University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2005 *Paper* 196

Application of a Variable Importance Measure Method to HIV-1 Sequence Data

Merrill D. Birkner* Mark J. van der Laan†

*Division of Biostatistics, School of Public Health, University of California, Berkeley, mbirkner@berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper196

Application of a Variable Importance Measure Method to HIV-1 Sequence Data

Merrill D. Birkner and Mark J. van der Laan

Abstract

van der Laan (2005) proposed a method to construct variable importance measures and provided the respective statistical inference. This technique involves determining the importance of a variable in predicting an outcome. This method can be applied as an inverse probability of treatment weighted (IPTW) or double robust inverse probability of treatment weighted (DR-IPTW) estimator. A respective significance of the estimator is determined by estimating the influence curve and hence determining the corresponding variance and p-value. This article applies the van der Laan (2005) variable importance measures and corresponding inference to HIV-1 sequence data. In this data application, protease and reverse transcriptase codon position on the HIV-1 strand are assessed to determine their respective variable importance, with respect to an outcome of viral replication capacity. We estimate the W-adjusted variable importance measure for a specified set of potential effect modifiers W. Both the IPTW and DR-IPTW methods were implemented on this dataset

1 Introduction

In many genomic cases, the prediction of a phenotypic outcome by genetic markers is of biological importance. In particular, variable importance measures can be calculated for individual genetic components when predicting an outcome. Variable importance measures can be used to determine the effect of each variable with respect to the outcome. In this article, we will apply the variable importance measure methodology, proposed by van der Laan (2005) to an HIV-1 sequence dataset. In this application the variable importance of each codon is assessed with respect to its relationship to viral replication. A variable importance measure for each codon will be recorded and a corresponding significance will be calculated with the use of the estimator's influence curve. We estimate the W-adjusted variable importance measure for a specified set of potential effect modifiers W. Both the inverse probability of treatment weighted estimator (IPTW) and the double robust inverse probability of treatment weighted estimator (DR-IPTW) will be calculated.

1.1 Biological Motivation

Sequencing a virus, such as HIV-1, could potentially give further insight into the genotype-phenotype associations of a virus. The replication ability of a virus is vital, especially in the case of HIV, where replication is proportional to the severity of disease. In many cases, genetic mutations on the viral strand are associated with a change in the replication capacity of the virus. This in turn can change the virulence of the virus and/or cause resistance to previously effective antiretrovial drugs. As mentioned above, the application of the variable importance measure methodology is applied to the HIV sequence data. The motivation behind this analysis in focused on determining the significant codons which are related to replication capacity. The data consists of codon positions which are coded as mutated or non-mutated. Therefore one is interested in determining which position specific mutations are related to viral replication. This is a biologically relevant question since antiretroviral medications are manufactured to target specific regions on the viral strand. Therefore determining the importance of specific regions or codons is vital when assessing the viral regions which must be targeted by antiretroviral medications. In this article, codons will be assessed by their respective variable importance measure.

2 Data Description

Studying sequence variation for the Human Immunodeficiency Virus Type 1 (HIV-1) genome could potentially give important insight into *genotype–phenotype associations* for the Acquired Immune Deficiency Syndrome (AIDS). In this context, a key phenotype is the replication capacity (RC) of HIV-1, as it reflects the severity of the disease. A measure of replication capacity may be obtained by monitoring viral replication in an ideal environment, with many cellular targets, no exogenous or endogenous inhibitors, and no immune system responses against the virus (Barbour et al., 2002a; Segal et al., 2004).

Genotypes of interest correspond to codons in the protease and reverse transcriptase regions of the viral strand. The protease (PR) enzyme affects the reproductive cycle of the virus by breaking protein peptide bonds during viral replication. The reverse transcriptase (RT) enzyme synthesizes double-stranded DNA from the virus' single-stranded RNA genome, thereby facilitating integration into the host's chromosome. Since the PR and RT regions are essential to viral replication, many antiretrovirals (protease inhibitors and reverse transcriptase inhibitors) have been developed to target these specific genomic locations. Studying PR and RT genotypic variation involves sequencing the corresponding HIV-1 genome regions and determining the amino acids encoded by each codon (i.e., each nucleotide triplet).

