Example: The pairwise distance differences for group 1 were simulated as $N(-0.001243, 0.0003316)$ with inter-subject correlation 0 and correlation 0, 0.25,

or 0.50 for intra-subject pairwise distances that share a sequence, and the to-consensus distances in group 1 were simulated as independent $N(-0.0044, 0.0001927)$. The group-size $M_1 = M_2$ was selected as 5, 10, or 15, and the number of sequences per subject K as 4, 8, or 12. The group 2 distances were simulated using the same models with mean-shifts Δ equal to $0.7, 0.6,$ and 0.4 standard deviations above the group 1 mean for $K = 4, 8,$ and 12, respectively. Power was estimated as the fraction of test statistics computed on 1000 simulated datasets that exceeded 1.96 in absolute value.

The estimated powers are shown in Figure 2. With 4 sequences per subject, T_{cons} is consistently more powerful than T_{poolmn} and $T_{poolmed}$. For low levels of correlation, T_{poolmn} and $T_{poolmed}$ achieve greater power than T_{cons} as the number of sequences per subject increases. The power of the pooled statistics T_{poolmn} and $T_{poolmed}$ decreases markedly with the degree of positive correlation, while the power of T_{cons} is independent of the correlation. Consequently, the test of to-consensus distances is considerably more powerful than the tests of pooled pairwise distances for highly correlated data, but the pooled tests are more powerful for lightly correlated data, with efficiency advantage increasing with the ratio of the number of sequences per subject versus the number of subjects. The pooled test of means is more powerful than the pooled median test for independent or lightly correlated distances; this result is related to the fact that the pairwise distances were simulated under a normal mean-shift model.

FIGURE 2 HERE

4. Example

We developed the testing procedures to analyze the HIV genetic distances dataset described in the Introduction; we now apply them to this dataset. First, a meaningful measure of genetic distance should be defined in the context of the goal of the study, which is to look

for differences between CTL epitope encoding regions and others. Sequence regions with CTL epitopes are of interest because the host immune system targets HIV-infected cells by recognizing epitope sequences on the cell's surface, and these immune responses have been shown to be the most potent form of control against HIV. Therefore, HIV vaccines under development are being specifically designed to elicit cell-mediated immune responses to help protect against HIV disease (cf., Nabel, 2001). The 200 sampled sequences were codon-aligned and partitioned into two sets of regions: those predicted to bind HLA encoded by the child, and regions outside the predicted epitopes. HLA binding strength was treated as a surrogate for CTL epitopes, and computed using published algorithms (De Groot et al., 1997; Schafer et al., 1998). The regions were separated, and each was concatenated, to form CTL and non-CTL sequence regions for each child. Genetic distances between pairs of HIV DNA sequences within a child were estimated under the Kimura 2-parameter model of evolution, separately for the two sets of regions. In addition, nonsynonymous and synonymous distances were estimated for each set of regions using the Jukes-Cantor correction for multiple substitutions as implemented in MEGA software (Kumar et al., 2001).

Given one of the aforementioned metrics, D^g_{kij} was taken to be the difference between the two pairwise distances computed on predicted CTL epitope and non-CTL epitope regions. In addition, majority consensus DNA sequences were derived and the to-consensus DNA distances were computed for CTL and non-CTL regions for each subject in each group, and the differences $D^g_{ki;cons}$ were constructed. Nonsynonomous and synonomous distances to consensus were not computed because of the difficulties associated with deriving consensus codon sequences. We hypothesized that the level of intra-individual diversity as measured by D^g_{kij}, or by $D^g_{ki;cons}$, would differ between slow/non-progressors and progressors.

Figure 3 shows boxplots of all intra-child DNA, nonsynonymous, and synonymous pairwise difference measures D^g_{kij} for the two groups. In addition, a boxplot is shown for the

to-consensus DNA distance differences. Descriptive analyses and naive two-sample t-tests did not suggest differences between the groups in the difference measures based on overall DNA genetic distance or on nonsynonymous distance (p > 0.10), but they did suggest a difference with respect to the synonymous distance (p $= 2.2 \times 10^{-6}$, sample means -0.0113 and 0.00713 for the slow/non-progressor and progressor groups). The naive t-tests compare the 387 total differences in the slow/non-progressor group with the 523 total differences in the progressor group, and require the assumption of independence for all difference measurements. However, there are only 21 individuals, and the positive correlation of pairwise distance differences within individuals (estimated correlations 0.55 and 0.61 for the slow/non-progressor and progressor groups) implies that the small p-value obtained for the synonymous difference comparison grossly overstates the significance. We re-assess the results using the valid methods outlined here, with emphasis on the synonymous difference measures. The slow/non-progressors and progressors had comparable numbers of sequences per child (average 9.1 and 9.8 sequences per child, respectively, and 20 of 21 children had between 9 and 11 sequences), and there was no apparent association between intra-individual sequence diversities and sequence number. Therefore, the cluster sizes were evidently non-informative, and the testing procedures can be validly applied to the data.

FIGURE 3 HERE

For DNA, nonsynonomous, and synonomous pairwise distance differences, respectively, the pooled mean diversity test gave results $T_{poolmn} = -1.04$ $(p = 0.30, \widehat{\mu}_1 = -0.0012, \widehat{\mu}_2 = 0.00015)$, $T_{poolmn} = -0.21$ $(p = 0.83, \widehat{\mu}_1 = 0.000052, \widehat{\mu}_2 = 0.00025)$, and $T_{poolmn} = -0.58$ $(p = 0.56, \widehat{\mu}_1 = -0.011, \widehat{\mu}_2 = 0.0069)$. The pooled median diversity test yielded $T_{poolmed} = -0.73$ ($p = 0.46$, medians -0.00090 and 0.00010), $T_{poolmed} = 1.27$ ($p = 0.20$, medians 0.0 and 0.0), and $T_{poolmed} = -0.39$ ($p = 0.70$, medians -0.012 and 0.0), and the subject-

