214 -TRANSCRIPTOME PROFILING AND DATA MINING: PREDICTIVE GENES OF HIV DISEASE PROGRESSION

Francisco Díez-Fuertes[1], Esther Calonge[1], María Pernas[2], Humberto Erick de la Torre-Tarazona[1], Isabelle Casademont[3], Anavaj Sakuntabhai[3], José Alcamí[1]

[1]AIDS Immunopathology Unit, Centro Nacional de Microbiología, Instituto de Salud Carlos III, Madrid, Spain; [2] Molecular Virology Unit, Centro Nacional de Microbiología, Instituto de Salud Carlos III, Madrid, Spain; [3]Functional Genetics of Infectious Diseases, Pasteur Institute, Paris, FRANCE

Introduction

Elite controller-long term non progression (EC-LTNP) phenotype can be considered as a promising model of functional cure in HIV-infected individuals. LTNPs achieve a persistent control of HIV infection without the need of therapy, maintaining high levels of CD4$^+$ T lymphocytes and controlling HIV-1 replication [1]. However, the EC-LTNP phenotype is far to be fully understood, partially due to the heterogeneous nature of EC and LTNP phenotypes. Transcriptome sequencing of PBMCs isolated from HIV-infected patients with different patterns of disease progression along with the employment of data mining techniques allowed the identification of several marker genes of HIV disease progression.

Methods

The study population included 8 EC-LTNPs and 8 viremic LTNPs (vLTNPs) from LTNP-RIS cohort (Spanish HIV/AIDS Research Network), and 7 HIV-infected patients with a typical pattern of disease progression before (preTP) and after (postTP) receiving antiretroviral therapy (ART) from CoRIS cohort (Figure 1) [2]. The predictive model of disease progression was created combining the transcript abundance estimation obtained from Cufflinks with a bias-corrected feature selection procedure based in leave one out cross-validation (LOOCV) and a hierarchical Bayesian classification [3] (Figures 2 and 3).

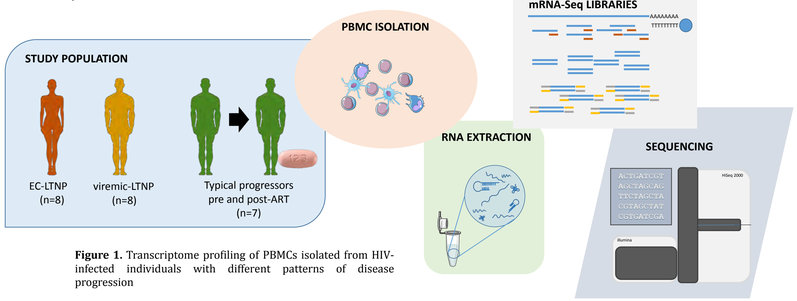


Figure 1. Transcriptome profiling of PBMCs isolated from HIV-infected individuals with different patterns of disease progression

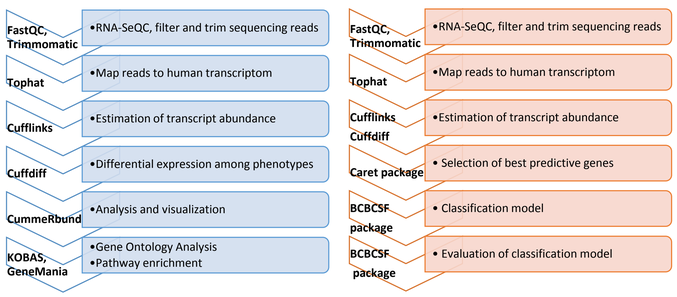


Figure 2. Typical pipeline of RNA-Seq experiments compared with the analysis workflow employed in this study and bioinformatics tools used in both approximations

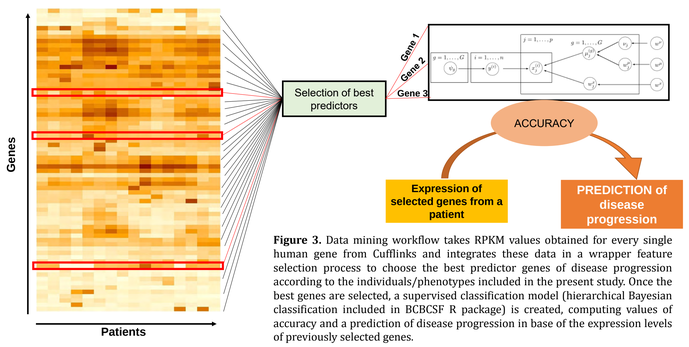


Figure 3. Data mining workflow takes RPKM values obtained for every single human gene from Cufflinks and integrates these data in a wrapper feature selection process to choose the best predictor genes of disease progression according to the individuals/phenotypes included in the present study. Once the best genes are selected, a supervised classification model (hierarchical Bayesian classification included in BCBCSF R package) is created, computing values of accuracy and a prediction of disease progression in base of the expression levels of previously selected genes.

