

# Predicting Demographic Group Structures Based on DNA Sequence Data

Jon P. Anderson,\* Gerald H. Learn,† Allen G. Rodrigo,†<sup>1</sup> Xi He,† Yang Wang,†  
Hillard Weinstock,‡ Marcia L. Kalish,‡ Kenneth E. Robbins,‡ Leroy Hood,\*<sup>2</sup>  
and James I. Mullins†

\*Department of Molecular Biotechnology, Health Sciences Center, University of Washington, Seattle; †Department of Microbiology, Health Sciences Center, University of Washington, Seattle; and ‡Centers for Disease Control and Prevention, Atlanta, Georgia

The ability to infer relationships between groups of sequences, either by searching for their evolutionary history or by comparing their sequence similarity, can be a crucial step in hypothesis testing. Interpreting relationships of human immunodeficiency virus type 1 (HIV-1) sequences can be challenging because of their rapidly evolving genomes, but it may also lead to a better understanding of the underlying biology. Several studies have focused on the evolution of HIV-1, but there is little information to link sequence similarities and evolutionary histories of HIV-1 to the epidemiological information of the infected individual. Our goal was to correlate patterns of HIV-1 genetic diversity with epidemiological information, including risk and demographic factors. These correlations were then used to predict epidemiological information through analyzing short stretches of HIV-1 sequence. Using standard phylogenetic and phenetic techniques on 100 HIV-1 subtype B sequences, we were able to show some correlation between the viral sequences and the geographic area of infection and the risk of men who engage in sex with men. To help identify more subtle relationships between the viral sequences, the method of multidimensional scaling (MDS) was performed. That method identified statistically significant correlations between the viral sequences and the risk factors of men who engage in sex with men and individuals who engage in sex with injection drug users or use injection drugs themselves. Using tree construction, MDS, and newly developed likelihood assignment methods on the original 100 samples we sequenced, and also on a set of blinded samples, we were able to predict demographic/risk group membership at a rate statistically better than by chance alone. Such methods may make it possible to identify viral variants belonging to specific demographic groups by examining only a small portion of the HIV-1 genome. Such predictions of demographic epidemiology based on sequence information may become valuable in assigning different treatment regimens to infected individuals.

## Introduction

The study of genetic variation in human immunodeficiency virus type 1 (HIV-1) is an ever evolving field because of the rapid genetic divergence of HIV-1 through nucleotide substitution, duplication, deletion, recombination, and selection (Howell et al. 1991; Bonhoeffer, Holmes, and Nowak 1995; Robertson, Hahn, and Sharp 1995; Robertson et al. 1995; Mansky 1996; Burke 1997; Salminen et al. 1997). Not only is there great variation between viruses of different subtypes, but even the quasispecies within an individual may contain high variation. Within the myriad of HIV-1 genomic mutations may reside important phenotypic information. The amount of diversity seen in the *env* gene has been shown to exceed 10% within an infected individual (Delwart et al. 1994; Wain-Hobson 1995). With such rapid evolution occurring within the HIV-1 population, the ability to identify and distinguish genetic structure within sequence data will be crucial in understanding the biology of HIV-1.

Phylogenetic and phenetic methods are often used in an attempt to identify structure in sequence data. Phylogenetic methods attempt to infer the evolutionary history that is most consistent with the observed data. These methods are used in such analyses as parsimony, maximum

likelihood (ML), and distance matrix. Phenetic methods, on the other hand, attempt to group sequences based on similarity alone. These phenetic methods are incorporated in the clustering analyses of the unweighted pair group method using arithmetic averaged (UPGMA), complete linkage, single linkage, and Ward's method. Phylogenetic analyses have typically been used to group the viruses involved in the HIV-1 pandemic into the main (M) group, an outlier (O) group, and the Non M, Non O (N) group. These techniques have been used in compartmentalization studies (Poss et al. 1998), detecting the origin of an epidemic (Zhu et al. 1998), comparing geographically distinct viral populations (Gao et al. 1996), and analyzing drug resistance (Leigh Brown and Cleland 1996), and have led to the establishment of ten distinct subtypes or clades (A through J) within the M group (Louwagie et al. 1993; Kostrikis et al. 1995; Myers et al. 1995; Gao et al. 1998).

Although the ability to distinguish viral subtypes from other diverse groups of sequences has been shown, there is little information to link the evolutionary structure of HIV-1 to the epidemiological information of a person, including risk and demographic factors. To correlate patterns of genetic diversity with risk and demographic factors, we studied an approximately 620 nt region from the second constant region (C2) to the fifth variable region (V5) of *env* in 100 HIV-1 subtype B infected individuals from the United States. The C2–V5 region encompasses most of the biologically significant sites in *env*, and most partial *env* sequences in the literature are included in this region. None of the individuals studied were linked by a series of known transmission events, and information about certain epidemiological characteristics was available for each person.

<sup>1</sup> Present address: School of Biological Sciences, University of Auckland, Auckland, New Zealand.

<sup>2</sup> Present address: The Institute for Systems Biology, Seattle, Washington.

Key words: HIV, multidimensional scaling, likelihood assignment, group prediction.

E-mail: jonand@u.washington.edu.

*Mol. Biol. Evol.* 20(7):1168–1180. 2003

DOI: 10.1093/molbev/msg128

*Molecular Biology and Evolution*, Vol. 20, No. 7,

© Society for Molecular Biology and Evolution 2003; all rights reserved.

Standard phylogenetic and phenetic analyses of these sequences showed little structural correlation with known epidemiological information, with only the geographic area of infection and the risk factor of men who engage in sex with men (MSM) showing any statistically correlated groupings. Although phylogenetic and phenetic methods may often distinguish between groups of structured sequences, they also may fail to identify more subtle relationships between the sequences. To search for this less pronounced structure in the sequence data, metric and nonmetric multidimensional scaling (MDS) was performed. The MDS method takes a set of pairwise dissimilarity measurements for each sample and projects the samples in two or more dimensions of Euclidean space such that the distance between each sample in Euclidean space matches the dissimilarities as closely as possible. Using metric and nonmetric MDS analysis on these data, we were able to show that the sequence information in the C2–V5 region of *env* correlated with the risk groups including men who engage in sex with men and individuals who engage in sex with injection drug users or use injection drugs themselves (IDU/IDX).

The MDS analysis results in the construction of canonical variables that can be further used to predict putative group memberships. A discriminant functions analysis (DFA) (Fisher 1936) can be applied to the canonical variables to give a set of classification functions that allow assignment of the sequences to different groups. The efficiency of these assignments can be tested using cross validation. Using DFA and tree topologies, we were able to predict group membership for the original 100 samples that we sequenced at a rate statistically better than would be expected by chance alone for the factors of MSM, geographical location (AREA), and IDU/IDX. Using a newly developed method of likelihood assignment to predict groups, we could predict the groupings for the factors of MSM and IDU/IDX at a statistically significant level. Finally, we were able to predict the MSM risk group memberships for a blinded set of 19 samples at a rate statistically better than by chance alone by using the likelihood assignment test.

We show that phylogenetic and phenetic analyses, along with the methods of MDS and likelihood assignment can be useful in identifying the underlying structure in sequences containing little signal. Furthermore, we can test the validity of the sample relationships by predicting group membership based on the underlying sequence structure that we detect in the data. Such methods may make it possible to identify viral variants belonging to specific demographic groups by examining only a small portion of the HIV-1 genome. With the effectiveness of antiviral drugs having been correlated to the genotype and phenotype of HIV-1 (Leigh Brown, Korber, and Condra 1999; Falloon et al. 2002), such predictions of demographic epidemiology based on a limited amount of sequence information may become valuable in assigning different treatment regimens to infected individuals. Gaining information on how different viral strains infect subgroups of the population may also prove valuable in future vaccine developments that target specific viral variants.