The HIV-1 sequence dataset consists of $n = 317$ records, linking viral replication capacity (RC) with protease (PR) and reverse transcriptase (RT) sequence data, from individuals participating in studies at the San Francisco General Hospital and Gladstone Institute of Virology (Segal et al., 2004). Protease codon positions 4 to 99 (i.e., $pr4$ – $pr99$) and reverse transcriptase codon positions 38 to 223 (i.e., $rt38$ – $rt223$) of the viral strand are studied in this analysis.

The outcome/phenotype of interest is the natural logarithm of a continuous measure of replication capacity, ranging from 0.261 to 151. The M covariates correspond to the $M = 282$ codon positions in the PR and RT regions, with the number of possible codons ranging from one to ten at any given location. A majority of patients typically exhibit one codon at each position. Codons are therefore recoded as binary covariates, with value of **zero** (or "wild-type") corresponding to the most common codon among the $n = 317$ patients and value of **one** (or "mutation") for all other codons. Previous biological research was used to confirm mutations and hence provide accurate PR

and RT codon genotypes for each patient (`hivdb.stanford.edu/cgi-bin/RTMut.cgi`). The data for each of the $n = 317$ patients therefore consist of a replication capacity outcome/phenotype Y and an M–dimensional covariate vector $X = (X(m) : m = 1, \ldots, M)$ of binary codon genotypes in the PR and RT HIV-1 regions.

3 Methods

The variable importance measure is based on n i.i.d. observations, a set of covariates, and an outcome. This section will outline the methods presented in the van der Laan (2005) paper. We refer the reader to van der Laan (2005) for a detailed description of the methods. This section will generalize the method to the case of the HIV mutation example. In this example $n = 317$ individuals and each individual has a vector of length 282 corresponding to the mutation status of the virus at each codon and a continuous measure of viral replication. The variable importance measure method is based on determining the individual importance of each codon. Before beginning this analysis, several codons were removed from the analysis for a variety of reasons. Firstly, a set W was constructed as a group of codons which were potential confounders of the individual codon effect on the outcome. In order to choose the set of W we applied the FDR procedure to the set of test statistics built from the marginal association of each codon against the outcome of replication capacity. In total, 16 mutations were chosen with an FDR adjusted p-value less than 0.05. Of these 16 mutations, there are three codons which are predicted perfectly by $pr29$ ($pr31$, $pr44$, and $pr52$). Therefore, $pr31$, $pr44$, and $pr52$ are removed as codons that we are interested in assessing since their significance should be identical to that of $pr29$. In addition, the remaining codons $(282 - 16)$ were assessed to determine the $P(A_i = 1)$, which corresponds to the probability that the codon A_i is mutated. We want to note that A will be defined as a single codon, whereas in the case when $A \in W$, W is the set of codons omitting A. Only those codons with $P(A_i = 1) \geq 0.1$ were chosen to obtain variable importance measures, since a $P(A_i = 1) \geq 0.1$ corresponds to the case where there is experimentation among the individuals. In conclusion, 37 codon positions were assessed to determine the subsequent variable importance measures and respective significance.

After the data was defined, the variable importance was calculated for

37 positions. We estimate the W-adjusted variable importance measure for a specified set of potential effect modifiers W. The inverse probability of weighted method (IPTW) estimator as well as the double robust IPTW method (DR-IPTW) estimator were applied. The various methods will be discussed below as well as a technique to estimate the respective inference. Again, the reader is referred to van der Laan (2005) for more detail regarding these methods.

3.1 IPTW Method

The IPTW estimator is an estimator which assesses the difference in the mean replication capacity between the mutated and non-mutated individuals, weighting the probability of each mutation given the set of confounders W, $P(A|W)$. We will initially define $D(O_i|\Pi_n)$ as:

$$D(O_i|\Pi_n) = Y_i \left(\frac{I(A_i = 1)}{\Pi_n(A = 1|W)} - \frac{I(A_i = 0)}{\Pi_n(A = 0|W)} \right)$$

The estimator Ψ_n is defined in terms of $D(O_i|\Pi_n)$ and can be written as:

$$\Psi_n = \frac{1}{n} \sum_{i=1}^{n} D(O_i|\Pi_n) = \frac{1}{n} \sum_{i=1}^{n} Y_i \left(\frac{I(A_i = 1)}{\Pi_n(A = 1|W)} - \frac{I(A_i = 0)}{\Pi_n(A = 0|W)} \right)$$

In this equation, $\Pi(A = a|W)$ corresponds to the probability obtained from fitting a logistic regression, which regresses the binary A values on the other W values, or potential confounders. In the specific HIV-1 example $A = (0, 1)$. In the case of this analysis the logistic regression was estimated with the POLYCLASS function in R.