Results

A mean of 33,670,437 ± 1,897,153 100 bp-reads was obtained for each library (91.7% ± 0.55% reads mapped to the human transcriptome). Multidimensional scaling showed a partial clustering of phenotypes, suggesting a high heterogeneity of the groups (Figure 4). The higher accuracy in terms of lower error rate (ER) and the average of minus log predictive probabilities (AMLP) was obtained employing the abundance estimation of 20 mRNAs as predictive variables (ER = 0.287; AMLP= 1.058) (Figures 5). The distribution of probabilities to be correctly classified obtained for each sample is showed in Figure 6. The model distinguishes LTNPs from TPs with an accuracy of 90% (Figure 6). A total of 13 genes mostly implicated in RNA binding (HELZ2, XRCC6, PARP12, PARP14, HERC5 and the components of the eEF1 complex EEF1G and EEF1B2), 6 pseudogenes (including RPL5P4, RPL4P5 and RPL4P4 ribosomal protein pseudogenes) and RP11.288L9.4 lncRNA (implicated in IFI6 mRNA silencing [4]) were selected as the best predictors of HIV disease progression (Figure 7). The functional annotation of these genes showed a statistical significant pathway for ISG15 antiviral mechanism (including HERC5, MX1 and RP11.288L9.4-IFI6; q=4.45x10^{-2}). Functional interaction analysis between selected genes is in Figure 8.

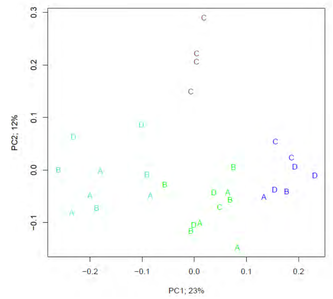


Figure 4. Multidimensional scaling (MDS) plot of the 30 samples based on the first two principal coordinates (PC, x and y axes). Labels A, B, C and D correspond to EC-LTNP, vLTNP, preTP and postTP phenotypes, respectively. Color code is based on k-means clustering results with N=4. The percentage of variability explained by each PC is indicated

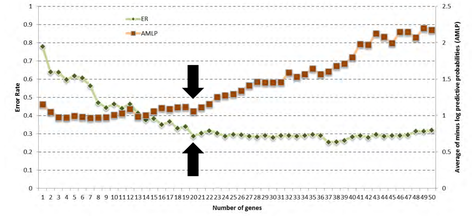


Figure 5. Identification of the best predictor genes of patient phenotype. Error rate (ER) and average minus log predictive probabilities (AMLP) were obtained in the feature selection process, evaluating the accuracy of 50 models employing the 1-50 best predictive genes. The model with 20 predictive genes was selected as the most accurate model.

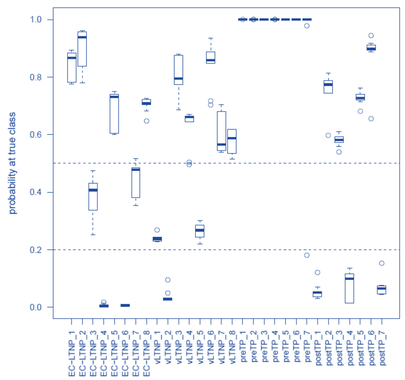


Figure 6. Probabilities to be correctly classified for each individual employing the 20 best predictive genes. A total of ten independent predictions were carried out with LOOCV and the distribution of these probabilities are showed as boxplots, indicating the first, second and third quartile values as well as the highest and lowest values (lines connected to the box through dashed lines) and the presence of outliers (open circles). The majority of the individuals (n=22, 73.3%) were correctly classified and 20 of them obtained p-values > 0.5 at true class for the 10 repetitions (and therefore p-values < 0.5 for the sum of the probabilities to be classified as any of the 3 other false classes). At the other extreme, six other individuals were repeatedly incorrectly classified with all the p-values < 0.2 for the 10 iterations (EC-LTNP 4 and 6, which were classified as vLTNPs; vLTNP 2, classified as EC-LTNP; and postTP 1, 4 and 7, which 2 of them were classified as EC-LTNP and the other as vLTNP). Simplifying the model to only two phenotypes (LTNPs and and TPs) an accuracy of 90% is achieved (27 out of the 30 patients were correctly classified).