## Materials and Methods

### Samples

Viral samples used in this study were acquired from 1994 to 1997 by the U.S. Centers for Disease Control and Prevention (CDC). Uncultured peripheral blood mononuclear cells (PBMCs) were collected from individuals located in five different cities in the United States along with demographic and risk factor information (unpublished data). Provirus DNA in PBMCs was isolated by using an Iso Quick Nucleic Acid Extraction kit from ORCA Research Inc. The HIV-1 envelope (*env*) C2–V5 region was amplified using primers ED31 and BH2, followed by second-round nested polymerase chain reaction (PCR) using primers DR7 and DR8. The PCR products, HIV-1 *env* C2–V5 region, were cloned into PCRII vector (Invitrogen, San Diego, Calif). The insert of approximately 700 bp of *env* C2–V5 region were identified by EcoRI restriction digestion, and sequenced by using fluorescent-labeled M13 forward and reverse on an ABI 377 sequencer. All proviral DNA isolations were processed in P3 laboratory facilities. The PCR amplifications were done with critical procedural safeguards.

### Risk and Demographic Factors

The groups/risk factors that were examined are given in table 1, along with the number of individuals in each group. All demographic and risk factors that were catalogued by the CDC were used in our analysis. In this report we will refer to the classification variables Age of Individual (AGE), Area of Infection (AREA), Injection Drug User (IDU), Men who Engage in Sex with Men (MSM), Sex with IDU (IDX), Race/Ethnicity of Individual (RACE), Status of Individual upon Initial Detection of HIV-1 (STATUS), IDU or IDX (IDU/IDX), and Gender of Individual (SEX) as *factors*. Each category within a factor (e.g., 15–29 within AGE, or Baltimore within AREA) will be referred to as a *group*.

### Phylogenetic and Phenetic Analyses

The phenetic method of UPGMA and the phylogenetic method of NJ were used to predict putative group structures for these sequences. Both methods were performed prior to obtaining the risk and demographic breakdowns for each individual. On the one hand, the UPGMA method is a hierarchical clustering technique that groups sequences based on similarity alone. The pairwise distance method of NJ, on the other hand, is a phylogenetic technique that groups sequences based on their inferred evolutionary history. Both the phenetic and phylogenetic methods rely on a pairwise distance matrix that describes the observed sequence data. A single pairwise distance matrix was constructed for use by both the phylogenetic and phenetic methods. The matrix was constructed under a general time reversible (GTR) (Lanave et al. 1984) model of substitution that estimated evolutionary distances between pairs of sequences under ML methods. The GTR model of evolution uses nucleotide substitution rates and a gamma distributed amount of site-to-site heterogeneity of

**Table 1**  
**Breakdown of Demographic/Risk Factors of the HIV-1 Infected Individuals Evaluated in This Study**

Factor	Sample Number	Factor	Sample Number
Age Group (AGE)		Area of Infection (AREA)	
15–19 years	7	Los Angeles	25
20–24 years	11	Miami	20
25–29 years	13	New Orleans	10
30–34 years	19	Houston	30
35–39 years	24	Baltimore	15
40–44 years	15		
≥45 years	11		
Injection Drug Use (IDU)		Man Having Sex with Men (MSM)	
Yes to risk	44	Yes to risk	57
No to risk	56	No to risk	43
Sex with IDU (IDX)		Race/Ethnicity (RACE)	
Yes to risk	41	White, not Hispanic	17
No to risk	59	Black, not Hispanic	58
		Hispanic	25
Gender (SEX)		Seroconversion Status (STATUS)	
Male	79	Prevalent positive	66
Female	21	Seroconverter	34
IDU or IDX (IDU/IDX)			
Yes to risk	58		
No to risk	42		

NOTE.—Each of the “factors” (classification variables)—e.g., AGE—is listed with the abbreviated name shown in parentheses. For each factor, the categories which make up the factor, “groups” (e.g., 15–29 years), are indicated along with the number of individuals sampled in each group.

rates that were previously estimated (Anderson et al. 2001). The computer program PAUP\* (Swofford 2002) was used to produce topologies for both the phenetic and phylogenetic analyses.

#### Slatkin/Maddison Test

To determine if there were any statistically significant associations between the UPGMA or NJ topologies and the epidemiological subgroups, we employed a method first described by Slatkin and Maddison (1989, 1990). For each factor, sequences were categorized according to group membership. The number of times sequences in different groups shared a common ancestor was counted. This value represents the most parsimonious number of times that sequences from one group “cross over” to another group: the lower this number the less traffic there is between groups, and the greater the evidence that sequences within a group are more closely related phylogenetically. To test whether this value is significantly smaller than one would expect by chance alone, 10,000 random trees with the same numbers of sequences in each group were constructed, and the numbers of cross overs was counted. The number of steps required by UPGMA or NJ was then compared to the distribution of number of steps produced by the random topologies. Given this distribution, the phylogenetic and phenetically derived topologies were tested to determine if they require fewer changes than 99% of the random trees, corresponding to an  $\alpha$  of 0.01. This analysis was performed using the program MacClade (Maddison and Maddison 1992).

#### Multidimensional Scaling

To search for less pronounced signal in the sequence data and to identify putative group structures, metric and nonmetric MDS analyses were performed. The MDS method tries to reduce the number of dimensions the data is represented from  $n - 1$  dimensions (with  $n$  being the number of samples), to as few as possible while still preserving the relationships between each pair of samples. For a good review of MDS, see Forrest W. Young (1987). With perfectly Euclidean data, all points can be represented perfectly in Euclidean space, so that the distance between any pair of points in Euclidean space is equal to the value in a distance matrix. The distances between sequences can be calculated using a Euclidean model; however, such constructions typically rely on very simple matching criteria, and they fail to take account of hidden substitutions. Consequently, Euclidean distances are unable to adequately explain the complexity of HIV-1 evolution. We used sequence distances that were calculated under a GTR model of substitution using ML methods. Because this is an estimation of sequence distances that accounts for variations in the rates of individual nucleotide substitutions and site-to-site heterogeneity, the matrix that is produced will likely not maintain a Euclidean relationship. Consequently, it becomes necessary to identify distances in Euclidean space that are as close as possible to those given in the distance matrix.

The metric MDS analysis uses the identical pairwise distance matrix that was produced for use by phenetic and phylogenetic analyses. Again, this distance matrix was produced using an ML GTR model of evolution with

substitution rates previously estimated (Anderson et al. 2001). The metric MDS analysis was performed using the program ViSta (Young 1996). For each possible dimension, MDS assigns a specific location in Euclidean space to every sample such that the distance between every pair of samples in Euclidean space matches the pairwise distance matrix as closely as possible. The nonmetric MDS analysis used the ranked order of the sequence distances to form the spatial relationships between samples. Again, the nonmetric analysis was performed using the program ViSta.

The MDS method identifies the subspace that best preserves the pairwise distances for the sequence data. The dimensions of the MDS analysis are produced with the first dimension accounting for the largest proportion of the data's variance and the last dimension accounting for the smallest proportion. After the assignment of spatial coordinates to every sample for each of the  $n - 1$  dimensions, the number of dimensions is reduced to those that are considered nontrivial and interpretable. The goal of dimension reduction is to determine the minimum number of dimensions that can be used to closely approximate the data.

Classically, a scree plot is used to identify the number of nontrivial dimensions. The scree plot indicates the amount of variance associated with each dimension. The number of dimensions to use may then be determined by eye, by looking for an elbow or bend in the curve, with all dimensions occurring before this bend being considered nontrivial and interpretable. In many data sets, however, the variance decreases smoothly with increasing dimensionality, making the choice of dimensionality difficult and ultimately subjective. A more objective method for determining the number of nontrivial dimensions is the broken-stick analysis (Jackson 1993). A broken-stick model is first produced, which represents the analysis of random data. The observed dimensions are then considered nontrivial and interpretable as long as the amount of variance explained by the given dimension exceeds the value generated by the broken-stick model. The broken-stick model of variances is determined by the equation:

$$b_k = \sum_{i=k}^p \frac{1}{i}$$

where  $p$  is the number of potential dimensions and  $b_k$  is the amount of variance explained for the  $k$ th dimension.

#### Discriminant Function Analysis

To predict group membership within the demographic factors of MSM and IDU/IDX, we employed the method of discriminant function analysis (DFA) (Fisher 1936). The DFA method finds distance formulas that describe the relationship of the multivariate means of the responses relating to each of the group categories. The analysis takes the multidimensional coordinates of the known (training set) samples and devises a set of equations that will predict classification functions based on the coordinates of an unknown sample. Thus DFA assumes that the sample observations are random, that each group is normally distributed, that the variance of each group is the same, and

that each of the observations in the training set is correctly classified.

