Additional file

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

André Altmann, Tobias Sing, Hans Vermeiren, Bart Winters, Elke Van Craenenbroeck, Koen Van der Borght, Soo-Yon Rhee, Robert W Shafer, Eugen Schülter, Rolf Kaiser, Yardena Peres, Anders Sönnerborg, W Jeffrey Fessel, Francesca Incardona, Maurizio Zazzi, Lee Bacheler, Herman Van Vlijmen and Thomas Lengauer

Antiviral Therapy **14**:273–283

Figure S1: ROC curves for expert algorithms, Pheno, Geno, and hybrid encodings using statistical learning on Stanford-California data (genotype-centric definition). Only part of the ROC curves is shown. AUC values are shown in brackets.

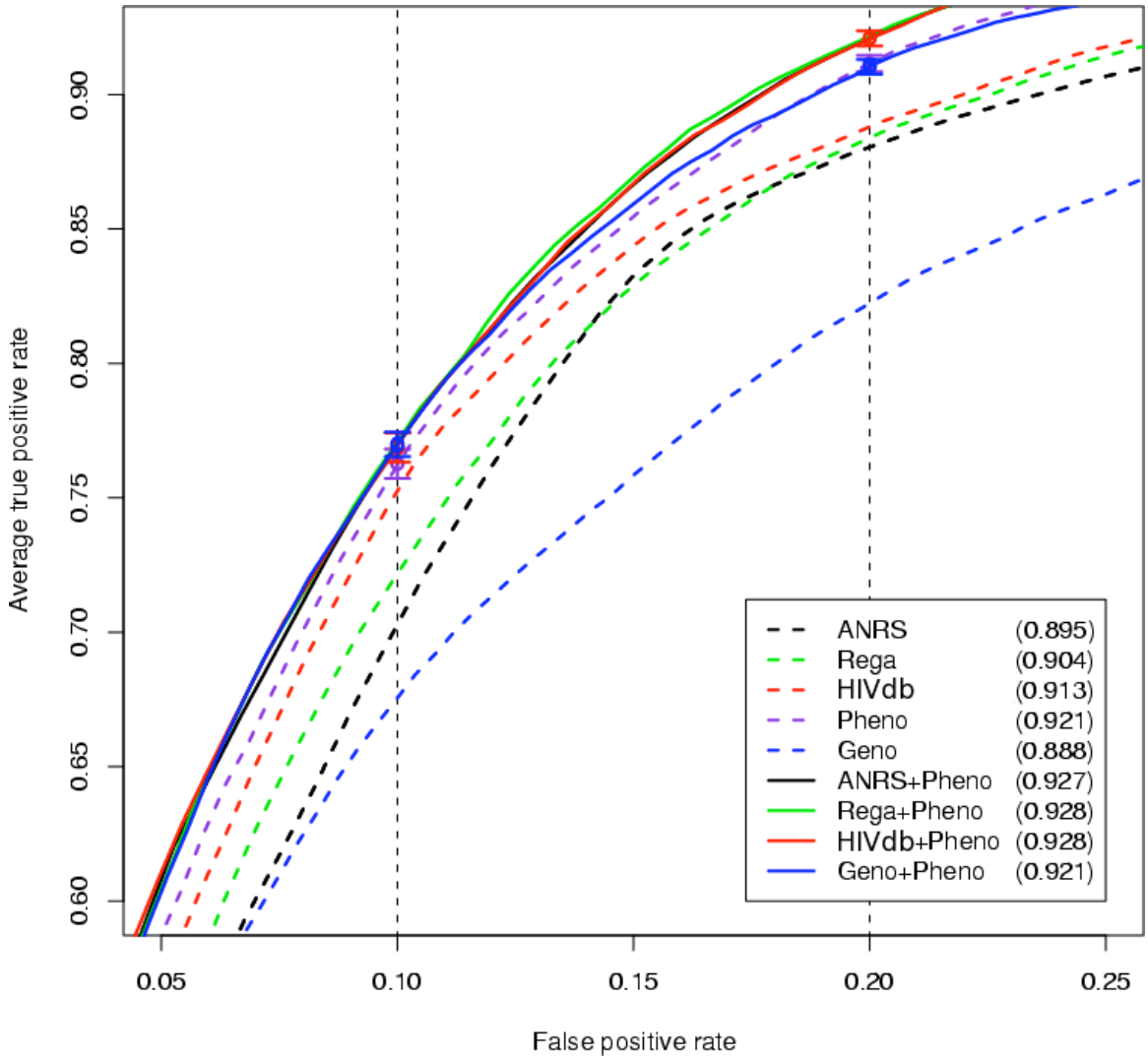

Figure S2: Comparison of variable importance in classic and genotype-centric definition.

The figure depicts the variable importance (in percent) for all drugs derived on datasets using ANRS (black), Rega (red), HIVdb (green), and Pheno (blue) encoding. The importance measures are grouped by standard datum definition. For more details see legend of Figure 3.

