Methods for Investigation of the Relationship between Drug-Susceptibility Phenotype and Human Immunodeficiency Virus Type 1 Genotype with Applications to AIDS Clinical Trials Group 333

Anne D. Sevin,[1] Victor DeGruttola,[1] Monique Nijhuis,[2] Jonathan M. Schapiro,[3] Andrea S. Foulkes,[1] Michael F. Para,[4] and Charles A. B. Boucher[2,a]

[1]*Harvard School of Public Health, Boston, Massachusetts;* [2]*University Hospital, Utrecht University, Utrecht, The Netherlands;* [3]*Stanford University School of Medicine, Stanford, California;* [4]*Ohio State University Medical Center, Columbus*

Use of human immunodeficiency virus (HIV) drug-resistance testing in therapeutic decision making may be aided by understanding the relationship between results of genotypic and drug-susceptibility phenotypic assays. We investigated this relationship by applying 3 different statistical methods—cluster analysis, recursive partitioning, and linear discriminant analysis—to results for 72 patients followed in the Adult AIDS Clinical Trials Group (ACTG) protocol 333. ACTG 333 was a multicenter, randomized trial comparing 2 formulations of saquinavir (SQV) to indinavir (IDV) in patients with extensive hard-gel SQV experience. Data include protease amino acid sequences and 50% inhibitory concentrations for SQV and IDV at baseline. The 3 methods give similar results showing the association of mutations at codons 10, 63, 71, and 90 with in vitro resistance to IDV and SQV. Recursive partitioning is especially useful because it can identify interactions among mutations at different codons and accommodates many types of data as well as missing observations.

Human immunodeficiency virus type 1 (HIV-1) resistance genotyping has recently been associated with improved virological response in patients switching antiretroviral therapy [1]. This success underscores a need to develop methods to relate HIV viral genotype to measures of viral susceptibility to drugs as new drugs are developed and new important mutations are implicated. Such methods will improve the ability to use more-complex genetic sequence information in making treatment decisions or in stratifying clinical trials by viral genotype. Such methods could also help clinicians determine whether virological failure results from resistance or some other factor, select salvage therapies, and identify drug resistance in patients with new HIV-1 infections. As expressed by Mayer [2, p. 2001], "A major effort is needed to standardize and validate [genotypic and phenotypic] assays, develop standardized reporting formats easily understood by practicing clinicians, develop better correlates between drug-resistance mutations and phenotypic susceptibility, and relate drug-resistance mutations or phenotypic drug-resistance levels to subsequent virological responses to combination drugs." Accomplishing these goals requires a system for classification of patients by viral genotype, in which a patient's classification predicts the success of available treatment options. This report describes several methods and their application to drug-resistance data from the Adult AIDS Clinical Trials Group protocol 333 (ACTG 333), a randomized clinical trial that compared 2 formulations of saquinavir (SQV) and indinavir (IDV) among patients with $\geq$1 year of prior exposure to SQV.

Analyses that relate viral genotype to drug-susceptibility phenotype provide several statistical challenges. First, there are a large number of possible mutations, the phenotypic effects of which must be considered. For example, the protease region of the HIV genome has 99 codons (297 nucleotides); and the reverse transcriptase (RT) region has ~560 codons (1680 nucleotides). In addition, the occurrence and effect of mutations at any given codon are influenced by the presence of mutations at other codons; therefore, it is necessary to detect interactions among mutations at various codons. Other considerations in the analysis of genotypic data include methods for handling mixtures of amino acids at a codon and for defining the distance between 2 nucleotide or amino acid sequences.

One promising statistical approach to this problem is recur-

Received 13 September 1999; revised 23 March 2000; electronically published 6 July 2000.

Presented in part: 2d International Workshop on HIV Drug Resistance and Treatment Strategies, Lake Maggiore, Italy, 24–27 June 1998 (abstract 58).

Written informed consent was obtained from subjects in accordance with guidelines of the local institutions where the study was conducted and the US Department of Health and Human Services. The study design was approved for implementation by the Adult AIDS Clinical Trials Group, the Community Constituency Group of the AIDS Clinical Trials Group, and all local institutional review boards at study sites prior to study initiation.

Financial support: This work was sponsored by the Adult AIDS Clinical Trials Group and supported by National Institutes of Health grants AI28076-09 and U01 AI38855. Additional support was provided by Merck & Co., Inc., and Hoffman-LaRoche, Inc.

[a] Present affiliation: Stanford University School of Medicine, Stanford, CA.

Reprints or correspondence: Dr. Victor De Gruttola, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave., Boston, MA 02115-6018 (victor@sdac.harvard.edu).

The Journal of Infectious Diseases 2000; 182:59–67

0022-1899/2000/18201-0009$02.00

Table 1. Baseline characteristics of 72 patients enrolled in AIDS Clinical Trials Group protocol 333 for whom baseline genotype and phenotype assays were done.