The DFA was performed using the program JMP (SAS 1995) on the first 23 dimensions of the metric MDS data and the first 12 dimensions of the nonmetric MDS data. Again, metric and nonmetric MDS analyses were performed using the program ViSta (Young 1996). For each possible dimension, metric MDS assigns a specific location in Euclidean space to every sample such that the distance between every pair of samples in Euclidean space matches the pairwise distance matrix as closely as possible. The nonmetric MDS analysis used the ranked order of the sequence distances to form the spatial relationships between samples. The number of dimensions used in the DFA analysis was previously determined by using stopping rules defined by the broken-stick model (Jackson 1993). The broken-stick model determines the number of dimensions that are considered nontrivial and interpretable for both the metric and nonmetric MDS analyses.

To test the ability of DFA to predict group membership of an unknown sample, we performed a  $U$  test (Sharma 1996) on the metric and nonmetric MDS data. The  $U$  test is a cross-validation procedure that can evaluate the predictive power of the DFA by using  $n - 1$  samples in the DFA training set, and using the remaining sample as the experimental unknown. This procedure is repeated, using each of the samples once as the unknown while using the other  $n - 1$  experimental samples as the new training set. Because the group designation is actually known for the experimental unknown samples in the  $U$  test, the ability of the DFA to predict group membership can be evaluated. A chi-square analysis of the predicted group category for each individual can be compared to the known category to determine if the group category assignments can be predicted significantly better than by chance alone. For our sample set, 99 sets of MDS coordinates were used as the training set, and a total of 100  $U$  tests were performed for both the nonmetric and metric MDS analyses.

#### Group Prediction Based on Tree Topology

To predict group membership within the factors of MSM and AREA, we analyzed the tree topologies that had previously been shown to form significant clustering of the groups within each of these factors. These topologies were produced by the phylogenetic method of Neighbor-Joining, and the phenetic method of UPGMA.

The predicted group membership of a sample was determined by the group membership of the most recent common ancestor to that sample. If the most recent common ancestor's group designation was ambiguous, then a prediction was not produced. To evaluate the predictive power of each topology, a  $U$  test was performed (Sharma 1996). A chi-square analysis was used to compare the predicted groups with the known groups and determine if the group predictions are significantly better than by chance alone. For each of our  $U$  tests, 99 samples were used as the training set, and a total of 100  $U$  tests were performed.

### Group Prediction Based on Likelihood Assignment

Another method used to predict group membership within the demographic factors was the likelihood assignment test. The likelihood assignment test compares the amino acid sequence of an unknown sample to that of the known samples, assigns a likelihood score to each group within a demographic factor, and determines which group is most similar to the unknown sample. This test determines the likelihood that amino acid sites along an unknown sequence correspond to sites within a subgroup of sequences that share a common feature.

The likelihood of an unknown sample belonging to a group was calculated for each group of a given demographic factor. The log-likelihood of belonging to a group,  $g$ , is given by:

$$\ln L(g) = \sum_{i=1}^s \ln \left( \frac{x_i}{n_g} \right)$$

where  $s$  is the number of sites along the unknown sequence,  $n_g$  is the number of sequences in group  $g$ , and  $x_i$  is the number of sequences in group  $g$  that have the same nucleotide/amino acid as the unknown sequence at site  $i$ . The unknown sequence is assigned to the group for which  $\ln L(g)$  is maximized. If for a given character the unknown sample did not share an amino acid with the group, then the log likelihood for that character was calculated to be  $\ln(1/(z + 1))$ , where  $z$  is equal to the number of members of the largest group for a given factor, thereby correcting any zero values in the denominator of the original equation. Rather than assigning a constant value for all data sets, we chose to have the zero denominator correction equation assign a likelihood that is slightly worse than would be achieved by having only a single group and sharing an amino acid with only one of the members. This analysis assumes that (1) each site in the sequence is independent of all other sites and (2) the observed or sampling frequency of an amino acid at a particular site in a given subgroup is the best estimate of the true frequency with which that amino acid is found at that site in an unknown sequence from that subgroup. Technically, the likelihood calculated is the conditional probability of obtaining the unknown sequence given subgroup membership and the frequencies of amino acids at each site. The latter are, however, nuisance parameters and have been excluded from the notation.

To test the ability of the likelihood assignment test to correctly predict group membership of an unknown sample, we performed  $U$  tests (Sharma 1996) for each of the demographic factors. A chi-square analysis of the predicted group category for each sample was compared to the known category to determine if the group category assignments can be predicted significantly better than by chance alone. For each of our  $U$  tests, 99 samples were used as the training set, and a total of 100  $U$  tests were performed.

### Prediction of Unknown Samples

A cross-validation analysis can help determine if our group prediction procedures are working correctly. However, this type of validation only looks at the set of

samples that were used to help design the prediction procedures in the first place. A true test of these methods requires a new, blinded set of unknown samples that can be used in an attempt to predict the group memberships. A total of 21 blinded samples were obtained from the CDC. Of these samples, 19 were amplified, cloned, and sequenced. For each of these 19 samples, group predictions were made for the factors of AREA, MSM, and IDU/IDX. The predictions were produced by using the methods of DFA, group prediction based on tree topology, and the likelihood assignment test. After these analyses, the true group designations were revealed and a Fisher's exact test, or chi-square analysis for the multicategorical group of AREA, was performed to determine if the group category assignments can be predicted significantly better than by chance alone.

### Results

#### Phenetic and Phylogenetic Reconstruction of Sequence Data

The phenetic methods of UPGMA and the phylogenetic method of NJ were used to predict risk and demographic factor associations for the 100 HIV-1 subtype B sequences from the United States. To assess the confidence level associated with the UPGMA and NJ topologies, bootstrapping (Felsenstein 1985) was performed on each method. The bootstrapping analyses on the UPGMA and NJ topologies show similar results with little supported structure and a few small group associations containing bootstrap support greater than 70% (figs. 1 and 2). The NJ and UPGMA analyses each identified several of the same groups, with the NJ analysis identifying two more pairs of sequences than were found in the UPGMA analysis. The groups in common between the two analyses include the following: 96US1548 & 96US2394; 94US7948, 94US8655 & 96US3398; and 95US1038 & 95US3551. Each analysis was only able to identify groups as large as three members with bootstrap support greater than 70%. Definitive groups of more members could not be identified with confidence using these methods.

#### Analysis of Risk and Demographic Group Reconstructions

To evaluate each of these methods in their ability to identify the risk and demographic groups, an analysis method by Slatkin and Maddison of tree randomization was performed (Slatkin and Maddison 1989, 1990). For each of the predetermined phenetic topologies, the least number of steps required to separate the members of a demographic group completely was determined. The number of steps required by each topology to separate each of the demographic groups was then compared to a distribution of number of steps produced by creating 10,000 random topologies. Given this distribution, the phenetically and phylogenetically derived topologies were tested to determine if they require fewer changes than 99% of the random trees, corresponding to an  $\alpha$  of 0.01 (fig. 3). The results show that UPGMA groups only the factor MSM at a significant level, whereas the NJ analysis

