**Figure, Supplemental Digital Content 2:**

**The *GABARAPL1* + *ATG9a* mRNA combination better segregates HIV positive and HIV negative individuals into 2 groups than *GABARAPL1* mRNA quantification alone.**

|  |
| --- |
|  |
| A-ROC analyses made with Stat-El software. 1st line: *GABARAPL1* mRNA data, 2nd line: *GALIG* mRNA data. 1st column: Receiver operating characteristic (ROC) curve on HIV-positive and HIV-negative PBMC data for *GABARAPL1* and *GALIG* mRNA copy numbers/ng of total RNA. 2nd column: distribution of effectives for HIV-positive (black line) and HIV-negative (grey line) PBMC mRNA rate (copy number/ ng of total RNA. B- ROC analyses made via the five-fold cross-validation method used to calculate the predictive accuracy of each Support Vector Machine with linear kernel (SVM) engine. “GABARAPL1”: ROC curves for GABARAPL1 alone prediction. “GABARAPL1+ATG9a”: ROC curves for the prediction of GABARAPL1 and ATG9a combination. “all genes”: ROC curves for the 8 genes included in the current study, GALIG, *BECN1*, *MAP1LC3B*, *ATG9a*, P62/SQSTM1, *GABARAP*, *GABARAPL1* and *GABARAPL2*. The area under curve (AUC) is specified for each ROC curve. |

**Methods**

The ability of a gene expression level to differentiate between those with or without HIV infection was evaluated using receiver operator characteristics (ROC) curve [1]. It graphically displays the sensitivity (true positive rate) and specificity (false positive rate) for various gene expression levels. It may identify the cut-points, i.e. the value of a given gene expression level where small gains in sensitivity result in large decreases in specificity (or vice versa). Classical ROC curves were made with Stat-El software.

The confusion matrix, used to evaluate the accuracy of the Support Vector Machine with linear kernel (SVM) algorithm allows to build ROC curves as well (true positive = specificity, true negative = 1-specificity) [2].

The area under the ROC curve (AUC) indicates the overall ability of the gene expression level to distinguish HIV-negative donors from cART-treated HIV-positive patients (AUC of 1 represents the perfect test and 0.5 a random test).

**References**

1 Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978; **8**:283–298.

2 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011; **12**:2825–2830.

**Table – Supplemental Digital Content 1: Predictive performance of 3 supervised machine learning algorithms**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Classified data** | **Machine learning algorithms** | ***GALIG*** | ***ATG9a*** | ***MAP1LC3B*** | ***GABARAP*** | ***BECN1*** | ***GABARAPL1*** | ***GABARAPL2*** | ***p62/SQSTM1*** |
| **One gene tested** | SVM1 | 0.8094(+/-0.317) | 0.598(+/-0.043) | 0.732(+/-0.224) | 0.555(+/-0.201) | 0.643(+/-0.237) | **0.911**(+/-0.139) | 0.568(+/-0.089) | 0.598(+/-0.043) |
| LR2 | 0.837(+/-0.207) | 0.612(+/-0.046) | 0.718(+/-0.217) | 0.643(+/-0.093) | 0.670(+/-0.233) | **0.911**(+/-0.139) | 0.568(+/-0.089) | 0.599(+/-0.126) |
| RF3 | 0.777(+/-0.153) | 0.536(+/-0.198) | 0.705(+/-0.280) | 0.646(+/-0.326) | 0.643(+/-0.321) | **0.910**(+/-0.111) | 0.476(+/-0.156) | 0.505(+/-0.141) |
| ***GABARAPL1*+ another gene** | SVM | 0.911(+/-0.139) | **0.955**(+/-0.074) | 0.896(+/-0.147) | 0.911(+/-0.139) | 0.941(+/-0.108) | skip | 0.896(+/-0.147) | 0.911(+/-0.139) |
| LR | 0.925(+/-0.091) | **0.925**(+/-0.091) | 0.911(+/-0.139) | 0.911(+/-0.139) | 0.941(+/-0.108) | 0.925(+/-0.091) | 0.896(+/-0.147) |
| RF | 0.910(+/-0.111) | **0.954**(+/-0.075) | 0.925(+/-0.091) | 0.910(+/-0.111) | 0.924(+/-0.169) | 0.912(+/-0.209) | 0.940(+/-0.061) |
| ***GABARAPL1* + *ATG9a*****+ another gene** | SVM | 0.940(+/-0.061) | skip | 0.925(+/-0.091) | 0.940(+/-0.061) | **0.955**(+/-0.074) | skip | 0.941(+/-0.108) | 0.925(+/-0.133) |
| LR | 0.940(+/-0.061) | 0.925(+/-0.091) | 0.925(+/-0.091) | **0.940**(+/-0.061) | 0.941(+/-0.108) | 0.955(+/-0.074) |
| RF | 0.940(+/-0.061) | 0.924(+/-0.097) | 0.910(+/-0.064) | **0.940**(+/-0.115) | 0.925(+/-0.091) | 0.924(+/-0.097) |

1: SVM: Support Vector Machine (linear kernel)

2: LR: Logistic Regression

3: RF: Random Forest

1, 2, 3: software *Scikit-learn* [1], a Python module integrating machine learning engines

4: Mean accuracy (+/- 95% standard deviation)

**Learning algorithm system and cross-validation method**

Data set encompassed the normalized expression level of 8 different genes (*GALIG*, *BECN1*, *MAP1LC3B*, *GABARAP*, *GABARAPL1*, *GABARAPL2*, *ATG9a*, *P62*/*SQSTM1*) in 27 HIV-positive cART-treated patients and 40 HIV-negative donors. We have used the software *Scikit-learn* [1], a Python module integrating state-of-the-art machine learning engines. Three classifiers were chosen to analyse the data set: Support Vector Machine with linear kernel (SVM), Logistic Regression (LR) and Random Forest (RF).

To examine the performance of the 3 prediction engines, the five-fold cross-validation method with feature normalization was used. The dataset is equally divided into 5 stratified parts. Four of them are used as training sets and the last one is used in turn, as the test set. Hence all samples in each part are selected as testing samples to test the classifier that is trained by samples in the other 4 parts. Each sample is tested once. To evaluate the performance of our prediction model, we calculated the predictive accuracy of each prediction. Using the confusion matrix, the accuracy is the ratio of all correct predictions (true positive + true negative) to the total number of cases evaluated (true positive + true negative + false positive + false negative).

**Reference**

1 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011; **12**:2825–2830.