Variable	Cluster 1 (n = 33)	Cluster 2 (n = 15)	Cluster 3 (n = 24)	All patients (n = 72)
Sex (male/female)	28/5	15/0	21/3	64/8
Median age, years	42	38	41.5	41 (range, 28–65)
Race				
White non-Hispanic	25	11	20	56
Black non-Hispanic	3	1	3	7
Hispanic, Latino	5	3	1	9
History of injection drug use				
No	29	10	21	60
Yes	4	5	3	12
Median CD4 cell count/μL	262.5	164.5	210	230 (range, 3–639)
Median $\log_{10}$ (HIV RNA)	3.8	4.9	4.4	4.3 (range, 2.7–5.3)
Median length of prior SQV use, weeks	92	92	110	99 (range, 53–196)
Median SQV IC_{50}, μM	.007	.029	.016	.009 (range, .005–.373)
Median IDV IC_{50}, μM	.030	.055	.042	.035 (range, .007–.230)

NOTE. HIV, human immunodeficiency virus; IDV, indinavir; SQV, saquinavir.

sive partitioning, also known as classification and regression–tree methodology [3, 4]. This methodology is designed to search for important patterns and relationships and to uncover hidden structure in highly complex and multidimensional data. A number of features of recursive partitioning make it well suited to genotypic data analysis. These features include its ability to select important predictor variables from among a large number of candidates and to identify interactions among predictors. Also of importance is the ability of recursive partitioning to handle all types of data (binary, categorical, and numerical) as well as missing data. Recursive partitioning is more flexible than many other statistical analyses in that prediction can depend on different subsets of the predictors for different classes of individuals. In this article, we illustrate how this statistical tool can provide useful insights into the relationship between HIV-1 drug-susceptibility phenotype and protease genotype for the data from ACTG 333. We also compare the results of recursive partioning to the results from 2 other better-known multivariate statistical techniques—cluster analysis and discriminant analysis.

The ultimate goal of such analyses is to improve the ability of physicians to select optimal therapy based on a patient's viral genotype. Although tables that relate the presence of specific mutations to drug resistance currently exist [5], they have arisen from investigations of mutations at specific codons rather than a systematic approach to examining the effects of all possible mutations and of their combinations. Such investigation must proceed in 2 steps: an exploratory step, in which all potentially important mutations and combinations of mutations are identified, and a confirmatory step, in which the ability of these mutations to predict treatment failure is tested in a statistically rigorous way. An example of the latter is the analysis conducted by the Resistance Collaborative Group (RCG), in which the prognostic value of viral genotype was investigated using results from 12 different studies of different combinations of antiretroviral therapy [6]. For each patient in each study, the number of drugs to which a patient was genotypically sensitive was associated with virological response; genotypic sensitivity was determined by the absence of specific mutations in the patient's HIV genetic sequence. Although this analysis demonstrated that the table of mutations developed by the RCG was useful for prediction, it did not prove that the table was optimal. Furthermore, the complexity of tables of mutations must grow as the number of drugs, and especially drug combinations, increases. The methodology we propose should aid in the development of increasingly sophisticated tables as AIDS therapeutic research evolves.

Methods

Clinical Trial

ACTG 333 was a multicenter, randomized phase II trial designed to determine whether, after long-term (>1 year) treatment with hard-gel SQV (SQV hc), recipients had a decrease of plasma HIV RNA following substitution of this therapy with IDV or soft-gel SQV capsules (SQV sgc). Eligibility criteria included laboratory documentation of HIV-1 infection, $\geq$1 year of exposure to SQV hc at 1800 mg/d, and no prior therapy with any protease inhibitor other than SQV. The data used in this analysis are the results of population genetic sequence analysis (GenBank accession numbers provided in Appendix) and phenotypic susceptibility assays of plasma-associated HIV-1 to IDV and SQV done at baseline.

Methods for Determination of Genotype and Drug-Susceptibility Phenotype

Viral RNA analysis. RNA was extracted from 100 μL of plasma according to the method described by Boom et al. [7]. After viral

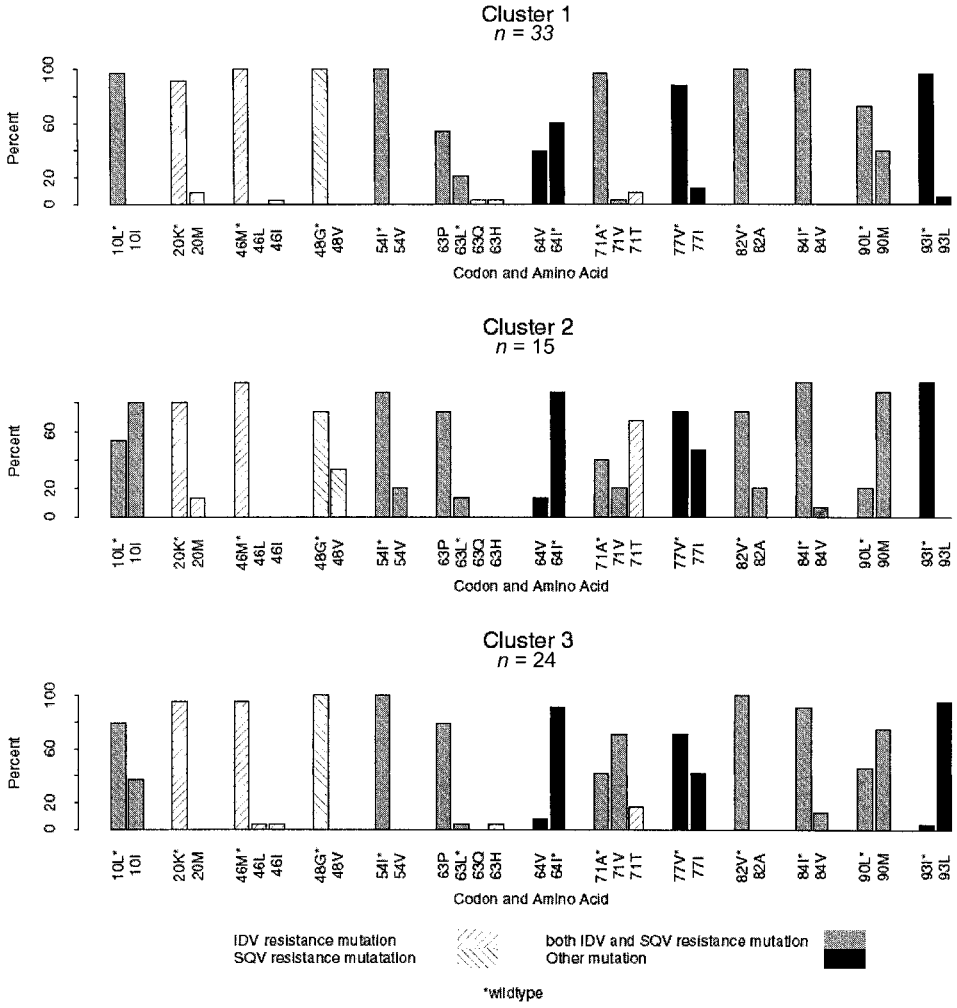

Figure 1. Cluster profiles of amino acids at selected sites of known resistance mutations to saquinavir (SQV) and/or indinavir (IDV)