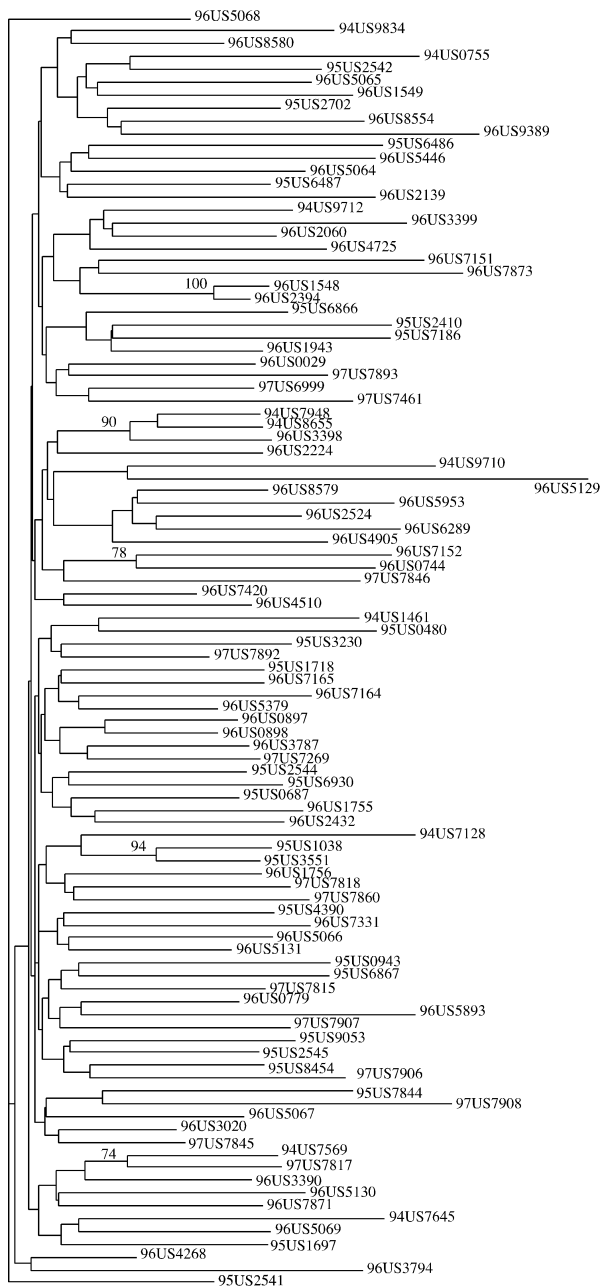


Fig. 1.—The phylogenetic relationship of 100 HIV-1 clade B *env* gp120 C2–V5 viral sequences is shown. The distance method of NJ was used to construct the tree topology. The NJ method used pairwise distances calculated under a maximum likelihood GTR model of evolution which has been previously described (Anderson et al. 2001). All bootstrap values of 70% or greater are indicated on the tree.

produces a topology which groups the factor of AREA significantly better than the random topologies.

### Multidimensional Scaling

Multidimensional scaling analyses were used to represent the relationships of the data set in n-dimensional Euclidean space in an attempt to identify putative group structures. Without information on the risk and demographic factors of the data, the MDS analysis was unable to identify definitive groupings, but it did provide a struc-

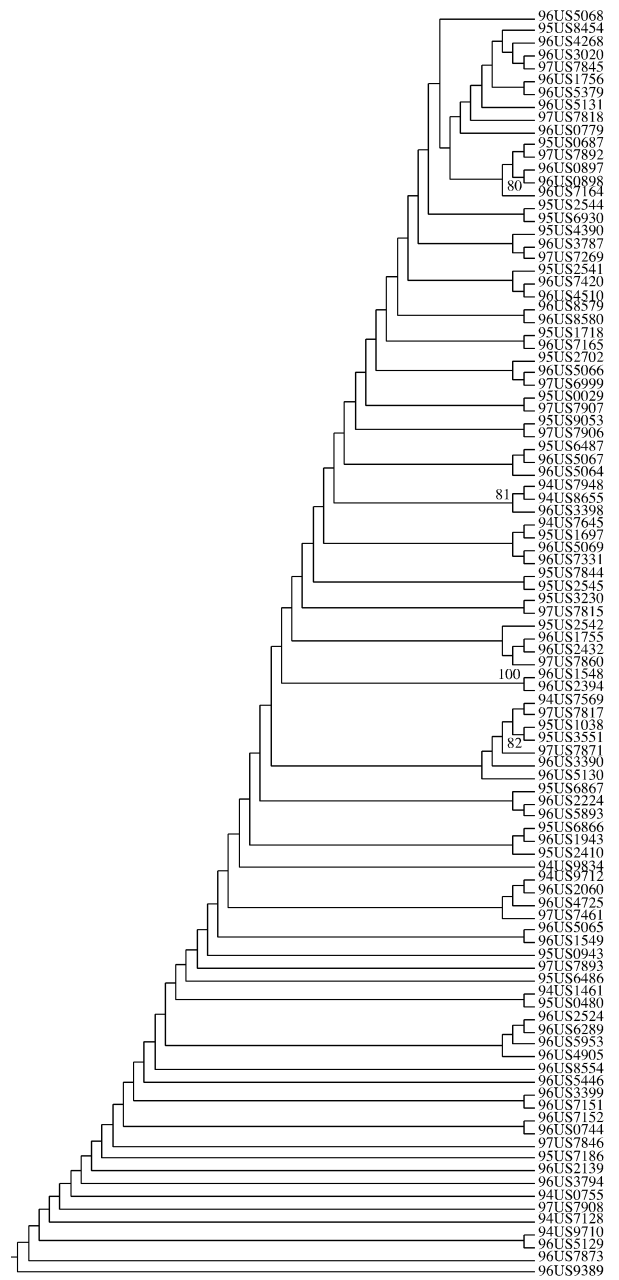


Fig. 2.—The phenetic relationship of the 100 HIV-1 clade B *env* gp120 C2–V5 viral sequences is shown. The hierarchical clustering method of UPGMA was used to construct the topology based on sequence similarity. The UPGMA method used the same maximum likelihood GTR pairwise distances as the NJ method. All bootstrap values of 70% or greater are indicated.

ture to represent the data in the form of 99 dimensional coordinates for each sample. To identify the number of nontrivial and interpretable dimensions, the amount of variance explained by each individual dimension was analyzed against a broken-stick model (Jackson 1993) (fig. 4). The number of nontrivial dimensions was determined to be 23 for the metric MDS and 12 for the nonmetric MDS. These nontrivial dimensional coordinates were used to identify factors that contained groups that correlated with the variations in the data set. The analysis was done using the computer program JMP (SAS 1995) by

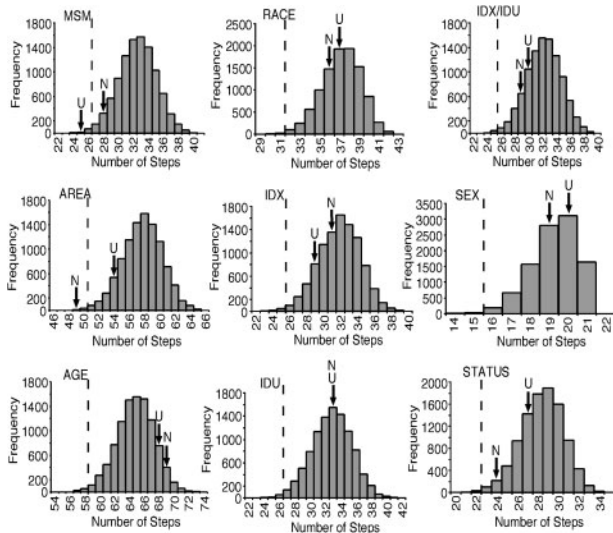


FIG. 3.—Evaluation of phenetic and phylogenetic methods used to infer groups. The method of Slatkin and Maddison (1989, 1990) was used to determine if the level of within-group clustering was significantly greater than one would expect by chance alone. For each factor, a histogram of 10,000 random topologies was produced, showing the number of times sequences in different groups shared a common ancestor. The histogram was compared to the corresponding numbers produced by the NJ (N) and UPGMA (U) topologies, which are indicated by arrows. Significant group clustering was determined when one of the derived topologies contained a fewer number of shared common ancestors than 99% of the random topologies. The dashed line represents this critical value of  $\alpha = 0.01$ .

performing a multiple analysis of variance (MANOVA) analysis on the dimensional coordinates generated from the MDS analyses (table 2). The MANOVA analyses compare the mean values of multiple sets of dependent responses that are generated from two or more independent factors. In other words, the MANOVA analysis determines if the positions of the centroids of groups in multidimensional Euclidean space are statistically separated. The MANOVA results indicate that the groups within the factors of MSM and IDU/IDX have significant correlation to the first 23 dimensions of the metric MDS data, and the first 12 dimensions of the nonmetric MDS data. To verify that the MANOVA analysis was not critically sensitive to the number of dimensions studied, we analyzed from 3 to 40 dimensions for the factors of MSM and IDU/IDX. The MANOVA results were statistically significant for dimensions 3 or 4 and higher for the factor of IDU/IDX using nonmetric and metric MDS, respectively. The factor of MSM showed significance with a MANOVA analysis of dimensions 18 or higher for metric MDS and dimensions 12 through 18 for nonmetric MDS (see Supplementary Material online).

