Master of Science in Omics Data Analysis

Master Thesis

Exploring HLA class II allele associations with markers of HIV control

by

Bruna Oriol Tordera

Supervisor: Alex Olvera Van der Stoep

Host Genetics and Cellular Immunity, AIDS Research Institute IrsiCaixa.

Co-supervisor: Christian Brander,

Host Genetics and Cellular Immunity, AIDS Research Institute IrsiCaixa.

Guarantor: Malu Calle Rosingana

Department of Systems Biology, University of Vic – Central University of Catalonia

Department of Systems Biology

University of Vic – Central University of Catalonia

[19.09.2016]

ACNOWLEDGEMENTS

I would like to thank all the people who, in some way, have helped me in this project. I will start thanking Alex Olvera for being my mentor during this project, for all his patience, help, corrections and suggestions as well as for all he has taught me. Following, I would also thank Christian Brander his predisposition to pursue this project and giving valuable feedback to get the most out of it. Next, thanks to Malu Calle not only for being my guarantor in this final master thesis and suggesting me how to improve it, but also as coordinator of the Msc in Omics Data Analysis that has provide me with all the knowledge I needed to carry on the study showed herein. I want also to thank all the teachers and mates of this course from who I have learnt a lot. A special thanks to all the people in the group "Host Genetics and Cellular Immunity" for welcoming me as one more from the very first day and all the people I have met in IrsiCaixa. I wish to express a warm gratitude to my parents who back me in all my decisions, my brothers and my family in general, as well as to Albert for putting up with all my worries and bad days. Finally, my thanks go to all my friends, especially to University friends and my most intimate circle.

ABSTRACT

Among HIV seropositive individuals, the progression of the HIV-related disease is quite heterogeneous, partially due to the genetic background of the patients. Different host genetic factors have been statistically associated with HIV disease control or progression, especially HLA class I (HLA-I) polymorphisms because of the direct role that these molecules play in the immune response against the virus. In recent years, the immune response against HIV based on $CD4^+$ T helper cells and HLA class II (HLA-II) restricted responses has gained importance to refine current vaccine approaches. Therefore, the main objective of the present study was to explore potential statistical associations of HLA-II with markers of HIV control, in particular HIV viral load and $CD4^+$ counts, in a Peruvian cohort of almost 400 individuals with existing HIV infection or at high risk for infection. Additionally, we also studied the associations of HLA-II with the presence of T cell responses to overlapping peptides (OLPs) covering the whole HIV proteome.

In order to achieve such objectives, different statistical approaches using the R statistical software were applied. Mainly, the Fisher's Exact Test was used to assess whether the expression of certain HLA-II alleles was associated with HIV infection or with the response to the different HIV epitopes. On the other hand, the Mann-Whitney Test was used to detect differences in viral loads or $CD4^+$ counts between patients with or without a certain HLA-II allele.

In our analyses, alleles HLA-DRB1*1201 and HLA-DRB1*1302 appeared to be significantly associated ($p<0.05$) with low and high viral loads, respectively. A number of additional alleles were significantly associated with $CD4^+$ counts. Most of the identified associations included the HLA-DRB1 locus, the most polymorphic of the HLA class II loci. Additionally, expression of HLA-DRB1*1201 was associated with T cell response to OLP 41 in the Gag protein, and HLA-DRB1*1302, with the response to OLP 82 in Nef protein, identifying dominant targets of the T cell response restricted by these two alleles.

Overall, the associations of HLA-DRB1*1201 and HLA-DRB1*1302 with viral load, support previous data suggesting a potential effect of the $CD4^+$ T cell responses on HIV disease control.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIDS	Acquired Immune Deficiency Syndrome
APC	Antigen Presenting Cell
CART	Classification And Regression Trees
$CD4^+$ T cells	$CD4^+$ T helper cells
CI	Confidence Interval
CTL	Cytotoxic T Lymphocytes
DC	Dendritic Cells
DGI	Decrease of Gini Impurity
DNA	Deoxyribonucleic Acid
EC	Ellite Controllers
FDR	False Discovery Rate
GWAS	Genome-Wide Association Studies
HESN	Highly Exposed Seronegative
HIV	Human Immunodeficiency Virus
HLA	Human Leukocyte antigen.
I(g)	Gini index
LD	Linkage Disequilibrium
LTNP	Long-Term Non Progressors
MDG	Mean Decrease Gini
MSE	Mean Squared Error
MSM	Men who have Sex with Men
OLP	Overlapping Peptides

OR	Odds Ratio
p	p-value
PBMC	Peripheral Blood Mononuclear Cell
RNA	Ribonucleic Acid
RT	Reverse Transcriptase
RF	Random Forest
TCR	T-Cell Receptor
VI	Variable Importance
VL	Viral Load
WHO	World Health Organization

1. INTRODUCTION

1.1. HIV Infection and AIDS

According to the World Health Organisation (WHO), HIV is still a most important public health concern, with around 36.9 million people infected with HIV by the end of 2014 and the estimation of, at least, 2 million new infections during that year (1).

Human Immunodeficiency Virus (HIV) is a RNA virus that targets immune cells, particularly $CD4^+$ T cells and impairs their functioning. The progression of HIV associated disease in infected people causes a decline of their immune response, which leads to the development of the acquired immunodeficiency syndrome (AIDS) if they are not adequately treated. AIDS is characterised by a major susceptibility of getting infected by opportunistic pathogens due to the weakness of the immune system, mainly consequence of this $CD4^+$ T cells depletion (2). Actually, HIV disease progression is determined by patient $CD4^+$ T cell counts and their viral load (Figure 1.1).

Natural course of HIV infection (Figure 1.1) is usually divided in 4 stages: Acute retroviral syndrome, asymptomatic phase, pre-AIDS syndrome and AIDS. The asymptomatic phase can last between 7 and 10 years during which the viral replication is remarkably stable maintained around a so-called viral set point. The viral load at set point predicts the likeliness of a patient to progress towards AIDS: high viral load in this set point is associated with a higher probability of disease progression (3, 4).

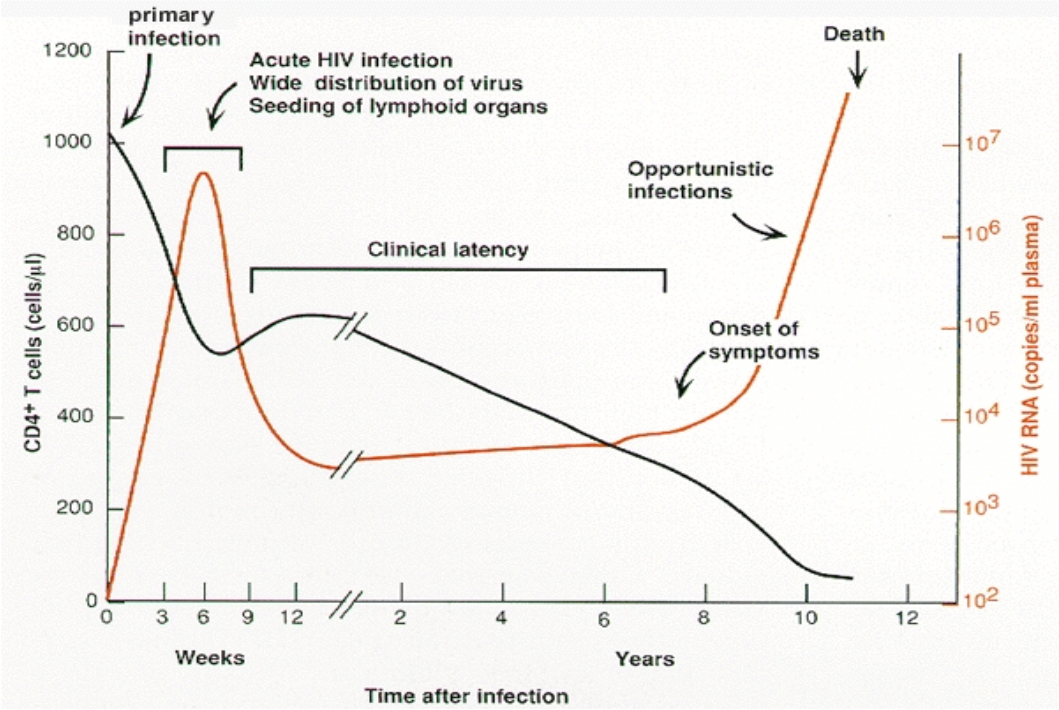


Figure 1.1 Natural History of HIV. Graph showing the different phases of HIV disease progression through weeks or years and how the parameters viral load and $CD4^+$ counts change during the different periods of time. There is an inverse correlation between these two parameters. *(Coffin, J., Hughes, S. and Varmus, H. Retroviruses, 1997. Cold Spring (NY): Harbor Laboratory Press.)*

Within the HIV infected human population, there are some individuals called Long-Term Non-Progressors (LTNP) that control viremia and present moderate $CD4^+$ T cell counts decline. Although they suffer a loss of these cells, the rate in which they are depleted is much slower than in normal infected individuals. Within the LTNP group, there are Elite Controllers (EC) subjects who maintain undetectable viral load (under 50 copies/ml) for a long period of time. Furthermore, there are some individuals that do not seem to become infected despite being highly exposed to HIV, for instance, sex workers or discordant couples. These subjects are often referred to as Highly Exposed Seronegatives (HESN). These subpopulations have been relevant to the detection of different biological markers, mainly host genetic factors, associated with the control of HIV disease progression. Filling the gaps in our knowledge of the pathogenesis of HIV might allow the refinement of the current HIV vaccine approaches (5).

1.2. HIV structure and genome

HIV is an enveloped retrovirus containing inside its capsid 2 copies of its RNA genome and the enzymes reverse transcriptase (RT), protease and integrase which are necessary to start the virus replication cycle (Figure 1.2). The envelope is formed by a host derived lipid bilayer where viral proteins gp41 and gp120 are embedded. These proteins interact with the host cells and prompt the membrane fusion between the targeted cell and the virus. Once HIV is inside the host cell, the RT enzyme converts the viral genome in proviral DNA, which is then integrated into the genome of the infected host cell by the integrase enzyme (2).

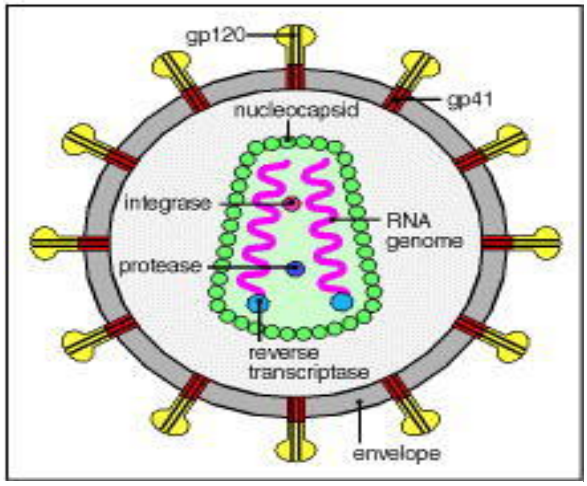


Figure 1.2 HIV structure. The external part of the virus, in contact with the external media, is formed by the lipid bilayer envelope containing envelope proteins gp120 and gp41. Inside the envelope there is the proteic nucleocapsid formed by the different Gag subunits. It contains the two copies of the RNA viral genome and the enzymes reverse transcriptase (RT), protease and integrase. *(Janeway CA Jr, Travers P, Walport M, et al. Immunobiology: The Immune System in Health and Disease, 2001. Garland Science (NY))*

The HIV genome (Figure 1.3) contains 9 genes flanked by long terminal repeats (LTR) regions. The 9 coding genes encode different proteins that can be divided in 3 groups: Structural proteins, Regulatory proteins and Accessory proteins. The structural genes include genes codifying for essential proteins common in all retroviruses: Gag, Pol and Env. Gag encodes viral core and matrix proteins; Pol, the enzymes Protease, Integrase, RNAseH and RT needed to start viral replication; and Env, the envelope glycoproteins gp120 and gp41. Regulatory

proteins Tat and Rev are fundamental in viral replication. Tat allows the viral RNA synthesis and Rev, the nuclear export to the host cell cytoplasm. Regarding the accessory proteins Nef, Vif, Vpr and Vpu, despite not being necessary for viral replication, they boost viral infection (3, 5, 6)

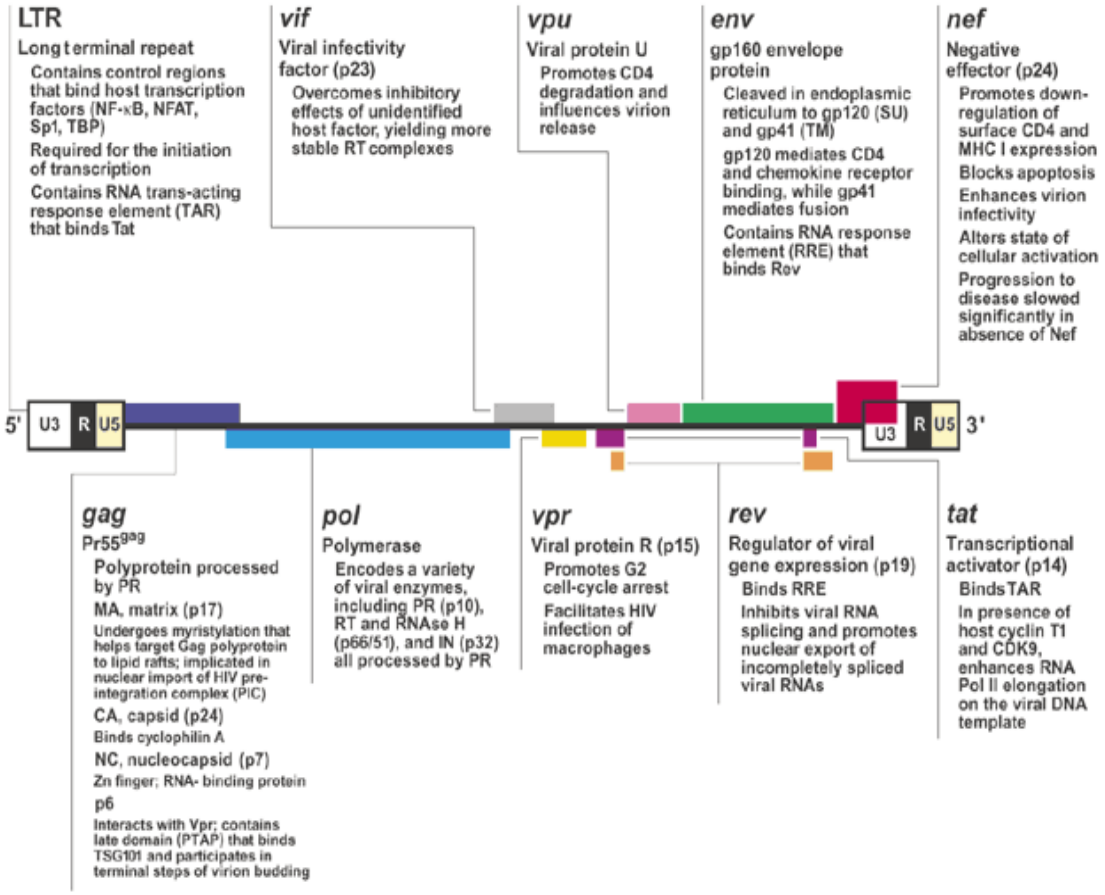


Figure 1.3 HIV genome. HIV genome consist of 9 genes that are translated in 15 different proteins. Genes Gag, Pol and Env encode for structural proteins; genes Tat and Rev for regulatory proteins and genes Vif, Vpu, Vpr and Nef, for accessory ones. *(Warner C. Greene,B. Matija Peterlin. Charting HIV's remarkable voyage through the cell: Basic science as a passport to future therapy. Nature Medicine. 8 (2002): 673 - 680)*

1.3. Host Genetics – The Human Leukocyte Antigen (HLA) System

Different host genetics factors have been related to HIV disease control, especially the HLA cell-surface receptors encoded in the HLA system. This is a highly polymorphic gene complex located in the short arm of chromosome 6.

There are 3 regions defined in this complex (Figure 1.4). Regarding HLA association studies, the interest is on regions I and II encoding the cell-surface molecules involved in antigen presentation. Region I contains the different loci for the HLA class I proteins: There are the classical HLA molecules (HLA-A, HLA-B and HLA-C), and the non-classical molecules (HLA-E, HLA-F and HLA-G) with some functions out of the scope of this study (7), although HLA-E has regained interest in adaptive immunity due to recent findings in CMV vaccinated monkeys (8). On the other hand, region II encodes the HLA class II molecules HLA-DM, HLA-DO, HLA-DP, HLA-DQ and HLA-DR. These molecules, the focus of the present study, are heterodimers of α and β chains encoded in different loci and referred as A or B, for instance, HLA-DQA and HLA-DQB. The HLA-II proteins that are involved in antigen presentation to T cells are the HLA-DP, HLA-DQ and HLA-DR molecules (7).

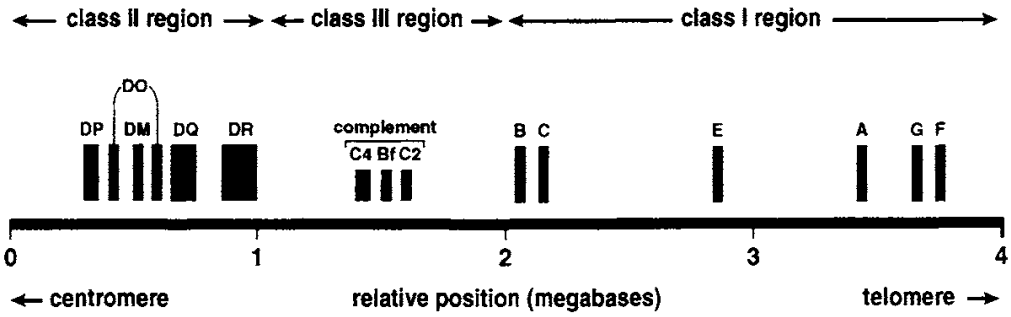


Figure 1.4 Simplified map of the HLA complex. This figure shows the short arm of chromosome 6 with the 3 regions defining the HLA complex. (*Steven G.E. Marsh, Peter Parham and Linda D. Barber. The HLA Facts Book, 1999. FactsBook series – Academic Press*).

Owing to the common function of HLA I and II molecules of binding antigenic peptides and presenting them to T cells, they show similar structures (Figure 1.5). For instance, both classes need a peptide binding cleft. However, these two receptors play different roles in the immune system. HLA-I molecules are found on the surface of almost all cell types in the human being, and bind intracellular processed peptides that are presented to CTLs. If the

peptide in question does not belong to the human being, the TCR of the CTL will recognise the cell as infected and the lymphocyte will eliminate it. In contrast, HLA-II molecules are preferentially found at the surface of antigen presenting cells (APCs). Their function is to present exogenous antigens that are recognised by the TCR of T-helper cells ($CD4^+$ T cells) which orchestrate an immune response against the pathogen (2, 7).