specific mean diversity test gave $T_{subj} = -1.09$ $(p = 0.28), T_{subj} = -0.70$ $(p = 0.49)$, and $T_{subj} = -1.91$ $(p = 0.057)$. Therefore, there are no significant differences by DNA or nonsynonomous pairwise distances. For to-consensus DNA distances, no differences were found by the t test, with $T_{cons} = -1.42$ $(p = 0.16, \widehat{\mu}_1^c = -0.0044, \widehat{\mu}_2^c = -0.0016)$, and a Wilcoxon test was borderline significant ($p = 0.038$). The subject-specific diversity test showed a trend towards larger synonomous pairwise distance differences in progressors than slow/non-progressors ($p = 0.057$), but the pooled mean and median diversity tests did not support a group difference; this discordant result could be related to the relatively low power of the pooled procedures when the degree of positive correlation is high (estimated at ≈ 0.60).

Next, for synonomous distances we applied the LCU test based on Z_{LCU}. Wilcoxon U-statistics were used, and in Steps 2 and 4 the weights $\widehat{w}_{ijkl}$ and $\widehat{w}_z^*$ were set to one. Linear combinations in Step 2 were taken over $N_{step2} = 100$ Wilcoxon statistics based on 100 random samples of maximal i.i.d. pairwise distances from the individuals i and j, and in Step 4 were taken over $N_{step4} = 100$ statistics Z_z based on 100 randomly sampled correspondence pairings of individuals. For the first 20 correspondence pairings CP_z, Figure 4 shows the histogram of the Z-statistics Z_{ij} calculated in Step 2, with the stratified Z-statistic Z_z of (10) indicated with a bold vertical segment. The 100 statistics Z_z ranged from -1.14 to -0.22 and the overall Z-statistic Z_{LCU} took value -0.539, with two-sided p-value 0.59. Therefore, this test gave a similar result as the pooled mean and median tests, and did not suggest differences in levels of synonymous substitutions within predicted CTL epitope relative to non-CTL epitope regions in progressors as compared to slow/non-progressors. The test was repeated five times using different random seeds, and the resulting Z-statistic was always within 0.05 of -0.539, demonstrating that the result was not sensitive to the particular choice of random samples.

Although a number of experimental and population genetic methodological issues are not addressed in this study and could have potentially contributed to the results of the tests, these results along with the documented significance of CTL reactivity in the control of virus replication support studies to more rigorously identify CTL epitopes at the level of individual or group followed by validation of these differences using population genetic approaches.

FIGURE 4 HERE

5. Discussion

In this article, we present several valid testing procedures for comparing intra-individual sequence diversity between two populations. The tests for comparing pooled mean diversities, pooled median diversities, and mean subject-specific diversities, are simple to use. The LCU test is more complicated but incorporates the most information and thus sometimes provides greater power. Given the computational burden of the LCU test, it currently requires coding in a fast language such as Fortran, and to ensure reliable results should only be used with at least 100 samples in each of Steps 2 and 4 of the procedure. The power of each test decreases with the degree of positive correlation of intra-individual pairwise distances that share a sequence. We compared the new tests to a standard t-test for comparing mean to-consensus distances between groups, and showed that the new procedures are more powerful when the correlation of distances is low-to-moderate. The power of t or Wilcoxon tests for to-consensus distances is independent of pairwise correlations; thus these procedures are relatively most powerful when the pairwise distances are highly correlated. In applications it may be useful to estimate the correlations to judge the relative power of the test procedures. We also found that the comparative power of the new methods based on pairwise distances versus the to-consensus method is greater for larger ratios of the number of sequences per subject to the number of subjects.

Although this article focuses on hypothesis testing, the pooled mean and mean subject-specific methods directly provide point and interval estimates of the group difference in pooled mean diversity and in mean subject-specific diversity, respectively. Moreover, the LCU test statistic can be inverted to obtain point and interval estimates of the location difference in diversities.

The purpose of this report is to provide valid empirical two-sample tests for comparing pairwise genetic distances. Studies of viral diversity are challenged by the facts that heterogeneous sequences derived from an individual share a common ancestor sequence, and viral diversity increases over time within individuals. The former issue implies a level of non-independence not addressed by the proposed methods, and the latter issue implies that diversity depends on the time between infection and sampling. The latter problem could be addressed by extending the current methods to allow adjustment for sampling time and other covariates. Furthermore, the methods we used to derive pairwise DNA as well as synonymous and nonsynonymous distances have a number of population genetic issues that are not addressed. For instance, we used simple models of substitution (Kimura 2-parameter and Jukes-Cantor) that do not account for rate heterogeneities and are at best limited approximations of the true models of evolution. Similarly, there is considerable uncertainty about the efficiency with which true CTL epitopes are identified by Epimatrix, or for that matter, any of the other available prediction methods. CTL epitopes in HIV have a strong tendency to cluster within conserved regions of the viral genome (Yusim et al., 2002) and such conservation had been hypothesized as an adaptation on the part of the host to constrain immune escape by the pathogen (da Silva and Hughes, 1998). However, it remains to be seen if the CTL epitopes documented in HIV literature represent a highly biased set since most laboratory studies have used laboratory or subtype consensus sequences while it has been shown recently that nearly 30% more CTL epitopes would be identified when reagents based on

autologous sequences, as opposed to the subtype consensus, were used (Altfeld et al., 2003).

Our basic approach to testing, to average pairwise distances or their ranks and to accommodate the correlation structure by U-statistic variance calculations, can be applied to many other analyses of sequence diversity. For example, because the Kruskal-Wallis test statistic for comparing K groups is a K-sample U-statistic, a testing procedure for comparing intra-individual sequence diversity between K groups can be constructed along the same lines. In addition, the one-sample t and Wilcoxon signed rank statistics are U-statistics, and valid linear combination of U-statistics tests could be constructed for assessing whether the mean or location center of intra-individual pairwise sequence distance is different from some fixed value.