The methods of MDS are useful in statistically analyzing data sets and identifying sequences that correlate with known group structures. However, visualizing the data in multiple dimensions can be difficult or impossible to perform. Even though the broken-stick model may identify several nontrivial dimensions, the first few dimensions, which contain the largest proportion of the data's variance, may begin to provide sufficient information to evaluate the relationships of the samples. The first two dimensions of

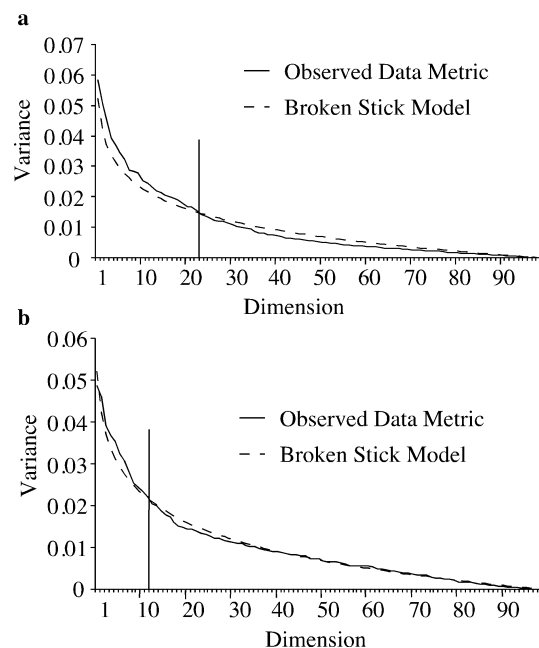


FIG. 4.—Broken-stick model evaluation of the (a) metric and (b) nonmetric MDS analysis. The amount of variance accounted for by each dimension of the MDS analysis is plotted against the broken-stick model. The point at which the two lines first intersect indicates the number of nontrivial dimensions. The intersection point is indicated by a vertical line. The y-axis shows the sample variance accounted for by each dimension, with the number of each dimension shown along the x-axis.

the metric MDS analysis for the risk factor MSM were plotted to search for relationships within the groups (fig. 5). The MDS plot shows that the groups within the factor of MSM are overlapping, but that they do maintain regions of separation.

#### Evaluation of Tree Construction Methods in Predicting Groups

Phylogenetic and phenetic tree construction methods have previously been shown to significantly cluster the groups of MSM and AREA (fig. 3). Each of these topologies can be used to predict the group designation of a sample by assigning the unknown sample the same designation as its most recent common ancestor. To evaluate the usefulness of these topologies in predicting group membership, a resampling test was performed on the tree construction data. The group designation for each of the samples in the data set was predicted by using the remaining  $n-1$  samples to establish the group membership of the most recent common ancestor. The predictions were compared to the known group designations using a chi-square analysis, which determines if the predictions are significantly better than by chance alone. The results of the resampling test show that the groups within each of the factors tested can be predicted significantly better than by chance alone (table 3).

#### Evaluation of Discriminant Function Analysis in Predicting Groups

To evaluate the ability of DFA to assign group membership correctly for the factors of MSM and IDU/IDX,

**Table 2**  
**Results Comparing (Top) the First 23 Dimensional**  
**Coordinates of the Metric MDS Analysis and (Bottom)**  
**the First 12 Dimensional Coordinates of the Nonmetric**  
**MDS Analysis Against the Group Designations for Each**  
**of the Various Factors**

Factor	<i>P</i> Value	$\alpha < 0.01$
AREA	0.0952	
AGE	0.9996	
STATUS	0.5798	
IDU	0.0116	
IDX	0.0998	
MSM	0.0023	*
RACE	0.8961	
SEX	0.2459	
IDU/IDX	<0.0001	*
AREA	0.1020	
AGE	0.8738	
STATUS	0.5334	
IDU	0.0276	
IDX	0.0663	
MSM	0.0094	*
RACE	0.4020	
SEX	0.0638	
IDU/IDX	0.0003	*

NOTE.—The MANOVA tests show that the factors MSM and IDU/IDX are significantly correlated to the MDS data ( $\alpha = 0.01$ ).

a *U* test (Sharma 1996) was performed on the metric and nonmetric MDS data. For each of the factors, the group designation of each sample was predicted while using the remaining 99 samples as the DFA training set. These predicted group designations were then compared to the known group designations by means of a chi-square analysis. The results of the DFA on the metric and nonmetric MDS data show that the demographic factors of MSM and IDU/IDX can be predicted significantly better than by chance alone (table 3). The ability of the DFA to identify the correct grouping assignments may provide information that is significantly better than by chance, but the analysis could not predict the group membership for any of the factors at a rate higher than 73% of the time.

#### Evaluation of the Likelihood Assignment Test in Predicting Groups

To evaluate the ability of the likelihood assignment test to correctly assign group membership for each of the factors, a *U* test (Sharma 1996) was also performed. For each of the factors, the group designation of each sample was predicted, while the remaining 99 samples were used as the likelihood assignment test training set. The predicted group designations were then compared to the known group designations by chi-square analysis. The results of the likelihood assignment test on the data show that the demographic factors of MSM and IDU/IDX can be predicted significantly better than by chance alone (table 3). The likelihood assignment test was able to predict the group designations for the factor of MSM 63% of the time and IDU/IDX 68% of the time.

#### Combining Test Results to Enhance Predictions

To attempt to increase the predictive ability of our analyses, we combined the results of the various prediction

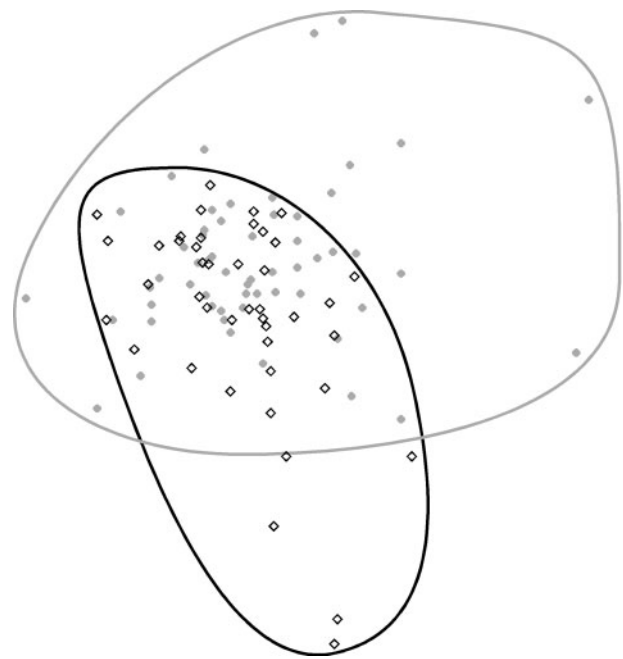


FIG. 5.—Metric multidimensional scaling scatter plot for the risk factor MSM (men who engage in sex with men). The first two-dimensional coordinates from the metric MDS analysis are plotted with each point representing an individual sample. Gray circles indicate individuals exposed to the risk factor MSM, and black squares indicate individuals not exposed to this risk factor. The large black and gray circles encompassing the samples are shown as references to highlight the relationships of the two groups.

methods and performed a *U* test on the combined results. The combination tests were performed on the factors of MSM and IDU/IDX, which were previously shown to have their groups separated significantly better than by chance alone (table 3). The factor of AREA could be separated only by the method of NJ tree topology assessment and thus was not applicable for a combined analysis.

Further *U* tests were performed that combined the results of the DFA analyses on the metric and nonmetric MDS data, or all analyses used in predicting the factors of MSM and IDU/IDX. For the DFA combination, the prediction was accepted if the predictions from each of the DFA analyses agreed. If the two DFA predictions disagreed, the prediction was classified as undeterminable. For a combination of all of the analyses, *U* tests were performed that classified groups when either all the analyses agreed or when a majority of the analyses agreed.

These combination tests resulted in the factors of MSM and IDU/IDX being predicted significantly better than by chance alone (table 3). The ability to predict the group membership for each of the factors was also increased. The use of all possible tests was found to perform better than using the combination of the DFA results alone.