RNA isolation an equivalent of 10 μL of plasma was used to reverse transcribe and amplify the protease gene (nucleotides 2252–2548). A one-tube RT–polymerase chain reaction (PCR) procedure, essentially as described by Nijhuis et al. [8] using 1 mM $MgCl_2$ and 10 pmol of primer 5′ prot-1 (5′-AGGCTAATTTTTTAGGGAAGATCTGGCCTTCC-3′ [nucleotides 2077–2108]) and primer 3′prot-1 (5′-GCAAATACTGGAGTATTGTATGGATTTTCAGG-3′ [nucleotides 2733–2702]) (Pharmacia Biotech, Roosendaal, The Netherlands), was done. Subsequent to this procedure, the amount of amplified product was further increased in a second (nested) amplification reaction, containing 12 pmol of primer 5′prot-2 (5′-TCAGAGCAGACCAGAGCCAACAGCCCCA-3′ [nucleotides 2135–2162]) and 11 pmol of primer 3 ′prot-2 (5′-AATGCTTTTATTTTTTCTTCTGTCAATGGC-3′ [nucleotides 2649–2620]).

The nested PCR products were sequenced using the *Taq* Dye Deoxy Terminator cycle sequencing kit (Applied Biosystems International, Foster City, CA) and oligonucleotides PR1 (5′-AGGAGCCGATAGACAAGG-3′ [nucleotides 2215–2232]) and PR2 (5′-CTTTTGGGCCATCCATTC-3′ [nucleotides 2609–2592]).

Phenotypic resistance analysis. Recombinant protease viruses were generated by introduction of viral protease sequences derived from plasma into a protease-deleted HIV-1 clone (HXB2 pro) [9] via homologous recombination. The nested PCR products were cotransfected with HXB2 pro (linearized with *Bst*EII) into SupT1 cells. The transfected cell cultures were subsequently monitored for the appearance of syncytia. When full-blown syncytia were observed, cell-free virus was harvested. The infectious virus titer (TCID50) was determined using end point dilutions in MT2 cells [10]. SQV and IDV susceptibility of the recombinant protease viruses were determined in duplicate using an MTT assay [11]. The reported IC_{50} values are the geometric means of the duplicate measurements (in μM).

Statistical Methods

This section describes the use of recursive partioning [3, 4], cluster analysis [12, 13], and discriminant analysis [14, 15] to investigate the relationship between HIV-1 drug-susceptibility phenotype and genotype. This investigation requires definition of a measure of

Table 2. Using the tree with 5 splits for prediction of week 0 IC_{50} values for indinavir (IDV) from week 0 genotype for 2 individuals who were missing week 0 phenotype.

Individual	Codon					Predicted IDV IC_{50}
	90	37	63	71	62	
1	wt (split L)	D (split R)	wt (split R)	—	—	.0473
2	M (split R)	—	—	wt (split L)	wt (split L)	.0435

NOTE. Wt, wild type; split L, split to the left; D, aspartic acid; split R, split to the right; M, methionine.

distance between any 2 amino acid sequences in the protease section of the HIV-1 gene; these distances can then be used to create clusters of subjects with similar genotypes. The association between these clusters and measures of drug-susceptibility phenotype provides information about the usefulness of the clusters. A recursive partitioning algorithm is used to determine which features of genotype (individual mutations or membership in a cluster) are most related to phenotype. These results are compared with another standard method, linear discriminant analysis.

Cluster analysis. Clustering sequences with similar genotypes permits investigation of the degree to which viruses with similar genotypes have similar drug-susceptibility phenotypes. To compute a distance between sequences, a set of indicator variables is created for each codon, by use of methods similar to those described by P. Cosman et al. (unpublished data). Each indicator variable is assigned a value of 1 if the amino acid it represents is present at the site alone or in a mixture; otherwise it is assigned a value of 0. Each sequence is represented by a vector of indicator variables; the distance between 2 sequences is defined as a Euclidean distance between the 2 corresponding vectors. The distance between 2 clusters is defined as the average of the distances between all pairs of members of the 2 clusters. Hierarchical clustering is used to choose the optimal number of clusters, and *k*-means clustering is used choose the final division of cases into this number of clusters [12, 13]. The optimal number is chosen as the number just greater than that which causes the within-cluster sums of squares (squares of the distances between each observation and the center of the cluster) to start to increase rapidly. To compare the median IC_{50} values among clusters, the Kruskal-Wallis test is used.

Recursive partitioning. Recursive partitioning [3, 4], an iterative technique for constructing a decision tree, starts by identification of the specific variable that best splits a population into 2 subpopulations, or nodes. It continues by identification of the variables that best split each of the resulting nodes into 2 more nodes until no more splits are reasonable. The best split of a nonterminal "parent" node is the one that minimizes the variability of the observations in the 2 "child nodes." The tree is then pruned back to an optimal number of splits, and a predicted value is assigned to the terminal node at the end of each branch of the resulting tree. A new case can be assigned a predicted value by following it through the decision tree to a terminal node. In addition to using the indicator variables for the presence of mutations as covariates, we can use cluster membership as well; this permits investigation of the combined effect of mutations that tend to occur together.