The methods developed here have special significance for HIV vaccine efficacy trials because of the unique features associated with HIV vaccines. Available evidence in human and animal studies suggest that leading HIV vaccine candidates may fail to prevent HIV infection, but could mitigate the disease course of HIV and diminish infectiousness, such as by reducing HIV viral load (Gilbert et al., 2003). Since many vaccine recipients may become infected in efficacy trials, it would be extremely useful to compare genetic changes within true CTL epitopes among individuals receiving vaccine to unvaccinated individuals. Such an analysis could help identify the immunologic and virologic bases for protection and the methods outlined here would be important in establishing these relationships.

ACKNOWLEDGEMENTS

The authors thank the Associate Editor and two referees for their helpful comments which led to the inclusion of three new testing procedures and many other improvements. This work is supported by NIH grants AI054165-01 and AI041870-01.

REFERENCES

Altfeld, M., Addo, M.M., Shankarappa, R., Lee, P.K., Allen, T.M., Yu, X.G., Rathod, A., Harlow, J., O'Sullivan, K., Johnston, M.N., Mullins, J.I., Rosenberg, E.S., Brander, C., Korber, B., and Walker, B.D. (2003). Enhanced detection of HIV-1-specific T cell responses to highly variable regions using peptides based on autologous virus sequences. *Journal of Virology* **77**, 7330-7340.

da Silva, J., and Hughes, A.L. (1998). Conservation of cytotoxic T lymphocyte (CTL) epitopes as a host strategy to constrain parasite adaptation: Evidence from the nef gene of Human Immunodeficiency Virus 1 (HIV-1). *Molecular Biology and Evolution* **15**, 1259-1268.

De Groot, A.S., Jesdale, B.M., Szu, E., Schafer, J.R., Chicz, R.M., and Deocampo, G. (1997). An interactive Web site providing major histocompatability ligand predictions: application to HIV research. *AIDS Research and Human Retroviruses* **13**, 529-531.

Gilbert, P.B., De Gruttola, V.G., Hudgens, M.G., Self, S.G., Hammer, S.M., and Corey, L.C. (2003). What constitutes efficacy for an HIV vaccine that ameliorates viremia: Issues involving surrogate endpoints in Phase III trials. *Journal of Infectious Diseases* **188**, 179-193.

Hoffman, E.B., Sen, P.K., and Weinberg, C.R. (2001). Within-cluster resampling. *Biometrika* **88**:1121-1134.

Kumar, S., Tamura, K, Jakobsen, I.K., and Nei, M. (2001). MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. http://www.megasoftware.net

Lee, A.J. (1990). *U-Statistics- Theory and Practice*. New York: Marcel-Dekker.

Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden Day.

Muse, S. (1996). Estimating synonymous and nonsynonymous substitution rates. *Journal of Molecular Biology and Evolution* **13**, 105-114.

Nabel, G.J. (2001). Challenges and opportunities for development of an AIDS vaccine. *Nature* **410**:1002-1007.

Nickle, D.C., Jensen, M.A., Gottlieb, G.S., Shriner, D., Learn, G.H., Rodrigo, A.G, and Mullins, J.I. (2003). Consensus and ancestral state HIV vaccines. *Science* **299**, author reply, 1515-1518.

Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929-936.

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

Posada, D. and Crandall, K.A. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817-818.

Schafer, J.R., Jesdale, B.M., George, J.A., Kouttab, N.M., and De Groot, A.S. (1998). Prediction of well-conserved HIV-1 ligands using a matrix-based algorithm, EpiMatrix. *Vaccine* **16**, 1880-1884.

Shankarappa, R., Margolick, J.B., Gange, S.J., Rodrigo, A.G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C.R., Learn, G.H., He, X., Huang, X.L., and Mullins, J.I. (1999). Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of Virology* **73**, 10489-10502.

Shankarappa, R., Rossini, A.J., Gilbert, P.B., Kazazi, F., Upchurch, D., Learn, G.H., DeGroot, A.S., Korber, B.T., and Mullins, J.I. (2002). Viral evolutionary changes in gag p17 within and outside potential CTL epitopes in HIV-1 infected children progressing at

different rates. *Technical report*, University of Washington, Seattle, Wa.

Swofford, D.L. (2002). PAUP* 4.0: Phylogenetic Analysis Using Parsimony (* and Other Methods)., 4.0b10 ed. Sinauer Associates, Inc., Sunderland, MA.

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

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computational Applied Biosciences* **13**:555-556.

Yusim, K., Kesmir, C., Gaschen, B., Addo, M.M., Altfeld, M., Brunak, S., Chigaev, A., Detours, V., and Korber, B.T. (2002). Clustering patterns of cytotoxic T-lymphocyte epitopes in human immunodeficiency virus type 1 (HIV-1) proteins reveal imprints of immune evasion on HIV-1 global variation. *Journal of Virology* **76**, 8757-8768.

Zanotto, P.M., Kallas, E.G., de Souza, R.F., and Holmes, E.C. (1999). Genealogical evidence for positive selection in the nef gene of HIV-1. *Genetics* **153**, 1077-1089.

