A Bayesian network approach to study host and viral genetic correlates of HIV-1 disease progression

Katholieke Universiteit Leuven
Rega Institute for Medical Research
Department of Microbiology and Immunology
Clinical and Epidemiological Virology

Joke Snoek[1,2], Concepción Casado[3], Sara Colombo[2], Andri Rauch[4], Soledad García[5], Raquel Martinez[2], Kristof Theys[1], Ricardo Khouri[1], Huldrych F. Günthard[6], Amalio Telenti[2] , Cecilio López-Galíndez[3] and Anne-Mieke Vandamme[1*]

[1]Clinical and Epidemiological Virology, Rega Institute, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium, [2]Institute of Microbiology, University Hospital Center, University of Lausanne, Lausanne, Switzerland, [3]Centro Nacional de Microbiología (CNM), Instituto de Salud Carlos III, Majadahonda. 28220 Madrid. Spain, [4]University Clinic of Infectious Diseases, University Hospital Bern and University of Bern, Bern, Switzerland, [5]Centro Sanitario Sandoval (CSS), IMSALUD, Comunidad Autónoma de Madrid, 28010 Madrid, Spain, [6]Division of Infectious Diseases and Hospital Epidemiology, University of Zurich, Zurich, Switzerland

Email contact: annemie.vandamme@uz.kuleuven.ac.be

Background

HIV disease progression is very variable among infected patients (Table 1). Using classical statistical methods based on a selected number of markers, Casado et al identified a number of host and viral genetic correlates for the clinical definitions of HIV-1 disease progression: elite controllers, long term non progressors including viremic controllers and clinical non progressors, regular progressors and rapid progressors (Fig. 1, 2 and 3).

Table 1. Clinical definitions of HIV-infected patients

Elite controllers	• Asymptomatic HIV Infection over 10 year after seroconversion • Longitudinal HIV RNA that includes a minimum of 3 determinations, in the absence of antiretroviral agents, which span at least a 12-month period. • Plasma HIV RNA levels without ART that are below the level of detection for the respective assay (e.g., < 75 copies/mL by bDNA or < 50 by ultrasensitive PCR). • Isolated episodes of viremia up to 1000 copies /mL as long as they are not consecutive and represent the minority of all available determinations.
Long term non progressors viremic controllers	• Asymptomatic HIV Infection over 10 year after seroconversion. • Longitudinal HIV RNA that includes a minimum of 3 determinations, in the absence of ART, which span at least a 12-month period. • Plasma HIV RNA levels without ART that are equal or below 2000 cop/ml. • Isolated episodes of viremia above 2000 copies /mL as long as such episodes represent the minority of all available determinations.
Clinical non progressors	• Asymptomatic HIV Infection over 10 year after seroconversion • Plasma HIV RNA levels without antiretroviral therapy above 2.000 copies in more than 50% of the samples.
Regular progressors	• Symptomatic infection or initiation of ART within 10 years after seroconversion • Longitudinal HIV RNA that includes a minimum of 3 determinations, in the absence of ART, with a viral set point level between 10,000-200,000 copies /ml
Rapid progressors	• >=2 CD4 T cell measurements below 350/mm^3 with no value >=350 afterwards in the absence of ART within 3 years after seroconversion. • And/or, ART initiated within 3 years after seroconversion, and at least one preceding CD4 < 350/mm^3. • And/or, AIDS or AIDS-related Death within 3 years after seroconversion and at least one preceding CD4<350/mm^3.

