

Rodrigo dos Santos Francisco

# **Unidades de Seleção nos Genes**

## ***HLA***

Units of Selection in *HLA* Genes

Versão corrigida, o original encontra-se disponível no  
Instituto de Biociências

São Paulo

2013

Rodrigo dos Santos Francisco

# **Unidades de Seleção nos Genes**

## ***HLA***

Units of Selection in *HLA* Genes

Tese apresentada ao Instituto de  
Biociências da Universidade de São  
Paulo, para a obtenção de Título de  
Doutor em Ciências, na Área de  
Genética.

Orientador(a): Diogo Meyer

São Paulo

2013

# Ficha Catalográfica

Francisco, Rodrigo dos Santos  
Unidades de Seleção nos Genes HLA- 97  
páginas

Tese (Doutorado) - Instituto de  
Biociências da Universidade de São Paulo.  
Departamento de Genética e Biologia Evolutiva.

1. MHC 2. Evolução 3. Seleção.  
Universidade de São Paulo. Instituto de  
Biociências. Departamento de Genética e Biologia  
Evolutiva.

## Comissão Julgadora:

---

Prof(a) Dr (a)

---

Prof(a) Dr(a)

---

Prof(a) Dr(a)

---

Prof(a) Dr(a)

---

Prof. Dr. Diogo Meyer  
Orientador

# Sumário

Introdução Geral .....	1
1.1. Genes <i>HLA</i> : Estrutura, Função e Evidências da atuação da Seleção Natural. ....	2
Capítulo 1: A classificação em supertipos e a seleção sobre aminoácidos específicos da fenda apresentadora de peptídeos dos genes HLA de classe I clássicos.....	7
1. Introduction.....	7
2. Materials and Methods .....	12
2.1. Population data.....	12
2.2. Supertype definition .....	13
2.3.1. Molecular Analysis .....	14
2.3.2. Test for supertype differentiation between populations .....	15
3. Results and Discussion.....	16
3.1. Supertype frequencies and differentiation .....	16
3.2. Heterozygosity, interpopulation differentiation and correlation with geography among alleles and supertypes.....	23
3.3. Simulation results. ....	28
4. Conclusions .....	32
Capítulo 2: Evolution of HLA genes in the American continent. ....	34
1. Introduction.....	34
2- Methods and Materials .....	37
2.1- Population sampling.....	37
2.2- Microsatellite dataset. ....	38
2.3- <i>HLA</i> typing.....	38
2.4- Data Analysis.....	40
3- Results And Discussion.....	41
3.1. Deviations from Hardy-Weinberg proportions. ....	41
3.2. Allele frequency distributions of <i>HLA</i> loci.....	42
3.3. Correlation between diversity indexes and the geographic distance from the Bering Strait.....	48
3.4. Neutrality and Equilibrium Test .....	53
4- Conclusions .....	55
Considerações Finais .....	70

## RESUMO

Os genes *HLA* (Antígenos leucocitários humanos) estão localizados no Complexo Principal de Histocompatibilidade humano (o MHC), e possuem os maiores níveis de variação do genoma, com milhares de alelos, altas taxas de heterozigose e diversidade nucleotídica. No presente estudo, nosso objetivo foi a identificação dos principais alvos da seleção natural nos genes *HLA*. Para isso, propusemos duas abordagens que resultaram na redação de dois manuscritos. Na primeira abordagem, nós testamos a hipótese de que os principais alvos da atuação da seleção natural nas moléculas HLA seriam os aminoácidos que compõem os sítios que ancoram os peptídeos antigênicos (os bolsões B e F da região de ligação de peptídeos (PBR)). Para isso, utilizamos um conjunto de dados de 6.435 e 6.409 indivíduos genotipados para os genes *HLA-A* e *-B* respectivamente, gerados para o 13º Workshop Internacional de Histocompatibilidade (IHW) e pertencentes a 55 populações espalhadas por todos os continentes. Nós estimamos a diversidade nucleotídica ( $\pi$ ) das sequências que codificam para os bolsões B e F e comparamos esses dados com os obtidos de outros bolsões da PBR. Concomitantemente, utilizamos a classificação de alelos dos locos *HLA-A* e *-B* em supertipos, que são agrupamentos alélicos com similaridades nos perfis de ligação de peptídeos, devido a semelhanças em aminoácidos específicos dos bolsões B e F. Nós descrevemos os padrões observados de variação dos supertipos e desenvolvemos um teste de hipótese onde comparamos os estimadores observados de diferenciação populacional ( $G_{st}$ ) e diversidade genética (taxa de heterozigose ( $H_e$ ) e número de alelos ( $k$ )) com os obtidos a partir de 10.000 réplicas constituídas por agrupamentos aleatórios de alelos.

O bolsão B foi a região que apresentou os maiores níveis de diversidade no gene *HLA-B* ( $p < 0,00001$ , teste de soma de ranques de Mann-Whitney) e boa parte de sua variação está estruturada entre os supertipos desse mesmo loco. Além disso, os padrões observados de variação nos supertipos de *HLA-B* não foram reproduzidos pelos agrupamentos aleatórios de alelos (com 98 % das simulações apresentando valores de  $G_{st}$  menores do que os observados, utilizando as amostras africanas, europeias e asiáticas). Esse resultado indicou que os supertipos e consequentemente as especificidades do bolsão B estão significativamente estruturadas entre as populações, um indicativo de adaptações locais aos patógenos específicos de diferentes regiões geográficas. Esses mesmos padrões não foram reproduzidos na análise do loco *HLA-A*, pois boa parte da variação no PBR desse gene não está localizada nos bolsões B e F. Além disso, as simulações envolvendo os supertipos de *HLA-A* reproduziram mais frequentemente os padrões observados de variação, indicando que os bolsões B e F não são os principais alvos da seleção nesse gene, ou que os níveis de seleção em *HLA-A* sejam menores dos atuantes em *HLA-B*.