APPENDIX: DERIVATION OF LINEAR COMBINATION OF U-STATISTICS (LCU) TEST

Let $S_i = [n_i^1/2]$ be the maximum size of an i.i.d. sample of pairwise distances from individual i in group 1, and similarly let $T_j = [n_j^2/2]$ for individual j in group 2; $[x]$ is the integer part of x. Let S be the maximum of the S_i for individuals in group 1, and $(n^1)^{pair}$ be the maximum of the $(n_i^1)^{pair}$ for individuals in group 1. Define T and $(n^2)^{pair}$ similarly for group 2. For measurements from individuals in group 1, let X_{isk} denote the distance between the s'th sequence pair from the k'th arrangement of unique sequence pairs from individual i ($i = 1, \cdots, M_1, s = 1, \cdots, S, k = 1, \cdots, (n^1)^{pair}$). For measurements from individuals in group 2, let Y_{jtl} denote the distance between the t'th sequence pair from the l'th arrangement of unique sequence pairs from individual j ($j = 1, \cdots, M_2, t = 1, \cdots, T, l = 1, \cdots, (n^2)^{pair}$). Let $X_{is} = (X_{isk} : k = 1, \cdots, (n^1)^{pair})'(i = 1, \cdots, M_1, s = 1, \cdots, S)$ and $Y_{jt} = (Y_{jtl} : l = 1, \cdots, (n^2)^{pair})'(j = 1, \cdots M_2, t = 1, \cdots, T)$ denote independent random samples with distribution functions F and G whose marginals are denoted by F_k and G_l, respectively ($k = 1, \cdots, (n^1)^{pair}, l = 1, \cdots, (n^2)^{pair}$). With M the maximum of $(n^1)^{pair}$ and $(n^2)^{pair}$, the null hypothesis to test is $H_0 : F(s_1, \cdots, s_M) = G(s_1, \cdots, s_M)$ for all $s_1, \cdots, s_M \in R^M$. The null hypothesis assumes exchangeability in that the $X'_{is}s$ and $Y'_{jt}s$ have equal marginal distributions. It also assumes that the distributions F_i and G_j are the same for all indices $i = 1, \cdots, M_1, j = 1, \cdots, M_2$. The data of some components of X_{is} and Y_{jt} will be missing for individuals with fewer sequences than the person with the most sequences in the respective groups. Set the indicator function δ_{isk} to 1 if X_{isk} is observed, 0 otherwise, and define ϵ_{jtl} similarly for Y_{jtl}. The indicators $\delta_{is} = (\delta_{isk} : k = 1, \cdots, (n^1)^{pair})'(i = 1, \cdots, M_1, s = 1, \cdots, S)$ and $\epsilon_{jt} = (\epsilon_{jtl} : l = 1, \cdots, (n^2)^{pair})'(j = 1, \cdots, M_2, t = 1, \cdots, T)$ are assumed to be independent random samples from possibly different populations, and to be independent of the underlying vectors X_{is} and Y_{jt}.

Consider the i'th individual in group 1 and the j'th individual in group 2. For each

$(SP_{ik}, SP_{jl}) \in (SP_i^1, SP_j^2)$, consider a two-sample U-statistic with kernel ϕ :

$$U_{ijkl} = \sqrt{Q}\,\{ST\}^{-1} \sum_{s=1}^{S} \sum_{t=1}^{T} \delta_{isk}\epsilon_{jtl}\,\{\phi(X_{isk}, Y_{jtl}) - \theta_{ijkl}\}$$

with $\theta_{ijkl} = E\{\phi(X_{sk}, Y_{tl})\}$ and $Q = S + T$. Under H_0, let $\theta_{ijkl} = \theta_{ijkl0}$, a known constant, and let $U_{ijkl} = U_{ijkl0}$. The distribution of intra-individual distances calculated from the set of sequence pairs $(s, t) \in (SP_{ik}, SP_{jl})$ differs between the two populations if $\theta_{ijkl} \neq \theta_{ijkl0}$.

For given distribution functions F and G, under the hypotheses

$$E\{\phi^2(X_{isk}, Y_{jtl})\} < \infty \quad ((s, t) \in (SP_{ik}, SP_{jl}))$$

and S/T converges to a constant $\rho \in (0, 1)$, the multivariate U-statistic $(U_{ijkl} : (SP_{ik}, SP_{jl}) \in (SP_i^1, SP_j^2))' \in R^{(n^1)^{pair}(n^2)^{pair}}$ converges to a mean-zero multivariate normal distribution. Let $\Lambda_{ij} = ((\sigma^2_{ijkk'll'}))$, $i = 1, \cdots, M_1, j = 1, \cdots, M_2, (SP_{ik}, SP_{jl})$, $(SP_{ik'}, SP_{jl'}) \in (SP_i, SP_j)$ be the limiting covariance matrix under H_0. If in addition

$$E\{\phi^4(X_{isk}, Y_{jtl})\} < \infty \quad ((s, t) \in (SP_{ik}, SP_{jl})),$$

then $\sigma^2_{ijkk'll'}$ is consistently estimated by $\widehat{\sigma}^2_{ijkk'll'} = (Q/S)\widehat{\sigma}^2_{1ijkk'll'} + (Q/T)\widehat{\sigma}^2_{2ijkk'll'}$. Here,

$$\widehat{\sigma}^2_{1ijkk'll'} = \{ST(T-1)\}^{-1} \\ * \sum_{1} \delta_{isk}\delta_{isk'}\epsilon_{jtl}\epsilon_{jt'l'}\{\phi(X_{isk}, Y_{jtl}) - \theta_{ijkl0}\}\{\phi(X_{isk'}, Y_{jt'l'}) - \theta_{ijk'l'0}\} \qquad (13)$$

and

$$\widehat{\sigma}^2_{2ijkk'll'} = \{TS(S-1)\}^{-1} \\ \sum_{2} \delta_{isk}\delta_{is'k'}\epsilon_{jtl}\epsilon_{jtl'}\{\phi(X_{isk}, Y_{jtl}) - \theta_{ijkl0}\}\{\phi(X_{is'k'}, Y_{jtl'}) - \theta_{ijk'l'0}\} \qquad (14)$$

where $\sum_1$ denotes summation over $s = 1, \cdots, S$ and $t \neq t' = 1, \cdots, T$; and $\sum_2$ denotes summation over $t = 1, \cdots, T$ and $s \neq s' = 1, \cdots, S$.

