Original article

Advantages of predicted phenotypes and statistical learning models in inferring virological response to antiretroviral therapy from HIV genotype

André Altmann[1†], Tobias Sing[1†], Hans Vermeiren[2], Bart Winters[2], Elke Van Craenenbroeck[2], Koen Van der Borght[2], Soo-Yon Rhee[3], Robert W Shafer[3], Eugen Schülter[4], Rolf Kaiser[4], Yardena Peres[5], Anders Sönnerborg[6], W Jeffrey Fessel[7], Francesca Incardona[8], Maurizio Zazzi[9], Lee Bacheler[10], Herman Van Vlijmen[2] and Thomas Lengauer[1]*

[1]Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Saarbrücken, Germany
[2]Virco BVBA, Mechelen, Belgium
[3]Division of Infectious Diseases, Stanford University, Stanford, CA, USA
[4]Institute of Virology, University of Cologne, Cologne, Germany
[5]Health Care and Life Sciences Group, IBM Research, Haifa, Israel
[6]Division of Infectious Diseases, Karolinska Institute, Stockholm, Sweden
[7]Kaiser–Permanente Medical Care Program, San Francisco, CA, USA
[8]Informa, Rome, Italy
[9]Department of Molecular Biology, University of Siena, Siena, Italy
[10]VircoLab, Inc., Durham, NC, USA

*Corresponding author: E-mail: altmann@mpi-inf.mpg.de
†These authors made an equal contribution to this work

Background: Inferring response to antiretroviral therapy from the viral genotype alone is challenging. The utility of an intermediate step of predicting *in vitro* drug susceptibility is currently controversial. Here, we provide a retrospective comparison of approaches using either genotype or predicted phenotypes alone, or in combination.
Methods: Treatment change episodes were extracted from two large databases from the USA (Stanford-California) and Europe (EuResistDB) comprising data from 6,706 and 13,811 patients, respectively. Response to antiretroviral treatment was dichotomized according to two definitions. Using the viral sequence and the treatment regimen as input, three expert algorithms (ANRS, Rega and HIVdb) were used to generate genotype-based encodings and VircoTYPE™ 4.0 (Virco BVBA, Mechelen, Belgium) was used to generate a predicted phenotype-based encoding. Single drug classifications were combined into a treatment score via simple summation and statistical learning using random forests. Classification performance was studied on Stanford-California data using cross-validation and, in addition, on the independent EuResistDB data.
Results: In all experiments, predicted phenotype was among the most sensitive approaches. Combining single drug classifications by statistical learning was significantly superior to unweighted summation ($P<2.2\times10^{-16}$). Classification performance could be increased further by combining predicted phenotypes and expert encodings but not by combinations of expert encodings alone. These results were confirmed on an independent test set comprising data solely from EuResistDB.
Conclusions: This study demonstrates consistent performance advantages in utilizing predicted phenotype in most scenarios over methods based on genotype alone in inferring virological response. Moreover, all approaches under study benefit significantly from statistical learning for merging single drug classifications into treatment scores.

Introduction

Modern antiretroviral combination therapy can substantially delay disease progression, prolong survival and maintain quality of life; but eradication of HIV infection remains out of reach. Therefore, research focuses not only on the search for novel drugs, but also on exploiting the currently available drug collection to

achieve the best possible effect through patient-tailored therapy. The main obstacle to ultimate treatment success is the ability of the virus to rapidly acquire mutations that confer resistance to specific drugs.

Currently, the genotype of relevant sections of the HIV genome at treatment failure is routinely determined by standardized nucleic acid sequencing protocols. The resulting data have led to the ongoing discovery of an increasing number of resistance-associated mutations, which tend to occur in diverse mutational patterns. The interpretation of a given pattern with respect to its effect on drug resistance and response to combination therapy is an extremely complex task. Consequently, a number of expert panels have been developing and refining interpretation algorithms based on expert rules or mutation scores. Supported by several clinical trials (reviewed in [1]), the benefit of using viral genomic information in addition to clinical and therapeutic data, as opposed to exclusively relying on the latter, is no longer controversial. However, the diversity and complexity of resistance-associated mutational patterns and their effect on the activity of drug combinations is far from being understood.

Although sequencing assays are the basic tool for detecting mutations in the viral population, they do not provide any information on the clinical effect of these mutations. A standard approach for distinguishing resistance-associated mutations from natural polymorphisms has been to compare mutation frequencies in untreated versus treated patients (reviewed in [2]). However, this approach cannot provide quantitative assessments of the phenotypical effect of particular mutations and is thus of limited utility in the development of prediction algorithms. In order to develop quantitative models, it is necessary to link genotypical and treatment information to virological response data reflecting the activity of a given regimen on a viral population with a given genomic make-up.

Such a link between a viral genotype, a drug combination and a measure of how effectively this particular drug combination works for this particular genotype is the fundamental information in datasets used for building models for virological outcome prediction. Adopting a terminology coined by the HIV Resistance Response Database Initiative (RDI), this basic data unit is commonly referred to as a treatment change episode (TCE). It is well know that other pieces of clinical and therapeutic information, such as baseline viral load or $CD4^+$ T-cell count, treatment and adherence history, can contribute to an improved prediction [3]. However, as this information is not always available, models assessing the benefit of a therapy solely on the basis of the viral genotype are more generally applicable.

Datasets comprising TCEs extracted from clinical databases can and have been used for model building without any further processing. Expert panels screen these data for mutations associated with treatment failure and can then integrate this information together with other sources of knowledge into their carefully handcrafted interpretation tables. As an alternative approach, an increasing number of research groups are trying to develop well-defined, algorithmic and strictly reproducible approaches to model building based exclusively on the foundation of statistics, statistical learning and bioinformatics [3–7].

Although TCEs can be used *as is* for model building, additional data can be derived from the TCE variables and used during model building to improve the genotype-based prediction models. Importantly, the emphasis is on using the additional features only during model building, as opposed to using them also during prediction, as would be the case in the example of using baseline viral load or treatment history as additional predictors. As a consequence, the final models still remain useful when only a genotype is available. Examples of this approach related to HIV include the use of X-ray protein structure data in combination with molecular modelling to improve to prediction of phenotypical drug resistance [8–10] or coreceptor usage [11], or the use of evolutionary modelling via the genetic barrier to resistance to improve the prediction of response to combination therapy [4,12].