Because the IC_{50} phenotypic data are skewed, all recursive partioning analyses are done using log10 (IC_{50}) values; for ease of interpretation, all means are reported as the geometric mean of IC_{50} values in the original scale. Measures of variability are reported on the log10 scale, however, because there is no simple transformation of these quantities back to the original scale. The predicted value for each terminal node of the tree is reported as the geometric mean of the IC_{50} values for the cases assigned to that node.

For the measure of variability in our trees, we used the within-node sum of squares, or deviance. The deviance is computed by finding the difference between each observed value and the node mean, squaring these differences, and then summing the squares of the differences over all individuals assigned to the node. Cross-validation and bootstrapping are used to determine the appropriate number of terminal nodes and to determine the reproducibility of the model.

Linear discriminant analysis. Linear discriminant analysis [14, 15] is also used to determine which genetic mutations best predict sensitivity to drugs, as defined by IC_{50}. We first define 2 groups, sensitive and resistant, on the basis of an IC_{50} cutoff. Then we use a linear discriminant function to predict IC_{50} category on the basis of genetic sequence. The linear discriminant function is just a linear combination of predictors (in our case the indicator variables for the amino acids at the 99 protease codons). To obtain a cutoff for definition of the groups, we calculated the median IC_{50} for patients who had no substitutions known to be associated with resistance for the drug of interest in available published literature. Patients whose IC_{50} values were $\geq$5-fold greater than this median were classified as resistant; those with a <5-fold increase were classified as sensitive.

All analyses were done with S-Plus 3.4 for Unix. Hierarchical and *k*-means clustering were done with the routines hclust and kmeans, respectively. The S-Plus RPART routines of Therneau et al. [16] were used for recursive partitioning. Linear discriminant analyses were done with the lda routine in Venables and Ripley's MASS library [17].

Results

Baseline Characteristics and Clustering

Eighty-nine subjects were enrolled in clinical trial ACTG 333, of whom 72 had baseline genotype and phenotype assay results. The baseline characteristics of these 72 patients are shown in the last column of table 1.

Using the methods described in the Statistical Methods section, we grouped the baseline genetic sequences into clusters. The most reasonable grouping was division into 3 clusters of 33, 15, and 24 individuals. Figure 1 compares the frequencies of mutations known to be associated with SQV and/or IDV resistance for the 3 clusters [18–25].

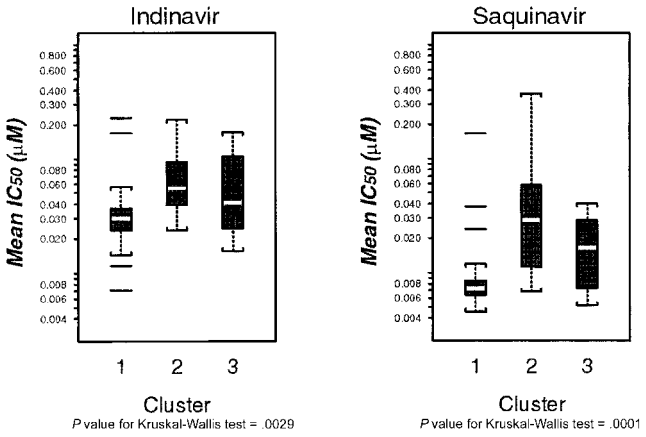

Figure 2. Baseline phenotypic resistance to saquinavir and indinavir by cluster.

Codons 10, 71, and 90, known to be associated with both SQV and IDV resistance, discriminated well among the 3 clusters. This was not surprising, in light of the extensive exposure to SQV of these patients. At protease codons 20, 46, 48, 54, 82, and 84, the majority of patients in all 3 clusters were wild type; all individuals with resistance mutations at codons 48, 54, and 82 were in cluster 2.

Baseline characteristics of subjects in the 3 clusters are shown in table 1, and the baseline phenotypic resistance to SQV and IDV, as measured by IC_{50}, is displayed in figure 2. Phenotypic resistances both to IDV and to SQV were lowest for cluster 1, the cluster with the highest proportions of patients with the consensus wild-type amino acids at codons 10, 71, 90, and 93. Phenotypic resistances both to IDV and to SQV were highest for cluster 2, the cluster with the highest proportions of patients with the L10I (80%), G48V (33.3%), A71T (66.7%), and L90M (86.7%) resistance mutations. Phenotypic resistance to both drugs was slightly lower for cluster 3, as were the proportions of patients with L10I (37.5%), G48V (0%), A71T (16.7%), and L90M (75%) resistance mutations; however, higher proportions of patients in cluster 3 had A71V (70.8%) resistance mutation as well as the L93I (95.8%) substitution. The median IC_{50} values for the 3 clusters were significantly different for both IDV ($P = .0029$) and SQV ($P = .0001$). Furthermore, the median IC_{50} for cluster 1 was significantly less than those for both clusters 2 and 3 for both protease inhibitors, but there was no significant difference between clusters 2 and 3.

Recursive Partitioning to Investigate the Relationship between Baseline Genotype and Phenotype

Variables considered as predictors of IC_{50} included the alphabetic codes for the amino acids at each of the 99 protease codons and membership in the 3 clusters. In figures 3 and 4, each codon is coded as a dash (—) if the consensus wild-type amino acid was present, as a single uppercase letter for a substitution (e.g., L for leucine), and as a pair of lowercase letters for a mixture (e.g., lm for a mixture of leucine and methionine). Within each node the figure provides the geometric mean IC_{50} (μM) and the deviance of the log10 (IC_{50}) for observations assigned to the node as well as the number of observations in the node.