For the statistic $V_{ij} = \sum_{(SP_{ik},SP_{jl})\in(SP_i,SP_j)} \widehat{w}_{ijkl}U_{ijkl0}$, the possibly data-dependent weight $\widehat{w}_{ijkl}$ is assumed to converge in probability, as $Q \to \infty$, to a deterministic quantity w_{ijkl} that is a function of the underlying distributions F and G under H_0. If Λ_{ij} is positive definite then under H_0 the statistic $Z_{ij} = V_{ij}(\widehat{w}'_{ij}\widehat{\Lambda}_{ij}\widehat{w}_{ij})^{-\frac{1}{2}}$ has a limiting standard normal distribution, as $Q \to \infty$, where $\widehat{w}_{ij} = (\widehat{w}_{ijkl} : (SP_{ik}, SP_{jl}) \in (SP_i^1, SP_j^2))' \in R^{(n^1)^{pair}(n^2)^{pair}}$ and $\widehat{\Lambda}_{ij} = ((\widehat{\sigma}^2_{ijkk'll'}))$.

To calculate the LCU statistic Z_{LCU}, it is necessary to estimate $Cov(Z_z, Z_{z'})$ for each $(CP_z, CP_{z'}) \in CP$. If $z = z'$, then the covariance is one. For $z \neq z'$, it equals

$$\begin{aligned} Cov(Z_z, Z_{z'}) &= d^{-1/2}Cov\left(\sum_{(i,j)\in CP_z} \widehat{Var}(V_{ij})^{-1}V_{ij}, \sum_{(i',j')\in CP_{z'}} \widehat{Var}(V_{i'j'})^{-1}V_{i'j'}\right) \\ &= d^{-1/2}\sum_{(i,j)\in CP_z}\sum_{(i',j')\in CP_{z'}} \widehat{Var}(V_{ij})^{-1}\widehat{Var}(V_{i'j'})^{-1}Cov(V_{ij}, V_{i'j'}), \quad (15) \end{aligned}$$

where

$$\begin{aligned} d &= \left(\sum_{(i,j)\in CP_z} \widehat{Var}(V_{ij})^{-1}\left[1 + \sum_{(i',j')\in CP_z} \widehat{Var}(V_{i'j'})^{-1}Cov(V_{ij}, V_{i'j'})\right]\right) \\ &\quad * \left(\sum_{(i,j)\in CP_{z'}} \widehat{Var}(V_{ij})^{-1}\left[1 + \sum_{(i',j')\in CP_{z'}} \widehat{Var}(V_{i'j'})^{-1}Cov(V_{ij}, V_{i'j'})\right]\right) \quad (16) \end{aligned}$$

and

$$Cov(V_{ij}, V_{i'j'}) = \sum_{(SP_{ik},SP_{jl})\in(SP_i,SP_j)}\sum_{(SP_{i'k},SP_{j'l})\in(SP_{i'},SP_{j'})} \widehat{w}_{ijkl}\widehat{w}_{i'j'k'l'}Cov(U_{ijkl}, U_{i'j'k'l'}). \quad (17)$$

The covariance $Cov(V_{ij}, V_{i'j'})$ can be estimated by 0 if the sets $\{i, j\}$ and $\{i', j'\}$ do not intersect, and by $\widehat{\sigma}^2_{ijj'kk'll'} = (Q/S)\widehat{\sigma}^2_{1ijj'kk'll'} + (Q/T)\widehat{\sigma}^2_{2iji'j'kk'll'}$ otherwise, with

$$\begin{aligned} \widehat{\sigma}^2_{1iji'j'kk'll'} &= \{ST(T-1)\}^{-1} \\ &\quad * \sum_1 \delta_{isk}\delta_{i'sk'}\epsilon_{jtl}\epsilon_{j't'l'}\{\phi(X_{isk}, Y_{jtl}) - \theta_{ijkl0}\}\{\phi(X_{i'sk'}, Y_{j't'l'}) - \theta_{i'j'k'l'0}\} \quad (18) \end{aligned}$$

and

$$\widehat{\sigma}^2_{2iji'j'kk'll'} = \{TS(S-1)\}^{-1}$$
$$* \sum_2 \delta_{isk}\delta_{i's'k'}\epsilon_{jtl}\epsilon_{j'tl'}\{\phi(X_{isk}, Y_{jtl}) - \theta_{ijkl0}\}\{\phi(X_{i's'k'}, Y_{j'tl'}) - \theta_{i'j'k'l'0}\} \quad (19)$$

where $\sum_1$ denotes summation over $s = 1, \cdots, S$ and $t \neq t' = 1, \cdots, T$; and $\sum_2$ denotes summation over $t = 1, \cdots, T$ and $s \neq s' = 1, \cdots, S$.

For a test based on Wilcoxon statistics, $\phi(x, y) = I(y > x)$ and $\theta_{ijkl0} = 1/2$, and for a test based on t-statistics, $\phi(x, y) = y - x$ and $\theta_{ijkl0} = 0$.

Figure Legends

Figure 1. (A) The left panel shows D^1_{115}, D^1_{123}, D^1_{146}, D^1_{179}, and D^1_{1810}, a maximal i.i.d. sample of pairwise sequence distances calculated from the 10 sequences of Patient 1 in the Slow/Non-Progressor Group 1. The right panel shows D^2_{726}, D^2_{738}, D^2_{745}, and D^2_{779}, a maximal i.i.d. sample of pairwise sequence distances calculated from the 9 sequences of Patient 7 in the Progressor Group 2. (B) For the 9 persons in the Slow/Non-Progressor Group 1 and the 12 persons in the Progressor Group 2, the figure illustrates a correspondence pairing that links each individual in Group 1 with a unique individual in Group 2.

Figure 2. The figure shows the estimated power in the simulation study of the test statistics T_{poolmn} (black solid lines), $T_{poolmed}$ (red dashed lines), and T_{cons} (green dotted lines), versus the correlation 0, 0.25, or 0.50 among the intra-individual pairwise distances that share one sequence. The left, middle, and right rows are for $K = 4, 8,$ and 12 sequences per subject, and the left, middle, and right columns are for $M = 5, 10,$ and 15 subjects per group.