POLYCLASS is an exploratory, data-adaptive, or black box regression technique used to predict categorical or binary outcomes. This classification method, uses forward addition and backward deletion, searches through a series of models defined by main effects, splines and cross-products to create a logistic regression model. The procedure uses cross-validation to choose the complexity (number of basis functions) of the model. This method therefore attempts to balance the variance/bias of the classification error. This data-adaptive logistic regression technique combines stepwise (hierarchical)

addition and deletion of variables and finds a linear combination of variables that provides a better predictor of the outcome event. For example, with respect to the addition steps, proposed new predictors are either 1) main effects not already in the model 2) knots to existing main effects creating linear spline terms or 3) any product of terms already in the model. For the deletion step, terms are removed hierarchically (e.g., a main effect term is not removed before it's corresponding interaction term).

3.2 DR-IPTW Method

The double robust method differs from the IPTW method in that it incorporates a regression of $E(Y|A,W)$, which is referred to as $\theta(A,W)$. In addition to this regression, the quantity $E(Y|A=1,W) - E(Y|A=0,W)$ is computed. This quantity is referred to as $\theta(A=1,W) - \theta(A=0,W)$, which corresponds to the difference in effect of $A=1$ and $A=0$. We will initially define $D(O_i|\Pi_n,\theta_n)$ as:

$$D(O_i|\Pi_n,\theta_n) = Y_i\left(\frac{I(A_i=1)}{\Pi_n(A=1|W)} - \frac{I(A_i=0)}{\Pi_n(A=0|W)}\right) -$$
$$\theta(A,W)\left(\frac{I(A_i=1)}{\Pi_n(A=1|W)} - \frac{I(A_i=0)}{\Pi_n(A=0|W)}\right) + \theta(1,W) - \theta(0,W)$$

The estimator Ψ_n is defined in terms of $D(O_i|\Pi_n,\theta_n)$ and can be written as:

$$\Psi_n = \frac{1}{n}\sum_{i=1}^{n} D(O_i|\Pi_n,\theta_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i\left(\frac{I(A_i=1)}{\Pi_n(A=1|W)} - \frac{I(A_i=0)}{\Pi_n(A=0|W)}\right)$$
$$-\theta(A,W)\left(\frac{I(A_i=1)}{\Pi_n(A=1|W)} - \frac{I(A_i=0)}{\Pi_n(A=0|W)}\right) + \theta(1,W) - \theta(0,W)$$

When calculating the DR-IPTW estimator, $\theta(A,W)$ is calculated in this example with the POLYMARS function in R. This method is similar to the POLYCLASS method with the exception that it is adapted to continuous outcomes, whereas the POLYCLASS method is based on the logit function and

therefore adapted to binary or categorical outcomes. The POLYMARS function is an adaptive regression procedure which uses linear splines to model the response. Therefore this method examines all main effects, interactions and splines to model the outcome by a set of predictor variables.

The advantage of the DR-IPTW method is that the estimator remains consistent if either Π_n or θ_n is modelled correctly. In this case, the estimator Ψ_n is consistent and asymptotically linear if either Π_n or θ_n converge to the truth of Π_0 or θ_0 respectively.