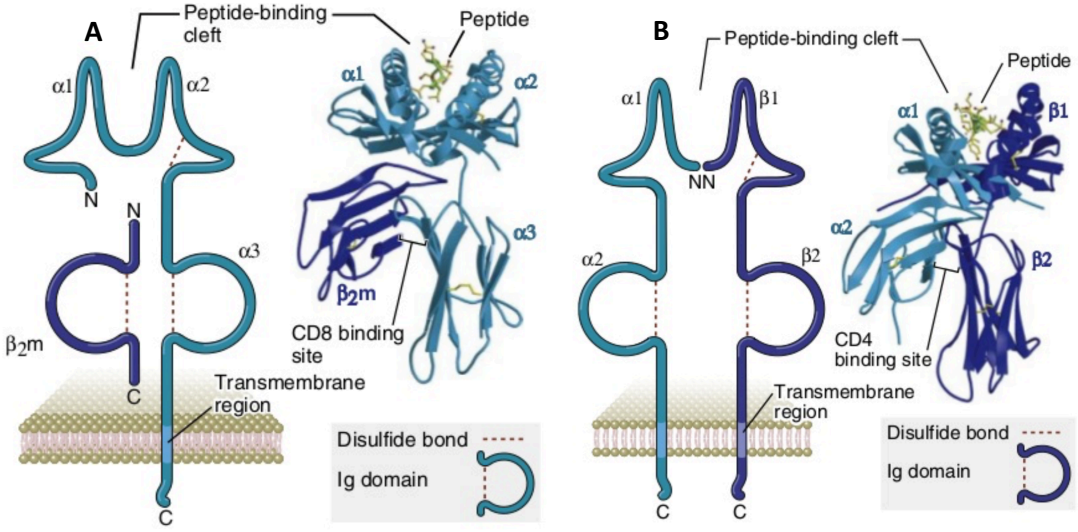


Figure 1.5 Structure of HLA cell-surface molecules. On the left (A) the structure of a HLA-I molecule is shown. On the right (B), the structure of a HLA-II protein formed by the α and β chains. Both molecules show similar structures and a peptide binding cleft. (*Abul K. Abbas & Andrew H. Lichtman. Basic Immunology. Functions and Disorders of the Immune System. 2011. Saunders – Elseiver).*

One particularity of HLA molecules is their co-dominance, which means that all the individuals will express the two copies of each of the genes in their genome. This fact provides humans with a wider capacity to present and recognise different antigens. At the population level, co-dominance, together with the fact that HLA complex genes are highly polymorphic, allow the presence of a huge variability of HLA haplotypes which is further increased with heterozygosity. This is beneficial not only at individual level by ensuring broad immune responses, but also at population level by blocking pathogens spread (7, 9).

The HLA polymorphisms are differently distributed among the distinct human populations (7), one allele might be highly frequent in Afroamericans but not in Caucasians or vice versa. This fact must be considered when designing therapies for different diseases and in the case of HIV, to get a vaccine applicable to worldwide populations.

1.4. T cell responses to HIV epitopes

The different HLA-I and HLA-II associations with HIV control might be consequence of the peptides presented by these molecules to T cells, which unleash different immune responses against HIV. These different responses have been mapped to the HIV genome, indicating that different alleles restrict the presentation of the different epitopes in a protein. For instance, in Figure 1.6 there is the map for p17 subunit from the HIV Gag protein (10).

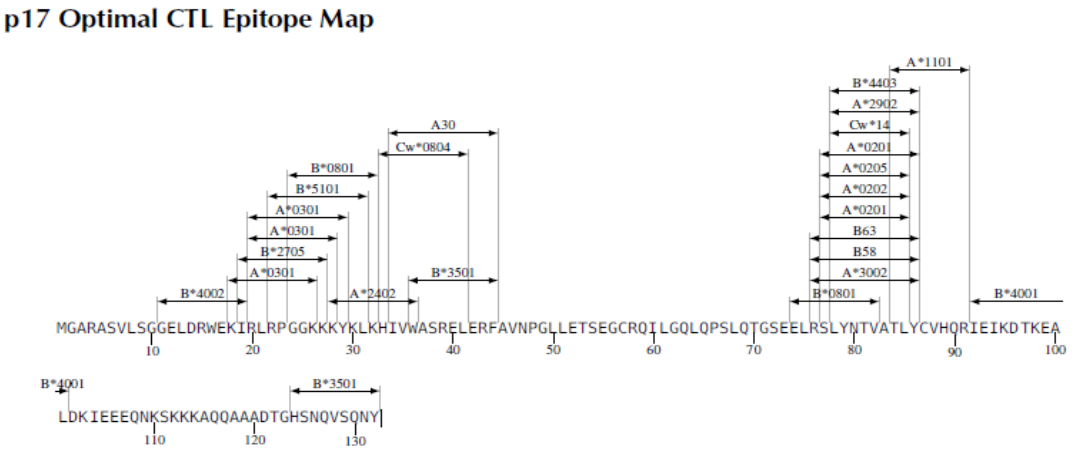


Figure 1.6. p17 Epitope Map. Aminoacid sequence of the Gag protein p17 with the HLA-I alleles that restrict each of the protein epitopes presentation. *(A. Llano, A.Williams, A.Olvera, S. Silva-Arrieta, C.Brander. Best-Characterized HIV-1 CTL Epitopes: The 2013 Update. HIV molecular immunology. 2013. Published by Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM. LA-UR 13-27758)*

To design an effective HIV vaccine, we need to select the best immunogen that covers those epitopes that are capable of triggering the most effective immune response. Accordingly, it is worth and necessary to study the T cell response of the different patients to the entire viral proteome. This can be accomplished by performing an IFNγ ELISpot assay where PBMCs of patients are stimulated with different *in vitro* synthesised overlapping peptides (OLPs). Normally, in the study of the HIV immune response, a set of 410 overlapping 18-residues-long peptides with an overlap of 10 or 11 amino acids is used to ensure the whole HIV proteome is covered. When these PBMCs are activated by one of the OLPs, they produce IFNγ that is then detected with biotinylated antibodies. This cytokine is mainly produced by $CD8^+$ T cells and, in less extent, by $CD4^+$ T cells (11, 12).

Such studies point out the regions of the HIV genome more frequently attacked by the human immune system. However, the selection of the most suitable peptide candidates to be included in a HIV vaccine is not trivial. The immune system needs to cope with the mutation capacity of some HIV regions that can facilitate the escape from T cell recognition and avoid the elimination of infected cells. Consequently, more conserved immunogens are likely to outperform immunogens covering more variable segments of the virus. For instance, Gag peptides, especially those in p24, are highly conserved, while Nef and Env peptides show a tremendous mutation rate (3, 5, 6).

Finally, each HLA molecule can recognise a wide range of peptides, but these molecules show preferences for the peptides they bound. With the information of the HLA haplotype of the patients, one can search statistical associations of HLA alleles with the T cell response to the different OLPs. The HLA alleles returning significant associations will be selected in immunodominance studies. This studies allow the determination of the HIV epitopes more frequently targeted by patients bearing a specific HLA allele. Therefore, it allows the identification of dominant epitopes whose presentation is restricted by these specific alleles.

2. OBJECTIVES

Only a few studies have been reported that focused on possible associations of HLA-II alleles with HIV control. However, there is a growing interest on the role of HLA-II and $CD4^+$ T cells in the immune response against HIV (13, 14, 15). For this reason, the central aim of the present project was the exploration of associations between the HLA-II alleles in the HLA-DQA, HLA-DQB, HLA-DRB1, HLA-DRB3, HLA-DRB4 and HLA-DRB5 loci, and different markers of HIV control and HIV infection risk in a Peruvian cohort of almost 400 individuals. This cohort was already studied for HLA-I associations (16), identifying individual HLA-I genes and gene combination that were associated with viral load and $CD4^+$ T cell counts. Both studies contribute with a better characterization of the immunogenetics in the Peruvian population, which has been little explored. In addition, the T cell responses of 218 HIV infected subjects in the cohort against a set of 410 overlapping peptides (OLPs) covering the HIV genome, were analysed with an IFNγ- ELISpot analysis (12).

The specific objectives of these analyses included:

1. Exploration of the associations between HIV infection risk and specific HLA-II alleles.
2. Exploration of the associations of HIV control and specific HLA-II alleles.
3. Determination of the effect of heterozygosity and population allele frequencies on HIV control.
4. Study of the associations of T cell responses to HIV epitopes and HLA-II alleles.

3. MATERIALS AND METHODS

3.1. Study Cohort

In the present project we worked with a Peruvian cohort of 392 subjects established and followed at the IMPACTA HIV clinics in Lima, Perú. All subjects were of mixed Amerindian ethnicity and showed high-risk behaviour and most of them were men who had sex with other men (MSM). Among them, there were 148 seronegative and 244 HIV-1 infected (seropositive) individuals. Of the 244 HIV infected patients, 11 were under antiretroviral treatment (16). These treated individuals were excluded from the analyses in the present study.

From the untreated seropositive individuals, we had information about $CD4^+$ T cell counts and viral load, measured by routine standard flow cytometry and qPCR respectively. The median viral load in the cohort was 37113 HIV copies/ml and the median CD4 counts was 384 $CD4^+$ T cells counts/µl.

From all patients in the cohort, we had information of their HLA-II typing for loci HLA-DQA, HLA-DQB, HLA-DRB1, HLA-DRB3, HLA-DRB4 and HLA-DRB5. These last three loci were studied as a single one, HLA-DRB345, due to their low allelic variability and their genome proximity. Additionally, for 218 of seropositive patients, T cell responses were measured by FNɣ-ELISpot assay using peripheral blood mononuclear cells (PBMCs) (11, 12). The protocol used was optimised for $CD8^+$ T cell response detection (12), notwithstanding, it is known that part of the IFNɣ detected can be produced by $CD4^+$ T cells.

3.2. HLA Nomenclature

Figure 3.1 offers a schematic view of HLA nomenclature. First of all, there is the prefix “HLA” and separated by a hyphen, the locus in question (in Figure 3.1, locus DQA that encodes the α chain of HLA-DQ molecule). Then, after an asterisk, the allele group is indicated, that generally corresponds to two digits typing in line with serological typing, an antibody based method that has been used for HLA typings since before DNA based methodologies (PCR and sequencing) emerged. Finally, a second set of digits is added (optionally after a colon)

identifying a specific HLA allele. These latter digits inform about the changes in the amino acid sequence and is determined by DNA based methodologies (16, 17).

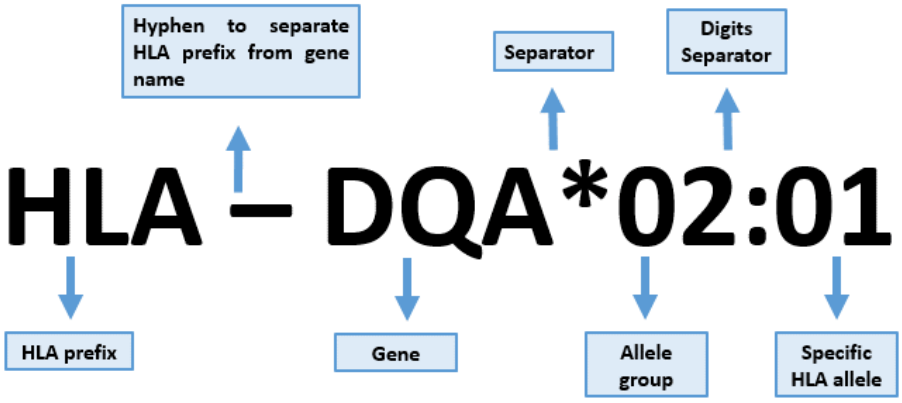


Figure 3.1 HLA Nomenclature. Scheme about how HLA molecules are designated. (Adapted from IPD-IMGT/HLA Database, http://hla.alleles.org/nomenclature/naming.html)

Focusing on the HLA-II typing in the cohort studied herein, most of the HLA-II alleles were non-ambiguously notated with the 4 digits resolution, but a few of them were designated as a group of alleles with the 2 digits resolution. In this project, we did not separate with a colon the first and the second set of digits. Therefore, we would have indicated the HLA allele in figure 3.1 as HLA-DQA*0201. As the prefix is always "HLA", we considered it avoidable and hence, we also referred here for instance to this allele as DQA*0201.

As mentioned before, the three loci DRB3, DRB4 and DRB5 were studied as a single locus HLA-DRB345. Therefore, when designating the polymorphisms of this region, we indicated the specific locus 3, 4 or 5 to which the allele belonged ahead of the 4 digits identifying the specific allelic variant. For instance, HLA-DRB345*30101 from locus DRB3. In this HLA-DRB345 region there were two exceptions. Firstly, there was the allele HLA-DRB345*40101G. G indicates a group of alleles with the same nucleotide sequence in the peptide binding domain but with synonymous changes or differences in untranslated regions (UTR) and introns (12). Secondly, some individuals did not express genes DRB3, DRB4 and DRB5 in one or both copies of the genome, which was indicated as HLA-DRB345*AgBlank.

3.3. Statistical Analysis

The frequency of the different HLA-II alleles in the cohort was calculated in two different manners. On the one hand, the cohort frequency was calculated as the number of individuals bearing a certain allele. On the other hand, the allele frequency as the number of times that an allele was present considering the two copies of each locus in each individual. This second approach results in more or less half the frequency when compared to the frequency of individuals expressing a specific gene, but reflects situations of homozygosity as well. In addition, for each patient, we calculated the cumulative frequency of their HLA haplotype as the sum of the cohort frequency for each of the alleles expressed.

In order to assess linkage disequilibrium (18, 19) between alleles in different loci, we used chi squared test with 1 degree of freedom (χ_1^2). To assess the strength of LD, the measure D' was preferred (Equation 3.1.).

[Equation 3.1]

$$D' = \frac{D}{Dmax}$$

Where:

$$D = p(AB) - p(A) * p(B)$$

If $D > 0$:

$$Dmax = \min\left(p(A)p(b), p(a)p(B)\right)$$

If $D < 0$:

$$Dmax = \min\left(-p(A)p(B), -p(a)p(b)\right)$$

A Fisher's Exact Test was used to assess associations between HLA-II alleles and HIV infection risk, as well as to determine associations between HLA-II alleles and T cell responses to HIV epitopes. False discovery rate (FDR) was used to control multiple comparisons.

To explore the associations of HLA-II alleles and the quantitative variables viral load and $CD4^+$ counts, we first tested normality with a Shapiro test. As none of the variables fulfilled normality assumption, a Mann-Whitney test was used to compare the median of viral load or $CD4^+$ counts among the individuals bearing or not a specific allele. To analyse differences

on several non-normally distributed continuous variables (viral load, $CD4^+$ counts and cumulative frequency) between more than 2 groups, the Kruskal-Wallis test was applied.

Additionally, supervised random forest (RF) methodology was used to construct a regression model for each of the HIV control predictors, viral load and $CD4^+$ counts. HLA-II alleles were used as the covariates of the model. Their importance in the model to predict changes in viral load or $CD4^+$ counts was measured with the mean decrease in node impurity. The more important variables showed higher values of this measure. In parallel, RFs were also used to construct classification models about the T cell responses to different OLPs. A classification model was constructed for each OLP. As before, the HLA-II alleles were used as explanatory variables. In this case, the importance of each variable was measured with the mean decrease Gini (MDG) and again, the higher the MDG the more important the variable was (19-22). We only computed MDG on the classification models with out-of-bag (OOB) estimates with an error rate smaller than 20%.

For the determination of relevant variables for the construction of the regression or classification models with RFs, we used 10-fold cross-validation procedure. Each time an increasing number of variables were selected based on their relative importance, and when the selected variables returned the minimum error of RF predictor, we were able to determine the most influential variables in the model.

All the statistical analyses were performed using the R statistical software (23) and the R studio platform (24). The random forests methodology was addressed with the R package *randomForest* (25, 26). The different plots were made with packages *graphics* (27)*, ggplot2* (28), and to represent linkage disequilibrium, package *corrplot* (29).

Additionally, an R package named *AnalysisHLA* was constructed (Appendix II) to store all the functions used in the present study to carry out the different analyses. For the construction of this package, we used the R packages *devtools* (30) and *roxygen2* (31).

3.3.1. Additional web-based tools

In the study of associations between HLA-II alleles and T cell responses to different epitopes, we used the specific R function we made with this purpose (*Hepitope_f* function stored in *AnalysisHLA* package). The design of this function was inspired by the Hepitope tool from the HIV molecular immunology database (32, http://www.hiv.lanl.gov/content/immunology/hepitopes/), and improved it in two ways:

Adding information of LD and retrieving the results in a more user-friendly way. Moreover, the use of the R software allows a major flexibility on the analysis. However, the web-based tool was used to corroborate the results.

Finally, the tools *QuickAlign* (33) and *Entropy* (34) from the HIV sequence database (http://www.hiv.lanl.gov/content/sequence/HIV/HIVTools.html) were used to determine how conserved an OLP sequence was. *QuickAlign* aligns a given OLP sequence against all the similar HIV sequences in the HIV database. Then, this alignment serves as the input for the *Entropy* tool that returns the Shannon entropy (S) value of each position in the alignment.

4. RESULTS

4.1 Association of HIV infection risk and specific HLA-II alleles

4.1.1 Description of HLA-II allele frequency

In the high-risk Peruvian cohort analysed in the present study, there were 12 DQA, 17 DQB, 43 DRB1 and 12 DRB345 HLA-II alleles. The frequency of these alleles in the cohort, understood as the percentage of individuals bearing each allele, is showed in Figure 4.1. The main observation is that less polymorphic loci showed a heterogeneous distribution of the alleles, the majority of the individuals in the cohort expressed the same allele, therefore there was one very frequent allele in these loci while the other ones were more uniformly distributed. Regarding the less polymorphic locus DQA, 60% of individuals expressed HLA-DQA*0301, the most common allele in this locus and 40% of them, the second more frequent allele HLA-DQA*0501. In locus DQB, more polymorphic than DQA, the more frequent allele HLA-DQB*0302 was present in 40% of the individuals in the cohort. The DRB345 region is also characterized by a low polymorphism, and, as expected, 60% of the subjects in the cohort carried the most common allele HLA-DRB345*40101G. Finally, in the most polymorphic locus DRB1, the most common allele HLA-DRB1*0901 was carried by less than the 25% of the individuals.