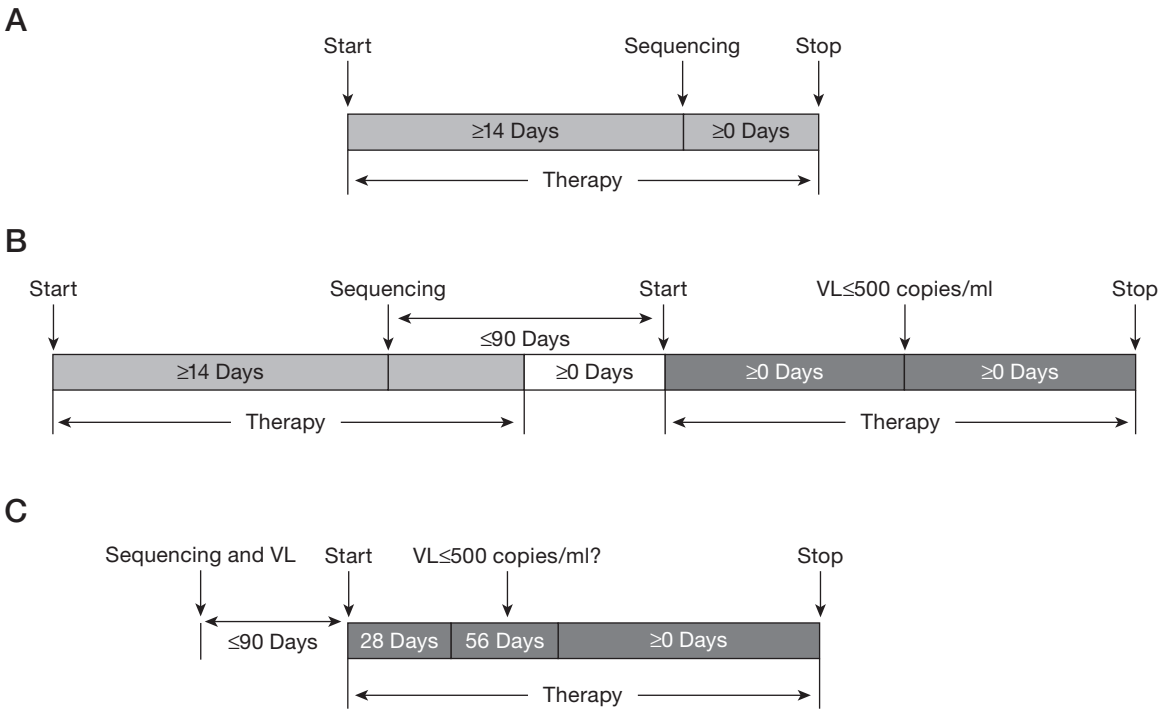
This study evaluates the role of phenotypical resistance data as a source of genotype-related additional information during model building. Phenotypical assays provide a quantitative measure of susceptibility to individual drugs. Although this is a more intuitive output compared with mutation patterns generated by genotypical tests, laboratory-based phenotyping is too complex and expensive for routine clinical use. Alternatively, sufficiently large paired genotype-phenotype datasets can be used to generate genotype-phenotype prediction models. These can then be used on a given genotype as an additional source of information in predicting virological outcome. However, the usefulness of phenotypical measurements in the clinical setting has been questioned because of inconsistent results in clinical trials [13–15] and limited ability to detect phenotypical effects of clinically relevant mutations [16]. In fact, HIV resistance experts have discouraged inferring response from genotype via the intermediate step of predicting the phenotype [3,17]. Thanks to the availability of a large number of TCEs, this work comparatively examined approaches exclusively based on predicted phenotypes, established genotypical interpretation algorithms and hybrid approaches.

Methods

Standard datum definitions

The virological response to a therapy was dichotomized as success or failure. Two different definitions for success

Figure 1. Genotype-centric and classic definition of failing or successful treatment change episodes



(A) In the genotype-centric definition, a treatment change episode (TCE) followed by sequencing is always considered to be a failure. **(B)** If the viral load (VL) decreases to undetectable levels during the treatment, the TCE is considered to be a success. The genotype-centric definition allows for one treatment to be part of a failing and a successful TCE at the same time. For example, a virus with genotype 1 (G_1) is treated with a therapy comprising lamivudine plus zidovudine plus efavirenz (3TC+AZT+EFV), which reduced the VL to ≤500 copies/ml. Hence, the pair G_1, 3TC+AZT+EFV is a successful TCE, as shown in (B). As the treatment continues, resistance mutations appear and a further genotype is obtained (G_2); thus the pair G_2, 3TC+AZT+EFV is a failing TCE, as shown in (A). **(C)** In the classic definition, success and failure are determined by the level of the follow-up VL. Baseline genotype and baseline VL have to be available within 90 days before treatment start.

versus failure were used. The first focused on a genotype-centric standard datum defined by the EuResist consortium [18], and which has been applied previously [4]. A diagrammatic representation is shown in Figure 1. Any available genotype was considered as evidence of a failing regimen because sequencing can only be performed if the virus is replicating to detectable levels, usually ≥1,000 HIV RNA copies/ml. Consequently, the drug combination at the time of sequencing had to be considered an ineffective therapy for that particular genotype, and therefore formed a failing TCE without requiring a baseline genotype. Successful regimens were defined as those achieving an undetectable follow-up viral load and with an available genotype test performed no earlier than 3 months before starting the therapy. This definition allows for a therapy to be a success in the beginning and a failure at a later time point, but with a different genotype (Figure 1). The second definition was based on the classical standard datum used in the EuResist project [18,19]. Briefly, in the classical definition (Figure 1), success or failure was defined on the basis of the viral load measurement that was closest to 8 weeks after onset of therapy among those available within 4–12 weeks after onset. If that value was ≤500 copies/ml, then the therapy was considered a success; if it was >500 copies/ml, then it was a failure. In addition, a reduction of ≥2 $\log_{10}$ copies/ml compared with a baseline measurement was also considered a success. Baseline genotype and baseline viral load had to be performed no earlier than 90 days before treatment start. In summary, the two TCE definitions differ mainly in the viral genotype associated with treatment failure: in the genotype-centric definition the genotype obtained during a failing treatment is taken (retrospective), whereas in the classic definition the genotype right before a failing treatment is considered (prospective).

Data

In order to ensure an unbiased comparison between expert algorithms and predicted phenotypes, we performed our analyses independently on two disjoint databases. The first dataset, henceforth called Stanford-California, consists of data from two Northern California clinic populations undergoing genotypical resistance testing at Stanford University and of data from

Table 1. Dataset summary

	Raw data				Genotype-centric				Classic			
Database	Therapies	Sequences	Patients	Viral load	Patients in dataset, *n*	Total TCEs, *n*	Failure	Success	Patients in dataset, *n*	Total TCEs, *n*	Failure	Success
Stanford-California	25,717	16,288	6,706	110,392	3,829	6,337	4,776[a]	1,583[a]	2,063	2,351	924	1,427
EuResistDB	44,220	16,243	13,811	158,125	3,034	5,224	4,320[a]	904[a]	904	1,064	450	614

[a]Using the genotype-centric definition, the number of therapies that received both labels (success and failure) in the Stanford-California dataset was 271 and in the EuResistDB dataset was 221. TCE, treatment change episode.