3.3 Inference of Ψ_n

Once the IPTW or DR-IPTW estimator is computed, one is often interested in the inference and therefore statistical significance of the respective variable importance measures. In order to determine the inference on the estimator the influence curve is used. In the case of the IPTW estimator we will define Ψ_n as follows:

$$\Psi_n = \frac{1}{n}\sum_{i=1}^{n} D(O_i|\Pi_n)$$

The asymptotic variance of $\sqrt{n}(\Psi_n - \Psi)$ can be conservatively estimated with:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(D(O_i|\Pi_n) - \Psi_n)^2$$

In the case of the DR-IPTW estimator, the θ_n portion of the estimator must also be incorporated in the construction of the variance. Therefore in this case, the estimator will be defined as:

$$\Psi_n = \frac{1}{n}\sum_{i=1}^{n} D(O_i|\theta_n, \Pi_n)$$

Again, the asymptotic variance of $\sqrt{n}(\Psi_n - \Psi)$ can be conservatively estimated with:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(D(O_i|\theta_n, \Pi_n) - \Psi_n)^2$$

In both the IPTW and DR-IPTW cases, the test statistic for the estimator is compared to a $N(0,1)$ distribution and is defined as:

$$T_n = \frac{\Psi_n}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \sim N(0,1)$$

3.4 Assessing the DR-IPTW Estimator

The DR-IPTW estimator was defined previously in terms of $D(O_i|\theta_n, \Pi_n)$. Each codon results in a summary Ψ_n value which corresponds to $D(O_i|\theta_n, \Pi_n)$. Another estimator can be defined to assess the variables in set W which best predict $D(O_i|\theta_n, \Pi_n)$ for each codon. We will define $D = E(D(O_i|\theta_n, \Pi_n)|W)$ as the appropriate estimator. In order to determine the confounding factors in the set W which predict $D(O_i|\theta_n, \Pi_n)$ we will build models using POLY-MARS to predict this outcome measure. In this case, the vector corresponding to $D(O_i|\theta_n, \Pi_n)$ for each codon position was regressed against the set W using this data adaptive regression method.

4 Results

The IPTW and DR-IPTW procedures produced variable importance measures and respective p-values for the codons illustrated in Table 2 and Table 3. These estimates are the W-adjusted variable importance measure for a specified set of potential effect modifiers W. The W-adjusted variable importance measure can be interpreted as the difference in mean replication capacity among strata W between those with a mutation at that specific codon versus those with no mutation at that specific codon. Therefore this variable importance measure is the average impact of the codon within strata of W. These two methods gave slightly different estimate values for several codons, but gave consistent measures of significance in the two cases. In the cases where the two values (IPTW and DR-IPTW) differ, the DR-IPTW variable importance measure is close in value to $\theta(A = 1, W) - \theta(A = 0, W)$,

Table 1: Codon Test Statistics and Marginal p-values

Codon	Test Statistic	p-value
$pr90$	-6.2378	<0.00001
$rt184$	-6.1626	<0.00001
$pr43$	-6.1184	<0.00001
$pr54$	-5.5391	<0.00001
$rt41$	-5.2253	<0.00001
$pr46$	-5.2239	<0.00001
$pr82$	-4.5206	<0.00001
$rt215$	-4.4794	<0.00001
$rt121$	-4.0703	0.00004
$pr10$	-3.7900	0.00015
$pr71$	-3.3939	0.00068
$rt102$	-3.0874	0.00201
$rt214$	2.1518	0.00314
$pr57$	1.5776	0.1146
$pr14$	-2.4362	0.0148
$rt196$	-0.1872	0.8514
$rt103$	-1.3612	0.1734
$pr13$	1.4431	0.1489
$pr15$	1.6171	0.1058
$pr36$	1.2917	0.1964
$pr64$	0.6550	0.5124
$rt83$	2.458	0.0139
$pr41$	-0.3640	0.7158
$rt177$	2.0754	0.0379
$pr35$	1.3597	0.1138
$pr62$	0.4136	0.6791
$rt200$	-1.3211	0.1864
$rt207$	0.6699	0.5029
$rt162$	1.1116	0.2663
$pr77$	-1.8474	0.0646
$pr37$	1.4975	0.1342
$rt122$	-1.3460	0.1782
$rt135$	-2.8680	0.0041
$pr93$	0.1422	0.8868
$pr63$	1.5628	0.1181
$rt211$	2.3058	0.0211
$rt178$	-1.8435	0.0652