ART: antiretroviral therapy

Materials and methods

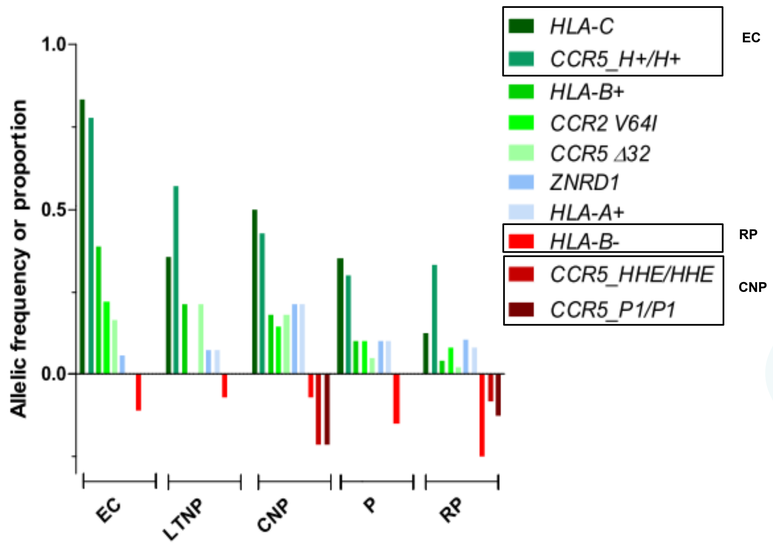
Host genetic and viral data for 64 patients as described by Casado et al were transformed to Boolean variables and used in a Bayesian Network (BN) learning approach, using the B-course software (http://b-course.cs.helsinki.fi/obc/) adapted by Deforche et al, 2006, scoring models by maximizing the posterior probability of the model. The predictive value of the network for classifying the clinical definitions, was investigated using exact Bayesian inference in the network. The network arcs that were weighing most in the predictive power of the network were uncovered by evaluating for each arc how predictive the network still was when omitting this arc. The stability of the dependency was assessed with a non-parametric bootstrap using 100 replicates. A higher frequency (bootstrap) of the arc means that the dependency is constant among the subgroups of patients of the dataset. All arcs with bootstrap support over 20% were considered important, arcs with bootstrap over 50% were considered stable.

References

Casado C, Colombo S, Rauch A, Martínez R, Günthard HF, Garcia S, Rodríguez C, Del Romero J, Telenti A, López-Galíndez C. **Host and viral genetic correlates of clinical definitions of HIV-1 disease progression.** PLoS One. 2010 Jun 11;5(6):e11079.

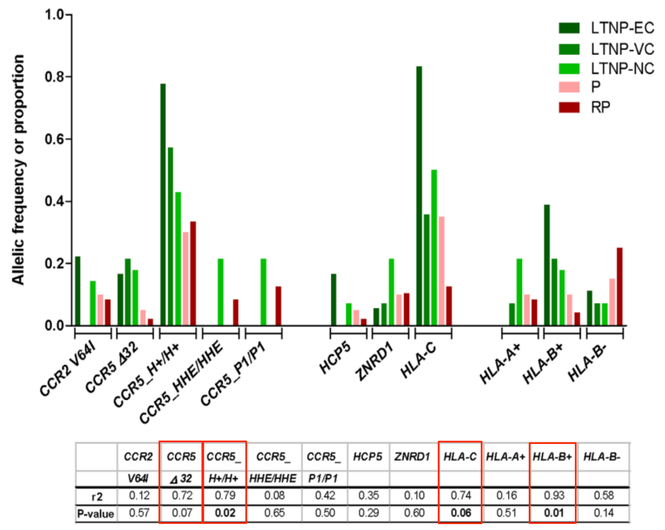
Deforche K, Silander T, Camacho R, Grossman Z, Soares MA, Van Laethem K, Kantor R, Moreau Y, Vandamme AM; non-B Workgroup. **Analysis of HIV-1 pol sequences using Bayesian Networks: implications for drug resistance. Bioinformatics.** 2006 Dec 15;22(24):2975-9. Epub 2006 Oct 4.