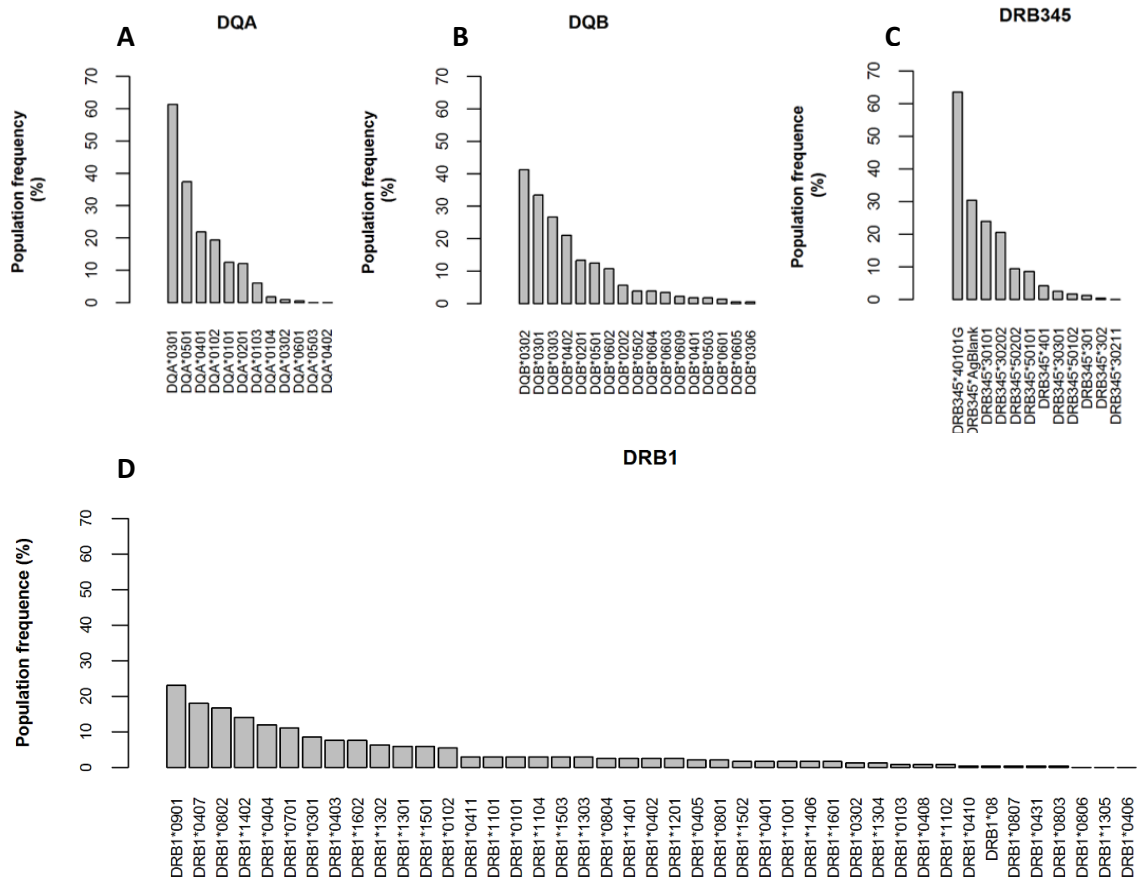


Figure 4.1 Frequency plots. Plots A, B, C and D show the percentage of individuals expressing each of the different alleles in the locus or regions DQA, DQB, DRB345 and DRB1 respectively.

The allele frequencies of the HLA-II alleles in the Peruvian cohort (Appendix I) were also calculated. These were compared with the allele frequencies in other Peruvian populations described in a limited set of studies in the Allele Frequency Net Database (35, http://www.allelefrequencies.net). Regarding locus DQA, in both the database and the cohort, the highest allele frequency was for HLA-DQA*03 followed by HLA-DQA*0501 and HLA-DQA*0401. In the case of the DQB gene, the three most frequent alleles described in the database were HLA-DQB*0301, HLA-DQB*0402 and HLA-DQB*0302. In the cohort, these alleles were ordered slightly different according the allele frequency: HLA-DQB*0302, HLA-DQB*0301, HLA-DQB*0303 and HLA-DQB*0402. The main difference regarding DQB locus was on the allele frequency of HLA-DQB*0303, which was reported as one of the less frequent alleles in the allele frequency database. In the DRB1 locus, the highest frequency in the cohort was found for the HLA-DRB1*0901 allele. Although this allele was in a similar frequency according the Peruvian populations described in the database (36, 37, 38), the more frequent alleles there were HLA-DRB1*0802 and HLA-DRB1*0403, which showed low

allelic frequencies in the studied cohort and which may be the consequence of demographic and geographic differences between the cohorts included in this study and the ones in the Allele Frequency Database. There was no information available about loci DRB3, DRB4 and DRB5 in the database.

Finally, based on the frequencies of the alleles in the cohort, we measured linkage disequilibrium (LD) between alleles in different loci. In Figure 4.2, the LD is shown between alleles in different loci measured with D' parameter (see section 3). There was a high degree of non-random associations between the HLA-II alleles in different loci. The fact that there was a high number of alleles always expressed together (D' = 1), implied the absence of other haplotypes (D' = -1), increasing the number of alleles in LD. In particular, the most frequent allele HLA-DQA*0301 was in linkage disequilibrium with all the most frequent alleles in the other loci (D' is near +1), indicating the predominance of the haplotypes DQA*0301-DQB*0302 and DQA*0301-DRB1*0901 and DQA*0301-DRB345*40101G.

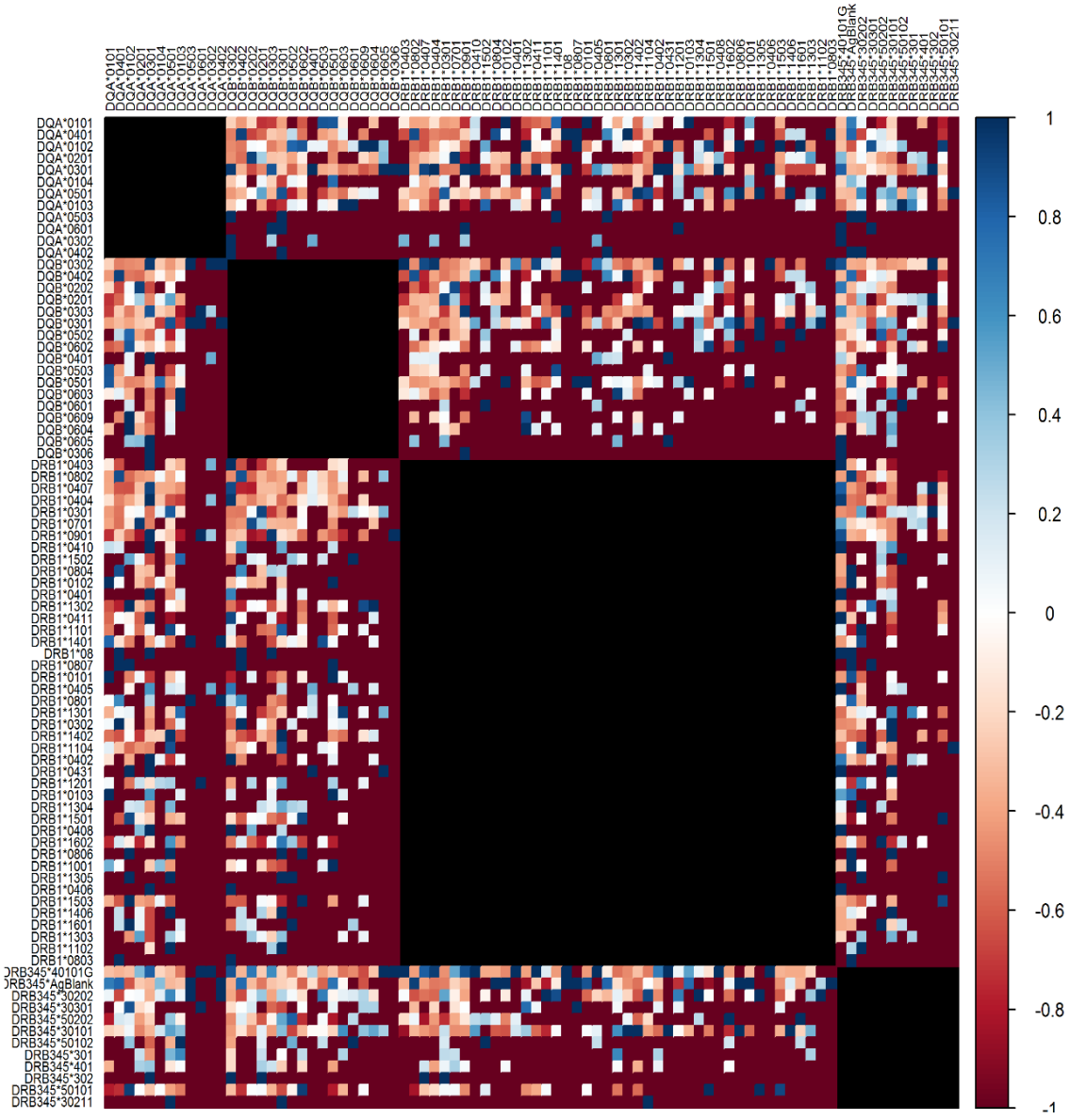


Figure 4.2 Linkage Disequilibrium between HLA-II alleles based on D' measure. Correlation-like plot based on D' measure. The different HLA-II alleles are indicated on the top and the left of the plot. On the right, there is the scale indicating the D' measure of linkage disequilibrium. Black boxes mark alleles in the same locus for which no LD was calculated.

4.1.2 Association of HIV infection and HLA-II alleles

To analyse associations of HLA-II alleles with HIV infection status, either referred to as seronegative or seropositive, we used a Fisher's Exact Test. The two alleles HLA-DRB345*401 (p = 0.035) and HLA-DQB*0202 (p = 0.048) showed significant associations with an increased and a lower risk of being seropositive, respectively. At first glance, in table 4.1, the odds ratio (OR) indicated that individuals highly exposed to HIV carrying the allele HLA-DRB345*401, had 7 times more risk to get infected than individuals without this allele.

On the contrary, allele HLA-DQB*0202 seemed to be a protective allele, as individuals bearing it showed a 2 fold reduced risk to be HIV infected than subjects without the allele. However, looking at the 95% confidence interval (CI) for these two alleles, it was not possible to identify any association of specific HLA-II alleles and the risk for HIV infection due to the large intervals containing the null value.

Table 4.1 Odds Ratio (OR) for the alleles with significant p-values from the Fisher's Exact Test in HIV infection. Each of the tables makes reference to an HLA-II allele: DBR345*401 and DQB*0202. The 0 or 1 means individuals without the allele and with it respectively. Then, it is indicated the Odds Ratio (OR) and the 95% confidence interval (CI).

DRB345*401	OR	CI (95%)
0	1.000	-
1	6.918	(0.986 – 300.328)

DQB*0202	OR	CI (95%)
0	2.149	(0.936 – 5.0233)
1	1.000	-

4.2 Association between HIV control and HLA-II genotype

The identification of associations between host genetics and markers of HIV control (viral load and $CD4^+$ counts), was the main goal of this study. Therefore, seropositive patients were stratified according to the presence or absence of each of the HLA-II alleles and then, their viral load and $CD4^+$ counts were compared using a Mann-Whitney test. The two HIV control markers are known to be inversely correlated (3) and the HLA-II associations with a p-value < 0.1 (arbitrary threshold to avoid the representation of all the HLA-II alleles in the cohort) for any of the two markers are presented in figure 4.3. Regarding viral load (Figure 4.3A), allele HLA-DRB1*1302 (p = 0.044) was associated with high viral load and so, with a lack of HIV control. On the contrary, allele HLA-DRB1*1201 (p = 0.015) was associated with low viral load and HIV control. Exploring the associations of $CD4^+$ counts with HLA-II alleles (Figure 4.3B), yielded 4 significant associations including: HLA-DRB1*0804 (p = 0.018), HLA-DRB1*0301 (p = 0.02), HLA-DRB345*30101 (p = 0.036) and HLA-DRB1*0401 (p = 0.038). All these alleles were associated with reduced counts of $CD4^+$ T cells, with the exception of HLA-DRB1*0804 which was related to higher $CD4^+$ counts. None of the associations remained significant after correcting for multiple comparisons with FDR (q-value > 0.2).

Previous studies observed (16, 39) that alleles expressed in a higher number of patients were associated with HIV progression compared to less frequent alleles. In Figure 4.3C, comparing the two alleles associated with differences in viral load, allele HLA-DRB1*1201, associated to low viral loads, was less frequent than HLA-DRB*1302 allele. For associations with $CD4^+$ T cell counts, the beneficial allele HLA-DRB1*0804 was less frequent than the two alleles associated with low $CD4^+$ counts, HLA-DRB345*30101 and HLA-DRB1*0301.

To analyse which HLA-II alleles (considering loci DQA, DQB, DRB1 and DRB345 together) better explained the levels of viral load and $CD4^+$ counts, we used the random forest methodology to obtain one regression model for each of the HIV control markers (dependent variables). Figure 4.4 shows the importance of the different covariates in the regression models ordered according to the importance measure mean decrease in node impurity (higher values means major importance). In both models a variable selection was made using 10-fold cross-validation. Regarding viral load, the HLA-II alleles selected as the more important predictors of viral load were HLA-DRB1*1001, HLA-DQA*0201, HLA-DRB345*40101G and HLA-DRB1*0407. In the case of $CD4^+$ T cell counts, HLA-II allele DQB*0302 appeared to be enough to predict changes on this parameter. However, these results were no reliable because the two models failed in explaining any variance on the dependent variable.

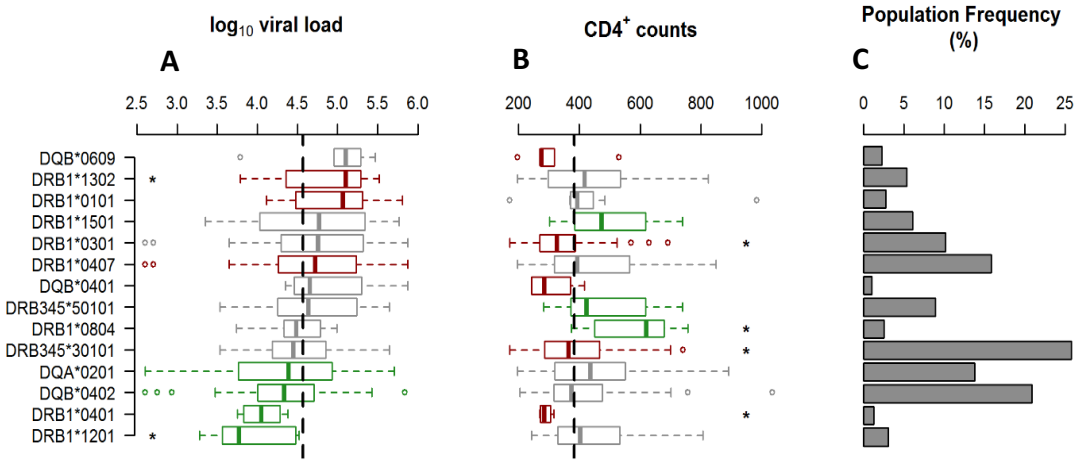
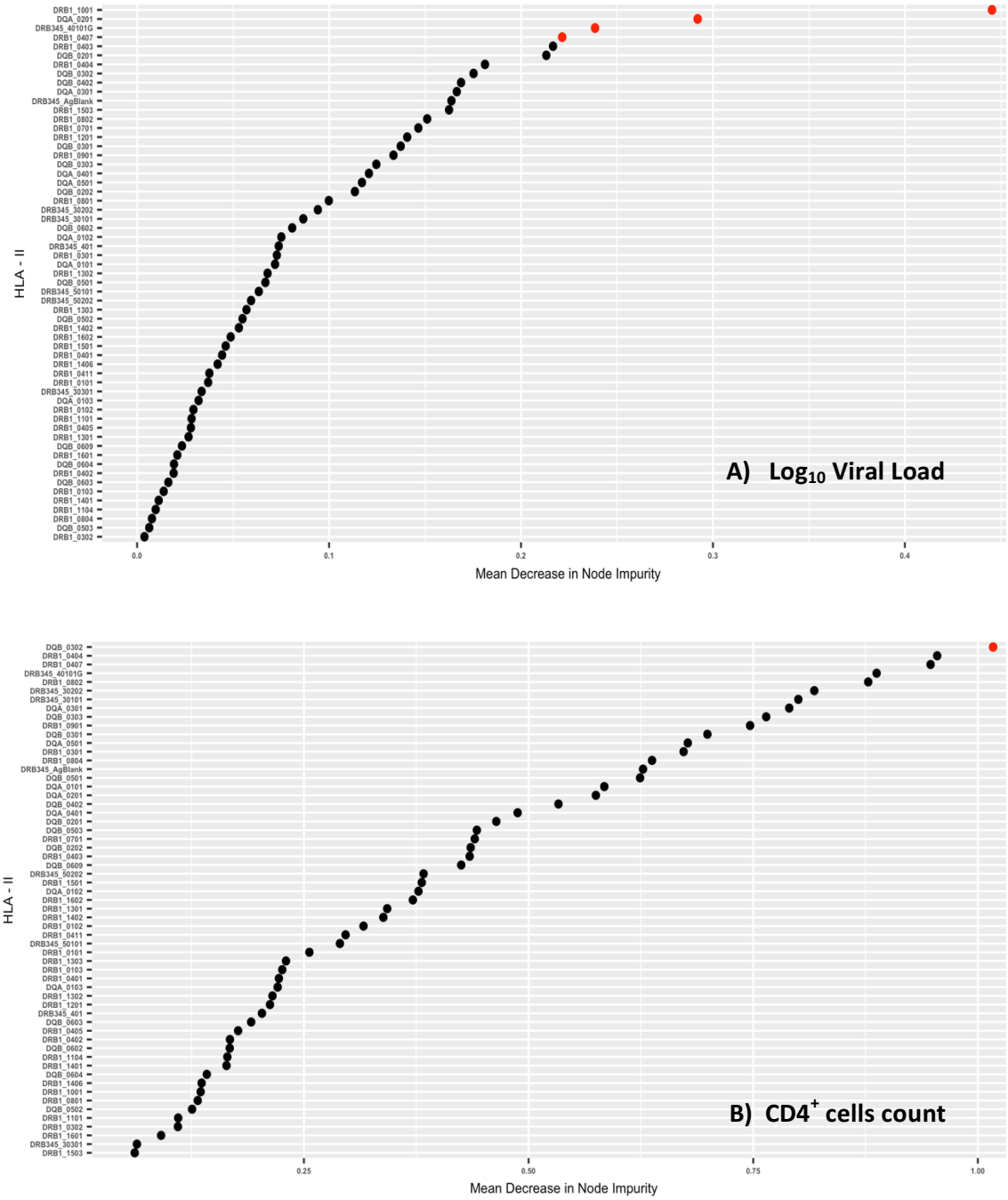


Figure 4.3 Associations of HLA-II molecules with viral load and $CD4^+$ counts. Alleles achieving a p-value < 0.1 for any of the two control markers are represented, the alleles associated (p<0.1) with one of the markers are highlighted in colour. Green means beneficial association and red, a deleterious association. The medians for log_{10} viral load and $CD4^+$ counts of the cohort are indicated by the by the black discontinuous line. Alleles with p-values < 0.05 are indicated with *. On the right part of the plot, there is the frequency of each of the selected alleles.