Table 2: IPTW Estimators and p-values

Codon	IPTW Ψ_n	p-value
$pr90$	-1.821	<0.00001
$rt184$	-2.119	<0.00001
$pr43$	-2.262	<0.00001
$pr54$	-2.838	<0.00001
$rt41$	-2.302	<0.00001
$pr46$	-2.435	<0.00001
$pr82$	-2.297	<0.00001
$rt215$	-1.173	0.00512
$rt121$	-2.252	<0.00001
$pr10$	-1.532	0.000915
$pr71$	-1.148	0.0190
$rt102$	-1.026	0.0691
$rt214$	-0.341	0.550
$pr57$	0.268	0.688
$pr14$	-0.896	0.0917
$rt196$	-0.616	0.288
$rt103$	-0.570	0.324
$pr13$	-0.460	0.367
$pr15$	0.240	0.675
$pr36$	0.314	0.617
$pr64$	-0.643	0.168
$rt83$	-0.531	0.261
$pr41$	-0.0488	0.921
$rt177$	0.0531	0.917
$pr35$	0.243	0.637
$pr62$	0.0954	0.853
$rt200$	-0.0210	0.964
$rt207$	0.111	0.821
$rt162$	0.022	0.962
$pr77$	-0.143	0.743
$pr37$	0.0297	0.948
$rt122$	0.118	0.785
$rt135$	-0.244	0.581
$pr93$	0.0464	0.918
$pr63$	-0.0663	0.878
$rt211$	0.231	0.594
$rt178$	0.101	0.814

Table 3: DR-IPTW Estimators and p-values

Codon	DR-IPTW Ψ_n	p-value
$pr90$	-0.519	<0.00001
$rt184$	-0.584	0.00006
$pr43$	-0.646	<0.00001
$pr54$	-0.322	0.00002
$rt41$	-0.288	0.00275
$pr46$	-0.544	<0.00001
$pr82$	-0.0772	0.249
$rt215$	0.0557	0.502
$rt121$	-0.561	<0.00001
$pr10$	-0.242	0.0170
$pr71$	0.238	0.00423
$rt102$	0.256	0.0164
$rt214$	0.366	0.00029
$pr57$	0.264	0.0577
$pr14$	-0.168	0.186
$rt196$	0.00734	0.955
$rt103$	-0.109	0.299
$pr13$	0.190	0.1668
$pr15$	0.227	0.1275
$pr36$	0.307	0.1455
$pr64$	0.0612	0.585
$rt83$	0.321	0.0361
$pr41$	-0.0633	0.577
$rt177$	0.256	0.0511
$pr35$	0.227	0.1521
$pr62$	0.0842	0.444
$rt200$	0.0151	0.876
$rt207$	0.134	0.158
$rt162$	0.108	0.233
$pr77$	-0.128	0.267
$pr37$	0.0904	0.478
$rt122$	0.0743	0.468
$rt135$	-0.129	0.187
$pr93$	-0.0142	0.891
$pr63$	0.0509	0.603
$rt211$	0.159	0.187
$rt178$	0.011	0.986

Table 4: FDR on IPTW Estimator's p-values

Codon	FDR p-value
$pr90$	<0.00001
$rt184$	<0.00001
$pr43$	<0.00001
$pr54$	<0.00001
$rt41$	<0.00001
$pr46$	<0.00001
$pr82$	<0.00001
$rt215$	0.0199
$rt121$	<0.00001
$pr10$	0.000396
$pr71$	0.06741
$rt102$	0.2246
$rt214$	0.9648
$pr57$	0.9648
$pr14$	0.2751
$rt196$	0.6605
$rt103$	0.7013
$pr13$	0.7538
$pr15$	0.9648
$pr36$	0.9648
$pr64$	0.4686
$rt83$	0.6350
$pr41$	0.9648
$rt177$	0.9648
$pr35$	0.9648
$pr62$	0.9648
$rt200$	0.9648
$rt207$	0.9648
$rt162$	0.9648
$pr77$	0.9648
$pr37$	0.9648
$rt122$	0.9648
$rt135$	0.9648
$pr93$	0.9648
$pr63$	0.9648
$rt211$	0.9648
$rt178$	0.9648

