Genetic basis of variation in tenofovir drug susceptibility in HIV-1

Robert J. Murray[a], Fraser I. Lewis[a,c], Michael D. Miller[b] and Andrew J. Leigh Brown[a]

Objective: To develop an improved model for the genetic basis of reduced susceptibility to tenofovir *in vitro*.

Methods: A dataset of 532 HIV-1 subtype B reverse transcriptase genotypes for which matched phenotypic susceptibility data were available was assembled, both as a continuous (transformed) dataset and a categorical dataset generated by imposing a cut-off on the basis of earlier studies of in-vivo response of 1.4-fold. Models were generated using stepwise regression, decision tree and random forest approaches on both the continuous and categorical data. Models were compared by mean squared error (continuous models), or by misclassification rates by nested crossvalidation.

Results: From the continuous dataset, stepwise linear regression, regression tree and regression forest methods yielded models with MSE of 0.46, 0.48 and 0.42 respectively. Amino acids 215, 65, 41, 67, 184 and 151 in HIV-1 reverse transcriptase were identified in all three models and amino acid 210 in two. The categorical data yielded logistic regression, classification tree and forest models with misclassification rates of 26, 24 and 23%, respectively. Amino acids 215, 65 and 67 appeared in all; 41, 184, 210 and 151 were also included in the classification forest model.

Conclusion: The random forests approach has yielded a substantial improvement in the available models to describe the genetic basis of reduced susceptibility to tenofovir *in vitro*. The most important sites in these models are amino acid sites 215, 65, 41, 67, 184, 151 and 210 in HIV-1 reverse transcriptase.

AIDS 2008, **22**:1113–1123

Keywords: antiretroviral therapy, antiviral drug resistance, classification trees, decision trees, HIV-1, machine learning, nucleoside reverse transcriptase inhibitors, reverse transcriptase, tenofovir, thymidine analog mutations

Introduction

Tenofovir disoproxil fumarate (TDF) is an oral prodrug of tenofovir, an acyclic nucleotide analogue, and is a widely used and highly potent antiretroviral demonstrating significant virological activity in both treatment-naïve and treatment-experienced HIV-1 infected patients [1–3]. Even among patients harbouring drug-resistant HIV-1 infection, treatment with TDF results in a decrease of approximately 0.6 $\log_{10}$ HIV-1 RNA copies/ml of plasma by week 24 [1]. Tenofovir is unique among the other FDA-approved nucleoside reverse transcriptase inhibitors (NRTIs) in showing continued activity against a wide variety of well characterized NRTI resistant-strains [4–6].

Resistance to NRTIs is mediated by mutations that impair the incorporation of nucleoside analogues into growing proviral cDNA, thereby preventing premature termination of chain elongation [7,8] and by mutations that excise incorporated analogues from prematurely

From the [a]Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, Scotland, UK, the [b]Gilead Sciences, Inc., Foster City, California, USA, and the [c]Veterinary Epidemiology Research Unit, SAC, Inverness IV2 4JZ, Scotland, UK.

Correspondence to Andrew J. Leigh Brown, PhD, University of Edinburgh, Ashworth Laboratories, West Mains Road, Edinburgh EH9 3JT, Scotland, UK.

E-mail: A.Leigh-Brown@ed.ac.uk
Received: 9 July 2007; revised: 13 February 2008; accepted: 19 March 2008.

terminated proviral DNA strands, allowing for continued cDNA synthesis [9,10]. The latter, thymidine analogue mutations (TAMs), include M41L, D67N, K70R, L210W, T215Y/F and K219Q/E. The occurrence of a single TAM is usually insufficient to confer significant levels of resistance, except for T215Y and zidovudine [11]. However, the accumulation of multiple TAMs leads to increasing levels of resistance and crossresistance between all NRTIs, including lamivudine [6,12,13]. Other mutations that confer multinucleoside resistance, by impairing the incorporation of nucleoside analogues, include mutations at residue T69, K65R, L74V and the so-called Q151M pathway [14–16].

Recently, much has been learned about the relationship between mutations in the HIV-1 genome and in-vitro phenotypic variation in susceptibility from statistical or machine learning analysis, including linear regression [17,18], linear discriminant analysis [19], classification and regression trees [20,21], artificial neural networks [22] and support vector machines [23]. These methods provide automated analytical routines that can process large volumes of sequence data, from multiple amino acid sites, matched with in-vitro drug susceptibility phenotype.

Although many models for NRTI in-vitro drug susceptibility phenotype currently exist, there are, however, very few models specifically for tenofovir and those that are available [18,24] are based on smaller datasets (up to 350 genotype–phenotype pairs) than those analysed for the other NRTIs (typically >500 isolates). In this study, we build and validate interpretable genotypic models predictive of tenofovir in-vitro drug susceptibility phenotype on a dataset of 532 genotype–phenotype pairs, using stepwise linear/logistic regression, decision tree and random forest analysis. We explore multiple methods to assess the extent of similarity between the performances of the resulting models and compare with earlier models of tenofovir in-vitro susceptibility.

Materials and methods

Genotype–phenotype dataset: collection

We obtained 237 HIV-1 viral sequences from the Stanford HIV-1 drug resistance database [25] for which the IC_{50} fold-change of tenofovir was available. Sequences were aligned using the HyPhy protein coding alignment tool [26] (HyPhy version 1.0 (http://www.hyphy.org/) to the reverse transcriptase gene of HIV-1 clone HXB2 (GenBank: accession number K03455) and were translated into amino acid sequences using an in-house sequence translator.

To this dataset we added a further 295 HIV-1 genotype–phenotype pairs for tenofovir. This dataset comprises lists of reverse transcriptase amino acid mutations (departures from consensus) from 105 clinical trial-derived plasma samples from highly treatment-experienced patients with multiple TAMs (median three) along with other NRTI, NNRTI and PI-associated mutations; 46 samples from treatment-naïve patients failing a regimen of TDF, lamivudine and efavirenz ($n = 16$ with K65R); and 144 HIV-1 samples from commercial NRTI resistance panels expressing a variety of NRTI mutations with or without additional drug resistance mutations.

Phenotypic data for tenofovir were available for all samples with either the antivirogram (Virco, Mechelen, Belgium) or the HIV-1 PhenoSense assay (Monogram Biosciences, South San Francisco, California, USA). Genotypes were grouped together, regardless of whether they were analysed by the antivirogram or PhenoSense phenotypic assays as the available information suggests that differences between assays are of the same order as the within assay variance [27,28,29].

Genotype–phenotype dataset: representation

Mutations were defined as departures from wild-type HIV-1 subtype B consensus and analysed as binary variables. Sites containing mixtures of amino acids were treated as mutant. Use of a binary classification instead of multiple categories resulted in a negligible loss in predictive power as previously found [30]. The most prominent example of a significant site, with multiple mutations having different effects, is position 215 in reverse transcriptase. T215F was present in 9% of genotypes in this dataset whereas T215Y was present in 40%. All sites where the mutant amino acids were present at frequencies of 4% or lower were excluded from analysis due to the lack of power to detect an effect.