Table 2. HIV subtype prevalence in the EuResistDB datasets

Subtype	Genotype-centric	Classic
A	0.66	1.07
A1	0.37	0.45
A2	0.05	0.00
A3	0.30	0.28
B	91.72	88.98
C	1.06	1.97
CRF01_AE	0.37	0.51
CRF02_AG	2.23	2.76
CRF06_CPX	0.18	0.17
CRF09_CPX	0.19	0.22
CRF10_CD	0.04	0.00
CRF11_CPX	0.05	0.11
CRF12_BF	0.12	0.00
CRF15_01B	0.46	0.67
D	0.14	0.79
F1	1.52	1.29
G	0.39	0.73
J	0.09	0.00

Prevalence of subtypes is indicated as percentages.

the Stanford HIV Drug Resistance Database (clinical studies ACTG 320, ACTG 364, GART and HAVANA). The second dataset, called EuResistDB, comprises the EuResist database (December 2006 update), which merges routine clinical data from the German database Arevir [20], the Italian database ARCA and the Swedish database maintained at Karolinska University Hospital. Information on the raw data in the databases and the datasets derived according to the two definitions is listed in Table 1. In addition, Table 2 lists the HIV subtype prevalence in the EuResistDB datasets. Subtype information was not available for the Stanford-California dataset. However, it can safely be assumed that the diversity is lower than in the European data [21]. Therapies considered included any combination of 16 drugs from the different classes. Specifically, nucleoside reverse transcriptase inhibitors (NRTIs) included zidovudine (AZT), didanosine (ddI), zalcitabine (ddC), stavudine (d4T), lamivudine, abacavir and tenofovir dipivoxil fumerate; non-NRTIs included nevirapine, delavirdine (DLV) and efavirenz (EFV); and protease inhibitors (PIs) included saquinavir, indinavir, nelfinavir (NFV), amprenavir, lopinavir and atazanavir, and PIs could be used with or without ritonavir as boosting dose (bRTV).

Inputs

For comparing predicted phenotypes with ratings based on expert genotypical algorithms and for investigating potential synergies by using hybrid approaches, a total of 13 input representations were explored. Three representations were simply based on ratings from expert algorithms, one on predicted phenotypes and another one on raw genotypes. Four representations were hybrid encodings combining either ratings from expert algorithms or raw genotype with predicted phenotype, and four representations were combinations of ratings from expert algorithms.

The expert algorithms used included ANRS (version 2006/07) [22], Rega (version 6.4.1) [23] and HIVdb (version 4.2.6) [24]. Drug activity predicted by these systems was scaled between 0 and 1, with 0 indicating fully inactive drugs and 1 indicating fully active drugs. The classification given by ANRS and Rega for 'resistant', 'intermediate' and 'susceptible' was mapped to 0, 0.5 and 1, respectively. The HIVdb five-level output of resistance, 'high-level resistance', 'intermediate resistance', 'low-level resistance', 'potential low-level resistance' and 'susceptible', was mapped to 0, 0.25, 0.5, 0.75 and 1, respectively. These mappings are herein termed 'expert encodings'.

The predicted phenotype scores (Pheno) were generated in a similar fashion. The VircoTYPE™ 4.0 system (Virco BVBA, Mechelen, Belgium) [25] was used to predict phenotypical resistance to the individual drugs for a given genotype. The VircoTYPE™ is a linear model with pairwise interaction terms that has been fitted on a dataset of matched genotype-phenotype pairs (median of 46,100 pairs per drug). For a given genotype, the prediction is the fold change in the 50% inhibitory concentration relative to a wild-type laboratory reference strain. The resistance factor returned by VircoTYPE™ was then normalized by linear interpolation to the

interval [0, 1] on the basis of the established upper and lower cutoffs [7], where values above the upper cutoff were set to 0 and values below the lower cutoff were set to 1. Thus, the expert and Pheno encodings each comprise 16 real valued variables and one binary indicator for bRTV.

The raw genotype (Geno) was encoded as a 0/1 representation of the genotype that indicates the presence (1) or absence (0) of specific mutations included in the International AIDS Society drug resistance mutations list (Fall 2006 [26]). This resulted in a total of 94 mutations, 62 in the protease and 32 in the reverse transcriptase (RT). An additional 17 binary variables indicated the presence (1) or absence (0) of the individual drugs plus bRTV in the treatment regimen. Thus, in total a genotype plus drug combination was represented by a (62+32+17) dimensional binary vector.

The remaining eight representations were concatenations of the individual encodings. These combinations were partitioned into two groups. The first group (hybrid) combined the Pheno encoding with either an expert encoding or the Geno encoding, resulting in three versions of a (17+16) dimensional vector and one version of a (17+94) dimensional vector, respectively. The second group combined up to three expert representations without any phenotypical information, resulting in a vector with (17+16) dimensions for a combination of two expert encodings (for example, ANRS+Rega) and a vector with (17+16+16) dimensions for a combination of all three expert encodings.

Combining individual drug scores into regimen scores

Traditionally, the activity of a drug combination has been determined from the activities of the individual components by simply adding activity scores of drugs used in the regimen (simple summation). This approach has been termed genotypical/phenotypical susceptibility (or sensitivity) score [27]. In this study, two different approaches of combining individual drug scores into regimen scores were compared. These methods included summation (for ANRS, Rega, HIVdb and Pheno) and statistical learning using random forest classification (for all inputs) [28]. The use of multivariate statistical learning procedures entailed the advantage that special handling of discordances between ratings in hybrid encodings was not required. The ratings were used *as is* as input to the learning algorithm that considered all ratings during the decision making process. This also held true for the Geno encoding, where the random forest algorithm determined which combinations of drugs and mutations would lead to a successful or failing treatment. The probability of observing a success predicted by the classifier was used as a continuous score normalized within the interval [0, 1].

Evaluation setup

Models were compared by nested 10-fold cross-validation [29] to ensure an unbiased comparison. The goal was to avoid overfitting of parameters by simple cross-validation.

The first set of analyses was based in the genotype-centric standard datum. In this first analysis, 1,000 TCEs were randomly removed from the full Stanford-California data (consisting of 6,337 TCEs) to determine the optimal number of trees in the random forest (ranging from 50 to 400). Standard 10-fold cross validation was performed on the remaining 5,337 TCEs to obtain an unbiased estimate of the model performance. The whole procedure was repeated 10× to account for randomness in determining the optimal number of trees.

Because the distribution of drug combinations used might be heavily skewed [30] depending on approval times, changes in treatment strategies or even regional differences from hospital to hospital [3], the models for the 13 different encodings were trained on the complete Stanford-California dataset with the parameters from the first analysis and used to predict the full EuResistDB data comprising 5,224 TCEs. This second analysis aimed at estimating how well the results generalized across completely different data collection efforts.

In addition, all analyses were repeated using the classic standard datum on the complete Stanford-California and EuResistDB set comprising 2,351 and 1,064 TCEs, respectively.