Figure 1



From LTNP-EC to RP, clinical definitions were associated with changes in the frequency (depletion) of protective host factors, in particular the CCR5 protective haplotypes (**CCR5_H+/H+**, proportion decreasing from f = 0.78 in EC to 0.33 in RP), the HLA-B protective alleles (B*2705, *5701, *5101, *1302, codified as HLA-B+, decreasing from f = 0.39 to 0.04), and the **HLA-C-35 rs9264942** variant (decreasing from f = 0.83 to 0.13). The inverse situation occurred with host markers related with rapid progression, in particular **CCR5_P1 homozygosity** (proportion increasing from 0 in EC to 0.13 in RP), and **HLA-B risk alleles** (B*1801, HLA-B*35Px alleles and B22 serogroup, increasing from f = 0.11 to 0.25).

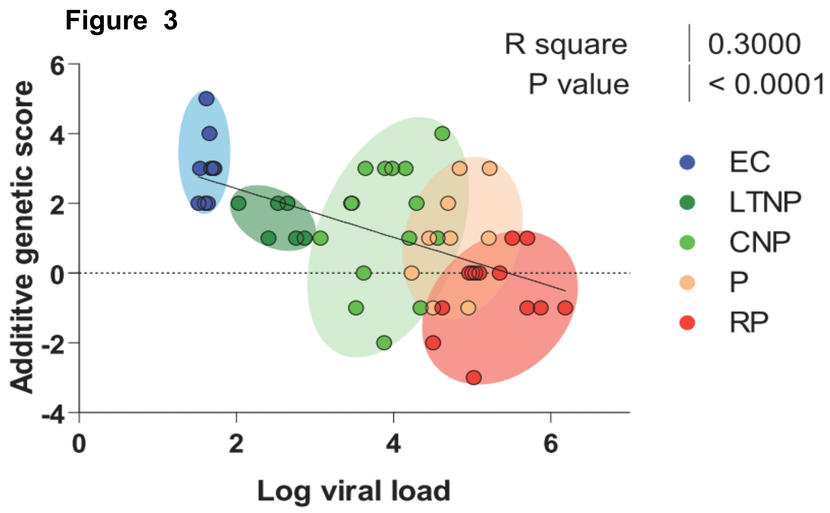
Figure 2



	CCR2 V64I	CCR5 Δ32	CCR5_ H+/H+	CCR5_ HHE/HHE	CCR5_ P1/P1	HCP5	ZNRD1	HLA-C	HLA-A+	HLA-B+	HLA-B-
r2	0.12	0.72	0.79	0.08	0.42	0.35	0.10	0.74	0.16	0.93	0.58
P-value	0.57	0.07	**0.02**	0.65	0.50	0.29	0.60	**0.06**	0.51	**0.01**	0.14

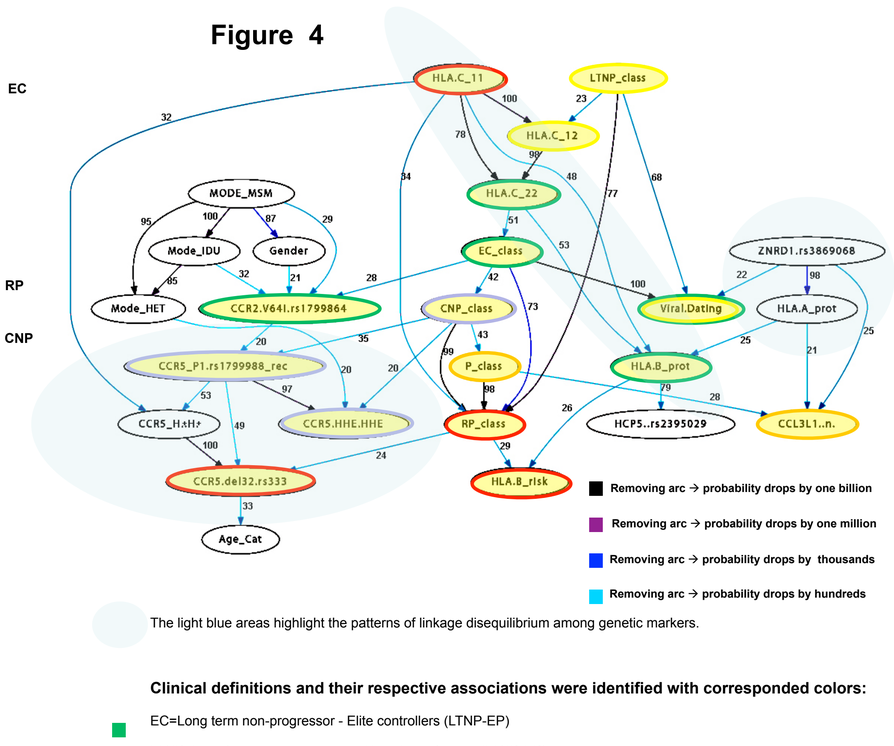
In regression analysis, several of the markers displayed significant statistical association in their frequency distribution across clinical definitions: protective **HLA-B+** (r2 = 0.93, p = 0.01) and **CCR5_H+/H+** (r2 = 0.79, p = 0.02), and trend association for **HLA-C-35** (r2 = 0.75, p = 0.06) and **CCR5 D32** (r2 = 0.72, p = 0.07)..