#### Predicting Unknown Sample Group Designations

A blinded set of 19 samples was analyzed in an attempt to predict the group designations for the factors of MSM, AREA, and IDU/IDX. The sequences of these samples were added to the alignment of the 100 known sample set, and predictions were made using the methods

**Table 3**  
Results of *U* Tests Run on the Various Methods of Predicting Group Designation

Test	Group	% Correct	$\chi^2$ Value	Number
Metric MDS	MSM	65%	0.0018	100
Metric MDS	IDU/IDX	73%	<0.0001	100
Nonmetric MDS	MSM	62%	0.0113	100
Nonmetric MDS	IDU/IDX	72%	<0.0001	100
Combined MDS	MSM	70.8%	0.0004	65
Combined MDS	IDU/IDX	80%	<0.0001	75
Tree Topology NJ	AREA	45.7%	0.0012	59
Tree Topology UPGMA	MSM	66.6%	0.0085	66
Likelihood Assignment	MSM	63%	0.0055	100
Likelihood Assignment	IDU/IDX	68%	0.0002	100
All Tests Agree	MSM	76.9%	0.0003	39
Majority Agree	MSM	70.7%	<0.0001	89
All Tests Agree	IDU/IDX	81.3%	<0.0001	59
Majority Agree	IDU/IDX	76%	<0.0001	100

previously described. Each of the methods predicted group designations for the unknown samples by analyzing each unknown separately in conjunction with the training set of 100 known samples.

After the predictions for the blinded samples were made, the true groupings were revealed and each method of prediction was evaluated using the Fisher's exact test or a chi-square analysis. For these 19 samples, the only method that provided a statistically significant number of correct group predictions at an  $\alpha$  of 0.05 was the method of likelihood assignment for the factor of MSM (table 4).

## Discussion

Phylogenetic and phenetic methods have often been used in HIV-1 research to establish subtype nomenclature, track the movement of the virus, and estimate the growth of the epidemic. The same methods have also largely been unable to correlate patterns of genetic diversity with risk and demographic factors in the analysis of a single subtype. Our own phylogenetic analysis of 100 HIV-1 subtype B *env* sequences from the United States resulted in a star-like topology with short interior branches leading to long terminal branches (fig. 1). Such star-like topologies are often produced when the population under study has experienced exponential growth. Initial observation of the phylogenetic NJ topology showed little structure and, with bootstrapping analysis, produced only a few supported sets of two to three sequences. A phenetic analysis using UPGMA produced a similar result, giving a topology that was not well supported by bootstrapping analysis and could only find small supported sets of two to three sequences (fig. 2).

Bootstrapping can identify the data's relative support for a given topology, but it is also important to determine if the topology contains any statistically significant associations to the groups within each factor. The Slatkin/Maddison test (Slatkin and Maddison 1989, 1990) was used to do this kind of determination by measuring the degree of group separation within each topology. The results indicated that the NJ topology contained strong associations to groups within the factor of AREA, with only 0.14% of the random topologies separating the

**Table 4**  
Results of Predicting the Group Designations for 19 Unknown Samples

Test	Group	% Correct	<i>P</i> Value	Number
Metric MDS	MSM	52.6%	0.6300	19
Metric MDS	IDU/IDX	57.9%	0.4443	19
Nonmetric MDS	MSM	36.8%	0.9451	19
Nonmetric MDS	IDU/IDX	31.5%	0.9851	19
Tree Topology NJ	AREA	20.0%	0.2929	10
Tree Topology UPGMA	MSM	64.3%	0.1538	14
Likelihood Assignment	MSM	78.9%	0.0149	19
Likelihood Assignment	IDU/IDX	52.6%	0.5557	19
Majority Agree	MSM	55.5%	0.4367	18
Majority Agree	IDU/IDX	47.4%	0.7801	19

NOTE.—*P* values were determined by the Fisher's exact test, except for the multicategory group of AREA, which was evaluated by a chi-square analysis.

groups as well as the NJ topology (fig. 3). The idea that viruses isolated from a given city are evolutionarily more closely related than viruses isolated from different cities should not be surprising. If a limited number of founder strains were established in each city, then samples from any given city should be more closely related to each other than to samples from different cities. In figure 6, we show the phylogenetic tree with sequence labels replaced by the cities from which these sequences were obtained. It is apparent that there is some degree of clustering with sequences from each of these cities.

The Slatkin/Maddison test was also performed on the UPGMA tree, and the results indicated that only the factor MSM showed significant clustering. Because a UPGMA tree is based on phenetic similarity, significant clustering indicates that sequences from the same group are more similar to each other than to sequences from other groups. This result suggests that the route of infection may be linked to the genotype of the virus, or that there is an underlying linkage between men who engage in sex with men that is not readily detected using phylogenetic techniques. To further explore this result, we compared the distribution of pairwise sequence distances within and between the groups of MSM. By looking at the intragroup and intergroup pairwise distributions, we could determine the extent of divergence between the two groups. If the intragroup and intergroup distances were completely separated, then group designation could be determined by the pairwise distance alone. As the overlap between the distributions increases, however, the ability to establish a group designation based solely on pairwise distance decreases. The results show that the intragroup and intergroup distributions greatly overlap, indicating that phenetic similarities may not be able to resolve the different groups easily (fig. 7).

Standard evolutionary methods can often distinguish group structure by looking for a strong phylogenetic or phenetic signal within the data. However, these methods may begin to fail when the signal is less pronounced. The MDS method, on the other hand, can be used to search for the less pronounced signals in the sequence data and to distinguish group structures based on these signals. Dimension-reduction nonhierarchical ordination methods such as MDS often are particularly useful when there is some a priori information that classifies the samples into

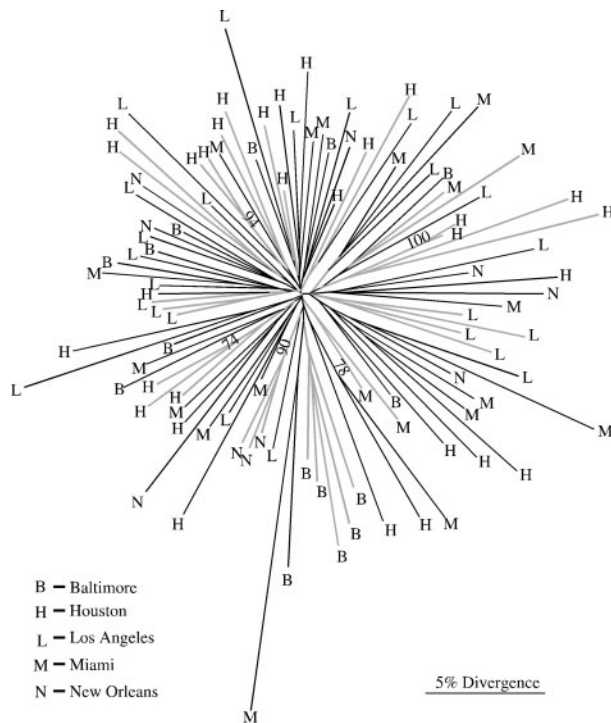


FIG. 6.—Phylogenetic relationship of the 100 HIV-1 subtype B sequences is shown. Each of the sequences has been labeled by the city from which the sample was obtained. Sequences from the same city that share a most recent common ancestor are indicated by gray branches. Bootstrap values of 70% or greater are shown.

different groups. When such information is available, it is frequently (but not always) the case that different regions of the data cloud are categorized into different groups. In this respect, it offers greater flexibility than hierarchical clustering methods when searching for structure, because it allows a “fuzzier,” and possibly more natural, means of identifying such structure. Metric MDS gives more weight to outlier samples, which may be greatly divergent from the main group of samples. Nonmetric MDS deemphasizes outliers, but it also helps to deconvolute a group of samples in order to find fine structures within the larger group.