Table 5: FDR on DR-IPTW Estimator's p-values

Codon	FDR p-value
$pr90$	0.00009
$rt184$	0.00037
$pr43$	0.00009
$pr54$	0.000148
$rt41$	0.01271
$pr46$	0.00009
$pr82$	0.3838
$rt215$	0.6191
$rt121$	0.00009
$pr10$	0.05718
$pr71$	0.01739
$rt102$	0.05718
$rt214$	0.00153
$pr57$	0.1524
$pr14$	0.3145
$rt196$	0.9815
$rt103$	0.4255
$pr13$	0.3145
$pr15$	0.31450
$pr36$	0.31450
$pr64$	0.6760
$rt83$	0.11130
$pr41$	0.6760
$rt177$	0.1454
$pr35$	0.31450
$pr62$	0.6084
$rt200$	0.9419
$rt207$	0.31450
$rt162$	0.3748
$pr77$	0.3951
$pr37$	0.6098
$rt122$	0.6098
$rt135$	0.3145
$pr93$	0.9419
$pr63$	0.6760
$rt211$	0.3145
$rt178$	0.9860

otherwise known as the likelihood based estimator (van der Laan, 2005). All of the codons which were claimed significant with a p-value less than 0.05 with the IPTW or DR-IPTW methods also had an unadjusted univariate p-value less than 0.05, refer to Table 1. In order to account for the multiple tests which were performed, FDR adjusted p-values are reported in Tables 4 and 5. The results produced in these tables correspond to biological significance and statistical significance outlined in Birkner et al. (2005).

4.1 Assessing the DR-IPTW Estimator

Table 6 displays the models assessing the DR-IPTW variable importance measure for each codon. For example, the first codon $pr90$, produces the following model: $D = -0.449 - 2.719(pr43) + 1.014(pr54)$. In this case, the codons $pr43$ and $pr54$ best predict the variable importance measure $D(O_i|\theta_n, \Pi_n))$. Therefore, the impact of $pr90$ is negative with regards to replication capacity among strata where there is a mutation in $pr43$. The effect is only positive among strata where there is a mutation at $pr54$ and no mutation at $pr43$. Another example is with regards to $rt178$. In this case the following model was obtained: $D = 0.1780 - 0.9007(rt41)$. Therefore the impact of $rt178$ is positive with regards to replication capacity among strata where $rt41$ is not mutated and negative when $rt41$ is mutated.

4.2 Biological Results

The procedures identified several codon positions as significantly associated with viral replication capacity. The models presented in Table 6 present several interactions of codons which predict $D(O_i|\theta_n, \Pi_n))$. In some cases, these interactions include protease and reverse transcriptase positions. The current HIV-1 literature does not mention interactions of this nature (protease and reverse transcriptase), so we cannot relate it to the known literature, but they might represent interesting findings. With regards to the individual codon mutations, these analysis included positions which are known to be associated with viral replication and antiretroviral resistance. In particular, protease positions $pr43$, $pr46$, $pr54$, and $pr90$, and reverse transcriptase positions $rt184$, and $rt215$, have been singled out in previous research as related to replication capacity and/or antiretroviral resistance (Birkner et al., 2004; Segal et al., 2004; Shafer et al., 2001a). The specific mutations observed in our dataset parallel those found in the literature. For example, $Mpr46I$,