Figure 3



The increase in viral load was significantly associated (p value,0.0001) with a simple additive score that included the most valuable genetic markers (**HLA-C -35**, **CCR5 D32**, **CCR2 V64I**, **CCR5_P1/P1**, and **HLA-B+** and **B- alleles**).

Figure 4



The light blue areas highlight the patterns of linkage disequilibrium among genetic markers.

Clinical definitions and their respective associations were identified with corresponded colors:

EC=Long term non-progressor - Elite controllers (LTNP-EP)

LTNP=Long term non-progressor – Viremic controllers (LTNP-EP)

CNP=LTNP non controllers – Non controllers (LTNP-NC)

P=Progressors

RP=Rapid progressors

Results

To visualize the various dependencies among clinical definitions, and host and viral factors, we used a Bayesian network representation of the data. A number of variables that were redundant (being part of the definition of the groups) or non-discriminative were left out for inferring the network: viral load, proviral load, predicted X4/R5 phenotype and nature and charge and ethnicity. The network correctly captured many known correlations in the data, for example, patterns of linkage disequilibrium among genetic markers (i.e., *HLA-A* and *ZNRD1*, *HLA-B* and *HLA-C* alleles, *CCR5* haplotypes), and provided numerical support to the various dependencies (Fig.4, light blue areas). Clinical definitions (class) were directly dependent on viral evolution expressed as “viral dating” (Fig.4, green/ yellow circle), protective *HLA-B* alleles (Fig.4, green and red circles), *HLA-C* (Fig.4, green, red and yellow circles) and CCR5 genotypes (Fig.4, red and purple circles). In addition, the association between CCR5 H+/H+ and clinical definitions is dependent on HLA-C genotypes, demonstrating that HLA-C is a better marker for HIV disease progression. Despite copy number of CCL3L1 did not discriminate among clinical groups (average copies per diploid genome; LTNP-EC, 2.1; LTNP-VC, 2.6; LTNP-NC, 2.1; P, 1.6; RP, 2.2) our model was able to identify the reduced copy number of CCL3L1 as marker of P group (Fig. 4, orange circles). The results were consistent with what Casado et al described in their paper.

Conclusions

We applied a Bayesian approach to the joint analysis of the diverse host and viral data. The resulting network provided a comprehensive and hierarchical structure of the various dependencies; it identified viral evolution and *HLA-B* and *HLA-C* alleles as key correlates of clinical definitions of disease progression, confirming the results obtained by classical statistics. We feel that the Bayesian approach is well suited to quickly explore large datasets, saving time by prioritizing subsequent statistical confirmation of the associations found. Such Bayesian networks could also be applied for predicting the clinical course of the individual.