4.4 Importance of the different HLA-II alleles to explain viral loads and CD4$^+$ counts in the regression models obtained with the Random Forest methodology. In both plots the importance variable measured as Mean Decrease in Node Impurity can be found in the coordinates, and the different HLA-II alleles, the covariates, in the ordinates. Highlighted in red are the HLA-II alleles that were identified as the more important ones to explain the differences on median viral loads and CD4$^+$ T cell counts using cross validation procedure. Plot A shows the results considering viral load as dependent variable, while plot B shows the results related to CD4$^+$ counts.

4.3. Effect of heterozygosity and rare HLA-II alleles on HIV control

In this section we analysed whether seropositive patients with HLA-II alleles in heterozygosity or with low cumulative HLA-II cohort frequency, showed a control of HIV as suggested by earlier studies (16, 39).

To study the effect of heterozygosity, different groups were made: Individuals in the heterozygous group were those with all the alleles in heterozygosis for the HLA-II regions DQ, DRB1 and DRB345. The rest of the individuals, having at least one locus in homozygosis, were considered as homozygous. Additionally, for each locus, individuals were divided according the homozygosity or heterozygosity of one specific gene. The results of the heterozygous effect are shown in Figure 4.5. The medians of viral load and $CD4^+$ counts from the different groups were compared with a Kruskal-Wallis test, no significant results appeared in any case ($p > 0.05$). Therefore, we did not find evidence that HLA-II heterozygosity meant an advantage for the HIV infected individuals in the Peruvian cohort.

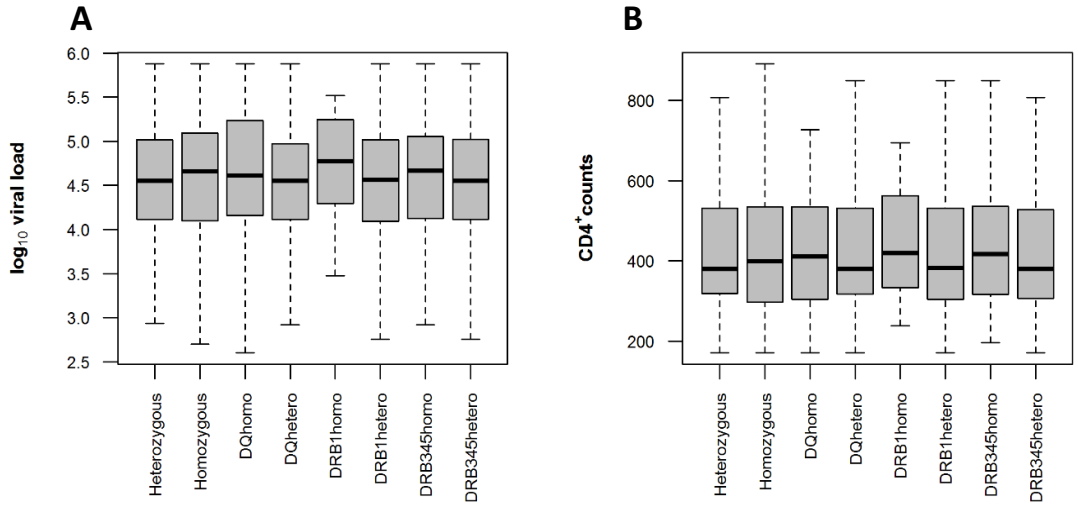


Figure 4.5 Comparison of viral loads and $CD4^+$ counts in HLA heterozygosity. Figure A is a boxplot with y-axis indicating the log_{10} viral load and the x-axis the different groups of patients made. Figure B shows similar results but in y-axis there are $CD4^+$ counts. The groups of patients are denominated as follows: DQhomo, DRB1homo and DRB345homo, that make reference to those individuals homozygous for the specific locus, and DQhetero, DRB1hetero and DRB345hetero, that refers to heterozygous individuals in these loci. Homozygous group contains these individuals with at least one locus in homozygosity, and the Heterozygous group, individuals totally heterozygous.

Next, we assessed whether rare alleles would provide any advantage on HIV control (Figure 4.6). To this end, individuals were divided in 4 groups according to the quartile of the levels of their viral load and $CD4^+$ T cell counts. Then, for each of the groups in each of DQ, DRB1 and DRB345 loci, we computed the cumulative frequency as the sum of the allele cohort frequency. It was expected that individuals with lower viral load or higher $CD4^+$ counts would show smaller cumulative frequencies. However, no significant differences in cumulative frequencies were detected with the Kruskal-Wallis test in any case ($p > 0.05$).

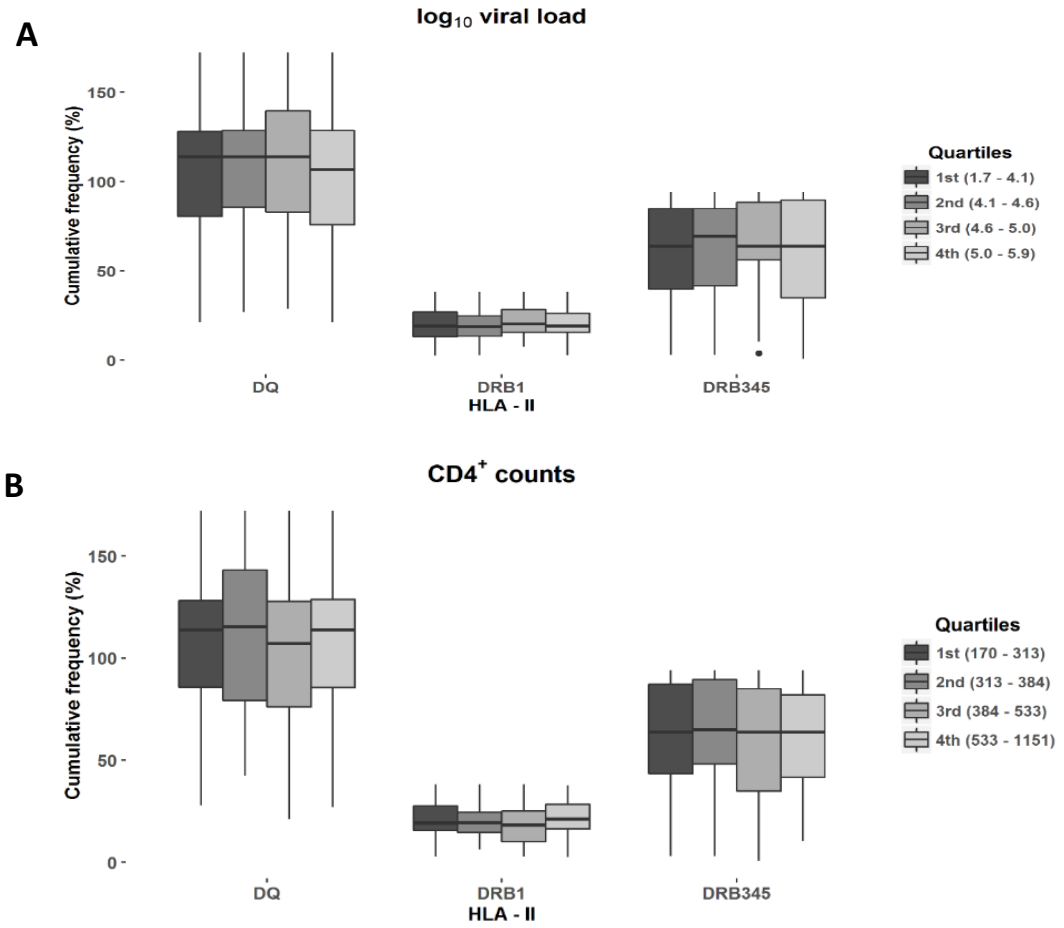


Figure 4.6 Comparison of median cumulative frequency among quartiles based on viral loads and $CD4^+$ cell counts. Figure A refers to viral load, and figure B, to $CD4^+$ counts. In both cases in the y-axis there is the cumulative frequency and in the x –axis the different HLA-II regions analysed: DQ, DRB1 and DRB345. For each of the regions, individuals are divided in 4 groups according the 4 quartiles of viral load and $CD4^+$ counts measures. There are no significant differences in the cumulative frequency from none of the regions neither for viral load nor for $CD4^+$ counts (Kruskal-Wallis test p-value > 0.05).

4.4. Association of T cell response and HLA-II

The T cell responses against HIV epitopes are the consequence of the HLA-I and HLA-II alleles molecules presenting different epitope peptides to virus-specific T cells. Here, we explored the associations between HLA-II alleles and T cell responses to overlapping peptides (OLPs) covering the whole HIV proteome. Reponses to the different OLPs were measured as IFNγ producing PBMCs in an ELISpot assay, bearing in mind that, part of the IFNγ producing cells are expected to be $CD8^+$ T cells and thus impacting the analyses of HLA-II restricted $CD4^+$ T cell responses.

A Fisher's Exact Test was used to determine independent associations between each of the HLA-II alleles in the cohort and the different OLPs to which the patients responded. Figure 4.7 maps the percentage of significant associations ($p < 0.05$) in each of the proteins or subproteins of the HIV genome. As might be expected, the two larger proteins and subproteins, reverse transcriptase and gp120 derived from the envelope protein precursor, were the ones with a major number of significantly associated OLPs mapped.

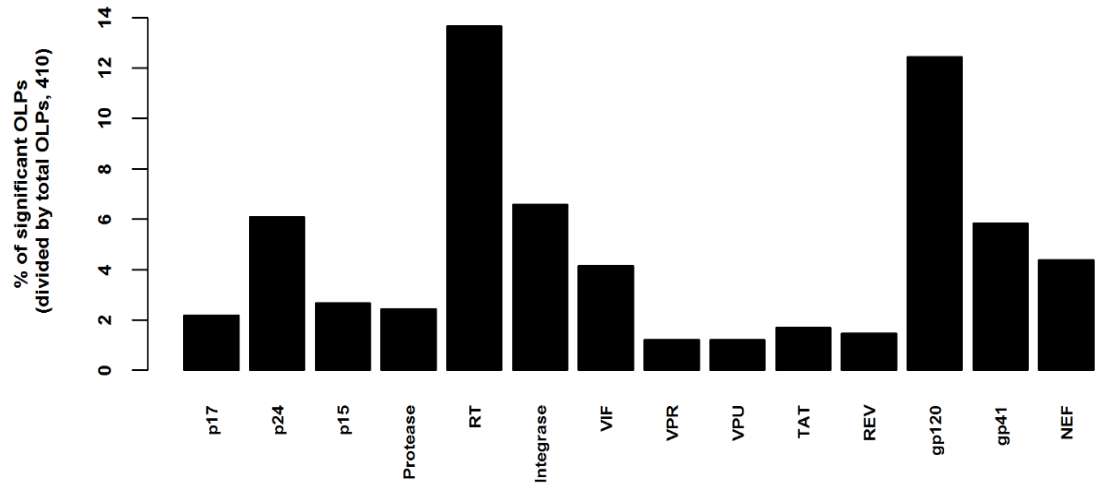


Figure 4.7 Map of OLP reactivity significantly associated with HLA-II alleles. Percentage of OLPs associated with at least one HLA-II allele ($p<0.05$) in each viral protein and subprotein of the HIV genome.

With these association results and the HLA-II genotype information of the Peruvian cohort, we determined the frequency of patients putatively responding to different OLPs. We only selected those OLPs showing a significant association ($p<0.05$) with at least one allele. These results are shown in figure 4.8. Only peptides against which at least 10% of the patients were predicted to respond are represented, lower frequencies were not considered to avoid

bias towards less reactive peptides. The highest frequency of patients was found on putative T cell responses against Gag, Pol and Nef proteins. The 20% of the HIV infected patients were predicted to react against OLP 41 in Gag protein and OLP 162, 190 and 207 in Pol protein. In addition, around 35 patients were predicted to respond to OLP 76 in Nef protein.

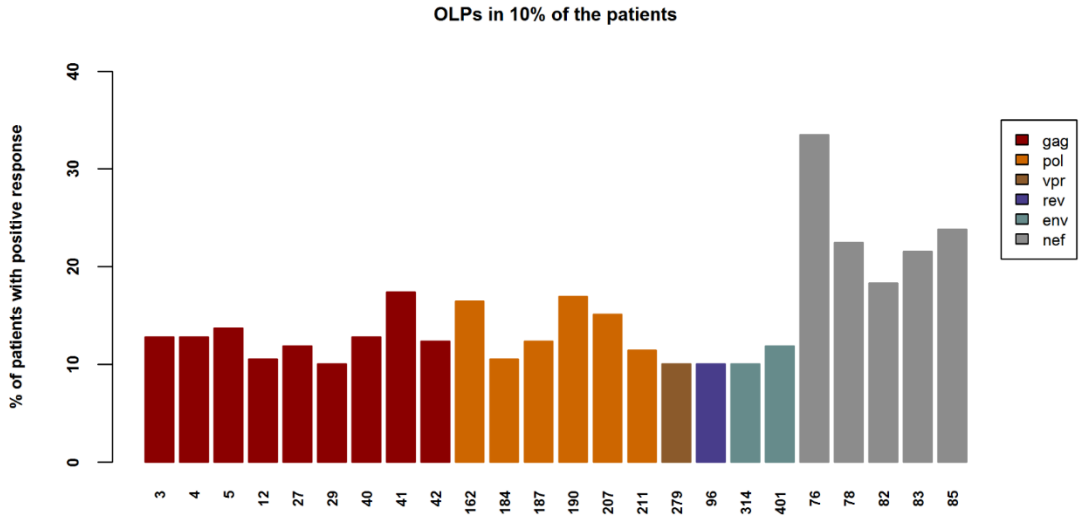


Figure 4.8 Percentage of patients with T cell responses to HIV Overlapping Peptides. We calculated the percentage of patients bearing HLA-II alleles significantly associated with the different OLPs. Only peptides predicted to show INFɣ responses in at least 10% of the patients are represented. The colours indicate the regions of the HIV genome.

4.4.1 Immunodominance on patients bearing HLA-DRB1*1201 or HLA-DRB1*1302 alleles

We finally studied the immunodominance patterns of patients expressing alleles HLA-DRB1*1201 or HLA-DRB1*1302 since they were related to lower and higher viral loads, respectively. Figure 4.9 shows the percentage of patients with HLA-DRB1*1201 or HLA-DRB1*1302 alleles responding to OLPs associated with T cell responses in at least 10% of the patients in the studied cohort. The immunodominance pattern in both cases is mainly mapped in the HIV genome regions Gag, Pol and Nef.

From the OLPs represented, only OLP 41 and OLP 82 appeared to be significantly associated with HLA-DRB1*1201 (p=0.038) and HLA-DRB1*1302 (p=0.036), respectively in the Fisher's Exact tests. These associations went in the same direction in the RF classification models

where the HLA-II alleles were ordered according to the MDG classification of the T cell responses to OLP 41 and OLP 82 (Figure 4.10). In the classification model for the T cell response against OLP 41, HLA-DRB1*1201 was selected as an important variable, and the same happened considering OLP 82 and HLA-DRB1*1302.

Comparing the immunodominance of these 2 groups of patients, the majority of HLA-DRB1*1201 patients exerted a T cell response against OLP 41, while the individuals expressing either one of these alleles showed comparable frequency of reactivity to OLP 82.

In order to further explore the contribution of the T cell response against these two OLPs on HIV progression, we computed the entropy of their peptide sequence using the Shannon Index measure. It has been reported (12) that more conserved sequences, showing a lower Shannon Index, are usually associated with the control of HIV disease progression and that peptides spanning more conserved regions, would more often match the patient's autologous sequence and thus more reliably detect potential responses than OLP covering more variable regions. In our results both sequences showed similar Shannon index values (data not shown). Therefore, we could not define the T cell responses against OLP 41 and OLP 82 as beneficial or prejudicial according the entropy of their sequences.