For regression analysis of continuous data, IC_{50} fold-change values were normalized by a Box–Cox power transformation [31]. Normality was retrospectively validated by the Shapiro–Wilk test [32]. For classification analysis, we adopted a threshold of 1.4-fold, which has been shown to be indicative of reduced virological response to TDF in treatment-experienced patients [33–35].

Linear and logistic regression

Forward stepwise selection was used to optimize the number of amino acid sites included in the logistic and linear regression models using change in the Akaike information criterion (AIC) to determine the total number of sites incorporated. The final models were further reduced on the basis of the change in deviance at each step.

Decision tree analysis

Decision trees were created by successively splitting the genotype–phenotype dataset until no further splits improved the accuracy [20]. For regression analysis, splits were determined using the least-squares deviation criterion and for classification analysis, splits were determined by the information gain metric [36,37]. To

avoid overfitting, trees were pruned using 10-fold crossvalidation (described below).

Random forest analysis

The method of random forests is an extension of decision tree learning [38] whereby splits are determined on a random subset of the available amino acid sites at each leaf. An ensemble, or forest, of such trees is created by growing individual trees on bootstrap samples of the data. The final predictions are the unweighted average of the predictions from the individual trees. The number of trees in the forest, the total number of amino acid sites randomly selected at each leaf and the maximum tree depth was optimized by 10-fold crossvalidation. For interpretation, the importance of each amino acid site to predict phenotype is determined. For this purpose, we used the 'permutation accuracy importance' measure, which estimates the difference in the accuracy of the forest before and after permuting the amino acids at each site [38].

Evaluation of the models: nested 10-fold cross-validation

To assess the performance of the models to predict phenotype in unseen genotypes we applied standard 10-fold cross-validation. For the regression models the generalization error is measured by the MSE and for the classification models the generalization error is measured by the percentage of genotypes misclassified and the model's sensitivity and specificity. To provide an unbiased estimate of the generalization error, we used a nested grid-search to optimize model parameters [39]. To generate classification models which maximize the trade-off between sensitivity and specificity we used a nested ROC analysis (not shown) to optimize the interpretation of probabilistic model outputs in terms of discrete 'susceptible' and 'resistant' classifications [40].

Evaluation of existing tenofovir models

To assess how well our models perform in comparison to existing models for tenofovir resistance, we obtained phenotype estimates for each genotype in the dataset from the ANRS-AC11 [41], Rega [42], Stanford HIV-*db* [43] and geno2pheno [44] systems. To standardize the predictions between systems and models, we converted the multiple prediction levels from the HIV*db*, ANRS-AC11 and Rega systems into either 'resistant' or 'susceptible' by imposing an 'ordered' binary partition over the levels. To select an optimal partition, we calculated an ROC curve using 10-fold crossvalidation and the partition which produced the 'best' tradeoff between sensitivity and specificity was chosen. Finally, this learning phase was nested within *N*-fold cross-validation (*N* is the total number of genotypes in the dataset) to obtain an unbiased 'out-of-sample' phenotype prediction for each genotype.

Results

Phenotype fold-change distribution

Analysis of the frequency distribution of tenofovir IC_{50} fold-change values revealed a low level of variation between samples (Fig. 1) with 75% of the samples having a susceptibility fold-change of 2.2 or less (1st quartile: 0.7-fold; median: 1.2-fold; 3rd quartile: 2.2-fold). This contrasts with most other NRTIs, where the range of fold-change values is much larger [20] and unlike tenofovir, the fold-change distribution is typically bimodal [20,45]. A straightforward separation of the strains into tenofovir-resistant and tenofovir-susceptible groups for classification analysis was not possible; hence, we adopted a 1.4-fold cut-off from previous clinical data [33–35]. Using this threshold resulted in 44% of our strains being classified as resistant.

Linear and logistic regression models

Forward stepwise logistic regression based on a 1.4-fold cut-off found a simple model containing only four previously known resistance-associated amino acid sites: 215, 65, 77, 67 ($P < 2.2e^{-16}$). Independently these sites were found to be highly significant ($P < 0.001$) in predicting tenofovir resistance (at the 1.4 IC_{50} fold-change cut-off). Mutations at site 65 lead to a 2.8-fold increase in the probability of resistance (compared with the intercept); similarly for mutations at site 77. Of the two TAM sites (215 and 67), mutations at 215 had the largest effect, leading to a 1.9-fold increase in the probability of resistance. Mutations at 67 increased the probability of resistance by approximately 1.6-fold.

Forward stepwise linear regression, based on AIC, resulted in a model with 23 amino acid sites. This model explained 51% of the total variation in IC_{50} fold-change ($\bar{R}^2$; R^2 adjusted for the total number of sites). To identify a more parsimonious model we compared the difference in deviance between the fully saturated model (which includes all amino acid sites with $\geq$4% variation) and the individual models identified at each stage of the stepwise procedure with the difference in value of the AIC for each model in the process. Although the AIC decreases with the inclusion of additional amino acid sites, the difference in deviance begins to plateau after the first nine amino acid sites have been added to the model – a clear indication that the AIC criterion has over-fitted (Fig. 2). We therefore favour a model containing only the nine amino acid sites: 215, 65, 67, 184, 210, 228, 41, 39 and 115 ($P < 2.2e^{-16}$). All of these sites were found to be significant at the 5% level ($P < 0.05$) in the re-fitted model. Although this model contains less than half the sites, it explains almost as much of the total variation in IC_{50} fold-change ($\bar{R}^2 = 47\%$ *vs.* 51% for the 23-site model).

Again, mutations at amino acid site 65 had the greatest impact on resistance, leading to a 7.8-fold increase in the normalized-IC_{50} fold-change (*n*FC). Mutations at amino acid site 39 had the smallest impact, leading to a 2.1-fold