IDV resistance. Cross-validation methods indicated that the best tree had 5 splits; this tree (fig. 3) had an ~20% reduction in deviance compared with the root node. Trees with anywhere from 1 to 6 splits, however, did almost as well. For the tree with just 1 split at codon 90 (the first split in fig. 3), the 38 cases with the wild-type amino acid, leucine either alone or in a mixture at codon 90, are in the left child node. The geometric mean IC_{50} for IDV in this node was 0.0287. The 34 cases with the known IDV resistance substitution M90L are in the right child node; their geometric mean IC_{50} for IDV was 0.0655, a 2.28-fold increase over that of the left child node.

In the full tree, each split represents an ~2-fold increase in phenotypic resistance to IDV (i.e., for each split the ratio of the geometric mean of the IC_{50}' values in the right "child" node to that in the left "child" node is ~2 [range 1.78–2.28]). The cases with the lowest IC_{50} values are in the leftmost terminal node of the tree, which has a geometric mean of 0.015. These cases had the wild-type amino acid L at codon 90; at codon 37, they had an A ($n = 3$), E ($n = 2$), or Q ($n = 2$) substitution or a mixture of N (wild type) and S ($n = 1$). The cases with the highest IC_{50} values are in the rightmost terminal node; the geometric mean of the IC_{50} values for this node is 0.102, a 6.7-fold increase over the cases in the leftmost terminal node. These cases were L90M, at codon 90, and A71V, either alone or a mixture, at codon 71. The other two splits in this tree were based on the amino acid at either codon 62 or codon 63.

Cross-validated trees were also grown on 25 bootstrap samples; of them, 23 trees had ≥ 1 split, and codon 90 was the first split for 15 of the 23. The other factors of importance in these trees were mutations at codons 10, 63, and 71 as well as cluster membership.

Table 2 illustrates how a tree can be used to predict baseline IC_{50} for IDV from a baseline genotype. The first case had the wild-type amino acid leucine (L) at codon 90 and therefore goes to the left at the first split. The next split is to the right because of the presence of aspartic acid (D) at codon 37. The final split is to the right, reflecting the presence of wild-type amino acid leucine (L) at codon 63. This case falls in a terminal node with predicted IC_{50} for IDV of 0.0473. Similarly, the path of the second individual can be followed through the tree to a terminal node with predicted IC_{50} for IDV of 0.0435.

SQV resistance. For analyses of SQV, cross-validation indicated, on the basis of the amino acid present at protease codon 10, that the best tree had just 1 split. Trees with 2–5 splits did almost as well, so we present the tree with all 5 splits in figure 4. Of 51 cases in the left child node, 49 had the wild-

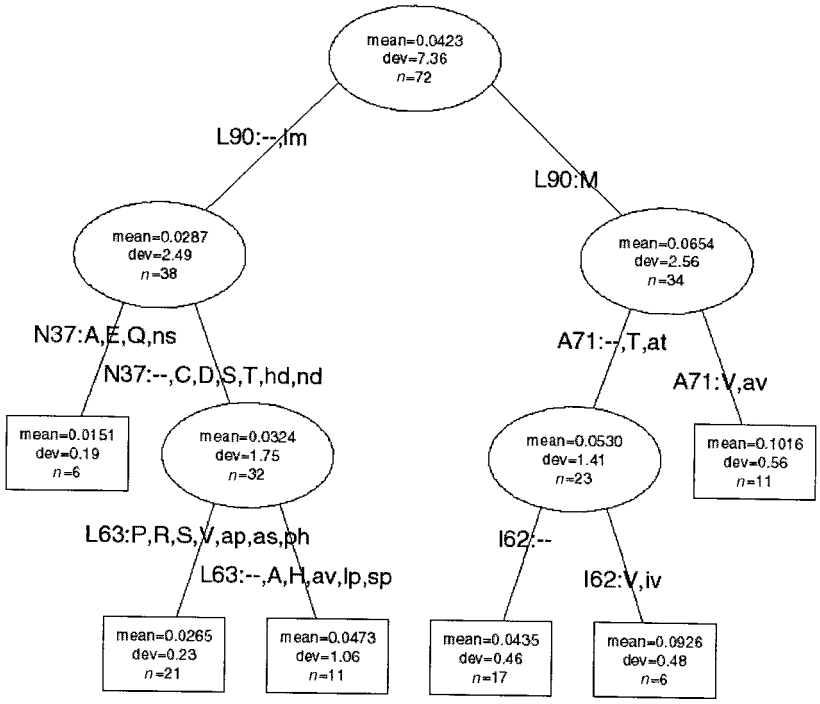

Figure 3. Pruned tree with 5 splits for predicting IC_{50} for indinavir from baseline genotype. *Rectangles* indicate terminal nodes. Numbers in each node are the geometric mean IC_{50} (μM) for the node, the deviance for the node on the log10 (IC_{50}) scale, and the number of cases in the node. Splits that would result in <5 cases in a node were not permitted. The left child node of the root node is labeled "L90: —,lm" indicating that the variable on which the cases split was codon 90 for which the wild-type amino acid is L (leucine) and that cases which were wild-type (—) or had a mixture of L and M (lm) went to the left. The right child node is labeled "L90:M" indicating that cases with the amino acid methionine (M) at codon 90 went to the right.

type amino acid, leucine, at codon 10, and the other 2 had mixtures of leucine and valine or isoleucine and valine. The 21 cases in the right child node had an L10I substitution ($n = 11$), an L10V substitution ($n = 1$), or a mixture of leucine (L) and isoleucine (I) ($n = 9$). Cases in the right node had a 2.86-fold increase in IC_{50} over the cases in the left child node. The variables on which the second through fifth splits occurred were codon 90, codon 62, codon 37, and membership in cluster 1. The pruned tree with 5 splits is displayed in figure 4. For this tree, each split represented an average 2.35-fold increase (range, 1.89–2.86) in phenotypic resistance to SQV for the right child node as compared with the left child node.