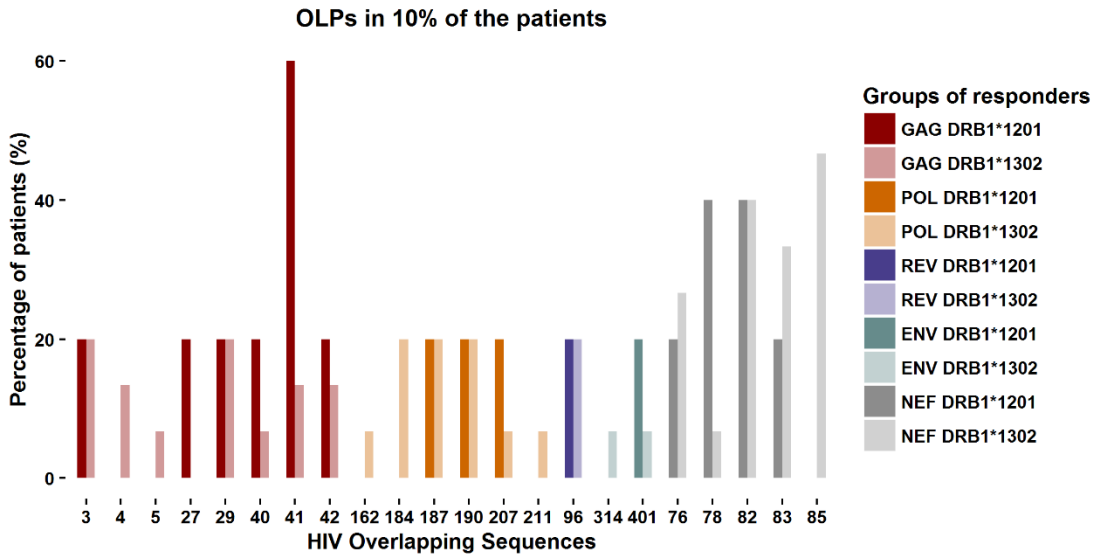
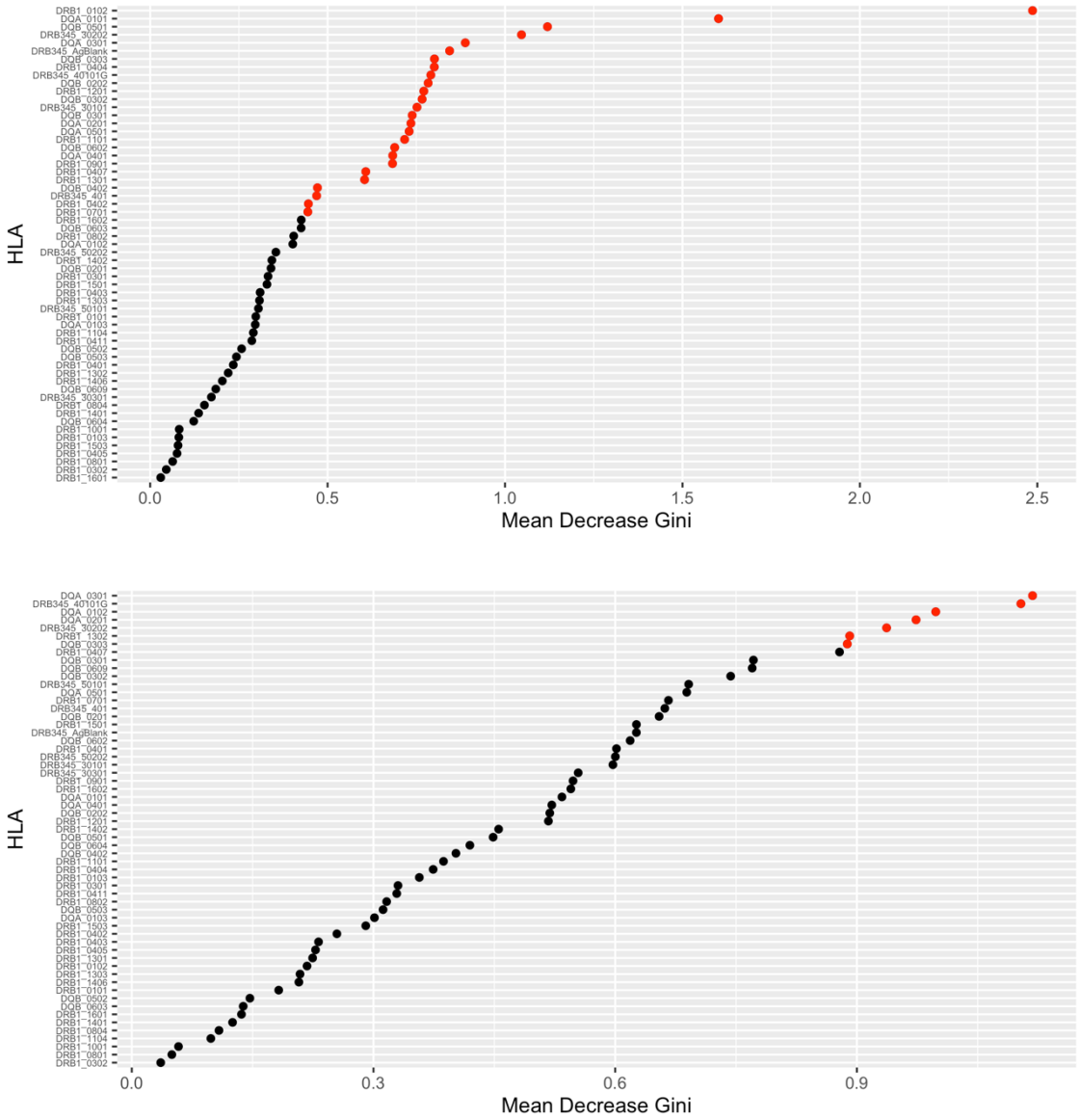


Figure 4.9 Immunodominance comparison between patients carrying HLA-DRB1*1201 or HLA-DRB1*1302. HLA-II molecules DRB1*1201 and DRB1*1302 have been found to be associated with differences in viral load. The OLPs associated ($p<0.05$) with at least one HLA-II allele and to which at least 10% of the HIV positive patients in the Peruvian cohort showed a response, are shown in the current plot (x-axis). The ordinates indicate the percentage of patients bearing DRB1*1201 or DRB1*1302 responding to each of the selected peptides.



4.10 Importance of the different HLA-II alleles to classify T cell responses to OLP 41 and OLP 82. In both plots the HLA-II alleles (y-axis) are ordered according the Mean Decrease Gini (y-axis). Plot A shows the classification of OLP 41 and B, the classification of OLP 82. The most important alleles to classify the responses to these OLPs are marked in red.

Finally, we searched if the OLPs against which at least a 10% of the patients responded in the Peruvian cohort had described T-helper epitopes in the HIV molecular database (40). The same database also contained the HLA-II alleles that had been described to restrict the T cell responses against these epitopes (Table 4.2). None of the mentioned OLPs were described to harbour an epitope presented by HLA-DRB1*1201. On the contrary, HLA-

DRB1*1302 was described to be associated with the T cell responses against OLP 4, 5 and 41 but no HLA-DRB1*1302 restricted epitope had been described for OLP 82.

Table 4.2. Epitopes described in the HIV molecular database that overlap with OLP recognized by a 10% of the HIV positive patients in the Peruvian cohort. In this table we show the peptide sequences of the OLPs to which the patients in the Peruvian cohort responded to and the HLA-II alleles that have been described to restrict T cell epitopes against epitopes contained in them. In red, the allele that was significantly associated with high viral load in our cohort (DRB1*1302) is marked.

EPITOPE	HLA	OLP	PEPTIDE
EKIRLRPGGKKKYKL		OLP 3	GAG
EKIRLRPGGKKKYKLKHI		OLP 3	GAG
RPGGKKKY?		OLP 3	GAG
GKKKYKLKHIVWASREL	DRB1**0101, DRB1**0701, DRB1**0801, DRB1**1101, **DRB1*1302**, DRB1*1303, DRB3**0101, DRB3**0301, DRB4**0101, DRB5**0101	OLP 4	GAG
KHIVWASRELERFAV	DRB1**0301, DRB1**1301, **DRB1*1302**, DRB1*1303, DRB3**0101, DRB5**0101	OLP 5	GAG
HIVWASRELER		OLP 5	GAG
HIVWASRELERFAVN?		OLP 5	GAG
AAEWDRLHPVHAGPIA	DRB1*0701	OLP 29	GAG
GPKEPFRDYVDRFYKTLR	DRB1*1301	OLP 40	GAG
YVDRFYKTLRAEQASQEV	DRB1**0101, DRB1**0401, DRB1**0405, DRB1**0701, DRB1**0801, DRB1**1001, DRB1**1101, DRB1**1301, **DRB1*1302**, DRB1*1303, DRB1**1501, DRB1**1502, DRB4**0101, DRB5**0101	OLP 41	GAG
RLHPVHAGPIA		OLP 29	GAG
GPKEPFRDYVDRFYK		OLP 40	GAG
PKEPFRDYV	DQ5	OLP 40	GAG
DRFYKTLRAEQASQ	DRB1*0401	OLP 41	GAG
RFYKTLRAEQAS	DRB1**0101, DRB1**0401, DRB1**0405, DRB1**0701, DRB1**1101, DRB11*501, DRB5*0101	OLP 41	GAG
FYKTLRAEQASQ	DRB1**0101, DRB1**0401, DRB1**1101, DRB5**0101	OLP 41	GAG
FYKTLRAEQASQE	DRB1**0101, DRB1**0401, DRB1**0405, DRB1**1101, DRB1**1501, DRB5**0101	OLP 41	GAG
YKTLRAEQA	DRB1*0101	OLP 41	GAG
LRAEQASQEVKNWMTETL		OLP 42	GAG
LAENREILKEPVHGV		OLP 207	POL
KTVRLIKFLYQSNPPPS		OLP 96	REV
VGFPVRPQ	DR1, DRw15(2)	OLP 76	NEF
YKAAVDLSHFLKEKGGL	DRB1**0701, DRB1**0804	OLP 78	NEF
LWVYHTQGYFPDWQNY	DRB1**0701, DRB1**1301, DRB1**1401, DRB3**0101, DRB4*0101	OLP 82	NEF

5. DISCUSSION

Several host genetic factors have been associated with different degrees of HIV infection control. Among them HLA-I, and mainly polymorphisms in the HLA-B locus, have been related with differences in HIV acquisition and control of HIV associated disease (16, 48, 49). However, the Major Histocompatibility Complex includes other classical (HLA-II) and non-classical (HLA-E, G, F) genes also involved in the immune response. Of them, HLA-II is one of the most studied and it is known to play a key role in the development of $CD4^+$ T cell and humoral immune responses. However, in association to HIV infection, it has not been studied to the same level as HLA-I polymorphisms. HLA molecules of class II are structurally similar to HLA-I ones, but their associations with HIV control may be subtler for several reasons: Firstly, because its function is related to the stimulation of the immune response instead of the direct killing of infected cells and secondly, because these molecules are formed by two subunits encoded in different loci (2, 9). Here, we took advantage of the access to a HLA-II typed cohort including almost 400 HIV-exposed and HIV infected individuals, to search for associations between HLA-II alleles and HIV infection control.

To our knowledge, this is the first study of HLA-II associations with HIV performed in Peru, a fairly understudied population although it has been part of several HIV vaccine trials in the past (41, 42). We contribute with 392 individuals to a better knowledge of the frequency distribution of the HLA-II alleles in the Peruvian population and, as far as the authors know, this is the first study that describes the frequency of genes DRB3, 4 and 5 in this population. Comparing the allele frequencies of the present cohort with the information of the three Peruvian populations described in the Allele Frequency Net Database (36, 37, 38), we observed similar allelic frequency distribution. After the linkage disequilibrium analysis, we could infer three predominant haplotypes DQA*0301-DQB*0302, DQA*0301-DRB1*0901, DQA*0301-DRB345*40101G.

In the present study, we did not find statistically robust associations between HLA-II alleles and risk for HIV infection. We did however identify several HLA-II alleles that were related to differences in the HIV control markers viral load and $CD4^+$ counts, but none of the associations remained significant after the correction for multiple comparisons by FDR. The HLA-II alleles were never significantly associated with both HIV control markers at the same time, which could be consequence of the fact that viral load is a reflection of viral set point

and not necessarily reflective of disease progression without knowing the time since infection, whereas $CD4^+$ T cell counts reflect more the clinical advancement of HIV disease. This is also reflected in the relatively weak correlation between viral load and $CD4^+$ counts in the present cohort ($r^2 = 0.114$). Hence, and in the absence of information on "time since infection" (43, 44), we focused on the HLA-II associations with viral load. The two main associations were with the HLA-II alleles DRB1*1201 and DRB1*1302. The former was significantly associated with low viral load, while HLA-DRB1*1302 was related with high viremia.

Other studies also found statistical associations with HLA-DRB1 alleles. Julg et al (45) identified HLA-DRB1*1303 to be significantly associated with low viral load in chronic HIV-1 clade C and B infected individuals. They transformed viral load for normality and applied a generalised linear model to compare the viral load mean of patients with or without each of the HLA alleles. However, as in the present study, the associations found by Julg et al did not reach significance after the correction for multiple comparisons. Similarly, Ranashinge et al (46) identified HLA-DRB1*1501 to be associated with low viremia and DRB1*0301, with high viremia. In this study the HLA-II associations with viral load maintained significance after Bonferroni correction for multiple comparisons, which is a more restrictive approach to deal with multiple comparisons issues than the false discovery rate. However, their finding may also be a consequence of the different statistical approach they used, as they divided the seropositive individuals in a group of controllers (VL < 2,000 copies/ml) and a group of progressors (VL > 10,000 copies/ml) and applied a logistic regression for each allele where the dependent variable viral load was converted into a binary variable: higher or lower to the viral load mean. To our knowledge, it is more robust to compare the median of continuous variables with Mann-Whitney test, the one applied in our analysis. Additionally, in the cohort division there were some individuals in the middle of the established cut-offs that were not considered, and which increased the differences between the two extreme groups.

None of the associations found in these earlier studies were reproduced herein, which is not surprising, considering they focused on cohorts of different ancestry and, consequently, with a different HLA-II distribution. However, some of their other findings were in agreement with our results; the HLA-II associations with viral load were overall weak and the majority of them were related with locus DRB1. Considering that the DRB1 gene is the

most polymorphic one of the HLA-II genes, this characteristic could be somehow related with the associations found. Actually, this hypothesis is reinforced by the fact that in different studies of HLA-I associations (16, 47, 48, 49), the most polymorphic gene HLA-B was also the most commonly related with changes in viral load. Both genes, HLA-DRB1 and HLA-B, are highly polymorphic and do not show any dominant allele, possibly reflecting ongoing host evolution after pathogen threat, further suggesting that the products of these genes are critically involved in the host defense. Therefore, the HIV antigens restricted by these low frequent alleles are less probable to suffer population-wide immune selection pressure and mutate. However, on the contrary of what Olvera et al (16) found regarding HLA-I alleles, no advantage could be attributed to HLA-II rare alleles. Additionally, we could not establish HLA-II heterozygosity as being advantageous for seropositive subjects.

In the present project, we also tried a random forest strategy to model the most relevant HLA-II alleles to predict differences in viral load or $CD4^+$ counts. However, the obtained regression models were inconclusive, no variance could be explained with them. Such an outcome might be consequence of the high linkage disequilibrium between the alleles, and although random forests can cope with a certain degree of collinearity, in this case it may have been too high and the importance of each of the covariates was diminished (50). Consequently, the associations with the independent variables were weakened. This fact, together with the indirect effect of HLA-II alleles on the immune response against HIV and the sparsity of the data, weakens such associations even more. Therefore, for the present analysis, the Mann-Whitney tests retrieved more reliable results. Nonetheless, larger sample sizes would clearly allow the determination of more robust associations.

It is thought the majority of univariate associations between HLA-II alleles and HIV progression are the consequence of the antigen they present to $CD4^+$ T cells. For these reason, we explored univariate associations between HLA-II alleles and the T cell response to the overlapping peptides (OLPs) covering the whole HIV proteome. As the majority of the associations were not significant after FDR, we selected those with a p-value < 0.05. Based on such selection, we calculated the percentage of patients that should respond to a given OLP according to their HLA genotype. We found that the majority of them were predicted to respond against peptides in proteins Gag, Pol and Nef. In other T cell response studies (51, 52, 53), these 3 proteins also appeared as the ones concentring the greatest number of patients responding against their peptides. In fact, Pol and Env are the largest proteins in

the HIV genome, and Gag and Nef are highly expressed during HIV infection (54, 55), consequently, the chances of the peptides derived from these proteins to be presented by different HLA alleles are increased, and a major number of patients were indeed able to respond against them.

Finally, we carried an immunodominance analysis of HLA-II restricted T cell responses (56). Immunodominance is defined as the percentage of patients with a concrete HLA allele that recognise a certain epitope or OLP (57, 58). As viral load is a better predictor of HIV control in comparison to $CD4^+$ T cells counts, we determined the T cell responses that could be driving the associations of the alleles HLA-DRB1*1201 and HLA-DRB1*1302 with low and high viral load. Comparing the T cell responses of the patients expressing each of the alleles, some differences on immunodominance were observed. However, not all the OLPs to which the patients responded were statistically associated (Fisher's Exact Test) with HLA-DRB1*1201 or HLA-DRB1*1302, therefore, such differences must be consequence of responses restricted by other alleles. Still, OLP 41 from Gag and OLP 82 from Nef, were significantly associated ($p<0.05$) with the presence of HLA-DRB1*1201 and HLA-DRB1*1302, respectively. Therefore, it was suspected OLP 41 could be a beneficial peptide and OLP 82, a deleterious one. In line with these associations, Gag responses have normally been related with beneficial HIV infection outcomes and Nef, to prejudicial ones (12, 51, 59, 60). Comparing the immunodominance of these two OLPs, the majority (60%) of the HLA-DRB1*1201 patients responded against OLP 41, establishing this as a dominant target of the HLA-DRB1*1201 restricted response to HIV and potentially providing some benefit for HIV control. There were no immunodominance differences in relation to OLP 82.

The inclusion of $CD4^+$ T cell epitopes in a vaccine against HIV would be highly desirable because it would provide the T-helper cells with the signal needed to trigger a strong cellular and humoral immune response. Therefore, the best epitopes to be included in a vaccine would be targets associated with HLA-II alleles that were, in turn, associated with lower viral loads, like HLA-DRB1*1201. Unfortunately, with our results and the cohort size at hand, we were not able to draw robust conclusions, it would be necessary a larger sample size and the preparation of an ELISpot assay optimized for the detection of $CD4^+$ T cells response. This could be achieved by the removal of $CD8^+$ cells from PBMCs and the increase of the incubation time to 40h (52). Despite these limitations, and the subtleness of the associations

between HLA-II and HIV progression, our results allow some glimpse on the potential effect that HLA-DRB1 alleles could have on HIV control.

6. CONCLUSIONS

1. In the present study no associations were found between HLA-II alleles and risk for HIV infection.
2. The present study revealed new associations of HLA-II alleles with HIV control markers. Regarding viral load, alleles HLA-DRB1*1201 and HLA-DRB1*1302 were associated with lower and higher HIV viral loads, respectively. In addition, HLA-DRB1*0301, HLA-DRB1*0401 and HLA-DRB345*30101 were associated with low $CD4^+$ cell counts, and HLA-DRB1*0804, with high $CD4^+$ cell counts. None of these associations were strong enough to maintain the significance after the adjustment of multiple corrections using a FDR cut-off of 0.2.
3. In this Peruvian high-risk cohort, neither the homozygosity of HLA-II alleles nor the cohort frequency of alleles were associated with low viral load or high $CD4^+$ cells counts.
4. Allele HLA-DRB1*1201 was significantly associated (Fisher's Exact Test $p<0.05$) with T cell responses to OLP 41 in the Gag protein and may identify a beneficial target of the HLA-II restricted $CD4^+$ T cell response to HIV.

REFERENCES

1. World Health Organisation (WHO). *Facts Sheets - HIV/AIDS*. November 2015 [Visited: 5th June 2016]. Available at: < http://www.who.int/mediacentre/factsheets/fs360/en/ >
2. Janeway CA Jr, Travers P, Walport M, et al. *Immunobiology: The Immune System in Health and Disease*, 2001. Garland Science (NY).
3. Rober R. Rich, Thomas A. Fleisher, et al. *Clinical Immunology. Principles and Practice,* 2012. Sunders (W.B)
4. Coffin, J., Hughes, S. and Varmus, H. *Retroviruses,* 1997. Cold Spring (NY): Harbor Laboratory Press.
5. B.Mothe, J.Ibarrondo, A.Llano and C.Brander. Virological, immune and host genetics markers in HIV infection. *Disease Markers*. 27 (2009): 105-120.
6. W.C. Greene, B.M. Peterlin. Charting HIV's remarkable voyage through the cell: Basic science as a passport to future therapy*. Nature Medicine.* 8 (2002): 673 - 680.
7. *Steven G.E. Marsh, Peter Parham and Linda D. Barber. The HLA Facts Book, 1999. FactsBook series – Academic Press.*
8. Hansen SG, Whu HL et al. Broadly targeted $CD8^+$ T cell responses restricted by major histocompatibility complex E. *Science*. 351 (2016): 714 - 720.
9. Abul K. Abbas & Andrew H. Lichtman. *Basic Immunology. Functions and Disorders of the Immune System.* 2011. Saunders – Elseiver.
10. A. Llano, A. Williams, A. Olvera, S- Silva Arrieta and C. Brander. Best-Characterized HIV-1 CTL Epitopes: The 2013 Update. *HIV molecular immunology*. 2013. Published by Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM. LA-UR 13-27758.
11. H. Streeck, N. Frahm, B.D. Walker. The role of IFN-ɣ Elispot assay in HIV vaccine research. *Nature Protocols*. 4 (2009): 461 – 469.
12. Beatriz Mothe, Anuska Llano, et al. Definition of the viral targets of protective HIV-1-specific T cell responses. *Journal of Translational Medicine*. 9 (2011): 208 – 228.
13. F.Porichis & D.E.Kaufmann. HIV-specific CD4 T cells and immune control of viral replication. *Current Opinion HIV AIDS*. 6 (2011): 174-180].
14. Eva Van Braeckel and Geert Leroux-Roels. HIV vaccines. Can $CD4^+$ T cells be of help? *Human Vaccines & Immunotherapeutics*. 8 (2012): 1795 – 1798.