Figure 3. Boxplots of the intra-individual pairwise (A) synonymous, (B) nonsynonomous, and (C) DNA difference measurements D^g_{kij}, for the 9 persons in the Slow/Non-Progressor Group 1 ($n = \sum_{k=1}^{9} n_k^1(n_k^1 - 1)/2 = 387$ total pairwise distances) and the 12 persons in the Progressor Group 2 ($n = \sum_{k=1}^{12} n_k^2(n_k^2 - 1)/2 = 523$ total pairwise distances). D^g_{kij} is the synonymous genetic distance between sequences i and j from person k in Group g computed on predicted CTL epitope regions minus this distance computed on predicted non-CTL epitope regions. (D) is a boxplot of the intra-individual to-consensus differences in DNA distances computed on predicted CTL epitope regions versus on non-CTL epitope regions.

Figure 4. For the first 20 of the 100 correspondence pairings (CP_z) of individuals in the Slow/Non-Progressor and Progressor groups used for applying the LCU test to the example data, the panel shows the histogram of the Z-statistics Z_{ij} calculated in Step 2. The stratified Z-statistic Z_{stratz} calculated for each correspondence pairing is indicated by a bold vertical line.

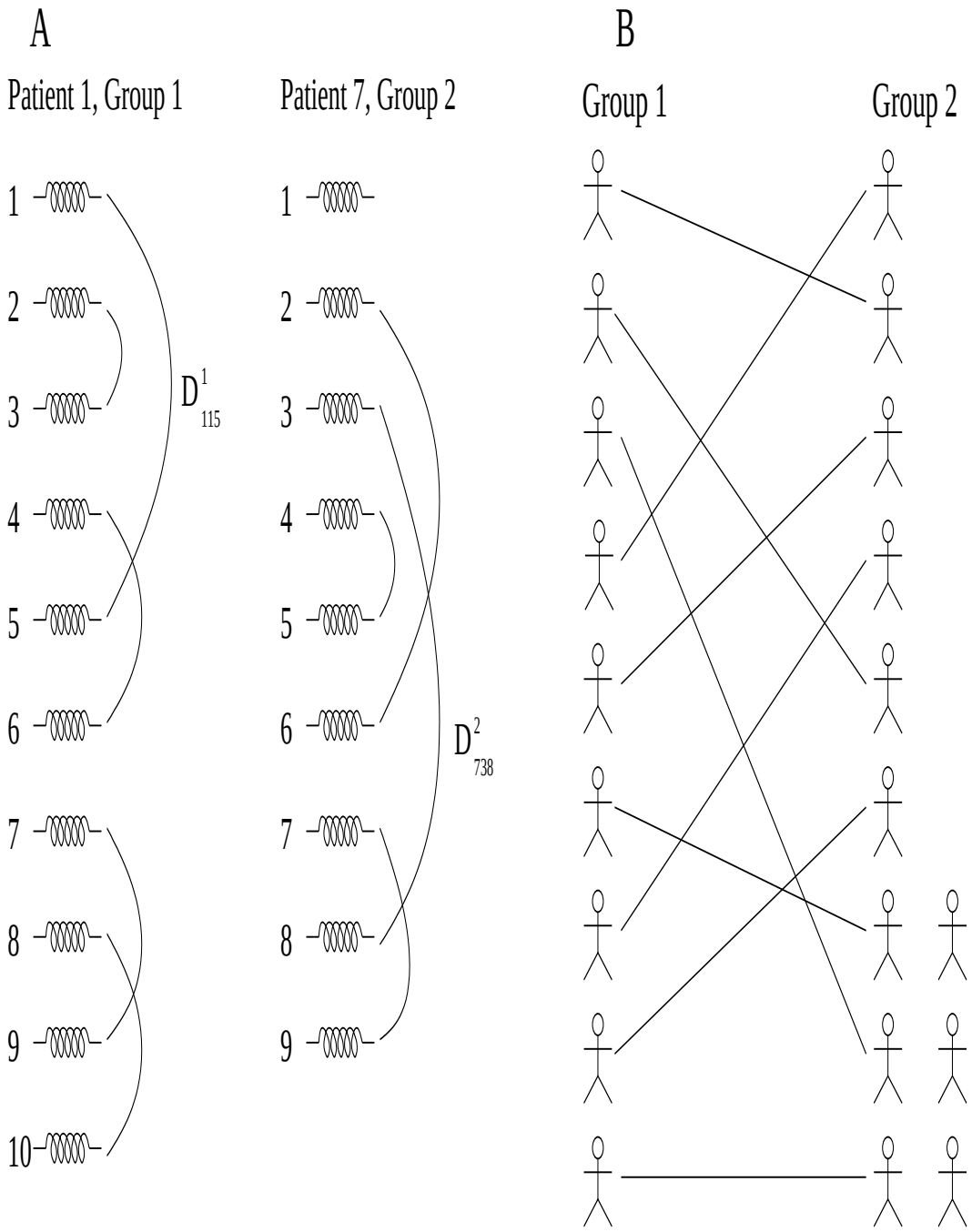
A
Patient 1, Group 1
Patient 7, Group 2
1
2
3
4
5
6
7
8
9
10
D^1_{115}
D^2_{738}
B
Group 1
Group 2

Figure 1.