Na segunda abordagem, o nosso principal objetivo foi a identificação de genes que contribuíram para a adaptação local das populações nativas das Américas, que teria ocorrido durante o recente processo de colonização desse continente. Nós sequenciamos os exons 2 e 3 dos loci de classe I *HLA-B* e *-C* e o exon 2 do loco de classe II *-DRB1* em 635, 524 e 568 indivíduos, respectivamente, pertencentes a 32 populações nativas do continente Americano. Os dados de sequência foram utilizados na estimativa das frequências alélicas, taxa de heterozigose ( $H_e$ ), grau de compartilhamento de alelos entre populações (medido pela distância de Prevosti) e desvios de neutralidade utilizando o teste D de Tajima. Nós também comparamos os padrões de variação das taxas de heterozigose obtidas a partir dos loci *HLA* ao longo do continente com os obtidos a partir de um conjunto de 61 microssatélites espalhados ao longo do genoma, permitindo-nos a diferenciação dos padrões provavelmente gerados pela história demográfica ou seleção natural.

O loco *HLA-B* apresentou o maior número de pares de populações em que não observamos compartilhamento de alelos (44 pares contra 4 e 6 para os loci *HLA-C* e *-DRB1*, respectivamente) sendo que a região leste da América do Sul (SAE) foi a que apresentou os menores níveis de compartilhamento de alelos com outras regiões das Américas (39 dos 44 pares de populações continham uma população SAE). Essa maior diferenciação do gene *HLA-B* nas populações SAE é uma consequência da presença de alelos exclusivos dessa região, originados por eventos de conversão genica e/ou recombinação envolvendo alelos presentes em outras regiões do continente. As populações SAE também apresentaram níveis elevados de variação para o gene *HLA-B*, resultado evidenciado pela falta de correlação entre a diminuição da taxa de heterozigose e o aumento da distância em relação ao Estreito de Bering ( $r^2 = -0,1117$ ,  $p > 0,05$ ), o que contrasta com a tendência geral observada nos microssatélites e genes *HLA-C* e *-DRB1* ( $r^2 = -0,1957$ ,  $-0,2261$  e  $-0,2637$ , respectivamente ( $p < 0,05$ )). Finalizando, as populações SAE apresentaram valores de *D* de Tajima maiores ( $p < 0,001$ , teste de soma de ranques de Mann-Whitney) e mais significativos ( $p < 0,0000005$ , aplicando um teste binomial exato) no loco *HLA-B*, quando comparadas às populações das outras regiões. Essas diferenças entre regiões geográficas não foram observadas nos genes *HLA-C* e *-DRB1*, corroborando a explicação seletiva para o aumento da frequência dos alelos de *HLA-B* originados por conversão gênica/recombinação em resposta aos novos desafios ambientais das regiões tropicais na América do Sul.

As conclusões obtidas a partir de ambas as abordagens do presente trabalho apontam o gene *HLA-B* como o principal alvo da seleção natural, uma vez que esse loco concentra as maiores evidências de atuação de seleção natural recente quando comparado aos demais genes *HLA* analisados. Nós também demonstramos com as análises intragênicas que o bolsão B do PBR de *HLA-B* concentra por boa parte das diferenças observadas entre as populações, implicando em diferenças nos perfis de apresentação de peptídeos entre essas mesmas populações, o que pode ser interpretado como um indicativo de adaptações locais aos conjuntos de patógenos presentes em distintas regiões geográficas.

## ABSTRACT

The Classical *HLA* genes (Human Leucocyte Antigens) are located in the human Major Histocompatibility Complex (the MHC) and present the highest levels of variation on the Human genome, with thousands of alleles associated with high levels of heterozygosity and nucleotide diversity. In the present study, our goal was the identification of the main targets of natural selection on the *HLA* genes. We proposed two different approaches to address this issue resulting in two manuscripts. At the first approach, we performed an intragenic analysis, verifying if the amino acids at the peptide-binding region (PBR) anchor positions (the B and F pockets) exhibit higher evidences of evolution under natural selection when compared with the remaining regions of the HLA molecules. To do so, we used a dataset generated for the 13<sup>th</sup> International Histocompatibility Workshop (IHW), composed by 6,435 and 6,409 individuals genotyped for *HLA-A* and *-B* respectively, belonging to 55 populations scattered along all the continents. We measured the levels of nucleotide diversity ( $\pi$ ) of the sequences coding for the B and F pockets and compared them with the remaining PBR pockets. Concomitantly, we applied the supertype classification which consists in groups of *HLA-A* and *-B* alleles which bind overlapping sets of peptides, as a consequence of sharing specific amino acids at B and F pockets and described the patterns of supertype variation in the observed data. Next, we developed a hypothesis test in which the observed patterns of population differentiation ( $G_{st}$ ) and variability (heterozygosity ( $H_e$ ) and number of alleles ( $k$ )), using the supertype definition, were compared with 10,000 replicates of random assigned groups of alleles.

At the *HLA-B* locus, the B pocket presented the highest levels of variation ( $p < 0.00001$ , Wilcoxon rank sum test) and concentrated most of the differences between superotypes. Our simulations results revealed that the reassignment of alleles into random groups could not reproduce the observed patterns of population differentiation (with 98% of the simulations presenting  $G_{st}$  values smaller than the observed, using the African, European and Asiatic samples), indicating that superotypes and more specifically the B pocket specificities are significantly structured among populations, which could be an indicative of adaptations to local pathogens. We did not observe the same patterns at the *HLA-A* locus which presented relative lower levels of variation at B and F pockets when compared with the remaining PBR regions, and simulated values of  $G_{st}$  and  $H_e$  which often reproduced the observed data.

At the second approach, our main objective was the identification of genes that contributed for local adaptation on Native American Populations because of the relatively recent colonization of the new American environments. We sequenced the exons 2 and 3 of the *HLA-B* and *-C* class I loci and the exon 2 of the *-DRB1* class II locus in 635, 524 and 568 individuals, respectively, belonging to 32 different Native American Populations scattered along all the Americas. We estimated the allele frequencies, expected heterozygosity ( $H_e$ ), degree of allelic sharing between populations (measured by the Prevost's Distance) and departure from neutral expectation using the Ewens-Watterson (EW) and Tajima's D test. Concomitantly, we used a dataset of 61 microsatellites scattered along the genome as a demographic control, comparing the degree of variation of the heterozygosity along the continent.