This tree provides an example of how recursive partitioning can identify different subsets of the predictors for different individuals. In this example, codon 62 is useful in predicting phenotypic resistance to SQV for individuals with the L10I mutation but not for individuals who are wild type at codon 10. Because the amount of information is limited, the reproducibility of splits after the first one requires further investigation.

Linear Discriminant Analyses of Baseline Data

Linear discriminant function analyses was conducted separately for SQV and IDV.

SQV resistance. To determine an IC_{50} cutoff for categorization of patients as sensitive or resistant, we first identified mutations associated with resistance, from sources other than ACTG 333. According to Larder et al. [26], reduced sensitivity to SQV is most often associated with the presence of either a G48V or an L90M mutation in the protease gene. For the 28 ACTG 333 patients who had neither of these mutations at baseline, the median IC_{50} for SQV was 0.007. Patients who had IC_{50} values <0.035 (a 5-fold increase over 0.007) were categorized as sensitive to SQV; the remainder were categorized as resistant. The protease codons that best discriminated between the 2 groups (contributed most to the linear discriminant function) were 10, 62, 63, 71, and 90. The cross-validated misclassification rate (a measure of reliability of the results) was 24.7%—results that are comparable to those of the recursive partitioning. Codons 10, 62, and 90 appeared in the regression tree for predicting IC_{50} for SQV that is displayed in figure 4. Cluster 1 also appeared in the regression tree, and codon 71 was one of the more influential codons in discriminating among the clusters, along with codons 10 and 90 (see figure 1).

IDV resistance. The following mutations in the protease gene have been associated with reduced susceptibility to indinavir [27]: L10I/R/V, K20M/R, L24I, V32I, M46I/L, I54V,

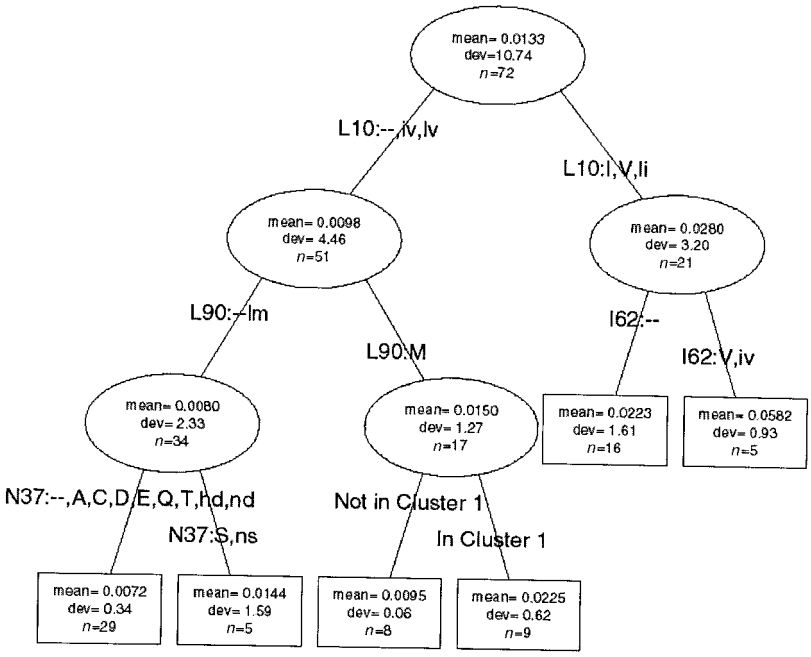

Figure 4. Pruned tree with 5 splits for predicting IC_{50} for saquinavir from baseline genotype. *Rectangles* indicate terminal nodes. Numbers in each node are the geometric mean IC_{50} (μM) for the node, the deviance for the node on the log10 (IC_{50}) scale, and the number of cases in the node. Splits that would result in fewer than 5 cases in a node were not permitted.

L63P, A71T/V, V82A/F/T, I84V, and L90M. For the 13 ACTG 333 patients who had none of these mutations, the median IC_{50} for indinavir was 0.024. Patients who had IC_{50} values <0.120 for IDV (a 5-fold increase over 0.024) were considered to be sensitive to IDV, and the remainder were considered to be resistant. The protease codons that best discriminated between the two groups were 10, 36, 63, 71, 73, 90, and 93. The cross-validated misclassification rate was 22.1%. Codons 63, 71, and 90 appeared in the regression tree for prediction of IC_{50} for IDV that is displayed in figure 3. Codons 10, 71, 90, and 93 were among the more influential codons in discriminating among clusters (see figure 1).

Discussion

This report investigates the use of standard statistical methods to relate HIV-1 genotype to drug-susceptibility phenotype. The clustering of viruses by genotype is of interest in itself because it permits determination of mutations that tend to occur together and investigation of the extent to which similarities in genotype are associated with similarities in phenotype. Clustering is also of interest because, by including cluster membership in recursive partitioning, we can determine whether the importance of a mutation is because of its direct effect or its tendency to occur with others. We have applied these statistical tools to relate genotype to phenotypic susceptibility of specific antiretroviral drugs, using results from ACTG 333. Recursive partitioning has also recently been used to study relationships between genotypic resistance, baseline HIV RNA viral load, and virological outcome [27].