15. Stephen A. Migueles and Mark Connors. The Role of $CD4^+$ and $CD8^+$ T Cells in Controlling HIV Infection. *Current Infectious Disease Reports*. 4 (2002): 461 – 467.
16. A. Olvera, S. Pérez-Álvarez et al. The HLA-C*04:01/KIR2DS4 gene combination and human leukocyte antigen alleles with high population frequency drive rate of HIV disease progression. *AIDS*. 29 (2015): 507 – 517.
17. IPD-IMGT HLA Database. *HLA Nomenclature*. April 2016 [Visited: 7th June 2016]. Available at: < http://hla.alleles.org/nomenclature/naming.html >
18. M. Slatking. Linkage disequilibrium —understanding the evolutionary past and mapping the medical future. *Nature Reviews – Genetics*. 9 (2008): 477-485.
19. R. Andrea S. Foulkes. Applied Statistical Genetics with R. For population-based genetic studies, 2009. Springer.
20. James, G., Witten, D., Hastie, T., Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*, 2013. Springer.
21. A.L Boulesteix, S. Janitza et al. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. WIREs Data Mining Knowledge Discovery. 2 (2012): 493 – 507.
22. L. Breiman. Random Forests. Machine Learning. 45 (2001): 5 – 32.
23. The R Foundation. *The R project for Statistical Computing*. [Visited: 29th July 2016]. Available at: < https://www.r-project.org >
24. R Studio. [Visited: 29th July 2016]. Available at: < https://www.rstudio.com/home/ >
25. Department of Statistics University of California, Berkeley. *Random Forest*. Leo Breiman and Adele Cutler. [Visited 7th July]. Available at: < https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm >
26. L.Breiman, A.Cutler, A.Liaw, M. Wiener. randomForest: Breiman and Cutler's Random Forests for Classification and Regression. *CRAN repository*. 2015.
27. Department of Statistics University of Auckland – New Zeland. *R Graphics Second Edition by Paul Murrell.* [Visited: 29th July 2016]. Available at: < https://www.stat.auckland.ac.nz/~paul/RG2e/index.html >
28. *ggplot2*. Hadley Wickham. [Visited: 29th July 2016]. Available at: < http://ggplot2.org >
29. CRAN: The Comprehensive R Archive Network. *Package corrplot: Visualization of a Correlation Matrix*. Taiyun Wei and Viliam Simko. [Visited: 29th July 2016]. Available at: < https://cran.r-project.org/web/packages/corrplot/index.html >

30. CRAN: The Comprehensive R Archive Network. *Package devtools: Tools to Make Developing R Packages Easier.* Hadley Wickham and Winston Chang [Visited: 29th July 2016]. Available at: < https://cran.r-project.org/web/packages/devtools/index.html >
31. CRAN: The Comprehensive R Archive Network. *roxygen2: In-Source Documentation for R*. Hadley Wickham, Peter Danenberg and Manuel Eugster [Visited: 29th July 2016]. Available at: < https://cran.r-project.org/web/packages/roxygen2/index.html >
32. HIV molecular immunology database. *Tools, Hepitope: HLA-Enriched Epitope.* [Visited: 5th June 2016]. Available at: < http://www.hiv.lanl.gov/content/immunology/hepitopes/ >
33. HIV Sequence database. *Tools, QuickAlign*. [Visited: 8th August 2016]. Available at: < http://www.hiv.lanl.gov/content/sequence/QUICK_ALIGNv2/QuickAlign.html >
34. HIV Sequence database. *Tools, Entropy*. [Visited: 8th August 2016]. Available at: < http://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy_one.html >
35. Allele Frequency Net database. *Allele Frequency Search*. [Visited: 9th August 2016]. Available at: < http://www.allelefrequencies.net >
36. Pablo R, Beraun Y, et al. HLA class I and class II allele distribution in the Peruvian population. *Tissue Antigens*. 56 (2000): 507 – 514.
37. Allele Frequency Net database. Publication details: The author of the Peru Lamas City allele distribution is A.Arntaiz-Villena. [Visited: 9th August 2016]. Available at: < http://www.allelefrequencies.net/pop6001c.asp?pop_id=1986 >
38. A.Arntaiz-Villena, V. Gonzalez-Alcos et al. HLA genes in Uros from Titikaka Lake, Peru: origin and relationship with other Amerindians and worldwide populations. *International Journal of Immunogenetics* 36 (2009): 159-167
39. A. Leslie, P.C. Matthews et al. Additive Contribution of HLA Class I Alleles in the Immune Control of HIV-1 Infection. *Journal of Virology*. 84 (2010): 9879 – 9888.
40. HIV molecular immunology database. *Epitope Variants and Escape Mutations - T Helper/CD4+ Epitope Variants and Escape Mutations.* [Visited: 7th June 2016]. Available at: < http://www.hiv.lanl.gov/content/immunology/variants/helper_variant.html >
41. Buchbinder SP, Mehrotra DV, Duerr A, Fitzgerald DW, Mogg R, Li D, et al. Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial. *Lancet*. 372 (2008): 1881–1893.
42. Goepfert PA, Elizaga ML, Seaton K, Tomaras GD, Montefiori DC, Sato A, et al. Specificity and 6-month durability of immune responses induced by DNA and recombinant

modified vaccinia ankara vaccines expressing HIV-1 virus-like particles. *Journal of Infectious Diseases*. 210 (2014) : 99 – 110.

43. Lefrere JJ, Roudot-Thoraval F et al. The risk of disease progression is determined during the first year of human immunodeficiency virus type I infection. *Journal of infectious disease*. 177 (1998): 1541-1548.
44. Pedersen C, Katzenstein T. Prognostic value of serum HIV-RNA levels at virologic steady state after seroconversion: relation to CD4 cell count and clinical course of primary infection. *Journal of Acquired Immune Deficiency Syndrome Human Retrovirology.* 16 (1997): 93-99
45. B. Julg, E.S.Moodley, et al. Possession of HLA Class II DRB1*1303 Associated with Reduced Viral Loads in Chronic HIV-1 Clade C and B Infection. *The Journal of Infectious Diseases*. 203 (2011): 803 – 809.
46. S. Ranasinghe, S. Cutler et al. Association of HLA-DRB1-restricted $CD4^+$ T cell responses with HIV immune control. *Nature Medicine*. 19 (2013): 930 – 933.
47. HLA-B35-PX and HLA-B-35-PY subtype differentiation does not predict observed differences in level of HIV control in a Peruvian MSM cohort*. AIDS Correspondence.* 28(2014): 2323-2325.
48. P.Kiepiela, A.J.Leslie, et al. Dominant influence of HLA –B in mediating the potential co-evolution of HIV and HLA. *Letters to Nature*. 432 (2004) :769 – 774.
49. The International HIV Controllers Study. The Major Genetic Determinants of HIV-1 Control Affect HLA Class I Peptide Presentation. Science. 10 (2010): 1551 – 1557.
50. K.L. Lunetta, L.B. Hayward, et al. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genetics*. 5(2004)
51. D.E. Kaufmann, P.M. Bailey, et al. Comprehensive Analysis of Human Immunodeficiency Virus Type-1 Specific CD4 Responses Reveals Marked Immunodominance Gag and Nef and the Presence of Broadly Recognized Peptides. *Journal of Virology*. 78 (2004): 4463 – 4477.
52. S. Ranasinghe, M. Flanders et al. HIV-Specific CD4 T Cell Responses to Different Viral Proteins Have Discordant Associations with Viral Load and Clinical Outcome. *Journal of Virology*. 2011. 277 – 283.
53. Janeway CA Jr, Travers P, Walport M, et al. *Immunobiology: The Immune System in Health and Disease*, 2001. Garland Science (NY).
54. John r. Mascola and Richard A.Koup. *Global HIV/AIDS Medicine*, 2008. Elseiver.

55. E.O. Freed. HIV-1 assembly, release and maturation. *Nature Reviews*. 13(2015): 484 – 496.
56. Ali Akram and Robert D. Inman. Immunodominance: A pivotal principle in host response to viral infections. *Clinical Immunology*. 143 (2012): 99 – 115.
57. Woodberry T, Suscovich TJ. Differential targeting and shifts in the immunodominance of Epstein-Barr virus-specific CD8 and CD4 T cell responses during acute and persistent infection. *Journal of Infectious Diseases*. 192 (2005): 1513 – 1524.
58. Bihl F, Frahm N. Impact of HLA-B alleles, epitope binding affinity, functional avidity, and viral coinfection on the immunodominance of virus-specific CTL responses. Journal of Immunology. 176 (2006): 4094 – 4101.
59. N. Erdmann, V.Y. Du et al. HLA Class-II Associated HIV Polymorphisms Predict Escape from CD4+ T Cell Responses. *Plos Pathogens*. 11 (2015).
60. Z.L.Brumme. HLA-Associated Immune Escape Pathways in HIV-1 Subtype B Gag, Pol and Nef Proteins. *Plos One*. 4 (2009)

APPENDIX I

Supplementary table

Table A1: Frequency table. Table with absolute and relative frequencies of the patients with a certain HLA-II allele in the Peruvian cohort. Columns HLA negative and HLA positive show the number of patients without or with a certain allele. From individuals expressing a certain allele there is the number of homozygotes and heterozygotes. The 4th column contain the cohort frequency, and the 5th, the allele frequency.

HLA – II alleles	HLA negative	HLA positive	Number Homozygotes	Number Heterozygotes	Cohort Frequency	Allele Frequency
DQA*0101	341	51	4	47	0.130	0.070
DQA*0401	308	84	7	77	0.214	0.116
DQA*0102	322	70	4	66	0.179	0.094
DQA*0201	338	54	4	50	0.138	0.074
DQA*0301	158	234	67	167	0.597	0.384
DQA*0104	382	10	0	10	0.026	0.013
DQA*0501	241	151	16	135	0.385	0.213
DQA*0103	369	23	0	23	0.059	0.029
DQA*0503	391	1	0	1	0.003	0.001
DQA*0601	391	1	0	1	0.003	0.001
DQA*0302	390	2	0	2	0.005	0.003
DQA*0402	391	1	0	1	0.003	0.001
DQB*0302	231	161	26	135	0.411	0.239
DQB*0402	310	82	4	78	0.209	0.110
DQB*0202	363	29	2	27	0.074	0.040
DQB*0201	337	55	4	51	0.140	0.075
DQB*0303	290	102	9	93	0.260	0.142
DQB*0301	263	129	11	118	0.329	0.179
DQB*0502	380	12	0	12	0.031	0.015
DQB*0602	352	40	2	38	0.102	0.054
DQB*0401	388	4	1	3	0.010	0.006
DQB*0503	384	8	1	7	0.020	0.011
DQB*0501	337	55	4	51	0.140	0.075
DQB*0603	375	17	0	17	0.043	0.022
DQB*0601	389	3	0	3	0.008	0.004
DQB*0609	383	9	0	9	0.023	0.011
DQB*0604	381	11	0	11	0.028	0.014
DQB*0605	390	2	0	2	0.005	0.003
DQB*0306	391	1	0	1	0.003	0.001
DRB1*0403	366	26	0	26	0.066	0.033
DRB1*0802	328	64	3	61	0.163	0.085
DRB1*0407	330	62	6	56	0.158	0.087
DRB1*0404	337	55	2	53	0.140	0.073
DRB1*0301	352	40	1	39	0.102	0.052
DRB1*0701	341	51	2	49	0.130	0.068

DRB1*0901	306	86	9	77	0.219	0.121
DRB1*0410	389	3	0	3	0.008	0.004
DRB1*1502	387	5	0	5	0.013	0.006
DRB1*0804	382	10	0	10	0.026	0.013
DRB1*0102	371	21	1	20	0.054	0.028
DRB1*0401	387	5	0	5	0.013	0.006
DRB1*1302	371	21	1	20	0.054	0.028
DRB1*0411	374	18	2	16	0.046	0.026
DRB1*1101	376	16	0	16	0.041	0.020
DRB1*1401	381	11	0	11	0.028	0.014
DRB1*08	391	1	0	1	0.003	0.001
DRB1*0807	391	1	0	1	0.003	0.001
DRB1*0101	381	11	1	10	0.028	0.015
DRB1*0405	386	6	0	6	0.015	0.008
DRB1*0801	385	7	0	7	0.018	0.009
DRB1*1301	368	24	0	24	0.061	0.031
DRB1*0302	386	6	0	6	0.015	0.008
DRB1*1402	341	51	3	48	0.130	0.069
DRB1*1104	379	13	1	12	0.033	0.018
DRB1*0402	383	9	0	9	0.023	0.011
DRB1*0431	391	1	0	1	0.003	0.001
DRB1*1201	380	12	0	12	0.031	0.015
DRB1*0103	386	6	0	6	0.015	0.008
DRB1*1304	389	3	0	3	0.008	0.004
DRB1*1501	368	24	1	23	0.061	0.032
DRB1*0408	390	2	0	2	0.005	0.003
DRB1*1602	361	31	1	30	0.079	0.041
DRB1*0806	391	1	0	1	0.003	0.001
DRB1*1001	383	9	0	9	0.023	0.011
DRB1*1305	391	1	0	1	0.003	0.001
DRB1*0406	390	2	0	2	0.005	0.003
DRB1*1503	380	12	0	12	0.031	0.015
DRB1*1406	387	5	0	5	0.013	0.006
DRB1*1601	387	5	0	5	0.013	0.006
DRB1*1303	382	10	0	10	0.026	0.013
DRB1*1102	390	2	0	2	0.005	0.003
DRB1*0803	391	1	0	1	0.003	0.001
DRB345*40101G	142	250	74	176	0.638	0.413
DRB345*AgBlank	273	119	19	100	0.304	0.176
DRB345*30202	309	83	7	76	0.212	0.115
DRB345*30301	383	9	0	9	0.023	0.011
DRB345*50202	355	37	1	36	0.094	0.048
DRB345*30101	291	101	18	83	0.258	0.152
DRB345*50102	388	4	0	4	0.010	0.005
DRB345*301	389	3	1	2	0.008	0.005

DRB345*401	380	12	6	6	0.031	0.023
DRB345*302	391	1	0	1	0.003	0.001
DRB345*50101	357	35	2	33	0.089	0.047
DRB345*30211	391	1	1	0	0.003	0.003

APPENDIX II

R package “AnalysisHLA”

Package 'AnalysisHLA'

September 16, 2016

Type Package

Title Analysis of different HLA associations

Version 1.0

Date 2016-09-13

Author Bruna Oriol

Maintainer Bruna Oriol <bruna.oriol@uvic.cat>

Description The present package allows the statistical univariate analysis of associations between HLA alleles and HIV progression markers.

License GPL (>= 2)

R topics documented:

`AnalysisHLA-package` *Analysis of different HLA associations*

Description

The present package allows the statistical univariate analysis of associations between HLA alleles and HIV progression markers.

Details

The DESCRIPTION file: This package was not yet installed at build time.
The present package allows the analysis of HLA and T cell responses to OLPs (peptides) with HIV disease progression markers in an univariate manner.

Index: This package was not yet installed at build time.

Author(s)

Bruna Oriol

Maintainer: Bruna Oriol <bruna.oriol@uvic.cat>

`allelic.cumulative` *Get the matrices to compute the allelic cumulative frequency*

Description

Given the allelic frequency of the different alleles for each HLA locus, it returns a matrix with patients in rows and HLA alleles in columns. Each cell contains a 0 if the patient does not have the allele, and the allelic frequency in the opposite case.

Usage

```
allelic.cumulative(table012, freq, n)
```

Arguments

`table012`	Matrix with patients IDs in rows and HLA alleles in columns. It must contain the alleles for all the isotypes. Each cell of the matrix contain 1 or 0 according if each patient presents or not the allele.
`freq`	Data frame with different columns. One of which must be the allelic relative frequency of each of the HLA alleles for an HLA isotype in the cohort.
`n`	Number correspoding to the position of the column where the relative frequences are found.

Details

It is needed the output of the funcion *fill.matrix2* for different HLA isotypes as input of the present function and a data frame containing a column with the allelic relative frequence of each HLA allele in the studied locus.

Value

It returns a matrix with patients in rows and the HLA alleles in columns, then if the patient is positive for a given allele it contains the allelic frequence of this allele in the cohort.

Author(s)

Bruna Oriol

See Also

fill.matrix2, CohortData

Examples

```
## The function is currently defined as
function (table012, freq, n)
{
   matrix <- matrix(nrow = length(rownames(table012)), ncol = length(colnames(table012)))
    rownames(matrix) <- rownames(table012)
    colnames(matrix) <- colnames(table012)
    for (i in 1:length(colnames(table012))) {
        patients <- NULL
        allele <- table012[, i]
        for (a in 1:length(allele)) {
            patient <- as.character(allele[a])
            if (patient == "1") {
                patient <- freq[i, n]
                patients <- c(patients, patient)
            }
            else {
                patients <- c(patients, 0)
            }
        }
        matrix[, i] <- patients
    }
    return(matrix)
  }

## EXAMPLES
data(CohortData)
patients <- CohortData[, c(1,3,4)]
HLA_types <- HLAtypes(CohortData[,c(3,4)])
table012 <- fill.matrix2(patients, HLA_types)
colnames(table012) <- paste("HLA1", HLA_types, sep = "_")
al_frqHLA <- apply(table012, 2, sum)
rel_frqHLA <- (al_frqHLA/(nrow(table012)*2))
freq <- data.frame(al_frqHLA, rel_frqHLA)
allelic.cumulative(table012, freq, 2)
```

CD4.function *Association between HLA alleles and CD4 counts*

Description

For each allele in a given HLA locus, the patients" CD4 counts are divided in 2 groups: a group with CD4 counts for patients not presenting the studied HLA allele, and another one, with the patients with the HLA allele. CD4 counts of these two groups are compared with a T-test and a Mann-Whitney test to see if there are differences between them. The problem of multiple comparisons is adressed with the false discovery rate (FDR) method.