The *HLA-B* locus showed the highest number of pairs of populations in which we did not observe any sharing of alleles (44 pairs against 4 and 6 for *HLA-C* and *-DRB1* loci, respectively) and the Eastern South American (SAE) region was the one presenting the smallest levels of allelic sharing with other American regions at this locus (39 out the 44 pair of populations contained a SAE population). The presence of exclusive gene conversion and/or recombination alleles accounts for the higher differentiation of SAE populations at the *-B* locus. The *-B* locus also exhibited a higher level of variation at SAE populations which was evidenced

by the lack of correlation between the decrease of variation and the increase of distance from Bering Strait ( $r^2 = -0.1117$ ,  $p > 0.05$ ), a general trend observed at the microsatellites marker and *HLA-C* and *DRB1* loci ( $r^2 = -0.1957$ ,  $-0.2261$  and  $-0.2637$ , respectively ( $p < 0.05$ )). Finally, The SAE populations presented the highest ( $p < 0.001$ , Wilcoxon rank sum test) and most significant ( $p < 0.0000005$ , applying an exact binomial test) positive Tajima's D values for the *-B* locus when compared with the other regions. We did not observe such differences between geographic regions at the *HLA-C* and *-DRB1* loci, corroborating the adaptive explanation for the increase on the frequency of the *HLA-B* gene conversion/recombination alleles as a response for the new environmental challengers of the tropical regions in South America.

The combined conclusions obtained from both approaches indicate that *HLA-B* locus seems to be the main target of natural selection, collecting the highest evidences of recent evolution when compared with the other *HLA* loci analyzed in the present study. We also demonstrated with the intragenic analysis that at *HLA-B* PBR, the B pocket seems to be the portion of the molecule whose variation accounts for most of the differences observed between populations, which can be direct associated with the immunological function of *HLA-B*, since this pocket was one of the more important containments for the peptide presentation.

## Introdução Geral

O entendimento da contribuição relativa dos processos estocásticos e da seleção natural na evolução e manutenção da variação genética é um dos principais objetivos da pesquisa em biologia evolutiva (Piertney and Oliver 2006). Dentre os *loci* que apresentam sinais de evolução por seleção natural, podemos citar os genes *HLA* (*Human Leukocyte Antigen*), que apresentam níveis de diversidade incompatíveis com o esperado de acordo com o modelo de evolução neutra em humanos. Além das evidências populacionais, os resultados de estudos de associação com doenças infecciosas indicam que os genes *HLA* estão evoluindo sob seleção natural dirigida por patógenos (Meyer and Thomson 2001).

Apesar do grande número de estudos, ainda há debate a respeito de diversos aspectos do processo seletivo que molda a variação dos genes *HLA*. Nosso objetivo no presente trabalho foi entender qual ou quais foram os principais alvos da atuação da seleção natural dentro desse sistema gênico. Alvos plausíveis incluem desde aminoácidos específicos, até genes inteiros. Esses dois níveis de análise, o intra e intergênico, motivaram a organização do presente trabalho em dois capítulos no formato de manuscritos. Porém, antes de entrarmos nas questões específicas exploradas em cada um desses capítulos, nós faremos uma breve contextualização do tema, descrevendo a estrutura e função dos genes *HLA* e apresentando as evidências de que esses genes estão evoluindo sob a ação da seleção natural.

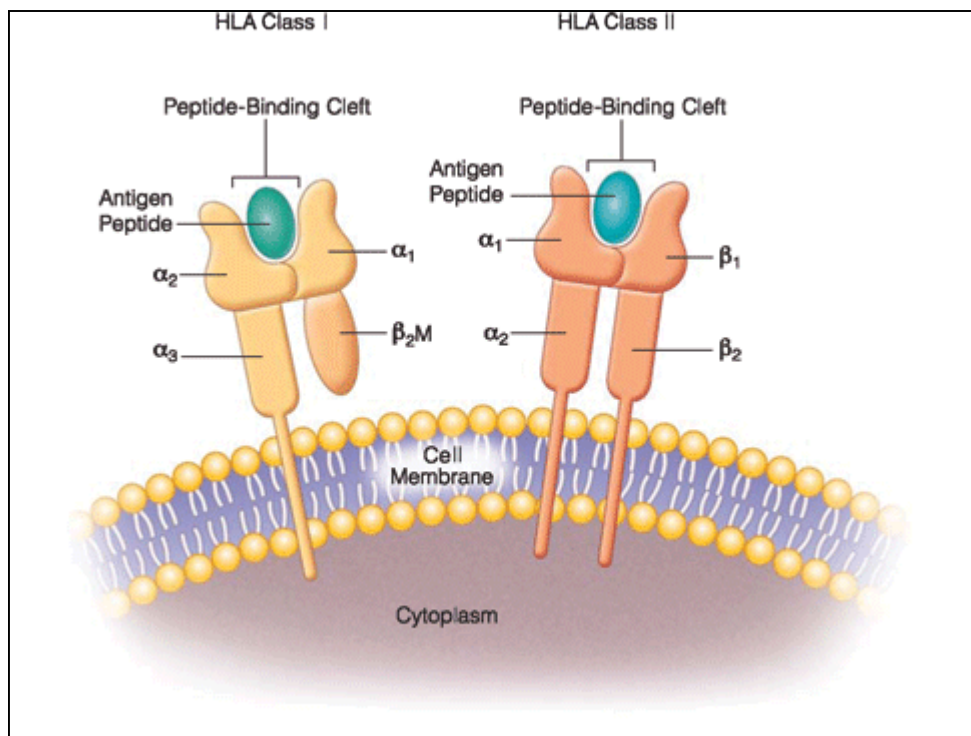
### 1.1. Genes *HLA*: Estrutura, Função e Evidências da atuação da Seleção Natural.

Os genes HLA encontram-se no MHC (o Complexo Principal de Histocompatibilidade), localizado no braço curto do cromossomo 6 (6p21.3) (Figura 1). O MHC é uma das regiões mais bem conhecidas do genoma humano, apresentando altos níveis de diversidade e uma alta concentração gênica (The MHC sequencing consortium 1999). Esse complexo é organizado em três regiões, definidas de acordo com a função dos genes localizados em cada uma delas: classe I, II e III.

Na região de classe I estão localizados os genes *HLA-A*, *-B* e *-C*, também conhecidos como genes *HLA* de classe I clássicos, que codificam respectivamente as moléculas HLA-A, -B e -C. As moléculas HLA de classe I são formadas por duas subunidades: a cadeia  $\alpha$  e a  $\beta$ 2-microglobulina. A cadeia  $\alpha$  é codificada pelos genes HLA de classe I e apresenta três domínios extracelulares:  $\alpha$ 1,  $\alpha$ 2 e  $\alpha$ 3. Os domínios  $\alpha$ 1 e  $\alpha$ 2 formam a fenda de ligação de peptídeos (Saper, Bjorkman, and Wiley 1991).