Measures of predictive performance

The output of both the simple summation, as well as the random forest classification, is given by a score rather than an actual class prediction. The analysis of such scoring classifiers is usually performed in the framework of receiver operating characteristic (ROC) analysis [31]. The analyses were performed using the classifier evaluation package ROCR [32]. We considered successes as positive and failures as negative samples and focused on three performance measures related to the ROC curve. The area under the ROC curve (AUC) provided a cut-off-independent measure of class separation by a scoring classifier. The true-positive rate (TPR; sensitivity or recall) was represented by TPR_{10} and TPR_{20} at a false-positive rate of 10% and 20%, respectively (equivalently, a specificity of 90% and 80%).

Evaluating the importance of variables

For every variable used, the random forest statistical learning method provides a measure of importance for the classification result. This importance measure is based on the mean decrease of the Gini index (used as a measure of impurity) when a new split at that variable is introduced in a tree [28]. Using this measure of variable importance we sought to find out whether

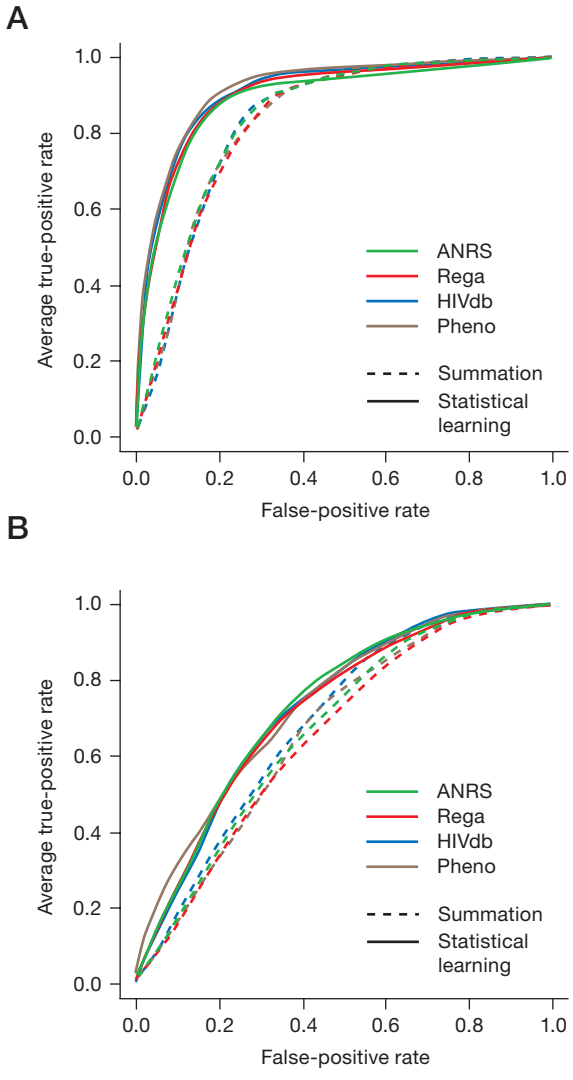
there were differences between the two standard datum definitions and whether one algorithm was better than another in predicting response to a specific drug. There is no fair method for comparing exact measures of variable importance across different datasets because the importance of a variable strongly depends on the dataset used for model building.

Results

Regimen scores from individual drug scores: summation versus statistical learning

The first analysis was aimed at comparing summation versus statistical learning as methods for combining scores for single drugs into regimen scores. Figure 2A shows averaged ROC curves (10 replicates of 10-fold cross-validation) based on the ANRS, Rega, HIVdb and Pheno input representation for the genotype-centric definition of success and failure. For each representation, regimen scores are either derived by simple summation or by statistical learning, as described in the Methods section. At nearly all practically relevant false-positive rates, the regimen score based on statistical learning drastically outperformed the summation-based regimen score. For all input representations and performance criteria, the improvements were consistent and highly statistically significant ($P<2.2\times10^{-16}$; Wilcoxon signed-rank test was used for all P-values). For example, at a false-positive rate of 10%, the improvements in TPR ranged from 25.7% (ANRS) to 37.1% (Pheno) points. Figure 2B shows averaged ROC curves (10 replicates of 10-fold cross-validation) for the same setting on the data from the classical standard datum. Although the performance was generally lower than for the genotype-centric definition, the qualitative differences were confirmed: statistical learning outperformed summation with the largest $P=8.479\times10^{-10}$. On the basis of these consistent results, all further experiments were performed with regimen scores derived by statistical learning.

Figure 2. ROC curves for expert algorithms and Pheno using summation and statistical learning on Stanford-California data



Models were tested using 10 replications of 10-fold cross-validation. **(A)** Genotype-centric definition. **(B)** Classic definition. ROC, receiver operating characteristic.

Method comparison

The second experiment was devoted to a detailed comparison of the five individual input representations and the eight combined input representations. Table 3 displays averaged performance measures and their SD for the 13 different inputs from 10 replicates of 10-fold cross-validation on the Stanford-California data. Remarkably, the Pheno representation was the best single representation overall, with the highest AUC, TPR_{10} and TPR_{20} (P-values ranged from 0.01154 to 10^{-11} compared with best single encoding). HIVdb, which uses five instead of three susceptibility levels, provided the best encoding derived from an expert algorithm (but the improvement with respect to Rega at TPR_{20} did not reach statistical significance; $P=0.0881$) and, according to all three performance measures, Geno was the worst individual encoding (P-values ranged from 10^{-5} to 10^{-16} compared with ANRS, the next worst single encoding). Moreover, no combination of expert algorithms outperformed HIVdb in terms of AUC and TPR_{10}. Only the difference between Rega+HIVdb and HIVdb at TPR_{20} reached statistical significance ($P=0.0312$). However, the combination of the Pheno representation with a single expert algorithm significantly improved the performance with respect to all three measures (P-values ranged from 10^{-3} to 10^{-14}) compared with Pheno alone. Because of poor performance of combinations of expert encodings, only hybrid approaches (that is, Pheno with