Table 6: DR-IPTW Models (Note: $D = E(D(O_i|\theta_n, \Pi_n)|W)$

Codon	Model
$pr90$	$D = -0.449 - 2.719(pr43) + 1.014(pr54)$
$rt184$	$D = -0.584$
$pr43$	$D = -0.098 - 2.160(pr90) - 1.041(rt102)$
$pr54$	$D = -0.324 - 0.4886(rt121) + 0.6677(pr90) - 0.100(pr46) - 0.225(pr10)$ $-9.3028(rt121 * pr10) - 1.65(pr43) + 4.05(pr43 * pr46) + 3.952(pr10 * pr43)$
$rt41$	$D = -0.339 + 0.982(rt184) + 0.475(pr82) + 0.9821(pr43) - 0.357(pr46) - 1.303(pr43 * pr46)$ $+2.008(rt184 * pr43) + 10.256(pr43 * pr82) - 0.239(rt215) - 10.619(pr43 * rt215)$
$pr46$	$D = -0.369 - 3.152(rt121) + 0.580(pr90) - 0.372(pr71) - 0.557(rt41)$ $-2.849(pr90 * rt41) + 3.586(rt41 * pr71)$
$pr82$	$D = -0.117 - 1.515(pr43) + 0.0381(rt121) + 0.611(pr90) - 0.351(pr54) + 0.334(rt41)$ $+1.719(pr43 * pr54) - 4.514(pr54 * rt121) - 2.632(pr90 * pr43) + 3.093(rt41 * rt121)$
$rt215$	$D = 0.175 - 1.713(pr43)$
$rt121$	$D = 0.0609 - 4.346(pr10) + 0.975(pr43) + 0.673(rt41)$
$pr10$	$D = -0.0000935 - 4.037(rt121)$
$pr71$	$D = 0.351 - 0.603(pr43) - 0.190(rt102) - 2.644(pr43 * rt102)$
$rt102$	$D = 0.425 - 2.442(pr43)$
$rt214$	$D = 0.192 + 0.469(rt41) - 0.0535(rt121) + 2.697(rt121 * rt41)$
$pr57$	$D = 0.2642$
$pr14$	$D = 0.0926 - 3.679(pr43)$
$rt196$	$D = 0.007347$
$rt103$	$D = -0.1091$
$pr13$	$D = 0.1334 + 0.1184(pr10) - 1.6556(pr43) + 3.648(pr43 * pr10)$
$pr15$	$D = 0.1359 + 0.4573(pr43) - 0.4048(rt121) + 4.414(pr43 * rt121)$
$pr36$	$D = -0.0613 + 1.0887(rt215) + 1.7347(rt121)$
$pr64$	$D = -0.07115 + 0.527(rt121) + 0.396(pr43) + 3.876(pr43 * rt121)$
$rt83$	$D = 0.238 + 0.335(pr43) + 0.0114(rt102) + 3.0613(pr43 * rt102)$
$pr41$	$D = -0.2120 + 0.571(pr46) + 0.1438(rt121) + 3.4048(pr46 * rt121)$
$rt177$	$D = 0.20437 - 0.1530(rt121) - 0.3593(pr43) + 4.546(pr43 * rt121)$
$pr35$	$D = 0.00792 + 1.4499(rt121) + 1.003(pr46)$
$pr62$	$D = 0.08422$
$rt200$	$D = 0.00647 - 0.69739(rt102) + 0.81504(rt41) - 0.14504(pr43) - 3.2498(pr43 * rt102)$
$rt207$	$D = 0.0768 + 0.5336(rt121) + 0.0735(pr43) + 3.2798(pr43 * rt121)$
$rt162$	$D = 0.00971 + 1.6486(rt121)$
$pr77$	$D = 0.0544 - 1.561(pr54)$
$pr37$	$D = -0.1186 - 0.1748(rt121) + 0.4457(rt215) + 3.5289(rt121 * rt215)$
$rt122$	$D = 0.1603 - 1.239(pr43)$
$rt135$	$D = -0.0495 - 1.155(pr43)$
$pr93$	$D = -0.0142$
$pr63$	$D = -0.2001 + 1.0782(pr46) + 0.9524(rt102)$
$rt211$	$D = 0.0559 + 0.4844(rt121) + 0.2141(pr43) + 3.1499(pr43 * rt121)$
$rt178$	$D = 0.1780 - 0.9007(rt41)$