Na região MHC de classe II estão localizados os genes *HLA-DPA1* e *-DPB1* que codificam a molécula HLA-DP; os genes *HLA-DQA1* e *-DQB1* que codificam a molécula HLA-DQ; e os genes *HLA-DRA* e o complexo dos genes *-DRB* (*HLA-DRB1*, *-DRB3*, *-DRB4* e *-DRB5*) que codificam as moléculas HLA-DR. As moléculas de classe II são também formadas por duas subunidades: as cadeias  $\alpha$  e  $\beta$ , entretanto, diferentemente das moléculas de classe I, ambas as subunidades são codificadas por diferentes genes *HLA* (ex: as cadeias  $\alpha$  e  $\beta$  da molécula HLA-DRB1 são codificadas pelos genes *HLA-DRA* e *HLA-DRB1*, respectivamente. Cada uma das subunidades das moléculas de classe II apresentam dois domínios: a cadeia  $\alpha$  com  $\alpha$ 1 e  $\alpha$ 2; e a cadeia  $\beta$  com  $\beta$ 1 e  $\beta$ 2. Os domínios  $\alpha$ 1 e  $\beta$ 1 formam a fenda de ligação do peptídeo (Figura 2).

**Figura 1. Mapa esquemático do MHC humano:** as caixas indicam os genes que foram estudados no presente trabalho *HLA-A*, *-B* e *-C* e *-DRB*. Fonte: (The MHC sequencing consortium 1999).



**Figura 2.** Imagem ilustrando as estruturas das moléculas HLA de classe I e classe II: (veja descrição no texto).

Inicialmente, acreditava-se que a função dos genes *HLA* estava correlacionada apenas com a compatibilidade tecidual observada durante transplantes (PONTES e PORTO 2007). A verdadeira função fisiológica desses genes somente foi compreendida posteriormente: as moléculas codificadas pelos genes *HLA-A*, *-B* e *-C* apresentam peptídeos citoplasmáticos na superfície da maioria das células nucleadas aos linfócitos T CD8<sup>+</sup> (Doherty and Zinkernagel 1975) e também são ligantes dos receptores KIR (Killer cell immunoglobulin-like receptors) expressos pelas células natural killers, ou NK (Parham 2005). Já as moléculas de classe II são constitutivamente expressas pelas chamadas células apresentadoras de antígeno profissionais (APC) (Watts 1997; Varney, Gavrilidis e Tait 1999) e apresentam peptídeos oriundos do exterior das células (que foram previamente fagocitados ou pinocitados) aos linfócitos T CD4<sup>+</sup> (Varney et al., 1999).

Normalmente as moléculas HLA são expressas associadas a peptídeos oriundos de proteínas próprias. Porém, na presença de proteínas não próprias, moléculas HLA com peptídeos antigênicos são levadas à superfície celular onde são reconhecidas por linfócitos T (Abbas e Lichtman 2005). A sobrevivência dos patógenos está atrelada à presença de variantes antigênicas que permitem a evasão da apresentação pelas moléculas HLA, enquanto que a sobrevivência dos hospedeiros está atrelada à capacidade de apresentação de antígenos dos patógenos.

A consequência do elo evolutivo entre hospedeiros e patógenos faria com que as moléculas HLA e antígenos coevoluiam. Os patógenos tenderiam a acumular mutações de escape, que são aquelas que evitam a apresentação de seus peptídeos pelas moléculas HLA mais comuns na população de hospedeiros. Por outro lado, novas moléculas HLA, capazes de apresentar de forma eficiente essas mutações de escape, seriam adaptativas e aumentariam de frequência na população de hospedeiros, recomeçando o ciclo. Portanto, os microrganismos patogênicos seriam a fonte da seleção atuante nos genes HLA. Essa hipótese evolutiva é conhecida como seleção dirigida por patógenos (Gillespie 1977).

Simulações computacionais mostraram que a seleção dirigida por patógenos explica de maneira satisfatória várias das características observadas nos dados reais para genes *HLA*, tal como o grande número de alelos, altas taxas de heterozigose e longa persistência dos alelos nas populações ao longo das gerações (Borghans, Beltman, and De Boer 2004). Esse mecanismo evolutivo também seria capaz de explicar as altas taxas de identidade por descendência descritas para genes os genes *HLA* (Albrechtsen, Moltke e Nielsen 2010).

Estudos de associação com doenças infecciosas corroboram a hipótese de seleção dirigida por patógenos. Por exemplo, foram descritas associações entre alelos *HLA* e resistência à malária em populações do oeste africano (Hill et al. 1997). Alelos de *HLA-B* foram associados

a uma maior ou menor progressão à infecção por HIV (Carrington et al. 1999; Kaslow et al. 1996). Além disso, Prugnolle et al. (2007), Qutob et al. (2011) e Sanchez-Mazas et al. (2012) mostraram que populações localizadas em regiões com grande diversidade de patógenos, possuem em média, uma maior diversidade nos loci *HLA*.

Todos esses resultados indicam que os genes *HLA* estão evoluindo sob ação da seleção natural e que a provável fonte da pressão seletiva atuante nesses loci vem da interação com os patógenos presentes no ambiente.

Baseando-se nessas premissas nós desenvolvemos as duas principais questões que motivaram o desenvolvimento do presente trabalho:

- 1) As regiões das moléculas HLA que interagem de forma direta com os antígenos patogênicos foram os alvos principais da atuação da seleção natural?
- 2) Quais foram os genes *HLA* mais impactados durante a ocupação de novos ambientes pelas populações humanas?

# Capítulo 1: A classificação em supertipos e a seleção sobre aminoácidos específicos da fenda apresentadora de peptídeos dos genes HLA de classe I clássicos

## 1. Introduction

The *HLA* (Human Leukocyte Antigen) classical class I genes are extremely polymorphic, with thousands of alleles, most of them coding for different proteins (1,612, 2,211 and 1,280 non-synonymous variants for *HLA-A*, *-B* and *-C loci*, respectively, (Robinson et al. 2011). *HLA* class I genes play a central role in the immune response, presenting processed peptides derived from proteins present in the intracellular environment (including foreign ones, derived from intracellular parasites such as viruses and some bacteria) to cytotoxic T lymphocytes, and also functioning as ligands for the *killer Immunoglobulin-like receptor (KIR)* of natural killer cells (Parham 2005).