Usage

```
CD4.function(HLA_types, tabProg)
```

Arguments

HLA_types	Output function *HLAtypes*. It is a vector containing all the possible alleles for a given HLA isotype or locus.
tabProg	Output function *fill.matrix* converted into a data frame (Patients in rows and alleles in columns) with additional columns of information correspoding to viral load or CD4 counts. It is important to give names to the columns of the data frame: the alleles names or the information each column contains. Viral load and CD4 counts columns name **must** be *VL* and *CD4* respectively, otherwise they will not be recognised by *VL.function* or *CD4.function*

Details

The present function *CD4.function* calls the functions *t.test*, *wilcox.test*, *median*, *mean* and *p.adjust* from the package **stats**.

Value

The output of this function is a data frame containing the different HLA alleles of a given HLA locus. The 2nd and 3rd column contain the mean for the CD4 counts of patients without or with the selected allele respectively. The 4th column contains the p-value for the T-test, and in the 5th column, there is this p-value corrected by FDR method. Analogously, the 6th and 7th columns contain the median of the 2 CD4 counts groups (individual not having or having the allele), and the 8th and the 9th, the p-value and adjusted p-value from the Wilcox test.

Note

Test if data fulfil normality assumption with function *shapiro.test* from **stats** to decide between T-test or Wilcoxon - Test results

Author(s)

Bruna Oriol

See Also

CohortData, HLAtypes, fill.matrix

Examples

```
## The function is currently defined as
function (HLA_types, tabProg)
{
    pval <- NULL
    pval_mw <- NULL
    mean0 <- NULL
    median0 <- NULL
    mean1 <- NULL
    median1 <- NULL
    hla_class <- NULL
    for (i in 1:length(HLA_types)) {
        dqa <- colnames(tabProg)[i]
        tabProg[, i] <- factor(tabProg[, i], levels = c(0, 1))
        hla_class <- c(hla_class, dqa)
        if (sum(tabProg[, i] == 1) > 1) {
```

```
            pv <- t.test(as.numeric(as.character(tabProg$CD4)) ~
                tabProg[, i])$p.value
            pvw <- wilcox.test(as.numeric(as.character(tabProg$CD4)) ~
                tabProg[, i])$p.value
            m0 <- subset(tabProg, tabProg[, i] == 0)
            m0 <- mean(as.numeric(as.character(m0$CD4)), na.rm = TRUE)
            m1 <- subset(tabProg, tabProg[, i] == 1)
            m1 <- mean(as.numeric(as.character(m1$CD4)), na.rm = TRUE)
            med0 <- subset(tabProg, tabProg[, i] == 0)
            med0 <- median(as.numeric(as.character(med0$CD4)),
                na.rm = TRUE)
            med1 <- subset(tabProg, tabProg[, i] == 1)
            med1 <- median(as.numeric(as.character(med1$CD4)),
                na.rm = TRUE)
            mean0 <- c(mean0, m0)
            mean1 <- c(mean1, m1)
            median0 <- c(median0, med0)
            median1 <- c(median1, med1)
            pval <- c(pval, pv)
            pval_mw <- c(pval_mw, pvw)
        }
        else {
            mean0 <- c(mean0, "NA")
            mean1 <- c(mean1, "NA")
            median0 <- c(median0, "NA")
            median1 <- c(median1, "NA")
            pval <- c(pval, "NA")
            pval_mw <- c(pval_mw, "NA")
        }
    }
    p.adj <- p.adjust(pval, method = "fdr")
    p.adj_mw <- p.adjust(pval_mw, method = "fdr")
    dd_CD4 <- cbind(hla_class, mean0, mean1, pval, p.adj, median0,
        median1, pval_mw, p.adj_mw)
    return(dd_CD4)
  }

## EXAMPLES
data(CohortData)
patients <- CohortData[, c(1,3,4)]
HLA_types <- HLAtypes(CohortData[,c(3,4)])
table01 <- fill.matrix(patients, HLA_types)
colnames(table01) <- paste("HLA1", HLA_types, sep = "_")
tabProg <- cbind(table01, CohortData$logVL, CohortData$CD4 )
tabProg <- data.frame(tabProg, stringsAsFactors = FALSE)
colnames(tabProg) <- c(colnames(table01), "VL", "CD4")
CD4.function(HLA_types, tabProg)
```

`cohort.cumulative` *Prepare matrices to compute the cohort cumulative frequency*

Description

Given the cohort frequency of the different alleles for each locus, it returns a matrix with patients in rows and HLA alleles in columns where each cell contains a 0 if the patient does not have the allele, or the frequency of the allele in the opposite case

Usage

```
cohort.cumulative(table01, freq, n)
```

Arguments

table01	Matrix containing all the patients in a cohort in rows, and all the HLA alleles in the cohort for all the locus we want to evaluate in columns. For each cell of the matrix we have a 1 or a 0 according if an allele is present or not in a patient.
freq	Data frame with at least a column containing the relative frequency of each of the HLA alleles for an HLA isotype in the cohort.
n	Number correspoding to the column position in the data frame where the relative frequence is found.

Details

It is needed the output of function *fill.matrix* for different HLA isotypes as input of the function.

Value

It retrieves a matrix with patients in rows and HLA alleles in columns. In each cell, if the patient is positive for a given allele, it contains the frequency of this allele in the cohort, otherwise, it contains a 0.

Author(s)

Bruna Oriol

See Also

fill.matrix, CohortData

Examples

```
## The function is currently defined as
function (table01, freq, n)
{
    matrix <- matrix(nrow = length(rownames(table01)), ncol = length(colnames(table01)))
    rownames(matrix) <- rownames(table01)
    colnames(matrix) <- colnames(table01)
    for (i in 1:ncol(table01)) {
        patients <- NULL
        allele <- table01[, i]
        for (a in 1:length(allele)) {
            pat <- as.character(allele[a])
            if (pat == "1") {
                patient <- freq[i, n]
                patients <- c(patients, patient)
            }
            else {
                patients <- c(patients, 0)
            }
        }
        matrix[, i] <- patients
```

```
    }
    return(matrix)
  }

## EXAMPLE
data(CohortData)
patients <- CohortData[, c(1,3,4)]
HLA_types <- HLAtypes(CohortData[,c(3,4)])
table01 <- fill.matrix(patients, HLA_types)
colnames(table01) <- paste("HLA1", HLA_types, sep = "_")
al_frqHLA <- apply(table01, 2, sum)
rel_frqHLA <- (al_frqHLA/nrow(table01))
freq <- data.frame(al_frqHLA, rel_frqHLA)
rownames(freq) <- rownames(table01)
cohort.cumulative(table01, freq, 2)
```

`CohortData` *Toy example of a HIV Cohort dataset*

Description

This is a toy example of what a HIV cohort to be analysed might contain: Patients identifiers (IDs), the HIV status, the pairs of HLA alleles for each gene or locus, the viral load and its logaritmic conversion, the CD4 counts and finally, the overlapping sequences (OLPs)

Usage

```
data("CohortData")
```

Format

A data frame with 6 observations on the following 10 variables.

`PatientID` a character vector

`HIV_Status` a character vector

`HLAII_1` a character vector

`HLAII_2` a character vector

`HLAIIa_1` a character vector

`HLAIIb_2` a character vector

`VL` a numeric vector

`logVL` a numeric vector

`CD4` a numeric vector

`OLPs` a character vector

`fill.matrix` *Presence/absence matrix of HLA alleles*

Description

This function generates a matrix with patients in rows and HLA alleles in columns. For each cell in the matrix it is given the number 0 or 1 according if the patient has or not the allele in question.

Usage

```
fill.matrix(patients, HLA_types)
```

Arguments

`patients`	Data frame with 3 columns. The first one correspond to the patient ID, and the other 2, to the HLA alleles of each patient. We can take these 3 columns subsetting them from the toy example *CohortData*.
`HLA_types`	A vector with all the alleles in the cohort of study for a given HLA locus, it can be obtained from the function *HLAtypes* also in the present package.

Details

Function *fill.matrix* uses the output of function *HLAtypes*. The output of the present function will be necessary for functions *LD*, *cohort.cumulative*, *allellic.cumulative*, *VL.function* and *CD4.function*.

Value

It returns a matrix with patients in rows and HLA alleles in columns and 0 or 1 in each cell according if a patient has or not each allele.

Author(s)

Bruna Oriol

See Also

`HLAtypes`, `CohortData`

Examples

```
## The function is currently defined as
function (patients, HLA_types)
{
    matrix <- matrix(ncol = length(HLA_types), nrow = nrow(patients))
    for (i in 1:nrow(patients)) {
        alleles <- NULL
        pat1 <- patients[i, ]
        for (h in 1:length(HLA_types)) {
            hla_allele <- HLA_types[h]
            ifelse((pat1[, 2] == hla_allele | pat1[, 3] == hla_allele),
                alleles <- c(alleles, 1), alleles <- c(alleles,
                  0))
```

```
        }
        matrix[i, ] <- as.matrix(alleles)
        rownames(matrix) <- patients[,1]
    }
    return(matrix)
  }

## EXAMPLE
data(CohortData)
patients <- CohortData[, c(1,3,4)]
HLA_types <- HLAtypes(CohortData[, c(3,4)])
fill.matrix(patients,HLA_types)
```

`fill.matrix2` *HLA alleles matrix: Homozygosis/Heterozygosis*

Description

This function generates a matrix with patients in rows and HLA alleles in columns, then, for each cell in the matrix it is given the number 0, 1 or 2 according if the patient does not present this allele, presents it in heterozygosis or in homozygosis.

Usage

```
fill.matrix2(patients, HLA_types)
```

Arguments

`patients`	Data frame with 3 columns. The first one correspond to the patient ID, and the other 2, to the HLA alleles of each patient. We can take these 3 columns subsetting them from the *CohortData* data frame given as sample data in the present package.
`HLA_types`	It is a vector with all the alleles in the cohort of study for a given HLA locus. This vector is obtained from the function in this package called *HLAtypes*

Details

The present function *fill.matrix2* uses the output of the function *HLAtypes*

Value

It returns a matrix with patients in rows and HLA alleles in columns and 0, 1 or 2 in each cell according if a patient does not present the allele, presents it in heterozygosis or in homozygosis respectively.

Author(s)

Bruna Oriol

See Also

`HLAtypes`, `CohortData`

Examples

```
## The function is currently defined as
function (patients, HLA_types)
{
    matrix <- matrix(ncol = length(HLA_types), nrow = nrow(patients))
    for (i in 1:nrow(patients)) {
        alleles <- NULL
        pat1 <- patients[i, ]
        for (h in 1:length(HLA_types)) {
            hla_allele <- HLA_types[h]
            if (pat1[, 2] == pat1[, 3]) {
                ifelse(pat1[, 2] == hla_allele, alleles <- c(alleles,
                  2), alleles <- c(alleles, 0))
            }
            if (pat1[, 2] != pat1[, 3]) {
                ifelse((pat1[, 2] == hla_allele | pat1[, 3] ==
                  hla_allele), alleles <- c(alleles, 1), alleles <- c(alleles,
                  0))
            }
        }
        matrix[i, ] <- as.matrix(alleles)
        rownames(matrix) <- patients[,1]
    }
    return(matrix)
  }

## EXAMPLE
data(CohortData)
patients <- CohortData[, c(1,3,4)]
HLA_types <- HLAtypes(CohortData[, c(3,4)])
fill.matrix2(patients,HLA_types)
```

`Hepitope_f` *Hopeful-epitopes or Hepitopes*

Description

Based on the patients that respond to a given syntesised peptide (named overlapping peptides or OLP) and their HLA haplotype, associations between the fact of responding to a given OLP and the expression of certain HLA allele are tested.

Usage

```
Hepitope_f(Hepitop_df, Hepitope_HLAII, LD, Peptides, alpha = 0.05)
```

Arguments

`Hepitop_df`	Data frame from a matrix with peptides (OLPs) in columns and patients in rows. 1 means the patient shows response to the selected peptide and 0, it does not. The matrix can be obtained from the function in this package *matrixOLP_f*.
`Hepitope_HLAII`	Matrix with information of which patients (rows) present (the cell will contain 1) or not (the cell will contain 0) each of the HLA alleles (columns). It is important to join the matrices for all of the HLA locus we want to evaluate in a single one.

LD	Output of function *LD* that computes linkage disequilibrium. It is a data frame containing 7 columns. The 2 first ones, contain the different HLA alleles that are compared. Then, there are 2 columns containing the p-value and the p-value adjusted for multiple comparisons (using false discovery rate method) from the Fisher's test results applied to see if 2 alleles are associated or not. Finally, there are 3 columns with different linkage disequilibrium measures: D, D', r^2
Peptides	Data frame that must contain a column corresponding to the number of OLP, the OLP name (the peptide), the subunit and the sequence.
alpha	The significance level alpha, by default it is 0.05

Details

This function needs as input matrices that are obtained from the output of functions *fill.matrix*, *matrixOLP_f* and *LD* that corresponds to: Hepitope_df, Hepitope_HLAII and LD. Moreover, it is needed the dataset *Pept_Seq* as Peptides. In order to perform the Fisher Test, it is necessary to use functions *table* from **base** and *fisher.test*, from **stats**. It is important to note that if there are significant results after applying the Fihser's' test, the function will return OLP-HLA allele associations showing a p-value < 0.05; otherwise, it will return all the associations evaluated and their p-value.

Value

This function returns a matrix containing the number of OLP (1,2,3...), the peptide to which the OLP is mapped, the subunits, the sequences, the HLA allele to which the OLP is associated, the p-values from the Fisher's Exact Test and adjusted p-values for multiple comparisons (with false discovery rate method) . Finally, there are different columns making reference to the HLA alleles (retrieving association to the same OLP) in linkage disequilibrium (meaning a corrected p-value for multiple comparisons < 0.05, and a D' nearby 1) with the allele studied.

Author(s)

Bruna Oriol

References

The present function *hepitope* is made based on the Hepitope tool from the HIV molecular immunology databaser www.hiv.lanl.gov/content/immunology/hepitopes.

Examples

```
## The function is currently defined as
function (Hepitop_df, Hepitope_HLAII, LD, Peptides, alpha = 0.05)
{
    pvals <- NULL
    hla_classII <- NULL
    epitope_num <- NULL
    for (olp in 1:length(colnames(Hepitop_df))) {
        olp.1 <- colnames(Hepitop_df)[olp]
        olptab <- Hepitop_df[, olp.1]
        for (dq1 in 1:length(colnames(Hepitope_HLAII))) {
            dq1.1 <- colnames(Hepitope_HLAII)[dq1]
            dq1tab <- Hepitope_HLAII[, dq1.1]
            if (1 %in% dq1tab) {
```

```
            taula <- table(olptab, dq1tab)
            ft <- fisher.test(taula)
            pv <- ft$p.value
            pvals <- c(pvals, pv)
        }
        else {
            pvals <- c(pvals, "NA")
        }
        hla_classII <- c(hla_classII, dq1.1)
        epitope_num <- c(epitope_num, olp.1)
    }
}
hepitope_result <- data.frame(epitope_num, hla_classII, pvals)
hepitope_result$p.adj <- p.adjust(pvals, method = "fdr")
hepitope_result <- subset(hepitope_result, hepitope_result$pvals !=
    "NA")
hepitope_sigres <- subset(hepitope_result, as.numeric(as.character(hepitope_result$pvals)) <
    alpha)
LD_sig <- subset(LD, LD$p.adj < alpha & LD$Ds > abs(0.5))
if (length(hepitope_sigres[, 1]) < 1) {
    hepitope_sigres <- hepitope_result
}
if (length(LD_sig[, 1]) < 1) {
    LD_sig <- LD
}
peptides <- hepitope_sigres[, 1]
peptides <- as.character(peptides)
peptides <- unique(peptides)
vec <- NULL
rep <- NULL
for (i in 1:length(peptides)) {
    pep <- peptides[i]
    alleles <- subset(hepitope_sigres, hepitope_sigres[,
        1] == pep)
    alleles <- as.character(alleles[, 2])
    for (h in 1:length(alleles)) {
        a <- alleles[h]
        b <- pep
        vec <- c(vec, a)
        rep <- c(rep, b)
    }
}
pept_allele <- data.frame(rep, vec, stringsAsFactors = FALSE)
colnames(pept_allele) <- c("OLP", "HLA")
LDs <- LD_sig[, 1:2]
sigpeptds <- unique(pept_allele[, 1])
mm <- matrix(nrow = 1, ncol = 20)
for (k in 1:length(sigpeptds)) {
    olp_h <- sigpeptds[k]
    olph <- subset(pept_allele, pept_allele[, 1] == olp_h)
    matrix_ld <- matrix(nrow = length(olph[, 2]), ncol = 20)
    for (i in 1:length(olph[, 2])) {
        olp <- olph[, 1][i]
        hl <- olph[, 2][i]
        hl_s <- subset(LDs, LDs[, 1] == hl | LDs[, 2] ==
            hl)
        ld1 <- which(hl_s[, 1] != hl)
```

```
            ld1 <- as.character(hl_s[, 1][ld1])
            ld2 <- which(hl_s[, 2] != hl)
            ld2 <- as.character(hl_s[, 2][ld2])
            ld <- c(ld1, ld2)
            ld <- unique(ld)
            ld <- ld[which(ld %in% olph[, 2])]
            l.ld <- length(ld)
            l.olp <- length(olph[, 2]) + 1
            l.mat <- 20 - 2 - l.ld
            if (l.mat > 0) {
                matrix_ld[i, ] <- c(olp, hl, ld, rep(" ", l.mat))
            }
            if (l.mat == 0) {
                matrix_ld[i, ] <- c(olp, hl, ld)
            }
        }
        mm <- rbind(mm, matrix_ld)
    }
    mm <- mm[-1, ]
    mm <- mm[, !apply(mm, 2, function(x) all(x == " ")) == TRUE]
    LDnames <- paste0("LD", seq(1:(ncol(mm) - 2)))
    colnames(mm) <- c("OLP", "HLA-II", LDnames)
    hh <- cbind(hepitope_sigres, mm[, -(1:2)])
    OLPvec <- NULL
    Subunit <- NULL
    OLPseq <- NULL
    for (i in 1:length(hh[, 1])) {
        ep.num <- as.character(hh[, 1][[i]])
        df <- Peptides[Peptides == ep.num, ]
        olp <- as.character(df[, 2])
        sub <- as.character(df[, 3])
        seq <- as.character(df[, 4])
        OLPvec <- c(OLPvec, olp)
        Subunit <- c(Subunit, sub)
        OLPseq <- c(OLPseq, seq)
    }
    hh$Peptide <- OLPvec
    hh$Subunit <- Subunit
    hh$Sequence <- OLPseq
    hh <- hh[, c("epitope_num", "Peptide", "Subunit", "Sequence",
        "hla_classII", "pvals", "p.adj", LDnames)]
    return(hh)
  }
  ## EXAMPLE
# Load both sample datasets
data(CohortData)
data(Pept_Seq)

Peptides <- Pept_Seq
Peptides$Num <- paste0("OLP_", Peptides$Num)

# Apply function matrixOLP_f
olp_rawdata <- CohortData[, c(1,ncol(CohortData))]
olp_rawdata$OLPs <- as.character(olp_rawdata$OLPs)
olp_rawdata$OLPs <-strsplit(olp_rawdata$OLPs, ";")
Hepitop_df <- matrixOLP_f(olp_rawdata)
Hepitop_df <- as.data.frame(Hepitop_df)
```

```
colnames(Hepitop_df) <- paste0("OLP_", colnames(Hepitop_df))

# Apply function HLAtypes and following fill.matrix
# to each isotype or gene and make a data frame
# of the results.
patients1 <- CohortData[, c(1,3,4)]
HLA_types1 <- HLAtypes(CohortData[,c(3,4)])
patients2 <- CohortData[, c(1,5,6)]
HLA_types2 <- HLAtypes(CohortData[,c(5,6)])
HLA1_table <- fill.matrix(patients1, HLA_types1)
colnames(HLA1_table) <- paste("HLA1", HLA_types1, sep = "_")
HLA2_table <- fill.matrix(patients2, HLA_types2)
colnames(HLA2_table) <- paste("HLA2", HLA_types2, sep = "_")

Hepitope_HLAII <- data.frame(cbind(HLA1_table, HLA2_table))

#Apply function LD
LDoutput <- LD(HLA1_table, HLA2_table)

# Apply the current function Hepitope_f
Hepitope_f(Hepitop_df, Hepitope_HLAII, LDoutput, Peptides)
```

`HLAtypes` *Alleles in a cohort*

Description

This function makes a vector with all the alleles for a given HLA locus in a cohort of patients.