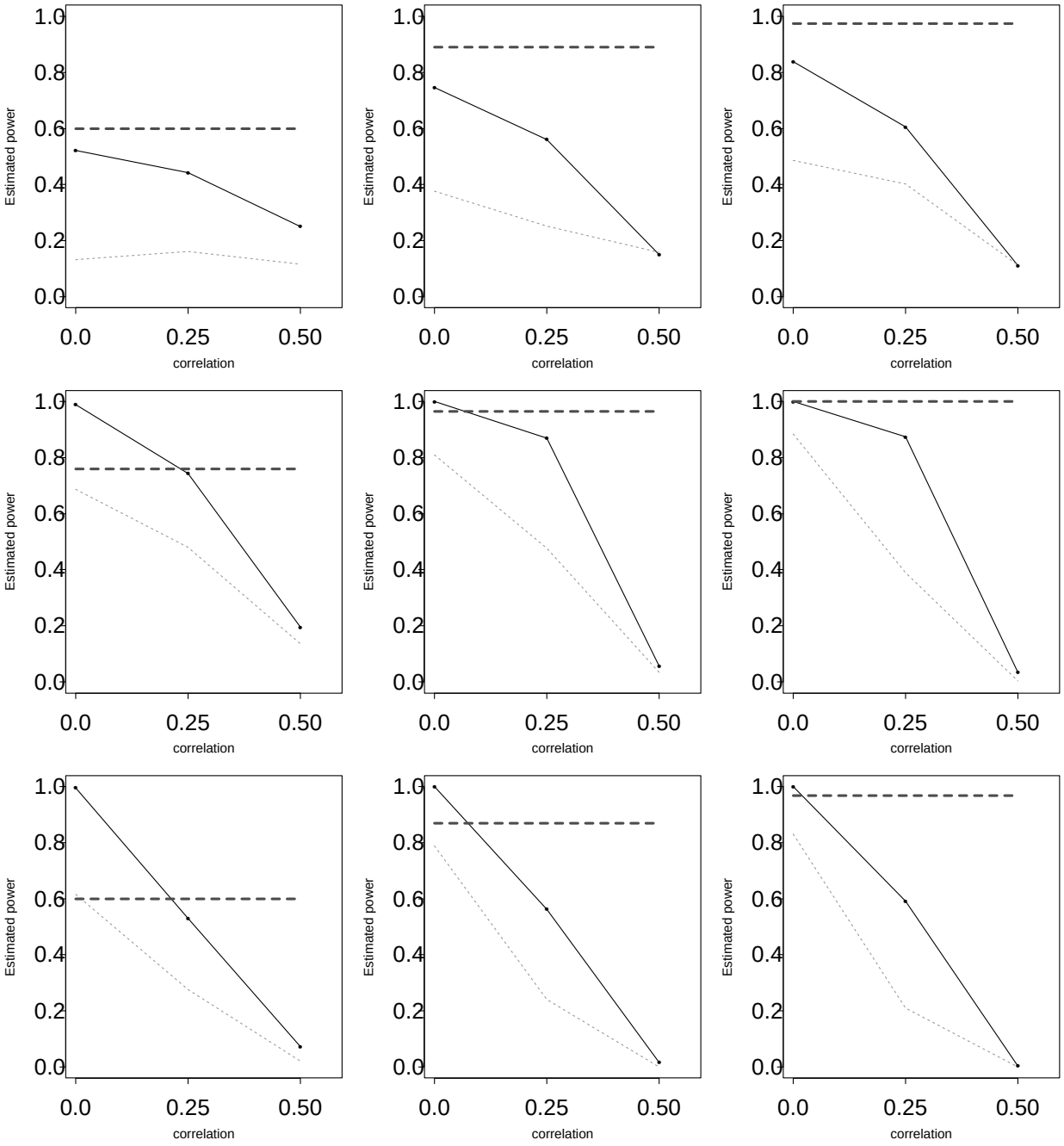

Figure 2.

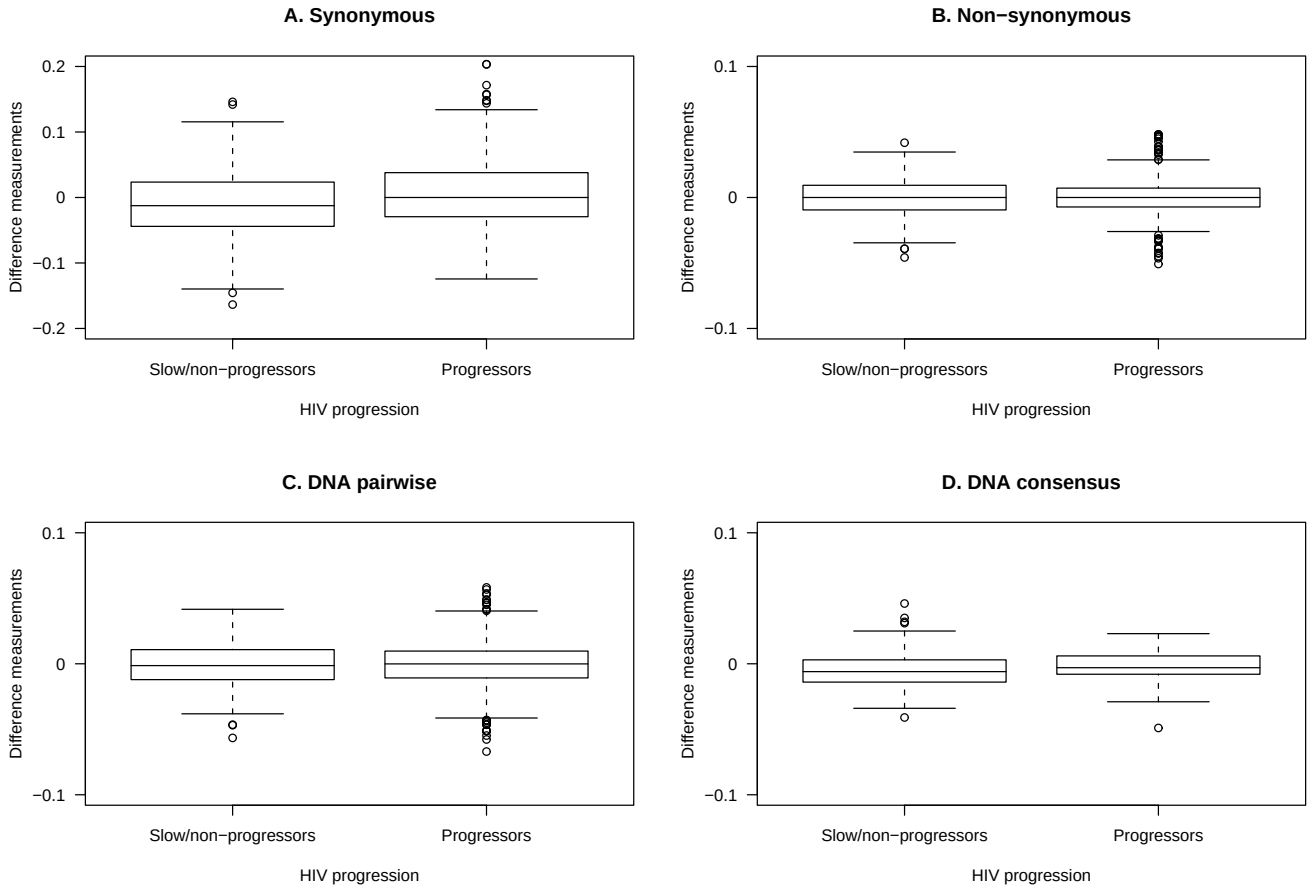
A. Synonymous
B. Non−synonymous
C. DNA pairwise
D. DNA consensus
Difference measurements
HIV progression
Slow/non−progressors
Progressors
0.2
0.1
0
−0.1
−0.2