Almost all of the *HLA* class I polymorphism is clustered in exons 2 and 3 that code for the  $\alpha 1$  and  $\alpha 2$  extracellular domains, which form a groove-like structure known as the peptide binding region (PBR), which engages the peptides (Saper, Bjorkman, and Wiley 1991). The PBR codons show several striking features regarding their diversity, including high heterozygosity levels (Parham et al. 1989; Lawlor et al. 1990; Hedrick et al. 1991) and high rates of non-synonymous substitutions (Hughes and Nei 1988, Takahata et al. 1992), contrasting with neutral expectations and supporting the hypothesis that balancing selection has kept variation at the sites of antigen recognition.

The high levels of variation of the sites involved in the peptide binding led to the proposition of the host-pathogen coevolution hypothesis (Apanius et al. 1997), which states that

the pathogenic microorganisms were the main evolutionary force shaping *HLA* variation (Gillespie 1977). Supporting this hypothesis, several studies have demonstrated a positive correlation between the diversity levels of both *HLA* genes and environmental pathogens (Prugnolle et al., 2005; Qutob et al. 2011; Sanchez-Mazas et al. 2012).

These results corroborates the idea that the codons making up the PBR constitute the main target of balancing selection within *HLA* genes, but the analyses performed to date generally treat this region as a homogeneous block. The PBR is composed by six pocket-like structures (A, B, C, D, E and F). Each pocket accommodates one of the nine amino acid residues of the bound peptide (1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup>, 6<sup>th</sup>, 7<sup>th</sup> and 9<sup>th</sup>, respectively)(Saper, Bjorkman, and Wiley 1991) and the binding affinity between an HLA molecule and a specific peptide is a function of the chemical properties provided by each of the PBR pockets (Saper, Bjorkman, and Wiley 1991).

The strongest interaction between the HLA molecules and the bound peptide is accounted by the interactions between the B and F pockets and the second and ninth amino acid residues of the peptide, respectively (Saper, Bjorkman, and Wiley 1991). As the amino acids composing B and F pockets play a central role in the peptide recognition by the HLA molecules, Sidney et al. (1996) proposed the supertype classification of *HLA* alleles. Supertypes are groups of alleles sharing chemical properties at the B and F pockets, and the logic behind such pooling is that alleles within supertypes would have largely overlapping peptide repertoires, and contrasts between supertypes would be those associated to important phenotypic differences.

Supertypes were originally defined by examining motifs from binding data, the sequencing of endogenously bound ligands, as well as analyses of the structure of the HLA molecules (Sette and Sidney 1999; Sidney et al. 1996, 2008). Four supertypes were described for *HLA-A* (A1, A2, A3 and A24) and five for *HLA-B* (B7, B27, B44, B58 and B62) and they were originally assigned to 31 *HLA-A* and 57 *HLA-B* alleles whose peptide binding specificities were

experimentally defined. These alleles were used to construct a reference panel for the B and F amino acid sequences. A set of 945 *HLA-A* and *-B* alleles with unknown binding specificities were checked for matches to the sequences in this panel (Sidney et al. 2008). Among the 945 previously unclassified alleles, 57% presented a full match in both B and F pockets with some allele from the reference panel, with known supertype status. Another 23.8% presented partial matches with residues found at the B and F pockets.

In line with the expectation that superotypes constitute a functionally relevant definition of *HLA* variation, several researchers have found that in association studies involving *HLA* loci it is useful to group alleles into superotypes, and test associations between the later and disease status (Chakraborty et al. 2013; Cordery et al. 2012; Gilchuk et al. 2013; Karlsson et al. 2012, 2013; Kuniholm et al. 2013; Trachtenberg et al. 2003; Xavier Eurico de Alencar et al. 2013).

On an evolutionary scale, the selection upon B and F pockets and, consequently, upon the genotype defined by the supertype, is expected to leave a visible signature. For example, under the assumption that pathogen driven selection shapes supertype frequencies, we expect genetic variation defined at the supertype level to show patterns of polymorphism and differentiation indicative of balancing selection.

The prediction that selection on supertype level variation would result in observable genetic signatures was raised by Sette and Sidney (1999), who argued, “Supertype frequencies were high and fairly conserved among different ethnicities, a pattern consistent with balancing selection acting upon supertype-level variation”. With a similar reasoning, Naugler and Liwski (2008) argued that “natural selection should favor maximization of the heterozygosity of allele superotypes instead of the heterozygosity of individual alleles”, making explicit the hypothesis that the superotypes, as defined by of B and F pocket variation, are targets of natural selection.

Such conservation of supertype frequencies between populations, and increased heterozygosity at the supertype level would lead to a pattern of low population differentiation when compared with that observed in analyses of variation and differentiation defined at the allelic level. Another expected effect of the action of balancing selection at the supertype level would be the maximization of the variation of B and F pockets compared with the remaining regions of the PBR, increasing the chances of the antigen recognition by the immune system.

The methodological challenge involved in testing these hypotheses lies in the direct comparison between differentiation and variability for genetic variation defined at the level of *HLA* alleles and superotypes. The problem arises because genetic variation at the allele level is nested within that at the supertype level (since superotypes are composed of sets of alleles) and heterozygosity at the supertype level is thus mathematically constrained to be lower or equal to that of the allele-level definition. Further, because genetic differentiation measured by statistics related to Wright's  $F_{st}$  is strongly determined by intrapopulation variability (Jost 2008), we similarly expect higher levels of differentiation at the supertype level, simply as a consequence of the decreased polymorphism associated to this level of allelic definition.

In the present study we investigate if, the supertype definition at *HLA* loci reduces interpopulation differentiation and increases heterozygosity, while controlling for the effect of the inherent difference in polymorphism associated to supertype and allele-level definitions. Our approach consisted of creating null distributions for heterozygosity and differentiation indexes by creating randomized sets of alleles (herein referred to as “random superotypes”), that match the true superotypes in sampling properties (number of superotypes, and number of alleles per supertype), but lack the biological criteria for the pooling of the alleles.