Table 3. Cross-validation results on Stanford-California

Encoding	Genotype-centric			Classic		
	AUC (±SD)	TPR_{10} (±SD)	TPR_{20} (±SD)	AUC (±SD)	TPR_{10} (±SD)	TPR_{20} (±SD)
Regimens scored by summation						
Single encodings						
ANRS	0.844 (0.015)	0.446 (0.066)	0.726 (0.048)	0.675 (0.033)	0.171 (0.045)	0.351 (0.058)
Rega	0.836 (0.016)	0.404 (0.071)	0.699 (0.045)	0.658 (0.034)	0.151 (0.048)	0.332 (0.065)
HIVdb	0.837 (0.014)	0.397 (0.091)	0.719 (0.079)	0.691 (0.035)	0.181 (0.060)	0.377 (0.079)
Pheno	0.834 (0.015)	0.392 (0.087)	0.717 (0.071)	0.670 (0.037)	0.159 (0.058)	0.334 (0.071)
Regimens scored by statistical learning						
Single encodings						
ANRS	0.895 (0.015)	0.703 (0.057)	0.880 (0.029)	0.734 (0.034)	0.250 (0.069)	0.481 (0.089)
Rega	0.904 (0.014)	0.722 (0.048)	0.884 (0.029)	0.725 (0.033)	0.254 (0.071)	0.471 (0.078)
HIVdb	0.913 (0.013)	0.753 (0.049)	0.888 (0.028)	0.731 (0.033)	0.241 (0.071)	0.472 (0.085)
Pheno	0.921 (0.013)	0.763 (0.055)	0.912 (0.030)	0.737 (0.034)	0.312 (0.065)	0.483 (0.073)
Geno	0.888 (0.017)	0.676 (0.048)	0.822 (0.038)	0.728 (0.037)	0.285 (0.066)	0.462 (0.074)
Hybrid encodings						
ANRS+Pheno	0.927 (0.012)	0.770 (0.052)	0.921 (0.027)	0.743 (0.034)	0.302 (0.068)	0.490 (0.078)
Rega+Pheno	0.928 (0.011)	0.771 (0.053)	0.922 (0.028)	0.740 (0.033)	0.305 (0.066)	0.483 (0.068)
HIVdb+Pheno	0.928 (0.012)	0.769 (0.055)	0.921 (0.028)	0.744 (0.033)	0.320 (0.065)	0.492 (0.069)
Geno+Pheno	0.921 (0.012)	0.770 (0.046)	0.910 (0.027)	0.742 (0.035)	0.311 (0.073)	0.486 (0.080)
Mixtures of expert algorithms						
ANRS+Rega	0.906 (0.015)	0.730 (0.056)	0.889 (0.032)	0.729 (0.033)	0.246 (0.065)	0.464 (0.085)
ANRS+HIVdb	0.911 (0.015)	0.747 (0.054)	0.897 (0.029)	0.724 (0.032)	0.242 (0.056)	0.468 (0.086)
Rega+HIVdb	0.912 (0.014)	0.739 (0.052)	0.898 (0.028)	0.721 (0.032)	0.257 (0.063)	0.479 (0.080)
ANRS+Rega+HIVdb	0.912 (0.014)	0.744 (0.052)	0.896 (0.030)	0.726 (0.032)	0.250 (0.032)	0.465 (0.082)

AUC, area under the receiver operating characteristic curve; TPR_{10}, true-positive rate at a false-positive rate of 10%; TPR_{20}, true-positive rate at a false-positive rate of 20%.

either an expert encoding or Geno) were considered in subsequent experiments. Figure S1 (see Additional file) depicts averaged ROC curves from 10 replicates of 10-fold cross-validation on the Stanford-California data for the five single and four hybrid encodings. Table 3 summarizes the cross-validation results on the Stanford-California data using the classic definition. Again, the Pheno encoding was the best performing single method with respect to all three measures. However, only the difference at TPR_{10} was statistically significant with respect to all expert-based encodings ($P<10^{-10}$). The Geno encoding was significantly better than all expert-based encodings at TPR_{10} ($P<10^{-4}$), but also significantly worse than ANRS in AUC and TPR_{20} ($P<0.03$). The benefit of the hybrid encodings was greatly reduced compared with the genotype-centric definition, but the best hybrid encoding, HIVdb+Pheno, still outperformed the Pheno with respect to all three measures (*P*-values ranged from 0.036 to 10^{-5}). Again, combinations of expert-based encodings failed to improve over the best single expert-based encoding in AUC and TPR_{20}. The improvement of Rega+HIVdb over Rega at TPR_{10} did not reach statistical significance ($P=0.6986$).

In order to investigate the degree to which classifiers learned from the Stanford-California data can be used for prediction on data collected from a different area, we used them for predicting the EuResistDB dataset, which was not used at all during training. Results are summarized in Table 4. Generally, the prediction performance of all methods decreased when they were used on the completely different dataset. For example, on the Stanford-California data using the genotype-centric definition, the best methods achieved a TPR_{10} of 77.1% and a TPR_{20} of 92.2% (in cross-validation) compared with 58.5% and 77.6% when trained on the Stanford-California data but evaluated on the EuResistDB data. However, combination of single-drug classifications to a treatment score by a method of statistical learning trained on different data compared with summation showed a clear benefit. For example, the TPR of ANRS at a false-positive rate of 10% increased by 18.3% and for Pheno it increased by 38.5%. Moreover, similar to the preceding analysis, the Pheno encoding was the single best encoding with respect to all three measures (only Geno was better at TPR_{10}) and the hybrid model Rega+Pheno significantly outperformed Pheno with respect to all three measures ($P<0.001$). Results obtained on TCEs from the classic definition confirm the benefit of hybrid encodings. Precisely, Rega+Pheno outperformed Pheno in AUC ($P=0.002$) and TPR_{10} ($P=0.08$), and ANRS+Pheno outperformed Pheno at TPR_{20} ($P=0.002$). All results remained qualitatively the same when obsolete regimens (that is, regimens

Table 4. Results on EuResistDB when trained on Stanford-California

Encoding	Genotype-centric			Classic		
	AUC (±SD)	TPR_{10} (±SD)	TPR_{20} (±SD)	AUC (±SD)	TPR_{10} (±SD)	TPR_{20} (±SD)
Regimens scored by summation						
Single encodings						
ANRS	0.759 (0.000)	0.245 (0.000)	0.523 (0.000)	0.697 (0.000)	0.229 (0.000)	0.455 (0.000)
Rega	0.748 (0.000)	0.201 (0.000)	0.437 (0.000)	0.675 (0.000)	0.193 (0.000)	0.379 (0.000)
HIVdb	0.741 (0.000)	0.163 (0.000)	0.448 (0.000)	0.708 (0.000)	0.249 (0.000)	0.455 (0.000)
Pheno	0.752 (0.000)	0.178 (0.000)	0.454 (0.000)	0.692 (0.000)	0.232 (0.000)	0.391 (0.000)
Regimens scored by statistical learning						
Single encodings						
ANRS	0.816 (0.002)	0.428 (0.008)	0.707 (0.009)	0.696 (0.003)	0.248 (0.008)	0.443 (0.015)
Rega	0.828 (0.002)	0.475 (0.004)	0.705 (0.006)	0.701 (0.002)	0.234 (0.008)	0.434 (0.009)
HIVdb	0.839 (0.002)	0.512 (0.006)	0.748 (0.003)	0.692 (0.001)	0.239 (0.011)	0.382 (0.011)
Pheno	0.854 (0.002)	0.563 (0.010)	0.765 (0.004)	0.710 (0.004)	0.285 (0.013)	0.450 (0.017)
Geno	0.836 (0.002)	0.569 (0.006)	0.717 (0.007)	0.687 (0.005)	0.251 (0.015)	0.407 (0.017)
Hybrid encodings						
ANRS+Pheno	0.862 (0.002)	0.578 (0.008)	0.771 (0.006)	0.717 (0.004)	0.285 (0.018)	0.472 (0.009)
Rega+Pheno	0.865 (0.003)	0.585 (0.005)	0.776 (0.006)	0.719 (0.003)	0.292 (0.008)	0.466 (0.015)
HIVdb+Pheno	0.865 (0.001)	0.570 (0.009)	0.775 (0.004)	0.708 (0.002)	0.270 (0.011)	0.455 (0.010)
Geno+Pheno	0.858 (0.003)	0.572 (0.009)	0.743 (0.006)	0.713 (0.005)	0.266 (0.011)	0.442 (0.020)

AUC, area under the receiver operating characteristic curve; TPR_{10}, true-positive rate at a false-positive rate of 10%; TPR_{20}, true-positive rate at a false-positive rate of 20%.

containing DLV, ddC, NFV or unboosted PIs) were removed from the EuResistDB datasets (see Table S1 in the Additional file).