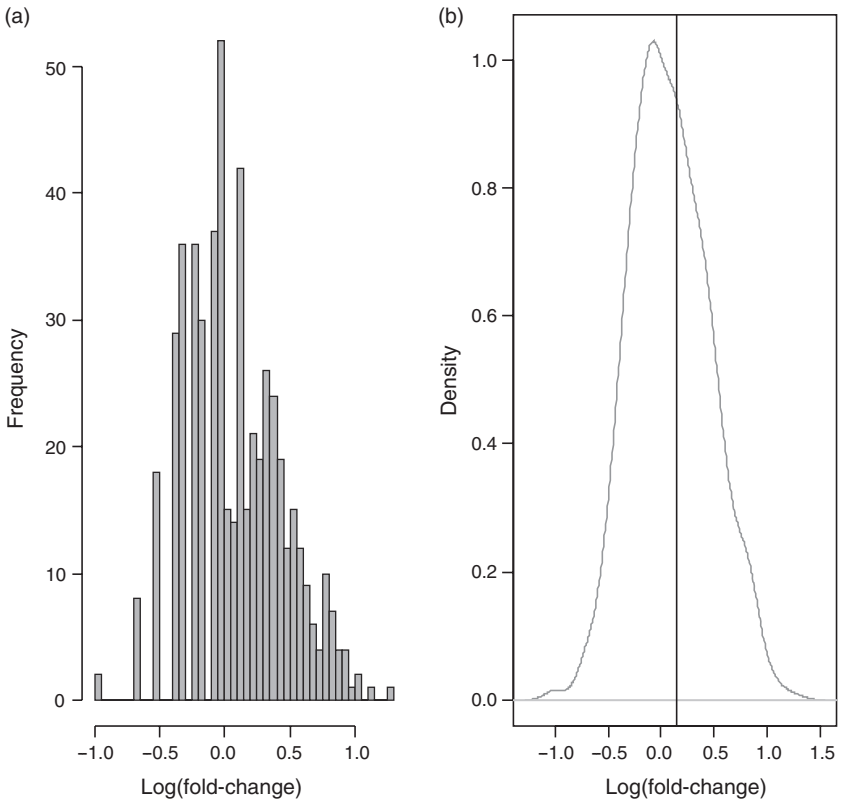

Fig. 1. Analysis of the frequency distribution of tenofovir IC_{50} fold-change values. Histogram (a) and density plot (Gaussian kernel). (b) The vertical line shows the location of the 1.4-fold cut-off: strains with a susceptibility of at least 1.4 -fold relative to the wild type control strain are classified as resistant.

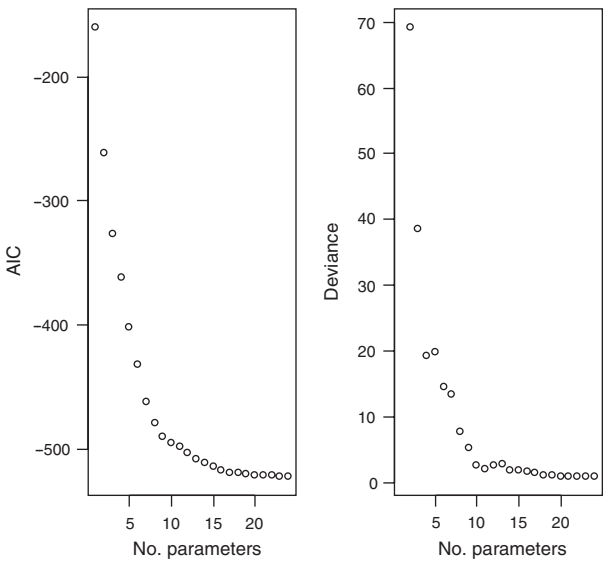

Fig. 2. Change in AIC (left) and Deviance (right) during forward stepwise linear regression. As more sites are added to the model, the value of AIC continues to decrease. However, after the first nine sites have been added, the difference in deviance between the fully saturated model (containing all sites) and the current model stabilizes.

increase in *n*FC. The impact of mutations at TAM sites (215, 67, 210 and 41) was similar for 215, 210 and 41 – leading to approximately a three-fold increase in *n*FC. Mutations at site 67 increased *n*FC by approximately 3.8-fold. Mutations at site 184 had a hypersensitizing effect, causing a reduction in *n*FC by approximately 2.6-fold.

Decision tree models

Regression tree

On the basis of 10-fold crossvalidation, the best tree for predicting IC_{50} fold-change (the regression tree) had eight splits (Fig. 3). Each split represents an approximate 1.6-fold differential in IC_{50}, such that the ratio of the geometric mean of the fold-change values in the right branches to that in the left branches is approximately 1.6 (range: 0.39–2.7; median: 1.5). The genotypes with the lowest IC_{50} fold-change (0.7) were assigned to the leftmost leaf of the tree: 177 strains with wild-type amino acids at sites 215, 65, 151 and 70. Genotypes with the highest IC_{50} fold-change (5.04) were assigned to the rightmost leaf of the tree: 43 strains with mutations at sites 215, 41 and 67 and wild-type at site 184. An advantage of tree-based models is their ability to capture interaction effects, in this case the interaction between amino acids 215 and 184, which has previously been described [4,34]. Thus, if 215 is mutant and site 184 is wild-type, then the predicted IC_{50} fold-change of TDF is 3.3 (data not shown); in contrast, if 215 is