Finally, we also analyzed the supertype variation at nucleotide level, partitioning the sequences into segments corresponding to the different pockets. Our hypothesis was that

because B and F pockets ,the ones defining supertypes, are the major determinants of the peptide binding specificities, they will therefore constitute the main targets of balancing selection, showing higher levels of diversity when compared with the other pockets.

## 2. Materials and Methods

### 2.1. Population data.

We used a database generated for the 13<sup>th</sup> International Histocompatibility Workshop (IHW)(Mack et al. 2006), excluding populations which presented allelic resolution lower than two sets of digits (because the two first set of digits are necessary for the identification of the unique protein-level alleles), genotypic ambiguities and deviation from Hardy-Weinberg expectations, resulting in a dataset of 6,435 and 6,409 individuals for *HLA-A* and *-B* respectively, from 55 populations (seven Sub-Sahara African (SSA), two North African (NAF), eight South East Asian (SEA), four European (EUR), 23 South East Asian (SEA), four Pacific islanders (PAC), four aborigine Australian (AUS), two North Asian (NEA) and two Native American (AME) populations) (table 1 - Supplementary Material 1(S1)).

We tested the samples for deviation from Hardy-Weinberg equilibrium using the Gene [rate] program that estimates the fitting of the observed genotypes with the hypothesis of population equilibrium (Nunes. 2006).

Our dataset has a bias towards populations from regions that are well documented to have experienced extreme founder events, with a total of 24 out of 55 populations fitting this definition (Oceania, Taiwan and the Americas, with 7, 15 and 2 populations, respectively) which represent groups that experienced extreme founder effects, which could contribute disproportionately to the overall patterns of differentiation, reducing the visibility of global trends in reduction of diversity associated to human migrations from Africa for instance. To deal with this effect we carried out analyses for two sets of populations: (a) a complete dataset,

comprising all 55 populations; (b) and a reduced dataset of 32 populations, which excludes Oceania, Taiwan and the Americas.

## 2.2. Supertype definition

We assigned all alleles to their specific supertype using the classification described in figures 1 and 2 from Sidney et al. (2008). The alleles not assigned to any specific supertype were maintained at their allele-level definition for analyses of population differentiation and pooled into groups of non-classified alleles (NCA and NCB for *HLA-A* and *-B*, respectively) for the description of variation. We included A\*29:02, A\*29:01, A\*29:03, A\*30:01, A\*30:08 and A\*68:06 in the NCA group because of their ambiguous supertype allocation (Sidney et al. 2008). We also included B\*08 alleles in the non-classified group of *HLA-B* alleles because of their unique PBR structure which make the peptide binding profile unpredictable (Sidney et al. 2008).

## 2.3 Data analyses

We wrote R scripts for the estimation of frequencies, summary statistics (the number of alleles ( $k$ ), and sample expected heterozygosity ( $H$ )) and genetic differentiation between pairs of populations (measured by  $G_{ST}$ ) (Nei and Chesser 1983). We calculated the geographic distances between the population pairs using a great circle routes approach. The Mantel test for the significance of correlations between genetic and geographic distances was carried out using the *ade4* R package and all graphs were generated using R as well. We performed a hierarchical analysis of molecular variance (AMOVA) for each supertype, considering all other alleles as a unique group. In this way, we obtained estimates for both the diversity among populations ( $F_{ST}$ ), the diversity among populations within geographic regions ( $F_{SC}$ ) and the diversity among geographic regions ( $F_{CT}$ ) for each supertype separately. The AMOVA was implemented using

the Arlequin 3.5 package (Excoffier and Lischer 2010).

### 2.3.1. Molecular Analysis

We analyzed the molecular variation at each pocket using the coding sequences of the six (A to F) pockets which make up the *HLA* class I peptide-binding region. The definition of these codons was taken from Saper et al. (1991)(table 1.1). The B and F pockets were analyzed individually because of their central role in engaging peptides and in defining supertype. The C, D and E pockets together form the central region of the PBR and are shorter compared to other pockets, justifying pooling them for the present analysis. The A pocket was also analyzed individually because of its position at one end of the PBR.

We estimated nucleotide diversity ( $\pi$ ) ( Nei. 1987) for each population, and refer to this as “total  $\pi$ ”. We also computed the within and between supertype’s  $\pi$ , the later calculated using the following formula:

$$\pi_{st} = \frac{\pi_{total} - \pi_{within}}{\pi_{total}} \quad (1)$$

The total, within and between supertype  $\pi$  were calculated in two ways: (a) excluding the non-classified alleles or (b) including them as a single group. The use of the dataset at two digits resolution presented a challenge since it lacks information about the synonymous mutations. We addressed this problem applying the same strategy described by Buhler and Sanchez-Mazas (2011), treating as a missing data the nucleotide positions which were previously described as synonymous (Robinson et al. 2011). We excluded sites with more than 5% missing data.

**Table 1.1.** Codon compositions of the PBR pockets.

Pockets	Codons	Total size in base pairs (bp)
A	5, 7, 59, 63, 66, 99, 159, 163, 167 and 171	30 pb
B	7, 9, 45, 63, 66, 67, 70 e 99	24 pb
C, D and E	9, 70, 73, 74, 97, 99, 114, 147, 152, 155, 156, 159 and 160	39 pb
F	77, 80, 81, 84, 116, 123, 143, 146 e 147	27 pb

### 2.3.2. Test for supertype differentiation between populations

To test whether the levels of genetic differentiation among population pairs differs from those expected under the null hypothesis that supertypes are equivalent to random sets of alleles, we randomized the assignment of alleles into supertypes and calculated  $H_e$  and  $G_{st}$ . The randomized assignment of alleles into supertypes was performed using two different approaches:

- 1) We created randomized supertypes with the same number of alleles per supertype as observed in the original dataset, considering the total number of populations;
- 2) We randomized the assignment of alleles to supertypes, but did not impose any constraint on the number of alleles associated to a specific supertype, also considering the total number of populations.

The randomizations were repeated 10,000 times and p-values were estimated empirically by determining the number of randomized datasets with  $G_{st}$  values lower or  $H_e$  values higher than observed for the true data.