Evaluating the importance of variables

Figure 3 shows the variable importance (as percentages) for all drugs grouped by interpretation algorithm. The importance of the EFV predictions differed the most when exchanging the standard datum definition. Minor changes were detected for all of the NRTI predictions given by the expert systems. Interestingly, only the importance of a few NRTIs (ddI, AZT and d4T) differed notably for the Pheno prediction. The importances of the PI predictions remained almost unchanged. Importance for DLV and ddC was usually low because only very few samples were present in the dataset.

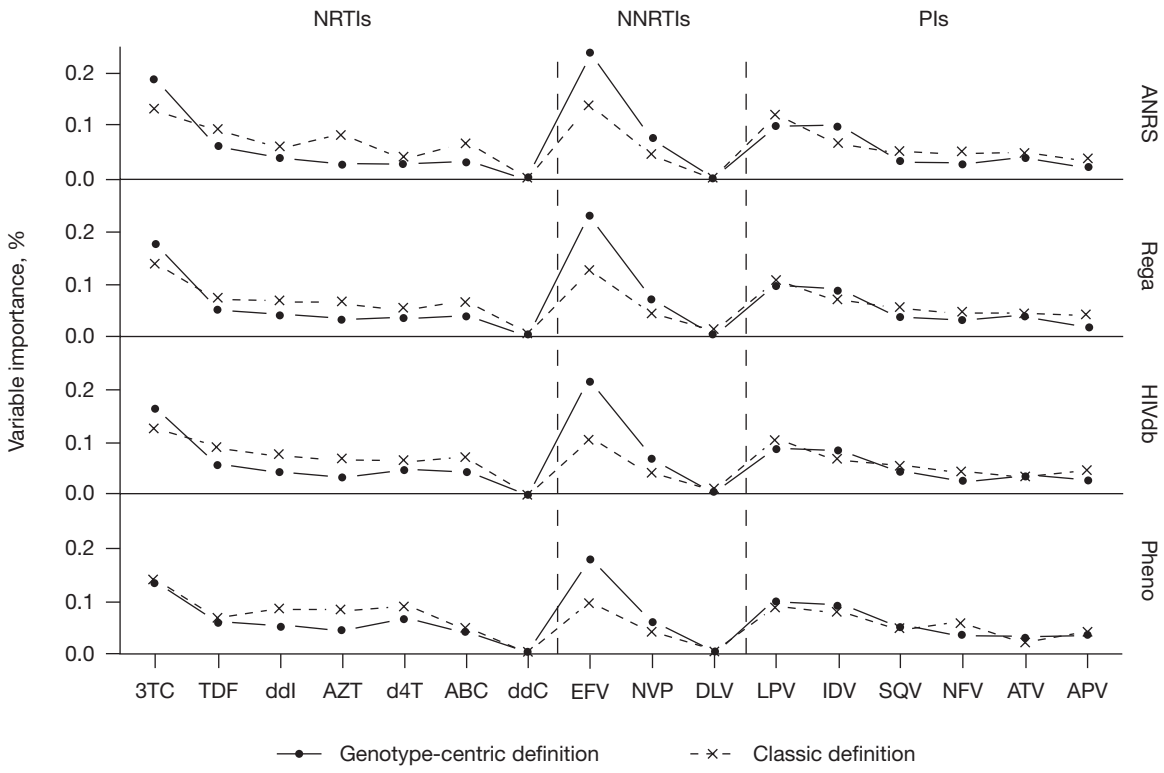
The variable importance in the hybrid models allowed for a direct comparison between predictions for a single drug by either an expert algorithm or the VircoTYPE™ prediction, because both predictions are part of the model. Thus, the importance of the expert prediction for a drug (for example, LPV) is assessed in the presence of the VircoTYPE™ prediction for the same drug and *vice versa*. Figure 4 shows scatter plots of the variable importances of drugs on the basis of the VircoTYPE™ prediction and expert algorithm predictions. This analysis showed that the prediction by VircoTYPE™ is equally or more important for all drugs than that of any expert algorithm for predicting the correct clinical outcome. This holds for the genotype-centric (Figure 4A) and even more so for the classic definitions (Figure 4B).

Discussion

In this large-scale study, we have compared a variety of input representations with respect to their ability to discriminate between treatment success or failure, on the basis of genotype and a chosen drug combination. Two different methods for combining scores predicted for individual drugs into scores for a regimen as a whole were evaluated, on the basis of either the traditional summation or on statistical learning using random forests. The use of statistical learning for score combination consistently outperformed the summation of scores in our experiments. We hypothesized that the superiority of the learning-based combination is a result of its intrinsic ability to assign specific weights to individual drugs in a regimen (possibly related to drug potency, ease of adherence and reliability of the interpretation algorithm) in contrast to the traditional unweighted summation, and for capturing drug–drug interaction effects, as far as they can be learned from the data. The benefits of weighting the contributions of individual drugs have previously been reported in the context of a linear model [33].

A major focus of our analysis was on comparing the performance of input representations directly based on the genotype with an input representation based on predicted phenotypes. This latter approach has been

Figure 3. Comparison of variable importance in classic and genotype-centric definitions



The variable importance for all drugs derived from datasets using the genotype-centric definition and the classic definition is shown. The importance measures are grouped by interpretation system. Variable importance corresponds to the influence of a drug rating during the decision making process of the random forest classifier. For example, lamivudine (3TC) and efavirenz (EFV) are the most important drugs because a correct interpretation with respect to drug resistance is crucial and they are part of many treatments. On the basis of this importance measure, 3TC and EFV should not be considered the most potent drugs; however, drug potency is one factor contributing to variable importance. ABC, abacavir; APV, amprenavir; ATV, atazanavir; AZT, zidovudine; ddC, zalcitabine; ddI, didanosine; DLV, delavirdine; d4T, stavudine; IDV, indinavir; LPV, lopinavir; NFV, nelfinavir; NNRTIs, non-nucleoside reverse transcriptase inhibitors; NRTIs, nucleoside reverse transcriptase inhibitors; NVP, nevirapine; PIs, protease inhibitors; SQV, saquinavir; TDF, tenofovir dipivoxil fumarate.

criticized previously [3,17], although there is little data available as a basis for judgment. By contrast to this criticism, none of the experiments reported here showed any limitation of the phenotype-based representation, as compared with the genotype-based representations. Rather, according to several performance measures (AUC, TPR_{10} and TPR_{20}), the phenotype representation is the best single input representation. This was confirmed by analysis of variable importance in hybrid models. However, it is important to keep in mind that the phenotype-based representation relies on normalization for realizing its full potential. This is most likely because of the widely differing ranges of resistance factors (fold change in 50% inhibitory concentration) observed for the different drugs and their non-uniform clinical relevance.