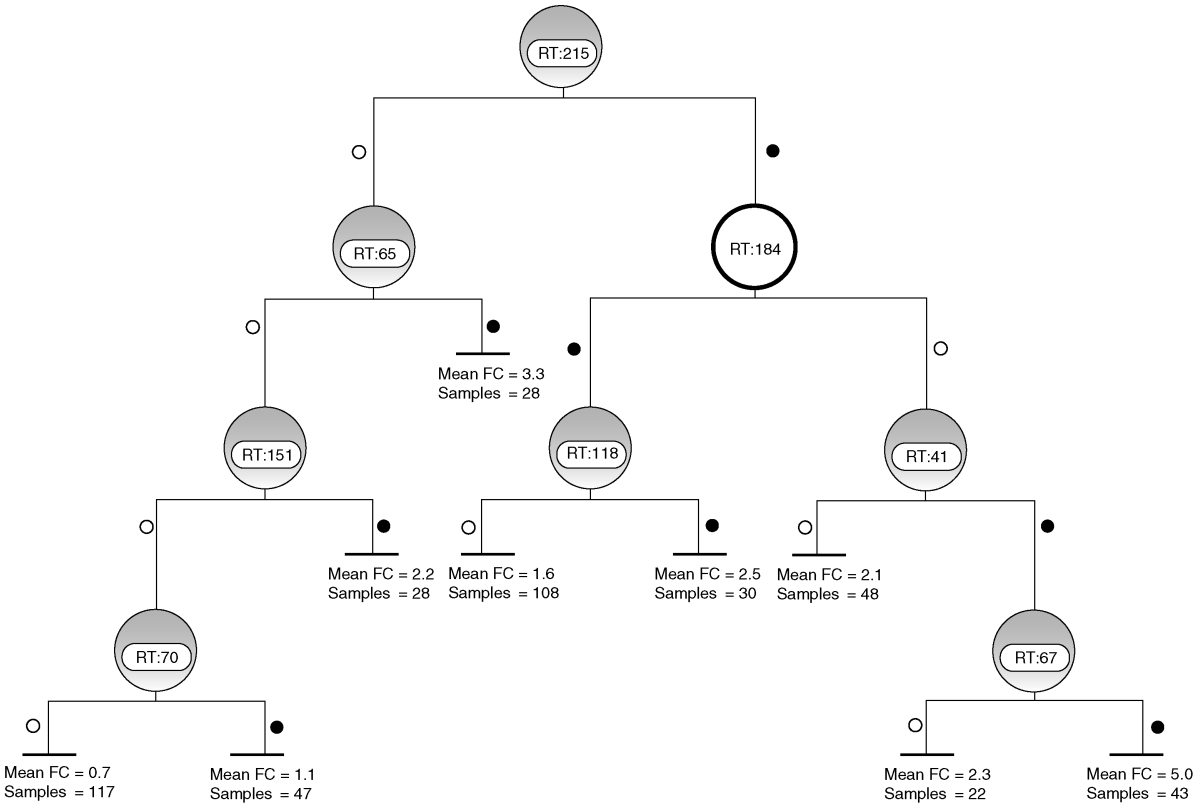

Fig. 3. Regression tree for tenofovir in-vitro drug susceptibility fold-change. Leaf nodes (⊥) give an estimate of IC_{50} fold-change (the mean IC_{50} fold-change of the genotypes in the training sample sorted to the leaf) for an arbitrary genotype sorted to the leaf. A genotype is sorted to a leaf as follows: start at the root node (labelled by RT:215) and test residue 215: if 215 is 'mutant' follow the branch labelled by the black dot; otherwise, if 215 is 'wild-type', follow the branch labelled by the open circle; repeatedly test for mutations at the residues specified by each descendent node until a leaf node is reached; and return the value at the leaf. Note that the RT:184 node is highlighted in bold because it has a hypersensitizing effect within the model. Following its mutant branch leads to an increase in susceptibility to tenofovir and following its wild-type branch leads to a decrease in susceptibility.

mutant and 184 is mutant, then fold-change is 1.6. In the latter case, an additional mutation is required before a larger reduction in tenofovir susceptibility is observed. Of 138 strains with 215 and 184, those with a mutation at 118 as well ($n = 30$) had a mean fold-change of 2.5, a substantial increase although still less than those that were wild type at 184. Cases that were wild type at 118 ($n = 108$) had a mean fold-change of 1.6.

Classification tree

Similar to the regression tree, the best tree for predicting the IC_{50} 1.4 fold-change cut-off had eight splits (Fig. 4). Again 215 was at the root with 65 and 151 on one branch and site 184 and others (including multiple TAMs) on the other. The hypersusceptibility effect of amino acid site 184 is also reflected in the classification tree: the probability of a genotype with a mutant 215 and wild-type 184 being resistant to tenofovir is 77%, while that of a genotype with a mutant 215 and mutant 184 being resistant is 51%.

Whereas the amino acid sites included in the regression tree (215, 184, 41, 67, 65, 151, 118 and 70) have all been previously associated with NRTI resistance, the classification tree includes two novel sites (211 and 207). A high proportion of the total dataset (49%) had a mutation at site 211, usually lysine. Overall, 211 does not distinguish resistant from susceptible genotypes – of the 258 genotypes with a mutation at 211, exactly half were resistant. However, among the 75 genotypes with mutations at 184 and 215, site 211 does appear to discriminate with 50 (67%) resistant. Replacing 211 with 210 had a similar, but slightly lesser effect (not shown). We also observed a novel association between mutations at site 207 and reduced tenofovir susceptibility. Of 133 genotypes with a mutation at 207 (usually glutamic acid) – 74 (56%) were resistant and 59 were susceptible. For the 398 cases that were wild-type at 207, 158 (40%) were resistant.

Random forest models: permutation accuracy importance

Permutation accuracy importance (PAI) was used to rank amino acid sites according to their impact on the capacity of the random forest models. The most important amino acid site by far was 215 (Fig. 5(a)). By permuting the arrangement of the mutations at this site we observed an approximately 14% increase in the MSE of the forest.

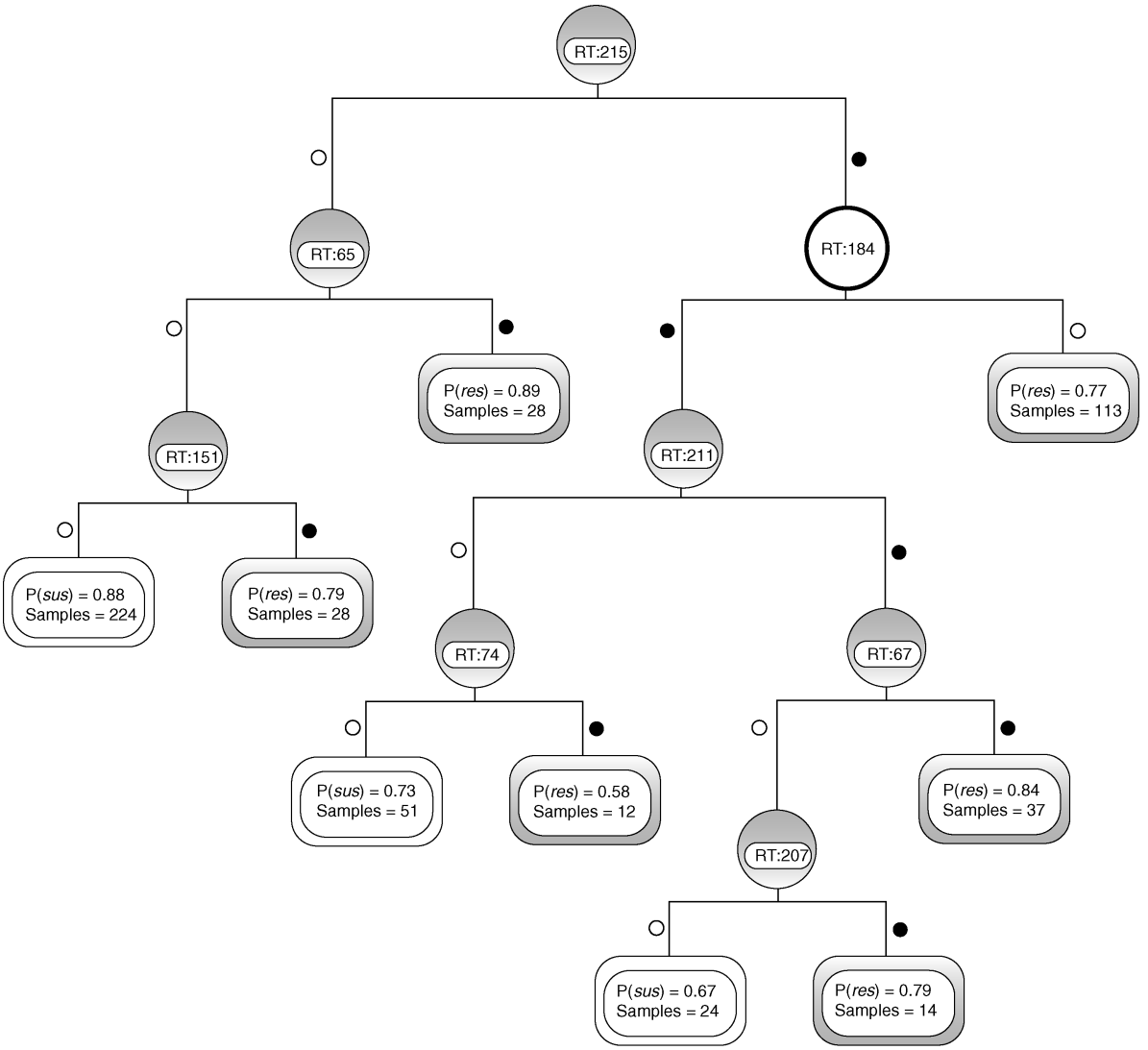

Fig. 4. Classification tree for tenofovir in-vitro resistance, where a 1.4-fold cut-off was adopted. Leaf nodes (⊥) give a resistance classification for a genotype allocated to the leaf-shaded rectangles denote 'resistant' classifications and clear rectangles denote 'susceptible' classifications. Genotypes are sorted to a leaf as described in the legend to Fig. 3 [33–35].