### 3. Results and Discussion

#### 3.1. Supertype frequencies and differentiation

In the only previous study to address differentiation at the supertype level, Sidney et al (1996) used five populations and reported a pattern in which all supertypes were present in all world regions, something not supported by our analysis. Some supertypes are completely absent in some populations and reach a frequency of more than 50% in others (figures 2A, 2D, 3 and 4). Among the *HLA-A* supertypes, A1 was the rarest, showing frequencies smaller than 9% in more than a half of the populations (figure 1.1A) and absent in five of them (figure 1.1D). A1 alleles were found with high frequencies (22% in average) in Africa, South West Asia and Europe (Figure 1.2), resulting in a significant geographic structure (with most of the variation between populations being found among geographic regions ( $F_{ct} > F_{sc}$ ; table 1.2). The A1 supertype was represented by a small number of alleles, with more than half of the populations exhibiting less than three alleles (figure 1.1B) and 14 of them showing a single one (figure 1.1C).

The A2 and A3 supertypes have a more even distribution, with half of the populations exhibiting frequencies ranging from 14 to 29% for A2 and 14 to 32% for A3 (figures 1.1A and 1.2). As consequence, A2 and A3 supertypes presented the lowest structure among *HLA-A* supertypes ( $F_{ct} < F_{sc}$ ; table 1.2). All populations presented at least one A2 allele (with eight showing just one A2 allele) but the A3 supertype was represented by a large number of alleles (figure 1.1B and 1.1C).

The A24 supertype was found in all populations (figure 1.1D), presenting frequencies ranging from 13 to 40% in half of them (figure 1.1A). Despite the broad distribution, the A24 supertype was usually represented by just two alleles, A\*23:01 and A\*24:02, with 26 and 10 populations showing just one or both of these alleles (figure 1.1B and 1.1C). This supertype was

found in higher frequencies (40% in average) in South East Asia (SEA), Pacific Islands (PAC), Australia (AUS), North East Asia (NEA) and Americas (AME) (figure 1.2). Despite the higher levels of population differentiation for this supertype ( $F_{st} = 11\%$ ,  $p < 0.0001$ ), most of the variation was found within geographic regions.

The frequencies of *HLA-A* non-classified alleles (NCA) varied greatly between populations, ranging from two to 14% in half of them (figure 1.1A). These alleles presented a large geographic structure, with 72.8% of the variation among populations being detected between geographic regions (table 1.2). The African and Australian populations were those presenting the highest frequencies for the NCA (averages of 16 and 43%, respectively) (figure 1.2).

The *HLA-B* supertypes fell into two main categories regarding the frequency distributions. On one hand, B7 and B44 showed a pattern resembling A2 and A3, with higher average frequencies (figures 1.1A and 1.3) and relatively lower levels of geographic structure (table 1.2). Half of the populations presented frequencies ranging from 18 to 31% and 21 to 32% for the B7 and B44 supertypes, respectively (figures 1.1A and 1.3). Both B7 and B44 were present in all populations (except for the Yami population, which lacks B7 alleles; figures 1.1D and 1.3), and exhibited large number of alleles per population (figure 1.1B and 1.1C).

On the other hand, B58 and B62 exhibited very low frequencies, with half of the population ranging from zero to 5.8% and 2.9 to 18%, respectively (figures 1.1A). The B62 supertype presented the highest level of population differentiation ( $F_{st} = 11.38\%$ ,  $p < 0.0001$ ; table 1.2), but without a clear geographic structure ( $F_{ct} < F_{sc}$ ; table 1.2). On the other hand, B58 presented the highest geographic structure among *-B* supertypes ( $F_{ct}=0.05594$ ; table 1.2), being found among Sub-Saharan (SSA) populations at an average frequency of 33% (ranging

from 23 to 60%, figure 1.3) against the 4.2% in the remaining regions (figure 1.3). B58 was also the supertype that was absent from the highest proportion of populations (18 of 53; figure 1.3).

The B27 supertype presented an intermediate pattern, with relatively lower frequencies (half of the populations exhibited frequencies from 7 to 19%; figure 1.1A), a higher level of population differentiation when compared with B7 and B44 ( $F_{st} = 7.5\%$ ,  $p < 0.0001$ ; table 1.2), but without geographic structure ( $F_{ct}$  very close to zero; table 1.2).

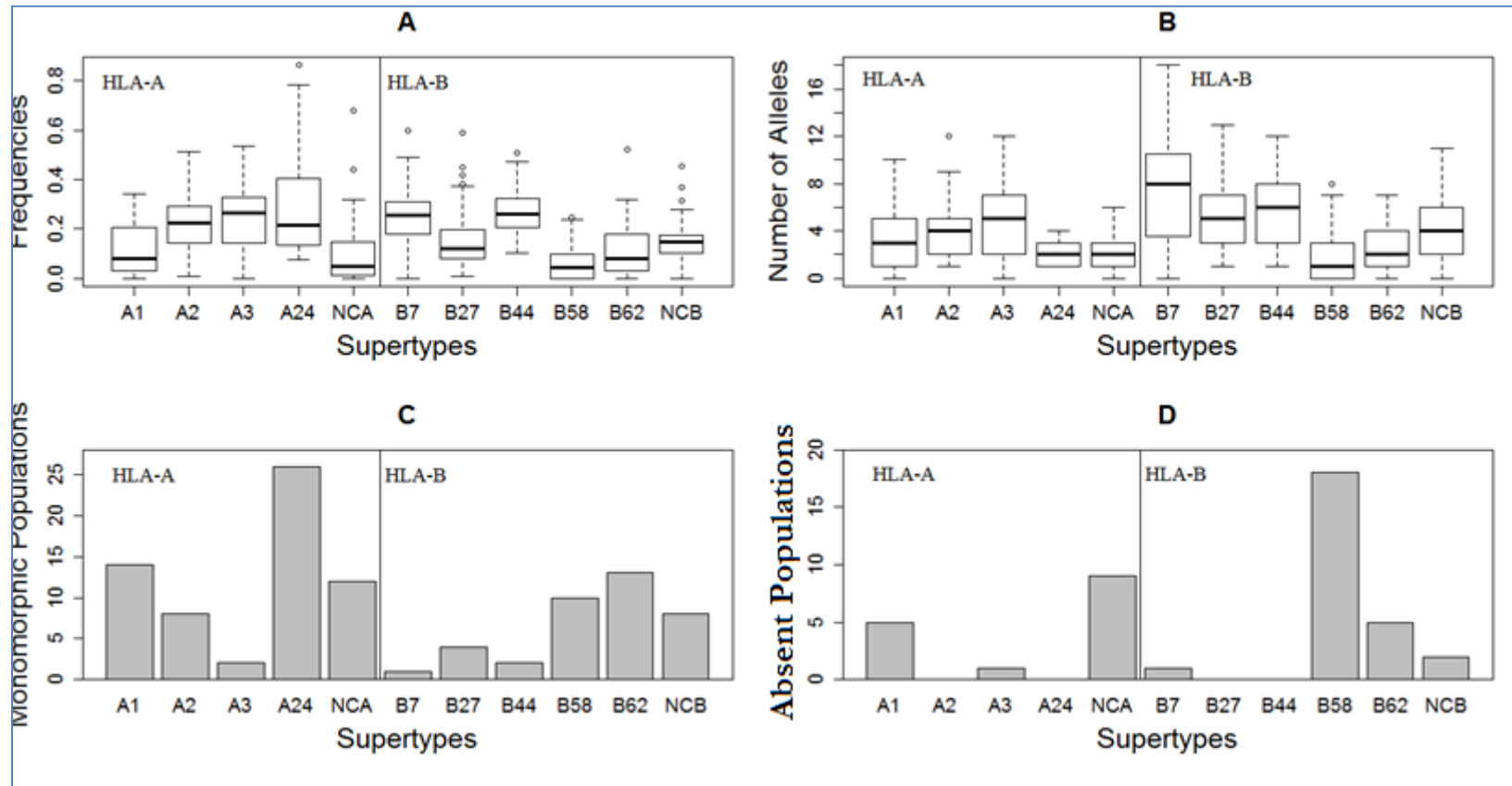
Contrasting with what was observed for the NCA, the non-classified alleles for *HLA-B* locus (NCB) were quite frequent, with half of the populations presenting frequencies ranging from 10 to 17% (figure 1.1A). More than 75% of the populations have at least two different NCBs (figure 2B) with just two populations lacking one of these alleles (figure 1.1D). We also observed a pattern of geographic structure, although it was not as strong as for NCA.