Drawbacks of (virtual) phenotypical drug resistance testing include the inability to weight RT 215 revertants (that is, 215D/C/N/S [16]), which have proved to affect response to NRTI treatment, and the inability to consider some mutations as likely indicators for the presence of other relevant but undetected mutations (for example, the presence of RT mutation L74V was reported to suggest K65R minorities [34]). However, despite the presence of 215 revertants in our datasets, this study demonstrated a performance of predicted phenotypes that is at least comparable with expert-based approaches. In the Stanford-California (EuResistDB) data, 6.1% (8.8%) of the classic TCEs contained RT sequences with 215 revertants; using the genotype-centric definition the prevalence decreased to 4.7% (7.7%). Moreover, services providing predicted phenotypes display a list of mutations along with the prediction results. Thus, information on mutations with clinical but no phenotypical effect is available to clinicians.

A third focus of our study was on assessing potential synergies between combinations of input representations. We hypothesize that representations derived from very different kinds of data might have a certain degree of complementarity, and approaches to combining different sources of knowledge might exploit the

complementarity and lead to more robust discrimination between successes and failures. Confirming this hypothesis, our experiments showed that combining genotypical representations with phenotypical representations incurs a significant benefit in cross-validation experiments. However, no combination of expert representations was consistently able to significantly improve the performance above the level of the single best expert encoding in the combination. The benefit observed in the hybrid inputs was qualitatively preserved when using classifiers trained on Stanford-California for predicting TCEs from EuResistDB.

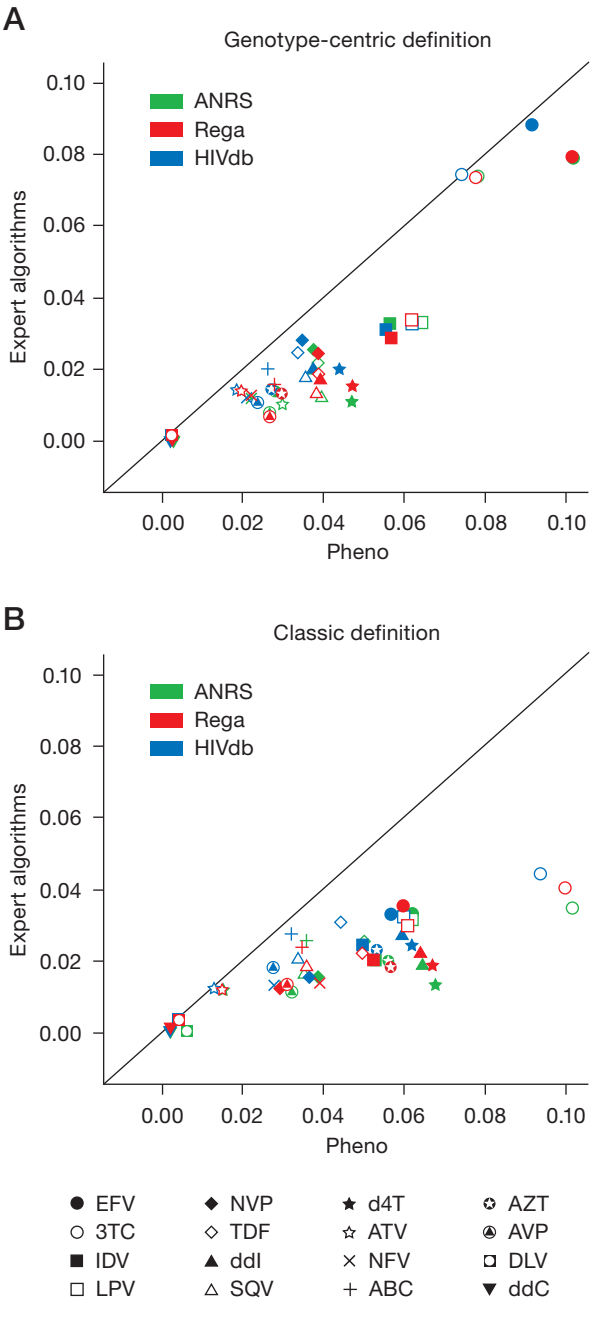
Finally, when using the classical instead of the genotype-centric definition, all performance measures were greatly decreased. However, analysis of the variable importance showed that, in general, the learned relations among drugs were similar. One reason for the decreased performance might be that, in contrast to the genotype-centric definition, the classic definition leaves a gap of 1–6 months between genotyping and detection of failure; that is, the genotype was taken at therapy start, or within 3 months before, and failure was detected after 1–3 months of treatment. This could be sufficient time for additional modifications in the viral genome. Another reason might be that, in the classic definition, a failure is defined as not reaching an undetectable viral load (or decrease of 2 $\log_{10}$ copies/ml) after 2 months, but in heavily pretreated patients a sufficient reduction of viral load might need more time. Indeed, 28% of samples defined as failure in the classic Stanford-California dataset showed a reduction ≤500 copies/ml later during the course of the therapy. In general, results obtained using the classic definition confirmed the picture obtained by the analyses using the genotype-centric definition: that is, statistical learning is better than summation, predicted phenotypes are the best single input representation and combining information on genotype and predicted phenotypes can improve predictive performance.

We have shown that the use of a phenotypical input representation is clearly competitive with genotype-based representations, the former often outperforming the latter. However, in our experiments, the magnitude of these synergies strongly depends on the dataset and on the standard datum definition. For a better understanding and definition of the potential benefits of hybrid approaches, these and similar experiments should be replicated with different datasets, standard datum definitions and possibly more sophisticated forms of hybrid representations.

Acknowledgements

The work at the Max Planck Institute for Informatics was partially supported by the EuResist project (EU grant IST-2004- 027173-STP). This study was presented, in part, at the *16th International HIV Drug Resistance Workshop*, 12–16 June 2007, Barbados, West Indies.

Figure 4. Comparison of variable importance of expert algorithms and predicted phenotype



The figure shows a scatter plot between the variable importance of a drug rated by either an expert algorithm (y-axes; ANRS, Rega or HIVdb) or the predicted phenotype (x-axes) derived from the hybrid models on **(A)** the genotype-centric and **(B)** the classic definitions. Points on the diagonal indicate equal importance, whereas points below or above the diagonal indicate that the phenotypical prediction is more or less important, respectively.