Amino acid site 65 had a lesser impact (∼8%), followed by 41 (∼6%), 67 (∼5.5%), 184 (∼5%) and 210 (∼4%). Changes to sites 219 and 151 lead to a marginal increase in MSE (∼2%). Similar results were found for the classification forest, predicting the IC_{50} 1.4 fold-change cut-off (Fig. 5(b)). Again the most important amino acid sites were 215 and 65, with changes at these sites leading to a ∼5.5 and ∼3.2% decrease in the percentage of correctly predicted samples, respectively. Site 41 had a similar impact to site 65, leading to a decrease of ∼3%. Site 67 had a lesser impact (∼2%), followed by 184 (∼1.5%), 151 (∼1.5%) and 210 (∼1%).

Model performance: comparison by nested cross-validation

The MSE was estimated for each regression model by nested 10-fold crossvalidation (Table 1). The regression forest model had the lowest MSE (0.42), and would thus be favoured over the nine-site linear regression model (0.46) or the regression tree model (0.48). For the classification models, the out-of-sample misclassification rates were also estimated by nested 10-fold crossvalidation. For these models, the percentage of samples misclassified was similar for the classification tree (24%) and the random forest (23%) and slightly higher for the logistic regression model (26%). The percentage of correctly predicted resistant samples (sensitivity) varied between models, being higher for the classification tree (81%) than either the classification forest (78%) or the logistic regression model (72%). On the other hand the classification forest and the logistic regression model (76%, respectively) were superior to the classification tree (68%) in correctly predicting susceptible samples (specificity). Overall, the classification forest represents the most balanced and accurate classifier (its sensitivity and specificity are high and similar). The threshold adopted throughout for classifying cases was 1.4-fold, for reasons

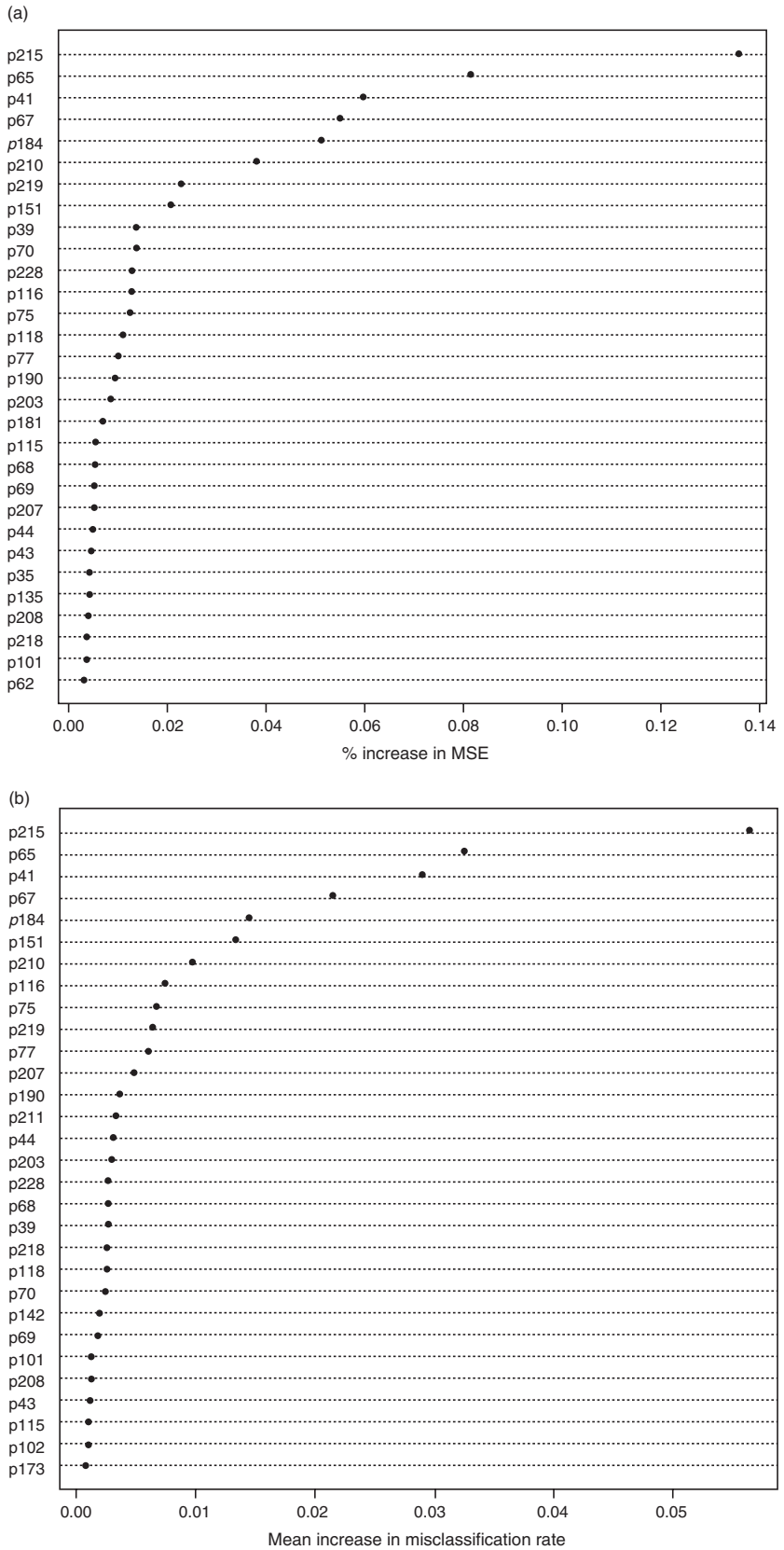

Fig. 5. Permutation accuracy importance (PAI) of amino acid sites in the random forest models. (a) Regression forest model. (b) Classification forest. Amino acid sites are ordered according to their impact on the performance of the model. RT:184 is in italics because it has a hypersensitizing effect within the forest [38].

Table 1. The estimated out-of-sample prediction accuracy of the individual models - estimated by nested 10-fold cross-validation.