The important but difficult task of relating the many possible genotypic patterns to phenotypic drug resistance has been the focus of many recent studies. Reports using other approaches, such as logistic regression, have produced anecdotal data highlighting the impact of a few specific mutations [28] or general rules regarding the total number of mutations and phenotypic resistance [29]. Very large databases including thousands of genotype-phenotype pairs have been compiled at great expense and effort [30]. Use of techniques like those described in this article should help in identification of patient characteristics, including genotype, that consistently predict drug-susceptibility phenotype.

In the present study, we were able to identify and relate the key mutations to the degree to which they affect resistance and also to find the relationships between the different mutations. Studies in which the mutations deemed relevant are specified before analysis [30] preclude discovery of important new, and perhaps poorly understood, mutations or relationships among mutations. For example, we found mutations often considered to be of only minor importance, such as those at protease positions 10 or 71, to play important roles in predicting drug resistance or phenotype in the ACTG 333 patients. This is consistent with the clinical observation that substitutions at positions 10, 71, and 90 were the most common in long-term

recipients of SQV, suggesting they are the more relevant in vivo substitutions [31]. The need for unbiased analysis is of major importance in the field of HIV resistance, because the biological implications of the individual mutations or groups of mutations have not yet been well characterized [2]. These statistical techniques also allow us to gain information by combining across different data sets. This feature will also allow researchers to gain valuable information from the multiple small studies in different populations that are often done in HIV research and not only from large multicenter trials.

The major advantage of recursive partitioning as an exploratory technique is the ability to identify potentially important mutations (or clusters of mutations) from among the many thousands that may occur in HIV-1 genome under pressure from antiretroviral drugs. In addition, it is possible to identify mutations whose importance is limited to a particular subgroup of patients, defined by the presence of other mutations or any other prognostic factor. By applying these techniques to the many studies examining genotypic and phenotypic resistance done to date and those currently under way, much insight into the complex relationship between genotype and drug-susceptibility phenotype may be gained.

These analyses confirm the associations of mutations at codons 10, 63, 71, and 90 with in vitro resistance to IDV and/or SQV. Mutations at codons 37 and 62, which appear in both trees, have not been previously associated with resistance to SQV or IDV. These results not only imply some well-known similarities in the mutations associated with resistance to IDV and to SQV, but also some degree of reproducibility of the methods. Furthermore, the results of the bootstrap investigation of the prediction of IDV resistance from baseline genotype also consistently showed the importance of the mutations mentioned above. Finally, the results of linear discriminant analyses support those found from recursive partitioning. Although both techniques may be useful, the latter allows the identification of mutations whose importance depends on the presence or absence of other mutations; this feature will be of special importance for larger databases. For these reasons, recursive partitioning should be considered in investigations relating genotype to drug-susceptibility phenotype.

Acknowledgments

We thank Bette T. Korber (Los Alamos National Laboratories) for helpful comments and suggestions and Marc Bijen and Albert van Wijk for technical support.

Appendix

The GenBank accession numbers are as follows: 041338a AF226078; 042298h AF226079; 042389e AF226080; 042393e AF226081; 043066f AF226082; 043067d AF226083; 043068b AF226084; 043075h AF226085; 050165h AF226086; 050690d AF226087; 050922a AF226088; 061452e AF226089; 061453c AF226090; 082261e AF226091; 082264k AF226092; 090939i AF226093; 090940f AF226094; 090941d AF226095; 090943l AF226096; 090949k AF226097; 090951c AF226098; 110054h AF226099; 110206d AF226100; 110522e AF226101; 140267b AF226102; 140681d AF226103; 140837d AF226104; 140857b AF226105; 140860d AF226106; 140868k AF226107; 140869i AF226108; 140967k AF226109; 141095g AF226110; 141096d AF226111; 141123d AF226112; 141125l AF226113; 171040c AF226114; 172016d AF226115; 172018l AF226116; 172019j AF226117; 172021b AF226118; 211647k AF226119; 220641k AF226120; 230035c AF226121; 230270e AF226122; 130333a AF226123; 230484d AF226124; 230516d AF226125; 230679a AF226126; 230815l AF226127; 230848b AF226128; 231025c AF226129; 231206j AF226130; 231449h AF226131; 240766d AF226132; 270788c AF226133; 271018l AF226134; 271469i AF226135; 271917g AF226136; 271925k AF226137; 350214j AF226138; 610049e AF226139; 610113g AF226140; 610137h AF226141; 610369i AF226142; 630081b AF226143; 630112e AF226144; 630326c AF226145; 630468e AF226146; 630469c AF226147; 630474a AF226148; 630476k AF226149.