Usage

```
HLAtypes(df)
```

Arguments

`df`	The input data frame consist on the 2 columns that contain HLA alleles for each HLA isotype or gene.

Details

The output of this function will be essential to go on with the analysis pipeline.

Value

It returns a vector with all the alleles in the studied cohort from a given HLA isotype or gene.

Author(s)

Bruna Oriol

See Also

`CohortData`

Examples

```
## The function is currently defined as
function (df)
{
    colnames(df) <- c("HLA1", "HLA2")
    A1 <- as.character(unique(df$HLA1))
    A2 <- as.character(unique(df$HLA2))
    HLAtypes <- union(A1, A2)
    return(HLAtypes)
  }


## EXAMPLE
data(CohortData)
HLAtypes(CohortData[, c(3,4)])
```

LD *Linkage Disequilibrium between alleles in two different HLA locus or isotypes*

Description

This function computes associations between alleles in two different HLA locus or isotypes using a chi-squared test with 1 degree of freedom as well as linkage disequilibrium measures: D, D' and r2

Usage

```
LD(HLA1_table, HLA2_table)
```

Arguments

HLA1_table	Matrix with patients in rows and HLA alleles for gene 1 in columns. Number 1 means the patient has the allele, and 0, it does not.
HLA2_table	Matrix with patients in rows and HLA alleles for gene 1 in columns. Number 1 means the patient has the allele, and 0, it does not.

Details

This functions needs as input the output of the function *fill.matrix* for the 2 studied HLA isotypes or locus. This function also calls the function *fisher.test* from the package **stats** in order to compute if the 2 alleles compared are associated or not. The problem of multiple comparisons is addressed with function *p.adjust* also from package **stats** using false discovery rate method (FDR). The output of this function will be necessary for function *Hepitope_df*.

Value

It returns a data frame containing 2 columns with the alleles compared each time, so in column 1 there are the alleles in HLA1, and in column2, the alleles in HLA2. The 3rd and 4th column correspond to the p-value and q-value, and finally there are 3 columns with different linkage disequilibrium measures: D, D' and r^2.

Author(s)

Bruna Oriol

See Also

CohortData, HLAtypes, fill.matrix

Examples

```
## The function is currently defined as
function (HLA1_table, HLA2_table)
{
    pvals <- NULL
    hla1 <- NULL
    hla2 <- NULL
    Obs <- NULL
    Exp <- NULL
    D <- NULL
    Ds <- NULL
    r2 <- NULL
    for (dq1 in 1:length(colnames(HLA1_table))) {
        dq1.1 <- colnames(HLA1_table)[dq1]
        dq1tab <- HLA1_table[, dq1.1]
        for (dq2 in 1:length(colnames(HLA2_table))) {
            dq2.1 <- colnames(HLA2_table)[dq2]
            dq2tab <- HLA2_table[, dq2.1]
            taula <- table(dq1tab, dq2tab)
            hla1 <- c(hla1, dq1.1)
            hla2 <- c(hla2, dq2.1)
            O <- taula[2, 2]/sum(taula)
            Obs <- c(Obs, O)
            pA = (taula[2, 1] + taula[2, 2])/sum(taula)
            pB = (taula[1, 2] + taula[2, 2])/sum(taula)
            E <- pA * pB
            Exp <- c(Exp, E)
            d <- O - E
            D <- c(D, d)
            if (d > 0) {
                Dmax = min(pA * (1 - pB), pB * (1 - pA))
                ds <- d/Dmax
                Ds <- c(Ds, ds)
            }
            if (d < 0) {
                Dmax = min(pA * pB, (1 - pA) * (1 - pB))
                ds <- d/Dmax
                Ds <- c(Ds, ds)
            }
            r <- (d^2)/(pA * pB * (1 - pA) * (1 - pB))
            r2 <- c(r2, r)
            chi = sum(taula) * r
            pv = 1 - pchisq(chi, 1)
            pvals <- c(pvals, pv)
        }
    }
    dd <- data.frame(hla1, hla2, pvals, Obs, Exp, D, Ds, r2)
```

```
    dd$p.adj <- p.adjust(pvals, method = "fdr")
    dd <- dd[c("hla1", "hla2", "pvals", "p.adj", "Obs", "Exp",
        "D", "Ds", "r2")]
    return(dd)
  }
## EXAMPLE
data(CohortData)

patients1 <- CohortData[, c(1,3,4)]
HLA_types1 <- HLAtypes(CohortData[,c(3,4)])
HLA1_table <- fill.matrix(patients1, HLA_types1)
colnames(HLA1_table) <- paste("HLA1", HLA_types1, sep = "_")

patients2 <- CohortData[, c(1,5,6)]
HLA_types2 <- HLAtypes(CohortData[,c(5,6)])
HLA2_table <- fill.matrix(patients2, HLA_types2)
colnames(HLA2_table) <- paste("HLA2", HLA_types2, sep = "_")

LD(HLA1_table, HLA2_table)
```

`matrixOLP_f` *Matrix of the patients respoding to ech of the overlapping sequences analysed (OLP)*

Description

Given a data frame with the patient IDs and a column with a vector of all the OLPs to which a patient responds to, it returns a matrix with all the OLPs in columns and the patients in rows where each cell contains 1 if the patient show a response to the OLP, or a 0, if it does not.

Usage

```
matrixOLP_f(olp_rawdata)
```

Arguments

`olp_rawdata`	A 2 columned data frame. In the first column, there are patient identifiers (Patient IDs) and in the second one, the vectors of the different OLPs (it must be of type *character*) to which each patient reponds to.

Details

This function is essential to run function *Hepitope_f* in this same package **AnalysisHLA**

Value

Given a data frame with patients, and for each patient a vector of the OLPs to which it responds, this function retrieves a matrix having the patients in rows and the HLA alleles in columns. Then, cells in the matrix are filled with 0 or 1 according the absence of response or positive response to a given overlapping sequence.

Author(s)

Bruna Oriol

See Also

Hepitope_f, PatientsHepitop_f, Pept_Seq, CohortData

Examples

```
## The function is currently defined as
function (olp_rawdata)
{
    colnames(olp_rawdata) <- c("PatientID, OLP")
    OLPs <- NULL
    for (i in 1:length(olp_rawdata$PatientID)) {
        patient <- olp_rawdata[i, 2]
        patient <- unlist(patient)
        for (h in 1:length(patient)) {
            olp <- patient[h]
            OLPs <- c(OLPs, olp)
        }
    }
    OLPs <- unique(OLPs)
    OLPs <- sort(as.numeric(as.character(OLPs)))
    OLPs <- unique(OLPs)
    matrixOLP <- matrix(ncol = length(OLPs), nrow = length(olp_rawdata$PatientID))
    rownames(matrixOLP) <- olp_rawdata$PatientID
    colnames(matrixOLP) <- OLPs
    for (j in 1:length(olp_rawdata$PatientID)) {
        epitop <- NULL
        patient <- olp_rawdata[j, 2]
        patient <- unlist(patient)
        for (k in 1:length(OLPs)) {
            ifelse(OLPs[k] %in% patient, epitop <- c(epitop,
                1), epitop <- c(epitop, 0))
        }
        matrixOLP[j, ] <- as.matrix(epitop)
    }
    return(matrixOLP)
  }



## EXAMPLE
data(CohortData)
olp_rawdata <- CohortData[, c(1,ncol(CohortData))]
olp_rawdata$OLPs <- as.character(olp_rawdata$OLPs)
olp_rawdata$OLPs <-strsplit(olp_rawdata$OLPs, ";")

matrixOLP_f(olp_rawdata)
```

PatientsHepitop_f *Patients Responding to OLPs*

Description

This function retrieves a data.frame with the following columns: OLP, PatientID and the HLA haplotype for each patient.

Usage

```
PatientsHepitop_f(hepnames, HLA, matrixOLP, num_isotypes)
```

Arguments

hepnames	Vector of class *factor* containing all the OLPs that give a response in the studied cohort.
HLA	It corresponds to the initial cohort data frame (see toy example CohortData). It contains a column with the PatientID, the HIV Status of the patient, the HLA-II isotype, the viral load and its logarithm and the CD4 counts. It will be useful to extract the HLA haplotype of the patient.
matrixOLP	It is a matrix with the patients in rows and the HLA alleles in columns. The cells contain then 1 or 0 according if the patient presents or not response to a given OLP.
num_isotypes	It corresponds to the total number of HLA isotypes or locus present in the studied cohort, in the toy case there are 2.

Details

This function needs as input the *CohortData* dataset, the OLPs taken from the output of the function *Hepitope_f* and the result of joining in a single matrix the output of function *matrixOLP_f* applied on the different HLA isotypes or locus.

Value

The output of this funcion consist on a data frame containing the OLP, PatientID and the HLA haplotype for each patient that is of especial interest for making plots of the hepitope results.

Author(s)

Bruna Oriol

See Also

Pept_Seq, CohortData,Hepitope_f

Examples

```
## The function is currently defined as
function (hepnames, HLA, matrixOLP, num_isotypes)
{
    matrixOLPnames <- matrixOLP
    colnames(matrixOLPnames) <- paste0("OLP_", colnames(matrixOLPnames))
    OLPvec1 <- NULL
    patients1 <- NULL
    for (i in 1:length(hepnames)) {
        ep.num <- as.character(hepnames[i])
        mat <- matrixOLPnames[, ep.num]
```

```
        mat1 <- mat[mat == 1]
        matnam <- names(mat1)
        l.names <- length(matnam)
        patients1 <- c(patients1, matnam)
        ep.num.t <- rep(ep.num, l.names)
        OLPvec1 <- c(OLPvec1, ep.num.t)
    }
    df_Patients <- cbind(OLPvec1, patients1)
    matrix_pat <- matrix(nrow = length(df_Patients[, 1]), ncol = 1 +
        2 * num_isotypes)
    for (i in 1:length(df_Patients[, 1])) {
        pat <- as.character(df_Patients[i, 2])
        y <- subset(HLA, HLA[, 1] == pat)
        y1 <- as.character(y[, 2 + seq(num_isotypes * 2)])
        matrix_pat[i, ] <- c(pat, y1)
    }
    mpat <- as.data.frame(matrix_pat)
    PatientsOLP <- cbind(df_Patients, mpat[-1])
    colnames(PatientsOLP) <- c("OLP", "Patient ID", colnames(HLA)[2 +
        seq(num_isotypes * 2)])
    return(PatientsOLP)
  }


## EXAMPLE
# Load both sample datasets
data(CohortData)
data(Pept_Seq)

Peptides <- Pept_Seq
Peptides$Num <- paste0("OLP_", Peptides$Num)

# Apply function matrixOLP_f
olp_rawdata <- CohortData[, c(1,ncol(CohortData))]
olp_rawdata$OLPs <- as.character(olp_rawdata$OLPs)
olp_rawdata$OLPs <-strsplit(olp_rawdata$OLPs, ";")
OLPmatrix <- matrixOLP_f(olp_rawdata)
Hepitop_df <- matrixOLP_f(olp_rawdata)
Hepitop_df <- as.data.frame(Hepitop_df)
colnames(Hepitop_df) <- paste0("OLP_", colnames(Hepitop_df))


# Apply function HLAtypes and following fill.matrix to each isotype or gene and make a data frame of the resul
patients1 <- CohortData[, c(1,3,4)]
HLA_types1 <- HLAtypes(CohortData[,c(3,4)])
patients2 <- CohortData[, c(1,5,6)]
HLA_types2 <- HLAtypes(CohortData[,c(5,6)])
HLA1_table <- fill.matrix(patients1, HLA_types1)
colnames(HLA1_table) <- paste("HLA1", HLA_types1, sep = "_")
HLA2_table <- fill.matrix(patients2, HLA_types2)
colnames(HLA2_table) <- paste("HLA2", HLA_types2, sep = "_")

Hepitope_HLAII <- data.frame(cbind(HLA1_table, HLA2_table))

#Apply function LD
LDoutput <- LD(HLA1_table, HLA2_table)
```

```
# Apply  function Hepitope_f and prepare vector hepnames
HepSigRes <- Hepitope_f(Hepitop_df, Hepitope_HLAII, LDoutput, Peptides)
hepnames <- unique(HepSigRes[,1])

# Apply the current function PatientsHepitop_f
PatientsHepitop_f(hepnames, Hepitope_HLAII, OLPmatrix, 2 )
```

Pept_Seq *Overlapping Sequences (OLP) from HIV genome*

Description

Data frame with 4 variables containing the OLP number, the peptide, the subunit and the amino acids sequence

Usage

```
data("Pept_Seq")
```

Format

A data frame with 410 observations on the following 4 variables.

`Num` a character vector

`OLP` a character vector

`subunit` a character vector

`sequence` a character vector

Details

Its is important to have these 4 columns to apply function *Hepitope_f*

VL.function *Association between HLA alleles and viral load*

Description

For each HLA allele in a given HLA locus, the patients" viral load (in logarithmic scale) is divided in 2 groups: a group with the viral load for patients not presenting the studied HLA allele, and another one, with the patients with it. The viral load levels of these two groups are compared with a T-test and a Mann-Whitney test to see if there are differences between them. The problem of multiple comparisons is adressed with the false discovery rate (FDR) method.

Usage

```
VL.function(HLA_types, tabProg)
```

Arguments

HLA_types	It corresponds to the output of the function *HLAtypes* and it is a vector containing all the possible alleles for a given HLA isotype or locus.
tabProg	Output function *fill.matrix* converted into a data frame (Patients in rows and alleles in columns) with additional columns of information correspoding to viral load or CD4 counts. It is important to give names to the columns of the data frame: the alleles names or the information each column contains. Viral load and CD4 counts columns name **must** be *VL* and *CD4* respectively, otherwise tehy will not be recognised for *VL.function* or *CD4.function*

Details

The present function *VL.function* calls the functions *t.test*, *wilcox.test*, *median*, *mean* and *p.adjust* from the package **stats**.

Value

The output of this function is a data frame containing the different HLA alleles of a given HLA locus. The 2nd and 3rd column contain the mean for the logarithm of the viral load value of patients without or with the selected allele respectively. The 4th column contains the p-value for the T-test, and in the 5th column, there is this p-value corrected by FDR (False Discovery Rate) method. Analogously, the 6th and 7th columns contain the median of the 2 viral load level groups (individual not having or having the allele), and the 8th and the 9th, the p-value and adjusted p-value from the Wilcox test.

Note

Test if data fulfil normality assumption with function *shapiro.test* from **stats** to decide between T-test or Wilcoxon - Test results.

Author(s)

Bruna Oriol

See Also

CohortData, HLAtypes, fill.matrix

Examples

```
## The function is currently defined as
function (HLA_types, tabProg)
{
    pval <- NULL
    pval_mw <- NULL
    mean0 <- NULL
    median0 <- NULL
    mean1 <- NULL
    median1 <- NULL
    hla_class <- NULL
    for (i in 1:length(HLA_types)) {
        dqa <- colnames(tabProg)[i]
        tabProg[, i] <- factor(tabProg[, i], levels = c(0, 1))
        hla_class <- c(hla_class, dqa)
```

```
        if (sum(tabProg[, i] == 1) > 1) {
            pv <- t.test(as.numeric(as.character(tabProg$VL)) ~
                tabProg[, i])$p.value
            pvw <- wilcox.test(as.numeric(as.character(tabProg$VL)) ~
                tabProg[, i])$p.value
            m0 <- subset(tabProg, tabProg[, i] == 0)
            m0 <- mean(as.numeric(as.character(m0$VL)), na.rm = TRUE)
            m1 <- subset(tabProg, tabProg[, i] == 1)
            m1 <- mean(as.numeric(as.character(m1$VL)), na.rm = TRUE)
            med0 <- subset(tabProg, tabProg[, i] == 0)
            med0 <- median(as.numeric(as.character(med0$VL)),
                na.rm = TRUE)
            med1 <- subset(tabProg, tabProg[, i] == 1)
            med1 <- median(as.numeric(as.character(med1$VL)),
                na.rm = TRUE)
            mean0 <- c(mean0, m0)
            mean1 <- c(mean1, m1)
            median0 <- c(median0, med0)
            median1 <- c(median1, med1)
            pval <- c(pval, pv)
            pval_mw <- c(pval_mw, pvw)
        }
        else {
            mean0 <- c(mean0, "NA")
            mean1 <- c(mean1, "NA")
            median0 <- c(median0, "NA")
            median1 <- c(median1, "NA")
            pval <- c(pval, "NA")
            pval_mw <- c(pval_mw, "NA")
        }
    }
    p.adj <- p.adjust(pval, method = "fdr")
    p.adj_mw <- p.adjust(pval_mw, method = "fdr")
    dd_VL <- cbind(hla_class, mean0, mean1, pval, p.adj, median0,
        median1, pval_mw, p.adj_mw)
    return(dd_VL)
  }



## EXAMPLES
data(CohortData)
patients <- CohortData[, c(1,3,4)]
HLA_types <- HLAtypes(CohortData[,c(3,4)])
table01 <- fill.matrix(patients, HLA_types)
colnames(table01) <- paste("HLA1", HLA_types, sep = "_")
tabProg <- cbind(table01, CohortData$logVL, CohortData$CD4 )
tabProg <- data.frame(tabProg, stringsAsFactors = FALSE)
colnames(tabProg) <- c(colnames(table01), "VL", "CD4")
VL.function(HLA_types, tabProg)
```

Index