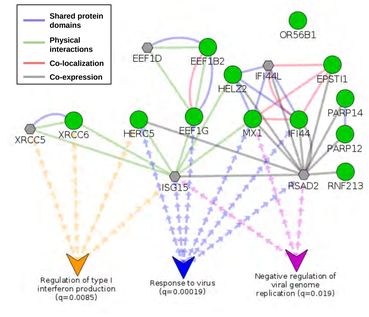


Figure 8. Gene-gene functional interaction network of genes selected by the classification model. Genes directly selected by the classification model are shown in green circles, whereas top five related genes according to Gene Ontology attributes are shown in gray hexagons. The network includes physical interactions between proteins (green edges), shared protein domains (blue edges), information about co-localization of proteins (red edges) and information about protein co-expression (gray edges). Statistically significant functions associated to this network are showed as colored arrow shapes, indicating the statistical q-value corrected by multiple comparisons as well as the genes associated to this function. The network was created using Cytoscape and GeneMANIA application.

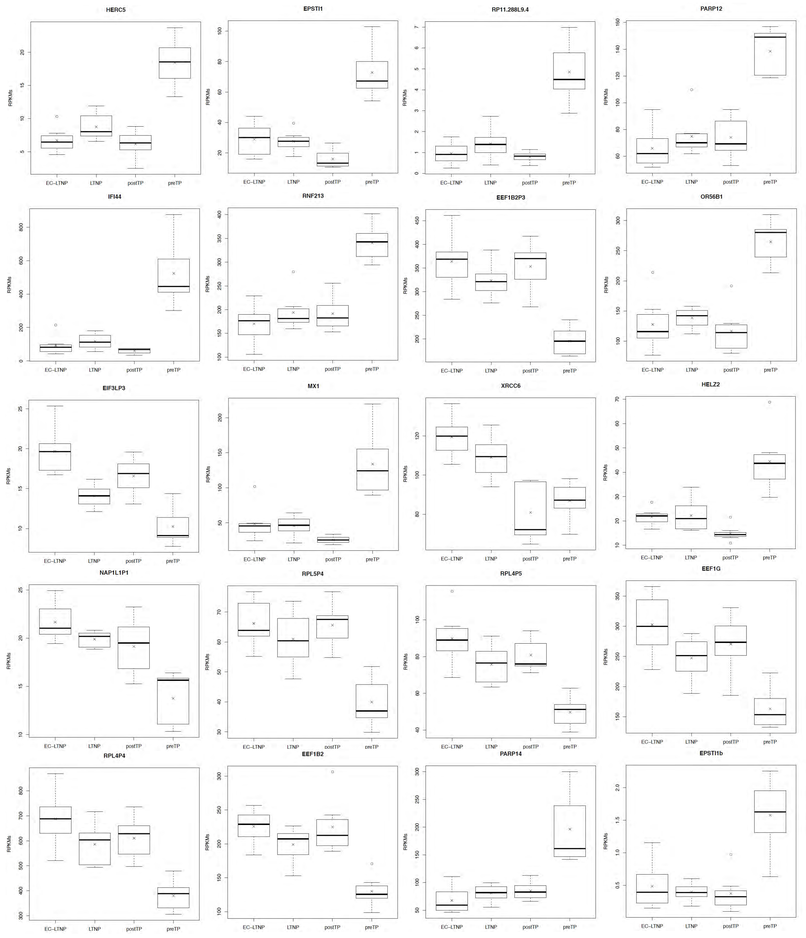


Figure 7. Best predictor genes of disease progression according to the hierarchical Bayesian classification model. The boxplots were generated in R and show the first and third quartile values for the RPKM distribution (upper and lower limits of the box), the median (the line splitting the box into two parts), the highest and lowest values (lines connected to the box through dashed lines), outlier values (open circles) and the mean value (crosses) for each phenotype.

Conclusions

- Supervised data mining classification methods combined with transcript abundance estimation was used to select 20 genes as the best predictors of HIV disease progression, obtaining a mathematical model able to distinguish between LTNPs (regardless of their HIV-control capacity) and TPs (without considering if they are on ART or not) with an accuracy of 90%.
- Selected markers of HIV disease progression point to the importance of the interferon-regulated ISG15 antiviral mechanism in preserving high CD4+ T cell counts and HIV-1 control capacity.
- Differential expression of genes and pseudogenes related with cell machineries of trancription/translation have been observed between EC-LTNPs, vLTNPs and typical progressors (preTP and postTP).

REFERENCES

[1] Casado C, et al. (2010). Host and viral genetic correlates of clinical definitions of HIV-1 disease progression. PLoS One 5: e11079
[2] García-Merino I, et al. (2009). The Spanish HIV BioBank: a model of cooperative HIV research. Retrovirology 6:27
[3] Longhai, L (2012). Bias-corrected hierarchical Bayesian classification with a selected subset of high-dimensional features. JASA 107:497
[4] Valadkhan S, et al. (2018). Regulation of the interferon response by lncRNAs in HCV infection. Front. Microbiol. 9:181.