Figure 3.

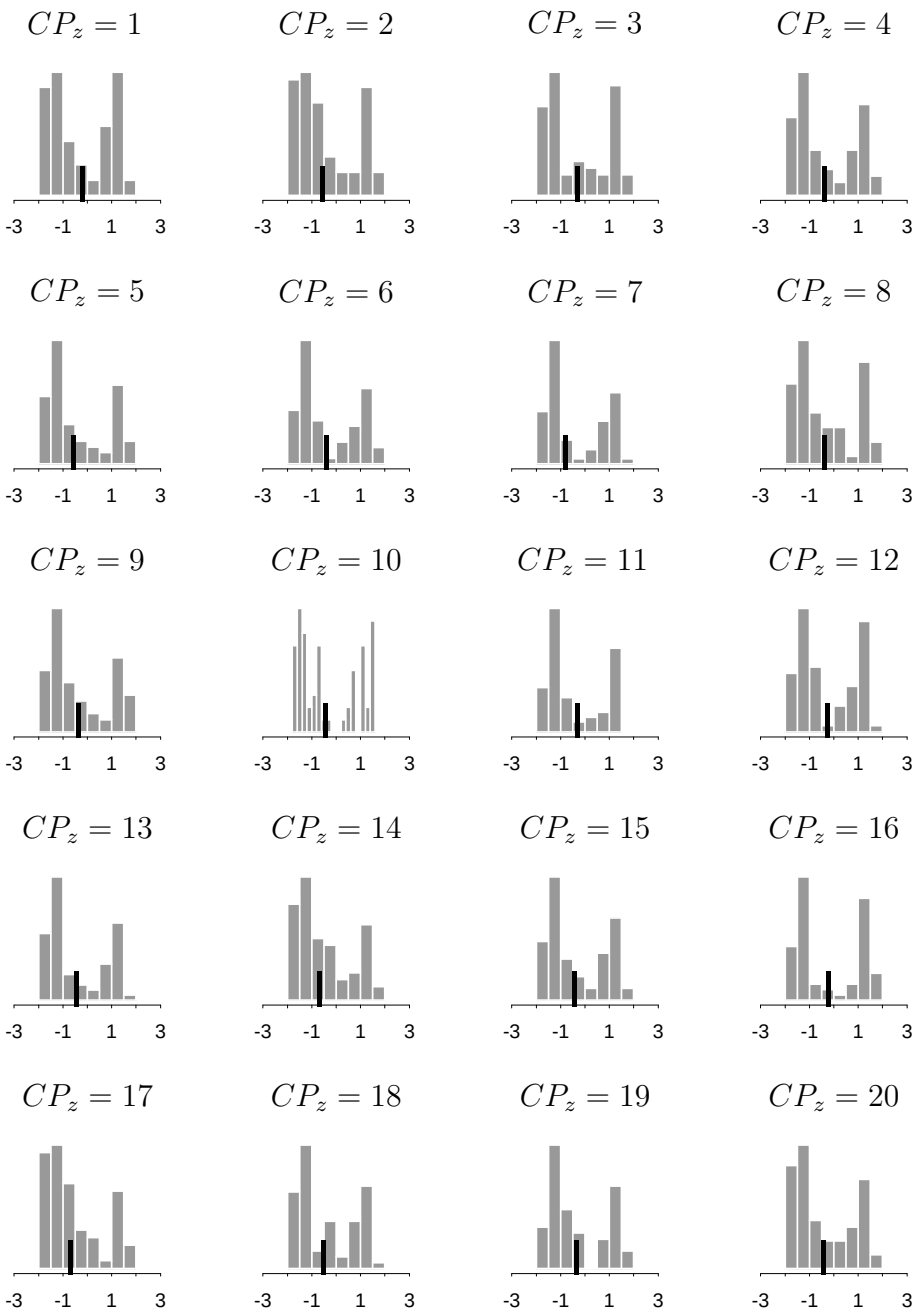
$CP_z = 1$
$CP_z = 2$
$CP_z = 3$
$CP_z = 4$
$CP_z = 5$
$CP_z = 6$
$CP_z = 7$
$CP_z = 8$
$CP_z = 9$
$CP_z = 10$
$CP_z = 11$
$CP_z = 12$
$CP_z = 13$
$CP_z = 14$
$CP_z = 15$
$CP_z = 16$
$CP_z = 17$
$CP_z = 18$
$CP_z = 19$
$CP_z = 20$
-3 -1 1 3

Figure 4.