Using the group information available for each of the different factors, we were able to identify several groups that correlated with regions of the data cloud. The number of nontrivial dimensions produced by the MDS analysis was determined by a broken-stick model analysis (Jackson 1993). This analysis indicated that the first 23 dimensions may hold information capable of distinguishing group associations within a given factor for metric MDS while the first 12 dimensions were considered nontrivial and interpretable for nonmetric MDS (fig. 4). MANOVA analyses on the 23 dimensional coordinates of the metric MDS and the first 12 dimensional coordinates of nonmetric MDS determined that the factors MSM and IDU/IDX contained greater within-group associations than would be expected by chance alone (table 2). These results may indicate that there is an underlying linkage between the group members of the factors of MSM and IDU/IDX that extends across the different cities used in

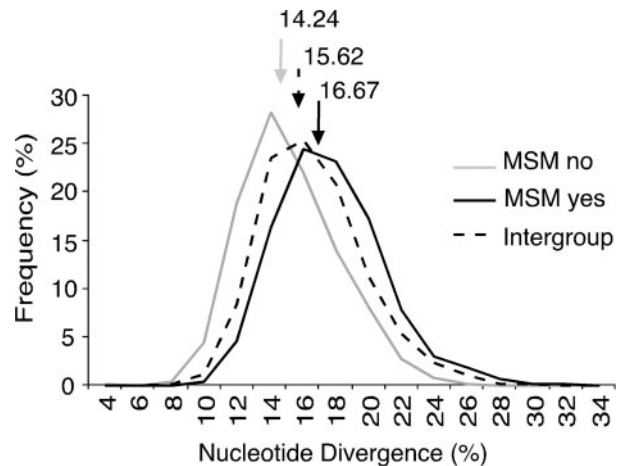


FIG. 7.—Frequency polygon showing the nucleotide diversity both between and within subgroup categories for the factor of MSM. The average diversity for each group is indicated by a number and corresponding arrow.

this study, but it is not readily detectable using standard phenetic and phylogenetic methods.

As we were able to identify significant signal within the factors of MSM, AREA, and IDU/IDX using various techniques, we next attempted to use this signal to predict putative group memberships within these factors. The ability to correctly predict group membership should help to validate our ability to identify the underlying sequence structure that we detect. To test the effectiveness of the various prediction methods that were used, we performed *U* tests (Sharma 1996) on each method and also analyzed 19 blinded unknown samples. The *U* test is an iterative analysis of the predictions made by each prediction method. We used  $n-1$  samples whose group designation was known as the training set for each methodology, with the final  $n$ th sample acting as the experimental pseudo-unknown. We then repeated the procedure, using each of the samples once as the pseudo-unknown while using the other  $n-1$  experimental samples as the new training set. In this way, the ability of each method to correctly predict group membership can then be evaluated, because the group designation is actually known for each pseudo-unknown sample used in the *U* test. A chi-square analysis was used to compare the predicted groups with the known groups and determine if the group predictions are significantly better than by chance alone.

Our analyses of a set of 100 HIV-1 subtype B *env* sequences from the United States produced star-like topologies that contained little bootstrap support, but still formed significant clusters for groups within the factors of MSM and AREA (figs. 1 and 2). This clustering effect indicated that there was some underlying signal within the data, even though it was not well supported by bootstrap analysis. The group membership of the most recent common ancestor to a sample proved to be a good predictor of the sample's true group membership. Using this approach, we show that this underlying signal can serve to correctly predict group membership significantly better than by chance alone (table 3). Using the *U* test, these methods could correctly predict the group

membership of samples significantly better than a random selection method; however, they could only predict the group designation at a rate of 66.6% for MSM and 45.7% for AREA (table 3). Unlike the other factors that were used for these group predictions, the factor of AREA contains a total of five groups. This increase in the number of possible group designations decreases the probability that a random selection method would choose the correct group. Therefore, even though the groups of AREA were correctly assigned at a rate of 45.7%, they were correctly assigned as a rate much greater than the 20% rate expected in a purely random selection process. However, when the 19 unknown samples were analyzed using this method, 64.3% of the MSM designations were made correctly, but only 20% of the AREA designations were correct (table 4). The MSM predictions for the unknowns were close to the same rate as identified in the *U* test, but the AREA predictions seem to fall in the realm of random chance.

These tree construction methods can produce meaningful topologies that cluster demographic groups together even though the sequence data do not contain strong phylogenetic signal. However, the lack of strong signal may still adversely affect these phylogenetic and phenetic methods. In contrast, MDS is used to search for the less pronounced signals in the sequence data and to distinguish group structures based on those signals. Therefore it may be possible for MDS to serve as a better predictor of group membership when studying sequences that contain less pronounced signals. Both metric and nonmetric MDS methods were used in the analysis of the sequence data. Metric MDS uses pairwise distance measurements from the sequence alignment to assign a specific location in Euclidean space to every sample such that the distance between samples in Euclidean space matches the pairwise distance. Nonmetric MDS uses the ranked order of the sequence distances to create the spatial relationships between all of the samples. The spatial coordinates found in the metric and non-metric MDS analyses were used by DFA to predict group designations. DFA produces a prediction for an unknown sample based on the information provided in the training set of samples that have been previously analyzed. The *U* test predictions based on the metric and nonmetric MDS analyses indicate that the groups within factors of MSM and IDU/IDX can be correctly assigned better than by chance alone. Again, like the tree construction methods, the methods of MDS could not correctly predict group designations better than 65% for MSM and 73% of the time for IDU/IDX (table 3). The analysis of the 19 unknown samples produced even fewer correct predictions, with IDU/IDX being predicted only 57.9% of the time and MSM 52.6% of the time (table 4).

To increase the DFA prediction rates for the factors of MSM and IDU/IDX, the predictions from the metric and nonmetric MDS analyses were compared. Predictions for group designation were accepted only if both the metric and nonmetric methods provided identical results. This more stringent method of group designation increased the correct *U* test prediction rate for both factors: MSM was predicted at a rate of 70.8% and IDU/IDX at a rate of 80% (table 3). However, this stringent method also provided no group designation for several samples because of a dis-

agreement between MDS predictions, providing predictions for only 65 of 100 MSM samples and 75 of 100 IDU/IDX samples (table 3).

Finally, a new method for predicting putative group memberships was developed in an attempt to better utilize the underlying signals within the samples and to use these signals to assign group memberships. The method, termed the *likelihood assignment test*, uses the amino acid frequency profile from each sample and assigns a group to an unknown sample based on the profiles of other known samples. *U* tests performed on this method showed that the likelihood assignment test could predict the group membership for the factors of MSM and IDU/IDX significantly better than by chance alone (table 3). The likelihood assignment test was also shown to correctly predict groups for the factor of MSM at a statistically significant level when analyzing the 19 unknown samples (table 4).

The underlying groups for the demographic factors of MSM and IDU/IDX were found to be significantly correlated to the sequence data by the phylogenetic and phenetic tree construction methods, the likelihood assignment test, and the methods of MDS. Because several different methods were used to correctly predict the group designations for these factors, we used the consensus predictions to try to increase our ability to correctly predict group memberships. The results from these consensus prediction methods show that the groups were correctly assigned at a significant rate for the *U* test analysis but were no better than chance alone for the 19 unknown samples (tables 3 and 4).

None of the analyses gave predictions for the blinded set of samples as well as they did on the previous *U* tests, even though in theory the two should give similar results. One possible explanation for the discrepancy lies in the time of sample isolation for the different sample sets. The 100 sample training set that was used for all of the *U* tests was collected in the years 1994 to 1997, while 17 of the 19 blinded samples were collected in 1998. Because of the high rate of mutation in the region analyzed (about 0.88% per year in the *env* C2–V5 region [Shankarappa et al. 1999]), the ability to make predictions on the more recent isolates may be hindered when using an older training set. Any recombination taking place within this region could also reduce the ability to predict group membership.

Phylogenetic and phenetic methods of analyzing sequence data often rely on strong signal within the data to construct highly supported topologies. However, these methods may begin to fail when the signal is less pronounced, and many of these analyses are unable to form highly supported topologies when working with unlinked HIV-1 samples from a single subtype. Topology-based analyses of our data could only identify small groups of two to three sequences with high bootstrap support, leaving a majority of the sequences grouped with only little support. Using the Slatkin/Maddison test as an exploratory analysis, we were able to show that the factors AREA and MSM contained risk/demographic groups that were separated to a significant degree by either the phenetic or phylogenetic topologies. Finally, we have shown for the first time a statistical correlation, linking individuals from cities across the United States who share

a particular HIV-1 risk factor, using the method of MDS. These methods indicated that patterns of genetic diversity can be explained to some extent by the risk factors of MSM and IDU/IDX.