$Ipr54V/L/T$, and $Lpr90M$, correspond to protease positions in which mutations increase the resistance to various protease inhibitors. An example of a protease mutation is position $pr10$, where $Lpr10I/F/V/R$, one of the most common mutations, is associated with resistance to all protease inhibitors when present with another mutation. Position $pr90$ has an impact on the substrate cleft of the virus and $L90M$ causes resistance to saquinavir when combined with various other mutations (Shafer et al., 1998). For example, the $Gpr48V/Lpr90M$ double mutation has shown delayed viral replication, whereas $Lpr90M$ alone had a higher replication capacity (Goudsmit et al., 1997). Position $Ipr54V/L/T$ also causes resistance to the other protease inhibitors when present with other mutations. Mutations at residues $pr54$ and $pr82$ produce resistance to Indinavir and Ritonavir. Additionally, mutations within $Vpr82A$, $Ipr84V$, and $Lpr90M$ have been associated with a median change in replication capacity (Barbour et al., 2002b; Shafer et al., 2001b). Mutations in several of the identified codons also have an impact on the replication capacity of the virus. Mutation $Mrt184V/I$ suppresses the wild-type activity of $Trt215Y$, thus decreasing AZT resistance (Shafer et al., 2001a). AZT, also known as Zidovudine, is a nucleoside reverse transcriptase inhibitor. It affects HIV's ability to replicate by producing faulty reverse transcriptase and hence inhibiting the transcription of RNA to DNA.

The results presented in this paper are consistent with previous research and other analyses of this HIV-1 dataset. The reader is referred to earlier articles by Birkner et al. (2004) and Segal et al. (2004) for alternative statistical analyses and biological discussion of a related HIV-1 dataset.

References

J. D. Barbour, T. Wrin, R. M. Grant, J. N. Martin, M. R. Segal, C. J. Petropoulos, and S. G. Deeks. Evolution of Phenotypic Drug Susceptibility and Viral Replication Capacity during Long-Term Virologic Failure of Protease Inhibitor Therapy in Human Immunodeficiency Virus-Infected Adults. *Journal of Virology*, 76(21):11104–11112, 2002a.

Jason D. Barbour, Terri Wrin, Robert M. Grant, Jeffrey N. Martin, Mark R. Segal, Christos J. Petropoulos, and Steven G. Deeks. Evolution of Phenotypic Drug Susceptibility and Viral Replication Capacity during Long-Term Virologic Failure of Protease Inhibitor Therapy in Human Immunodeficiency Virus-Infected Adults. *Journal of Virology*, 76(21):11104–11112, 2002b.

M. D. Birkner, S. E. Sinisi, and M. J. van der Laan. Multiple testing and data adaptive regression: An application to HIV-1 sequence data. Technical Report 161, Division of Biostatistics, University of California, Berkeley, 2004. URL `www.bepress.com/ucbbiostat/paper161`.

M. D. Birkner, S. E. Sinisi, and M. J. van der Laan. Multiple testing and data adaptive regression: An application to HIV-1 sequence data. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005. URL `http://www.bepress.com/sagmb/vol4/iss1/art8`.

Jaap Goudsmit, Anthony de Ronde, Ester de Rooij, and Rob de Boer. Broad Spectrum of in Vivo Fitness of Human Immunodeficiency Virus Type 1 Subpopulations Differing at Reverse Transcriptase Codons 41 and 215. *Journal of Virology*, 71(6):4479–4484, 1997.

M. R. Segal, J. D. Barbour, and R. M. Grant. Relating HIV-1 Sequence Variation to Replication Capacity via Trees and Forests. 3(1):Article 2, 2004. URL `www.bepress.com/sagmb/vol3/iss1/art2`.

R. W. Shafer, K. M. Dupnik, M. A. Winters, and S. H. Eshleman. A Guide to HIV-1 Reverse Transcriptase and Protease Sequencing for Drug Resistance Studies. In *HIV Sequencing Compendium*, pages 83–133. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, 2001a.

Robert W. Shafer, Mark A. Winters, Sarah Palmer, and Thomas C. Merigan. Multiple Concurrent Reverse Transcriptase and Protease Mutations and

Multidrug Resistance of HIV-1 Isolates from Heavily Treated Patients. *Annals of Internal Medicine*, 128(11):906–911, 1998.

Robert W. Shafer, Kathryn M. Dupnik, Mark A. Winters, and Susan H. Eshleman. A Guide to HIV-1 Reverse Transcriptase and Protease Sequencing for Drug Resistance Studies. In *HIV Sequencing Compendium*, pages 83–133. Theoretical Biology and Biophysics Group at Los Alamos National Laboratory, 2001b.

Mark J. van der Laan. Statistical Inference for Variable Importance. Technical Report 188, Division of Biostatistics, University of California, Berkeley, 2005. URL `www.bepress.com/ucbbiostat/paper188`.