Model	Regression MSE	Classification		
		Misclassification rate	Sensitivity	Specificity
Linear regression	0.46 (0.07)			
Logistic regression (%)		26 (4)	72 (10)	76 (10)
Regression tree	0.48 (0.05)			
Classification tree (%)		24 (7)	81% (7)	72% (9)
Regression forests	0.42 (0.09)			
Classification forests (%)		23 (4)	78 (8)	76 (8)

The standard deviation of the estimates, over the 10 individual folds, is given in parenthesis. MSE: mean squared error.

given earlier. We extended the study to investigate different thresholds, looking at cut-offs of 1.2-fold and 1.6-fold. The models obtained were very similar overall to those presented, both in predictive performance (4% or less variation in misclassification rate) and in structure (data not shown). In all models the key sites were 215, 65, 184 and 41, with 215 the most significant.

Comparison of different models

The Rega (32%), ANRS-AC11 (30%) and geno2pheno (28%) systems all have a similar proportion of misclassified samples when tested on this dataset that is clearly higher than the model obtained here, whereas that for the HIV*db* system (24%) is similar (Table 2). While the geno2pheno system achieved a higher sensitivity (92%) than the HIV*db* (85%), ANRS-AC11 (75%) and the Rega (52%) systems, the Rega system had a higher specificity (81%) than the HIV*db* (69%), ANRS-AC11 (66%) and the geno2pheno (57%) systems. Overall, the classification forest provides a more balanced classifier (misclassification rate: 23%; sensitivity: 78%; specificity: 76%) than any existing system.

Discussion

A number of statistical and machine learning models for HIV-1 drug resistance have been proposed, including linear regression, linear discriminant analysis, decision trees, artificial neural networks and support vector machines. The accuracy of the models to correctly predict in-vitro drug resistance from genotype is typically very high, explaining 65–89% of the total variation in IC_{50} fold-change for all drugs [23], with misclassification rates ranging between 9.5–13.5% for most drugs [17,20]. Concise and easily interpretable decision tree models are available for NRTI drug resistance [20]. These models were able to identify many well known NRTI resistance-associated mechanisms, including the association between M184V and high-levels of resistance to lamivudine, mutations associated with the Q151M pathway (Q151M, V75I, F77L), L74V and multiple TAMs (M51L, D67N, K70R, L210W, T215Y/F). There are, however, very few models for tenofovir *in vitro* drug susceptibility. The most recent model, based on a dataset of approximately 350 matched genotype–phenotype pairs [18] performed substantially less well than those obtained for other drugs in that study.

Table 2. The estimated predictive accuracy of the existing tenofovir models.

Model	Misclassification rate (%)	Sensitivity (%)	Specificity (%)
ANRS-AC11 [41]	30	75	66
Rega [42]	32	52	81
HIV*db* [43]	24	85	69
geno2pheno [44]	28	92	57
Classification Forests	23	78	76

The performance of our 'best' performing model (classification forest) is also shown.

We have obtained a series of genotypic models to predict tenofovir in-vitro drug susceptibility phenotype, using stepwise linear/logistic regression, decision tree and random forest analysis. On the basis of a large dataset ($N = 532$), while permitting identification of the genetic determinants, they provided accurate predictions of resistance and susceptibility in approximately 77% of cases, a substantial improvement on the ANRS-AC11, Rega and geno2pheno systems. The percentage of cases misclassified was similar for the random forest model (23%) and for the HIV*db* system (24%). However, the true misclassification rate of the HIV*db* system may be underestimated. Approximately half the samples used here came from that source and are likely to have contributed to the data used to develop the Stanford model causing its misclassification rate to be biased downwards.

Both statistical and machine learning methods successfully predicted tenofovir resistance for 'unfamiliar' genotypes. The optimal linear regression model identified nine amino acid sites: 215, 65, 67, 184, 210, 228, 41, 39 and 115. The logistic regression model included the sites 215, 65, 77 and 67. The difference in the complexity of the linear and logistic regression models can be explained in terms of the response, as the linear regression model explains the entire range of the phenotype fold-change distribution, while the logistic regression model explains the phenotypic variation around the 1.4-fold cut-off only. Similar sites to those present in the linear and logistic regression models were also identified in the decision tree models (Figs 3–5). Overall, seven reverse transcriptase amino sites were identified in multiple models (215, 65, 41, 67, 184, 151 and 210).

Most of the sites identified in the models have been previously associated with resistance to the other NRTIs [46]. Of particular interest is the occurrence of multiple

TAMs: 41, 67, 70, 210 and 215. Response to TDF is significantly reduced among patients with HIV-1 containing at least 3 TAMs, inclusive of the M41L or L210W mutations [33,35]. We also know that tenofovir selects for the K65R mutation *in vitro* and in about 2–4% of antiretroviral-experienced patients *in vivo*; this mutation results in a three-fold to four-fold decrease in tenofovir in-vitro susceptibility and it is correlated with impaired virological response to TDF [2,4]. Although the K65R mutation is relatively rare in clinical cohorts, its frequency has increased since the introduction of tenofovir into clinical practice: from ~0.4% in 1998 to 3.6–7.3% in 2003–2005 [47,48]. The potential for further increases in K65R prevalence has implications for NRTI-based therapy, since zidovudine is the only NRTI to retain activity against K65R-mutant strains [47].

In contrast to the rather rare appearance of K65R in patients, M184V and T215Y/F are the most common NRTI resistance mutations [48]. Both mutations appear in all models identified in this study, with 215 at the root of all trees. This is consistent with the results from a smaller study that included tree-based models for tenofovir in-vitro resistance [24]. Thus, T215Y/F appears to be highly significant for predicting tenofovir resistance. This is a key finding because T215Y/F can appear in as many as 42% of patients receiving antiretroviral therapy [49].

Unlike the other NRTI-associated mutations, the presence of M184V enhances the susceptibility of HIV-1 to tenofovir, consistent with previous in-vitro studies [4,50,51]. This hypersusceptibility effect is explicit in both our regression and classification trees (Figs 3 and 4). There is a more complex aspect to the relationship between 215 and 184, whereby a mutation at 118 counteracts the hypersusceptibility conferred by M184V/I in the regresion tree model. Although this site does not often appear in NRTI models, these observations are consistent with previous results on the effect of mutations at 118 on lamivudine susceptibility [52]. These interaction effects are explicit in the tree models, however, when the 215/184 interactions was specified as a covariate in stepwise linear and logistic regression analysis it was not included in the final model, and when added to the regression models, it gave no improvement. Finally, none of our models included the K65R/M184V interaction [53], but only 2% of strains in our dataset had mutations at both site 65 and 184.

Linear regression identified two novel amino acid sites: 39 and 228. These mutations have been previously associated with NRTI resistance and drug experience in database analyses [54]. Two further novel sites were identified in the classification tree: 207 and 211. All four of these sites were also identified by the classification forest and sites 39, 207 and 228 were identified by the regression forest. Sites 39, 207 and 211 are polymorphic with substitutions commonly present among treatment-naïve patients [54], however, the impact of these sites for predicting resistance was marginal (Fig. 5). It appears that the inclusion of 211 is through association effects: mutations at 211 were not significant for all but the classification tree model and mutations at 210 (a known TAM site) were significant for most of the other models. When we replaced 211 in the classification tree with 210 we observed a negligible loss in predictive power. Furthermore, when constructing multiple classification trees from bootstrap samples of the data (i.e. the classification forest model) the effect of site 211 was removed almost entirely (Fig. 5(b)). This highlights the difficulty in deriving standardized genotypic models to predict drug susceptibility and the importance of comparing several different models.