In investigating the relationships between viruses from across the United States, we developed a new method that uses a likelihood assignment test to look for correlations between sequences. Other assignment tests have been developed in population genetics that use genotype information to assign individuals to a population (Paetkau et al. 1995, 1997). To date, the only type of membership assignment test being used in HIV research is VESPA (Korber and Myers 1992; Ou et al. 1992). VESPA, however, only looks for the presence of signatures within the sequence data and does not take the frequencies of all mutations into account when making predictions about group membership. The use of our likelihood assignment test should aid the study of the genetic relationships of HIV and of other organisms that quickly mutate.

The methods of the likelihood assignments test, DFA and the phylogenetic and phenetic tree construction analyses have been shown to successfully predict demographic group designations for HIV-1 subtype B samples based solely on *env* C2–V5 sequences. The ability to correctly predict various demographic groups based on only a limited amount of sequencing could enable the use of custom treatment regimens or vaccine formulations based on the specific viral variant affecting the population. The classification of various demographic groups may also aid in HIV-1 research dealing with viral migration patterns and population dynamics. Overall, the ability to identify structure and predict group designations in sequences that contain little signal is an important step in better understanding the dynamics of the HIV-1 pandemic.

### Acknowledgments

This work is supported by grants from the Centers for Disease Control and Prevention (CDC), The University of Washington Center for AIDS Research (CFAR), and the U.S. Public Health Service.

### Literature Cited

- Anderson, J. P., A. G. Rodrigo, G. H. Learn, Y. Wang, H. Weinstock, M. L. Kalish, K. E. Robbins, L. Hood, and J. I. Mullins. 2001. Substitution model of sequence evolution for the human immunodeficiency virus type 1 subtype B gp120 gene over the C2-V5 region. *J. Mol. Evol.* **53**:55–62.
- Bonhoeffer, S., E. C. Holmes, and M. A. Nowak. 1995. Causes of HIV diversity. *Nature* **376**:125.
- Burke, D. S. 1997. Recombination in HIV: an important viral evolutionary strategy. *Emerg. Infect. Dis.* **3**:253–259.
- Delwart, E. L., H. W. Sheppard, B. D. Walker, J. Goudsmit, and J. I. Mullins. 1994. Human immunodeficiency virus type 1 evolution in vivo tracked by DNA heteroduplex mobility assays. *J. Virol.* **68**:6672–6683.
- Falloon, J., M. Ait-Khaled, D. A. Thomas et al. (11 co-authors). 2002. HIV-1 genotype and phenotype correlate with virological response to abacavir, amprenavir and efavirenz in treatment-experienced patients. *AIDS* **16**:387–396.
- Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*. **39**:783–791.
- Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**:179–188.
- Gao, F., D. L. Robertson, C. D. Carruthers et al. (12 co-authors). 1998. A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. *J. Virol.* **72**:5680–5698.
- Gao, F., D. L. Robertson, S. G. Morrison et al. (10 co-authors). 1996. The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. *J. Virol.* **70**:7013–7029.
- Howell, R. M., J. E. Fitzgibbon, M. Noe, Z. Ren, D. Gocke, T. A. Schwartz, and D. T. Dubin. 1991. In vivo sequence variation of the human immunodeficiency virus type 1 *env* gene: evidence for recombination among variants found in a single individual. *AIDS Res. Hum. Retroviruses* **7**:869–876.
- Jackson, D. A. 1993. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* **74**:2204–2214.
- Korber, B., and G. Myers. 1992. Signature pattern analysis: a method for assessing viral sequence relatedness. *AIDS Res. Hum. Retroviruses* **8**:1549–1560.
- Kostrikis, L. G., E. Bagdades, Y. Cao, L. Zhang, D. Dimitriou, and D. D. Ho. 1995. Genetic analysis of human immunodeficiency virus type 1 strains from patients in Cyprus: identification of a new subtype designated subtype I. *J. Virol.* **69**:6122–6130.
- Lanave, C., G. Preparata, C. Saccone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* **20**:86–93.
- Leigh Brown, A. J., and A. Cleland. 1996. Independent evolution of the *env* and *pol* genes of HIV-1 during zidovudine therapy. *AIDS* **10**:1067–1073.
- Leigh Brown, A. J., B. T. Korber, and J. H. Condra. 1999. Associations between amino acids in the evolution of HIV type 1 protease sequences under indinavir therapy. *AIDS Res. Hum. Retroviruses* **15**:247–253.
- Louwagie, J., F. E. McCutchan, M. Peeters et al. (10 co-authors). 1993. Phylogenetic analysis of gag genes from seventy international HIV-1 isolates provides evidence for multiple genotypes. *AIDS* **7**:769–780.
- Maddison, W. P., and D. R. Maddison. 1992. MacClade version 3.01. Sinauer Associates, Sunderland, Mass.
- Mansky, L. M. 1996. Forward mutation rate of human immunodeficiency virus type 1 in a T lymphoid cell line. *AIDS Res. Hum. Retroviruses* **12**:307–314.
- Myers, G., B. Korber, B. H. Hahn, K.-T. Jeang, J. W. Mellors, F. E. McCutchan, L. E. Henderson, and G. N. Pavlakis. 1995. *Human retroviruses and AIDS 1995*. A compilation and analysis of nucleic acid and amino acid sequences. Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, N.M.
- Ou, C. Y., C. A. Ciesielski, G. Myers et al. (11 co-authors). 1992. Molecular epidemiology of HIV transmission in a dental practice. *Science* **256**:1165–1171.
- Paetkau, D., W. Calvert, I. Stirling, and C. Strobeck. 1995. Microsatellite analysis of population structure in Canadian polar bears. *Mol. Ecol.* **4**:347–354.
- Paetkau, D., L. P. Waits, P. L. Clarkson, L. Craighead, and C. Strobeck. 1997. An empirical evaluation of genetic distance statistics using microsatellite data from bear (Ursidae) populations. *Genetics* **147**:1943–1957.
- Poss, M., A. G. Rodrigo, J. J. Gosink, G. H. Learn, D. de Vange Panteleeff, H. L. Martin, Jr., J. Bwayo, J. K. Kreiss, and J. Overbaugh. 1998. Evolution of envelope sequences from the genital tract and peripheral blood of women infected with clade A human immunodeficiency virus type 1. *J. Virol.* **72**:8240–8251.

- Robertson, D. L., B. H. Hahn, and P. M. Sharp. 1995. Recombination in AIDS viruses. *J. Mol. Evol.* **40**:249–259.
- Robertson, D. L., P. M. Sharp, F. E. McCutchan, and B. H. Hahn. 1995. Recombination in HIV-1. *Nature* **374**:124–126.
- Salminen, M. O., J. K. Carr, D. L. Robertson et al. (11 co-authors). 1997. Evolution and probable transmission of intersubtype recombinant human immunodeficiency virus type 1 in a Zambian couple. *J. Virol.* **71**:2647–2655.
- SAS. 1995. JMP Statistical Discovery Software. Version 3.1. SAS Institute, Inc., Cary, N.C.
- Shankarappa, R., J. B. Margolick, S. J. Gange et al. (11 co-authors). 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* **73**:10489–10502.
- Sharma, S. 1996. Applied multivariate techniques. John Wiley, New York.
- Slatkin, M., and W. P. Maddison. 1989. A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics* **123**:603–613.
- . 1990. Detecting isolation by distance using phylogenies of genes. *Genetics* **126**:249–260.
- Swofford, D. L. 2002. PAUP\*: phylogenetic analysis using parsimony (\*and other methods). Version 4.0b2a. Sinauer Associates, Sunderland, Mass.
- Wain-Hobson, S. 1995. Virological mayhem. *Nature* **373**:102.
- Young, F. W. 1987. Multidimensional scaling—history, theory, and applications. Lawrence Erlbaum Associates, Hillsdale, N.J.
- . 1996. ViSta: The Visual Statistics System. Version 5.05. University of North Carolina, Chapel Hill, N.C.
- Zhu, T., B. T. Korber, A. J. Nahmias, E. Hooper, P. M. Sharp, and D. D. Ho. 1998. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature* **391**:594–597.

Mike Hendy, Associate Editor

Accepted March 24, 2003