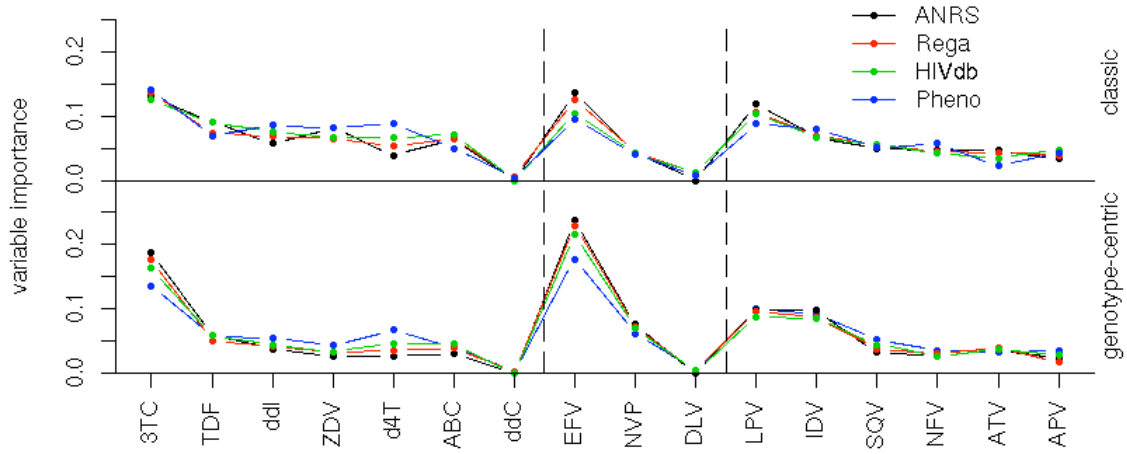

Table S1: Results on EuResistDB without obsolete treatments when trained on Stanford.

Obsolete treatments (i.e. regimens containing DLV, ddC, NFV, or unboosted PIs) have been removed from the EuResistDB datasets, resulting in 3,523 (822) TCEs using the genotype-centric (classic) definition. Columns one, two, and three (four, five, and six) display the area under the ROC curve, the true positive rate at a false positive rate of 10%, and the true positive rate at a false positive rate of 20%, respectively, using the genotype-centric (classic) definition for different encodings. Standard deviation for each measurement is indicated in parentheses. The treatment scores are derived either by summation (rows two to five) or by statistical learning (rows seven to fifteen). Rows two to eleven represent single encodings, and rows twelve to fifteen are hybrid encodings.

	Genotype-centric			Classic		
	AUC	TPR10	TPR20	AUC	TPR10	TPR20
Regimens Scored by Summation						
ANRS	0.774 (0.000)	0.248 (0.000)	0.541 (0.000)	0.708 (0.000)	0.242 (0.000)	0.481 (0.000)
Rega	0.769 (0.000)	0.209 (0.000)	0.566 (0.000)	0.695 (0.000)	0.200 (0.000)	0.400 (0.000)
HIVdb	0.760 (0.000)	0.157 (0.000)	0.454 (0.000)	0.719 (0.000)	0.261 (0.000)	0.482 (0.000)
Pheno	0.773 (0.000)	0.178 (0.000)	0.516 (0.000)	0.703 (0.000)	0.216 (0.000)	0.410 (0.000)
Regimens Scored by Statistical Learning						
ANRS	0.830 (0.001)	0.447 (0.008)	0.731 (0.009)	0.706 (0.005)	0.245 (0.014)	0.467 (0.016)
Rega	0.844 (0.002)	0.495 (0.008)	0.732 (0.004)	0.715 (0.002)	0.250 (0.013)	0.444 (0.015)
HIVdb	0.859 (0.001)	0.552 (0.007)	0.770 (0.006)	0.710 (0.001)	0.253 (0.012)	0.397 (0.008)
Pheno	0.872 (0.001)	0.580 (0.008)	0.803 (0.006)	0.723 (0.003)	0.279 (0.014)	0.462 (0.002)
Geno	0.854 (0.002)	0.612 (0.007)	0.755 (0.011)	0.695 (0.005)	0.253 (0.015)	0.409 (0.020)
ANRS+Pheno	0.882 (0.002)	0.616 (0.009)	0.807 (0.006)	0.726 (0.004)	0.268 (0.019)	0.472 (0.016)
Rega+Pheno	0.884 (0.002)	0.622 (0.013)	0.811 (0.005)	0.727 (0.003)	0.264 (0.012)	0.466 (0.014)
HIVdb+Pheno	0.883 (0.002)	0.606 (0.007)	0.804 (0.004)	0.719 (0.003)	0.261 (0.014)	0.451 (0.013)
Geno+Pheno	0.879 (0.003)	0.628 (0.007)	0.792 (0.009)	0.725 (0.005)	0.261 (0.014)	0.442 (0.021)