In conclusion, using multiple models we have been able to identify a subset of mutations in HIV-reverse transcriptase that appear to be most significant for tenofovir resistance: 215, 65, 41, 67, 184, 151 and 210. We conclude that other mutations are of marginal significance, as they appear in only a subset of the models tested and the cross-validated performance of the models was similar. This study illustrates the power of amalgamative models (such as random forests) over single models (such as linear/logistic regression tree models and decision trees) because they are less likely to be significantly influenced by the idiosyncrasies of individual datasets. These models offer an improvement over the best performing models to date.

Acknowledgements

R.J.M compiled the datasets, perfomed analysis and wrote the first draft of the manuscript; F.J.L advised on statistical methods; M.D.M contributed data; A.J.L.B conceived the project and prepared the final version of the manuscript.

We are very grateful to Professor C.K.I. Williams for many helpful discussions. This work was supported by the Biotechnology Biological Sciences Research Council (BBSRC) and the Wellcome Trust.

References

1. Squires K, Pozniak AL, Pierone G Jr, Steinhart CR, Berger D, Bellos NC, *et al*. **Tenofovir disoproxil fumarate in nucleoside-resistant HIV-1 Infection: a randomized trial.** *Ann Intern Med* 2003; **139**:313–320.
2. Margot NA, Isaacson E, McGowan I, Cheng AK, Schooley RT, Miller MD. **Genotypic and phenotypic analyses of HIV-1 in antiretroviral-experienced patients treated with tenofovir DF.** *AIDS* 2002; **16**:1227–1235.

3. Schooley RT, Ruane P, Myers RA, Beall G, Lampiris H, Berger D, *et al*. **Tenofovir DF in antiretroviral-experienced patients: results from a 48-week, randomized, double-blind study.** *AIDS* 2002; **16**:1257–1263.
4. Wainberg MA, Miller MD, Quan Y, Salomon H, Mulato AS, Lamy PD, *et al*. **In vitro selection and characterization of HIV-1 with reduced susceptibility to PMPA.** *Antivir Ther* 1999; **4**:87–94.
5. Srinivas RV, Fridland A. **Antiviral activities of 9-R-2-phosphonomethoxypropyl adenine (PMPA) and bis(isopropyloxymethylcarbonyl)PMPA against various drug-resistant human immunodeficiency virus strains.** *Antimicrob Agents Chemother* 1998; **42**:1484–1487.
6. Miller MD, Margot NA, Hertogs K, Larder B, Miller V. **Antiviral activity of tenofovir (PMPA) against nucleoside-resistant clinical HIV samples.** *Nucleosides Nucleotides Nucleic Acids* 2001; **20**:1025–1028.
7. Huang H, Chopra R, Verdine GL, Harrison SC. **Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance.** *Science* 1998; **282**:1669–1675.
8. Larder BA, Stammers DK. **Closing in on HIV drug resistance.** *Nat Struct Biol* 1999; **6**:103–106.
9. Arion D, Kaushik N, McCormick S, Borkow G, Parniak MA. **Phenotypic mechanism of HIV-1 resistance to 3′-azido-3′-deoxythymidine (AZT): increased polymerization processivity and enhanced sensitivity to pyrophosphate of the mutant viral reverse transcriptase.** *Biochemistry* 1998; **37**:15908–15917.
10. Boyer PL, Sarafianos SG, Arnold E, Hughes SH. **Selective excision of AZTMP by drug-resistant human immunodeficiency virus reverse transcriptase.** *J Virol* 2001; **75**:4832–4842.
11. Kellam P, Boucher CA, Larder BA. **Fifth mutation in human immunodeficiency virus type 1 reverse transcriptase contributes to the development of high-level resistance to zidovudine.** *Proc Natl Acad Sci U S A* 1992; **89**:1934–1938.
12. Shafer RW, Winters MA, Jellinger RM, Merigan TC. **Zidovudine resistance reverse transcriptase mutations during didanosine monotherapy.** *J Infect Dis* 1996; **174**:448–449.
13. Wainberg MA, White AJ. **Current insights into reverse transcriptase inhibitor-associated resistance.** *Antivir Ther* 2001; **6 (Suppl 2)**:11–19.
14. Iversen AK, Shafer RW, Wehrly K, Winters MA, Mullins JI, Chesebro B, *et al*. **Multidrug-resistant human immunodeficiency virus type 1 strains resulting from combination antiretroviral therapy.** *J Virol* 1996; **70**:1086–1090.
15. Harrigan PR, Stone C, Griffin P, Najera I, Bloor S, Kemp S, *et al*. **Resistance profile of the human immunodeficiency virus type 1 reverse transcriptase inhibitor abacavir (1592U89) after monotherapy and combination therapy: CNA2001 Investigative Group.** *J Infect Dis* 2000; **181**:912–920.
16. Winters MA, Merigan TC. **Variants other than aspartic acid at codon 69 of the human immunodeficiency virus type 1 reverse transcriptase gene affect susceptibility to nuleoside analogs.** *Antimicrob Agents Chemother* 2001; **45**:2276–2279.
17. Wang K, Jenwitheesuk E, Samudrala R, Mittler JE. **Simple linear model provides highly accurate genotypic predictions of HIV-1 drug resistance.** *Antivir Ther* 2004; **9**:343–352.
18. Rhee SY, Taylor J, Wadhera G, Ben-Hur A, Brutlag DL, Shafer RW. **Genotypic predictors of human immunodeficiency virus type 1 drug resistance.** *Proc Natl Acad Sci U S A* 2006; **103**:17355–17360.
19. Sevin AD, DeGruttola V, Nijhuis M, Schapiro JM, Foulkes AS, Para MF, *et al*. **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.** *J Infect Dis* 2000; **182**:59–67.
20. Beerenwinkel N, Schmidt B, Walter H, Kaiser R, Lengauer T, Hoffmann D, *et al*. **Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype.** *Proc Natl Acad Sci USA* 2002; **99**:8271–8276.
21. Leigh Brown AJ, Frost SDW, Good B, Daar ES, Simon V, Markowitz M, *et al*. **Genetic basis of hypersusceptibility to protease inhibitors and low replicative capacity of Human Immunodeficiency Virus type 1 strains in primary infection.** *J Virol* 2004; **78**:2242–2246.
22. Wang D, Larder B. **Enhanced prediction of lopinavir resistance from genotype by use of artificial neural networks.** *J Infect Dis* 2003; **188**:653–660.
23. Rabinowitz M, Myers L, Banjevic M, Chan A, Sweetkind-Singer J, Haberer J, *et al*. **Accurate prediction of HIV-1 drug response from the reverse transcriptase and protease amino acid sequences using sparse models created by convex optimization.** *Bioinformatics* 2006; **22**:541–549.
24. Wolf K, Walter H, Beerenwinkel N, Keulen W, Kaiser R, Hoffmann D, *et al*. **Tenofovir resistance and resensitization.** *Antimicrob Agents Chemother* 2003; **47**:3478–3484.
25. Shafer RW, Jung DR, Betts BJ, Xi Y, Gonzales MJ. **Human immunodeficiency virus reverse transcriptase and protease sequence database.** *Nucleic Acids Res* 2000; **28**:346–348.
26. Kosakovsky Pond SL, Frost SDW, Muse SV. **HyPhy: hypothesis testing using phylogenies.** *Bioinformatics* 2005; **21**:676–679.
27. Ross L, Boulmé R, Fisher R, Hernandez J, Florance A, *et al*. **A direct comparison of drug susceptibility to HIV-1 type 1 from antiretroviral experienced subjects as assessed by the Antivirogram and PhenoSense assays and by seven resistance algorithms.** *AIDS Res Hum Retroviruses* 2005; **21**:933–939.
28. Zhang J, Rhee SY, Taylor J, Shafer RW. **Comparison of the precision and sensitivity of the Antivirogram and PhenoSense HIV drug susceptibility assays.** *J Acquir Immune Defic Syndr* 2005; **38**:439–444.
29. Petropoulos CJ, Parkin NT, Limoli Y, *et al*. **A novel phenotypic drug susceptibility assay for human immunodeficiency virus type 1.** *Antimicrobial Ag Chemo* 2000; **44**:920–928.
30. DiRienzo AG, DeGruttola V, Larder B, Hertogs K. **Nonparametric methods to predict HIV drug susceptibility phenotype from genotype.** *Stats In Med* 2003; **22**:2785–2798.
31. Box GEP, Cox DR. **An Analysis of Transformations.** *J Roy Stat Soc B* 1964; **26**:211–252.
32. Royston JP. **An extension of shapiro and Wilk-W test for normality to large samples.** *J Roy Stat Soc C* 1982; **31**:115–124.
33. Masquelier B, Tamalet C, Montes B, Descamps D, Peytavin G, Bocket L, *et al*. **Genotypic determinants of the virological response to tenofovir disoproxil fumarate in nucleoside reverse transcriptase inhibitor-experienced patients.** *Antivir Ther* 2004; **9**:315–323.
34. Miller MD. **K65R, TAMs and tenofovir.** *AIDS Rev* 2004; **6**:22–33.
35. Miller MD, Margot N, Lu B, Zhong L, Chen SS, Cheng A, *et al*. **Genotypic and phenotypic predictors of the magnitude of response to tenofovir disoproxil fumarate treatment in antiretroviral-experienced patients.** *J Infect Dis* 2004; **189**:837–846.
36. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees*. Belmont, CA: Wadsworth International; 1984.
37. Quinlan JR. *C4.5: Programs for machine learning*. San Francisco: Morgan Kaufman; 1993.
38. Breiman L. **Random forests.** *Machine Learning* 2001; **45**:5–32.
39. Varma S, Simon R. **Bias in error estimation when using cross-validation for model selection.** *BMC Bioinformatics* 2006; **7**:91.
40. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. **The use of receiver operating characteristic curves in biomedical informatics.** *J Biomedical Informatics* 2005; **38**:404–415.
41. Agence Nationale de Recherches sur le SIDA, Groupe AC11 (ANRS-AC11), Paris, France. ANRS-AC11 genotypic interpretation guidelines – version 16. Electronic citation, 2007. Available at: http://www.hivfrenchresistance.org/. [Accessed December 2007].
42. Vercauteren J, Van Laethem K, Deforche K, *et al*. REGA genotypic resistance interpretation system – version 6.4 Electronic citation, 2004. Available at: http://www.rega.kuleuven.be/cev/. [Accessed December 2007].
43. Shafer RW, Rhee SY, Zioni R, *et al*. HIV drug resistance database, Stanford University – version 4.2 Electronic citation, 2007. Available at: http://hivdb.stanford.edu/pages/algs/HIVdb.html. [Accessed: June 2007].
44. Beerenwinkel N, Schmidt B, Walter H, *et al*. geno2pheno genotypic resistance interpretation system. Electronic citation, 2007. Available at: http://www.geno2pheno.org/. [Accessed June 2007].
45. Beerenwinkel N, Sing T, Lengauer T, *et al*. **Computational methods for the design of effective therapies against drug resistant HIV strains.** *Bioinformatics* 2005; **21**:3943–3950.
46. Johnson VA, Brun-Vezinet F, Clotet B, Kuritzkes DR, Pillay D, Schapiro JM, *et al*. **Update of the drug resistance mutations in HIV-1: Fall 2006.** *Top HIV Med* 2006; **14**:125–130.