Disclosure statement

The authors declare no competing interests.

Additional file

The additional file 'Supplementary material' can be accessed at www.intmedpress.com

References

1. Descamps D, Brun–Vezinet F. Benefits of resistance testing. In Geretti AM (Editor). *Antiretroviral resistance in clinical practice*. London: Mediscript 2006; pp.73–78.
2. Sabin C. Database analyses of predictors of resistance. In Geretti AM (Editor). *Antiretroviral resistance in clinical practice*. London: Mediscript 2006.
3. Larder B, Wang D, Revell A, *et al*. The development of artificial neural networks to predict virological response to combination HIV therapy. *Antivir Ther* 2007; **12:**15–24.
4. Altmann A, Beerenwinkel N, Sing T, *et al*. Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance. *Antivir Ther* 2007; **12:**169–178.
5. Prosperi M, Di Giambenedetto S, Trotta MP, *et al*. A fuzzy relational system trained by genetic algorithms and HIV-1 resistance genotypes/virological response data from prospective studies usefully predicts treatment outcomes. *Antivir Ther* 2004; **9:**U89.
6. Prosperi M, Zazzi M, Perno CF, *et al*. 'Common law' applied to treatment decisions for drug resistant HIV. *Antivir Ther* 2005; **10:**S62.
7. Winters B, Montaner J, Harrigan PR, *et al*. Determination of clinically relevant cutoffs for HIV-1 phenotypic resistance estimates through a combined analysis of clinical trial and cohort data. *J Acquir Immune Defic Syndr* 2008; **48:**26–34.
8. Draghici S, Potter RB. Predicting HIV drug resistance with neural networks. *Bioinformatics* 2003; **19:**98–107.
9. Jenwitheesuk E, Samudrala R. Prediction of HIV-1 protease inhibitor resistance using a protein-inhibitor flexible docking approach. *Antivir Ther* 2005; **10:**157–166.
10. Shenderovich MD, Kagan RM, Heseltine PN, Ramnarayan K. Structure-based phenotyping predicts HIV-1 protease inhibitor resistance. *Protein Sci* 2003; **12:**1706–1718.
11. Sander O, Sing T, Sommer I, *et al*. Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage. *PLOS Comput Biol* 2007; **3:**e58.
12. Beerenwinkel N, Daumer M, Sing T, *et al*. Estimating HIV evolutionary pathways and the genetic barrier to drug resistance. *J Infect Dis* 2005; **191:**1953–1960.
13. Cohen CJ, Hunt S, Sension M, *et al*. A randomized trial assessing the impact of phenotypic resistance testing on antiretroviral therapy. *AIDS* 2002; **16:**579–588.
14. Haubrich RH, Kemper CA, Hellmann NS, *et al*. A randomized, prospective study of phenotype susceptibility testing versus standard of care to manage antiretroviral therapy: CCTG 575. *AIDS* 2005; **19:**295–302.
15. Wegner SA, Wallace MR, Aronson NE, *et al*. Long-term efficacy of routine access to antiretroviral-resistance testing in HIV type 1-infected patients: Results of the clinical efficacy of resistance testing trial. *Clin Infect Dis* 2004; **38:**723–730.
16. Garcia-Lerma JG, Nidtha S, Blumoff K, Weinstock H, Heneine W. Increased ability for selection of zidovudine resistance in a distinct class of wild-type HIV-1 from drug-naive persons. *Proc Natl Acad Sci U S A* 2001; **98:**13907–13912.
17. Brun-Vezinet F, Costagliola D, Khaled MA, *et al*. Clinically validated genotype analysis: guiding principles and statistical concerns. *Antivir Ther* 2004; **9:**465–478.
18. Rosen-Zvi M, Altmann A, Prosperi M, *et al*. Selecting anti-HIV therapies based on a variety of genomic and clinical factors. *Bioinformatics* 2008; **24:**i399–i406.
19. Altmann A, Rosen-Zvi M, Prosperi M, *et al*. Comparison of classifier fusion methods for predicting response to anti HIV-1 therapy. *PLoS One* 2008; **3:**e3470.
20. Roomp K, Beerenwinkel N, Sing T, *et al*. Arevir: a secure platform for designing personalized antiretroviral therapies against HIV. *Lecture notes in computer science: data integration in the life sciences* 2006; **4075:**185–194.
21. Hemelaar J, Gouws E, Ghys PD, Osmanov S. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *AIDS* 2006; **20:**W13–W23.
22. Meynard JL, Vray M, Morand-Joubert L, *et al*. Phenotypic or genotypic resistance testing for choosing antiretroviral therapy after treatment failure: a randomized trial. *AIDS* 2002; **16:**727–736.
23. Van Laethem K, De Luca A, Antinori A, Cingolani A, Perna CF, Vandamme AM. A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients. *Antivir Ther* 2002; **7:**123–129.
24. Rhee SY, Gonzales MJ, Kantor R, Betts BJ, Ravela J, Shafer RW. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res* 2003; **31:**298–303.
25. Vermeiren H, Van Craenenbroeck E, Alen P, *et al*. Prediction of HIV-1 drug susceptibility phenotype from the viral genotype using linear regression modeling. *J Virol Methods* 2007; **145:**47–55.
26. Johnson VA, Brun-Vezinet F, Clotet B, *et al*. Update of the drug resistance mutations in HIV-1: fall 2006. *Top HIV Med* 2006; **14:**125–130.
27. DeGruttola V, Dix L, D'Aquila R, *et al*. The relation between baseline HIV drug resistance and response to antiretroviral therapy: re-analysis of retrospective and prospective studies using a standardized data analysis plan. *Antivir Ther* 2000; **5:**41–48.
28. Breiman L. Random forests. *Machine Learning* 2001; **45:**5–32.
29. Ruschhaupt M, Huber W, Poustka A, Mansmann U. A compendium to ensure computational reproducibility in high-dimensional classification tasks. *Stat Appl Genet Mol Biol* 2004; **3:** Article 37.
30. Beerenwinkel N. Computational analysis of HIV drug resistance data. Aachen: Shaker 2004.
31. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett* 2006; **27:**861–874.
32. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics* 2005; **21:**3940–3941.
33. Swanstrom R, Bosch RJ, Katzenstein D, *et al*. Weighted phenotypic susceptibility scores are predictive of the HIV-1 RNA response in protease inhibitor-experienced HIV-1-infected subjects. *J Infect Dis* 2004; **190:**886–893.
34. Svarovskaia ES, Margot NA, Bae AS, *et al*. Low-level K65R mutation in HIV-1 reverse transcriptase of treatment-experienced patients exposed to abacavir or didanosine. *J Acquir Immune Defic Syndr* 2007; **46:**174–180.

Accepted for publication 24 November 2008