References

1. Durant J, Clevenbergh P, Halfon P, et al. Genotyping: the Viradapt study. Lancet **1999**; 353:2195–9.
2. Mayer DL. Drug-Resistant HIV-1: The virus strikes back [editorial]. JAMA **1998**; 279:2000–2.
3. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Belmont, CA: Wadsworth, **1984**.
4. Zhang H, Singer B. Recursive partitioning in the health sciences. New York: Springer, **1999**.
5. Winters MA, Baxter JD, Mayers DL, et al. Frequency of antiretroviral drug resistance mutations in HIV-1 strains from patients failing triple drug regimens. The Terry Beirn Community Programs for Clinical Research on AIDS. Antivir Ther **2000**; 5:57–63.
6. De Gruttola V, Dix L, D'Aquila, et al. The relation between baseline HIV drug resistance and response to antiviral therapy: re-analyses of retrospective and prospective studies using a standardized data analysis plan. Antivir Ther **2000**; 5:43–50.
7. Boom R, Sol CJA, Salimans MMM, Jansen CL, Wertheim-van Dillen PME, Noordaavan der J. Rapid and simple method for purification of nucleic acids. J Clin Microbiol **1990**; 28:495–503.
8. Nijhuis, Boucher CAB, Schuurman R. Sensitive procedure for the amplification of HIV-1 RNA using a combined reverse-transcription and amplification reaction. Biotechniques **1995**; 19:178–80.
9. Maschera B, Furfine E, Blair ED. Analysis of resistance to human immunodeficiency virus type 1 protease inhibitors by using matched bacterial expression and proviral infection vectors. J Virol **1995**; 69:5431–6.
10. Reed LJ, Muench H. A simple method for estimating fifty percent endpoints. Am J Hyg **1983**; 493–7.
11. Boucher CAB, Keulen W, van Bommel T, et al. Human immunodeficiency virus type 1 drug susceptibility determination by using recombinant viruses generated from patient sera tested in a cell-killing assay. Antimicrob Agents Chemother **1996**; 40:2404–9.
12. Hartigan JA. Clustering algorithms. New York: Wiley, **1975**.
13. Everitt BS. Cluster analysis. 3d ed. New York: Wiley, **1993**.
14. Fisher RA. The use of multiple measurements in taxonomic problems. Ann Eugenics **1936**; 7:179–88.

15. Panel on Discriminant Analysis, Classification, and Clustering. Discriminant analysis and clustering. Stat Sci **1989**; 4:34–69.
16. Therneau TM, Atkinson EJ. An introduction to recursive partitioning using the RPART routines. Rochester, MN: Mayo Clinic, **1997**.
17. Venables W, Ripley B. Modern applied statistics with S-PLUS. New York: Springer, **1997**.
18. Condra JH, Holder DJ, Schleif WA, et al. Genetic correlates of in vivo viral resistance to indinavir, a human immunodeficiency virus type 1 protease inhibitor. J Virol **1996**; 70:8270–6.
19. Craig C, Race E, Sheldon J, et al. HIV protease genotype and viral sensitivity to HIV protease inhibitors following saquinavir therapy. AIDS **1998**; 12: 1611–8.
20. Jacobsen H, Hanggi M, Ott M, et al. In vivo resistance to a human immunodeficiency virus type 1 proteinase inhibitor: mutations, kinetics and frequencies. J Infect Dis **1996**; 173:1379–87.
21. Ruiz L, Nijhuis M, Boucher C. Efficacy of adding indinavir to previous reverse transcriptase nucleoside analogues in relation to genotypic and phenotypic resistance development in advanced HIV-1 infected patients. J Acquir Immune Defic Syndr Hum Retrovirol **1998**; 19:19–28.
22. Shafer RW, Hsu P, Patick AK, Craig C, Brendel V. Identification of biased amino acid substitution patterns in human immunodeficiency virus type 1 isolates from patients treated with protease inhibitors. J Virol. **1999**; 73: 6197–202.
23. Tisdale M, Myers RE, Maschera B, Parry NR, Oliver NM, Blair E. Cross-resistance analysis of human immunodeficiency virus type 1 variants individually selected for resistance to five different protease inhibitors. Antimicrob Agents Chemother **1995**; 39:1704–10.
24. Winters MA, Schapiro JM, Lawrence J, Merigan TC. Human immunodeficiency virus type 1 protease genotypes and in vitro protease inhibitor susceptibilities of isolates from individuals who were switched to other protease inhibitors after long-term saquinavir treatment. J Virol **1998**; 72: 5303–6.
25. Zhang Y-M, Imamichi H, Imamichi T, et al. Drug resistance during indinavir therapy is caused by mutations in the protease gene and in its gag substrate cleavage sites. J Virol **1997**; 71:6662–70.
26. Larder B, Richman D, Vella S, eds. HIV resistance and implications for therapy. Atlanta: MediCom, **1998**.
27. Katzenstein D, Bosch R, Shafer R, Albrecht M, Winters M, Hammer S. Classification-tree analysis of resistance mutations and virus load and failure in nelfinavir- and efavirenz-based therapies [abstract 66]. In: Programme and Abstracts of the 3rd International Workshop on HIV Drug Resistance and Treatment Strategies (San Diego). London: International Medical Press, **1999**:44.
28. Gingeras T, Ehm MG, Shortino DD, et al. Construction and analysis of an HIV-1 genotypic/phenotypic database confirm amino acid changes that co-occur and precede resistance [abstract 56]. In: Programme and Abstracts of the 2d International Workshop on HIV Drug Resistance and Treatment Strategies (Lake Maggiore, Italy). London: International Medical Press, **1998**:38.
29. Patick AK, Zhang M, Hertogs K, et al. Correlation of virologial response with genotype and phenotype of plasma HIV-1 variants in patients treated with nelfinavir in the US expanded access program [abstract 57]. In: Programme and Abstracts of the 2d International Workshop on HIV Drug Resistance and Treatment Strategies (Lake Maggiore, Italy). London: International Medical Press, **1998**:39.
30. Larder B, de Vroey V, Dehertogh P, Kemp S, Bloor S, Hertogs K. Predicting HIV-1 phenotypic resistance from genotype using a large phenotype-genotype relational database [abstract 59]. In: Programs and abstracts of the 3rd International Workshop on HIV Drug Resistance and Treatment Strategies (San Diego). London: International Medical Press, **1999**:40.
31. Para M, Collier A, Coombs R, et al. ACTG 333: relationship of baseline genotype to RNA response in ACTG 333 after switching from long term saquinavir (SQVhc) to SQV soft gelatin capsule (SQVsgc) or indinavir (IDV) [abstract 511]. In: Program and abstracts of the 5th Conference on Retroviruses and Opportunistic Infections (Chicago). Alexandria, VA: Foundation for Retrovirology and Human Health, **1998**:175.