47. Parikh UM, Bacheler L, Koontz D, Mellors JW. **The K65R mutation in human immunodeficiency virus type 1 reverse transcriptase exhibits bidirectional phenotypic antagonism with thymidine analog mutations.** *J Virol* 2006; **80**:4971–4977.
48. Rhee SY, Liu T, Ravela J, Gonzales MJ, Shafer RW. **Distribution of human immunodeficiency virus type 1 protease and reverse transcriptase mutation patterns in 4,183 persons undergoing genotypic resistance testing.** *Antimicrob Agents Chemother* 2004; **48**:3122–3126.
49. Mendoza CD, Garrido C, Corral A, *et al*. **Changing rates and patterns of drug resistance mutations in antiretroviral-experienced HIV-infected patients.** *AIDS Res Hum Retroviruses* 2007; **23**:879–885.
50. Miller MD, Anton KE, Mulato AS, Lamy PD, Cherrington JM. **Human immunodeficiency virus type 1 expressing the lamivudine-associated M184V mutation in reverse transcriptase shows increased susceptibility to adefovir and decreased replication capability in vitro.** *J Infect Dis* 1999; **179**:92–100.
51. Frankel FA, Invernizzi CF, Oliveira M, Wainberg MA. **Diminished efficiency of HIV-1 reverse transcriptase containing the K65R and M184V drug resistance mutations.** *AIDS* 2007; **21**:665–675.
52. Hertogs K, Bloor S, De V, V, van Den Eynde C, Dehertogh P, van Cauwenberge A, *et al*. **A novel human immunodeficiency virus type 1 reverse transcriptase mutational pattern confers phenotypic lamivudine resistance in the absence of mutation 184V**. *Antimicrob Agents Chemother* 2000; **44**:568–573.
53. Whitcomb JM, Parkin NT, Chappey C, *et al*. **Broad nucleoside reverse- transcriptase inhibitor cross-resistance in human immunodeficiency virus type 1 clinical isolates.** *J Infect Dis* 2003; **188**:992–1000.
54. Gonzales MJ, Wu TD, Taylor J, Belitskaya I, Kantor R, Israelski D, *et al*. **Extended spectrum of HIV-1 reverse transcriptase mutations in patients receiving multiple nucleoside analog inhibitors.** *AIDS* 2003; **17**:791–799.