Based in the observed data we can allocate supertypes into two main categories: A2, A3, B7, B27 and B44 fits the classical view that supertypes are evenly distributed (figures 1.1A, 1.2 and 1.3 and table 1.2), poorly geographically structured (table 1.2) and represented by a large number of alleles (figure 1.1B, 1.1C and 1.1D). On the other hand, A1, A24, B58 and B62 presented a higher frequency variation among populations (figures 1.2, 1.3 and table 1.2) in some cases with a significant geographic structure (in the case of A1 and B58, both being very common inside Africa) and represented by a few number of alleles.

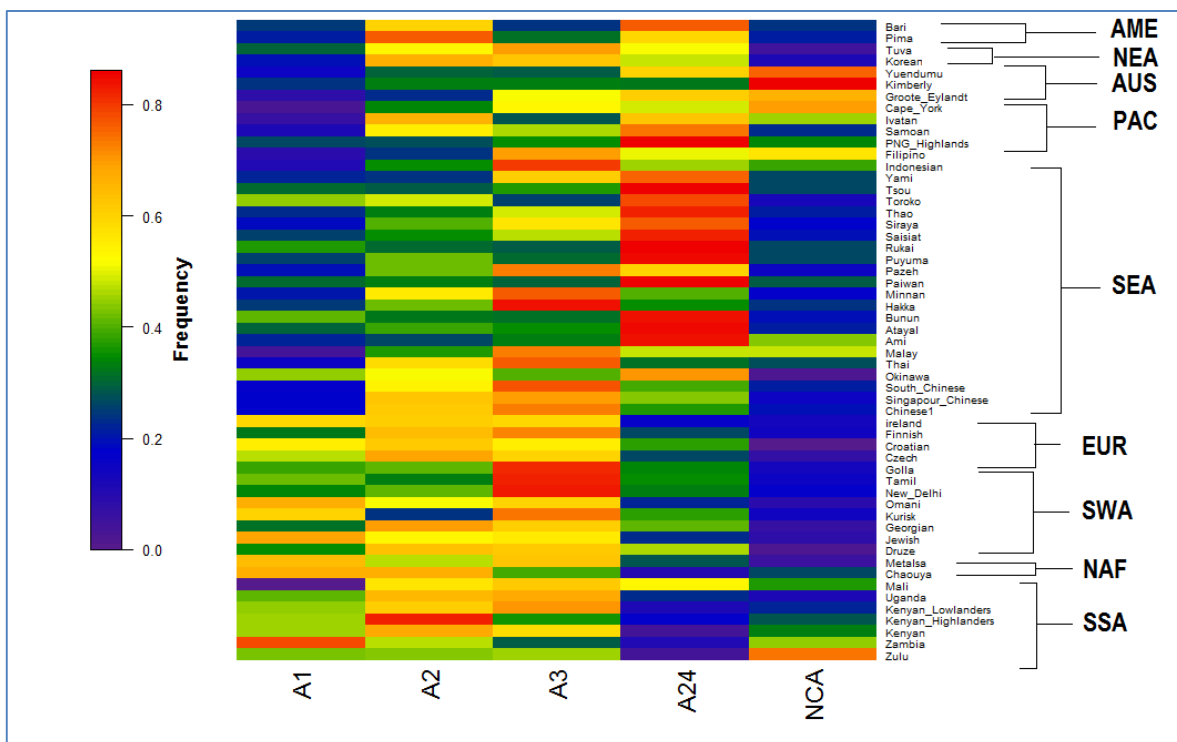
**Table 1.2.** Supertype differentiation indexes among populations (Fst), within geographic regions (Fsc) and among geographic regions (Fct).

Supertypes	Fst	Fsc	Fct
A1	0.09952***	0.02674***	0.07478***
A2	0.04851***	0.03391***	0.01511*
A3	0.06476***	0.06478***	-0.00003
A24	0.11144***	0.06662***	0.04802***
NCA	0.15897***	0.04899***	0.11565***
B7	0.05115***	0.03208***	0.01970*
B27	0.07543***	0.07105***	0.00471
B44	0.03208***	0.01725***	0.01509**
B58	0.08337***	0.02906***	0.05594**
B62	0.11382***	0.07355***	0.04347*
NCB	0.07018***	0.02902***	0.04240***

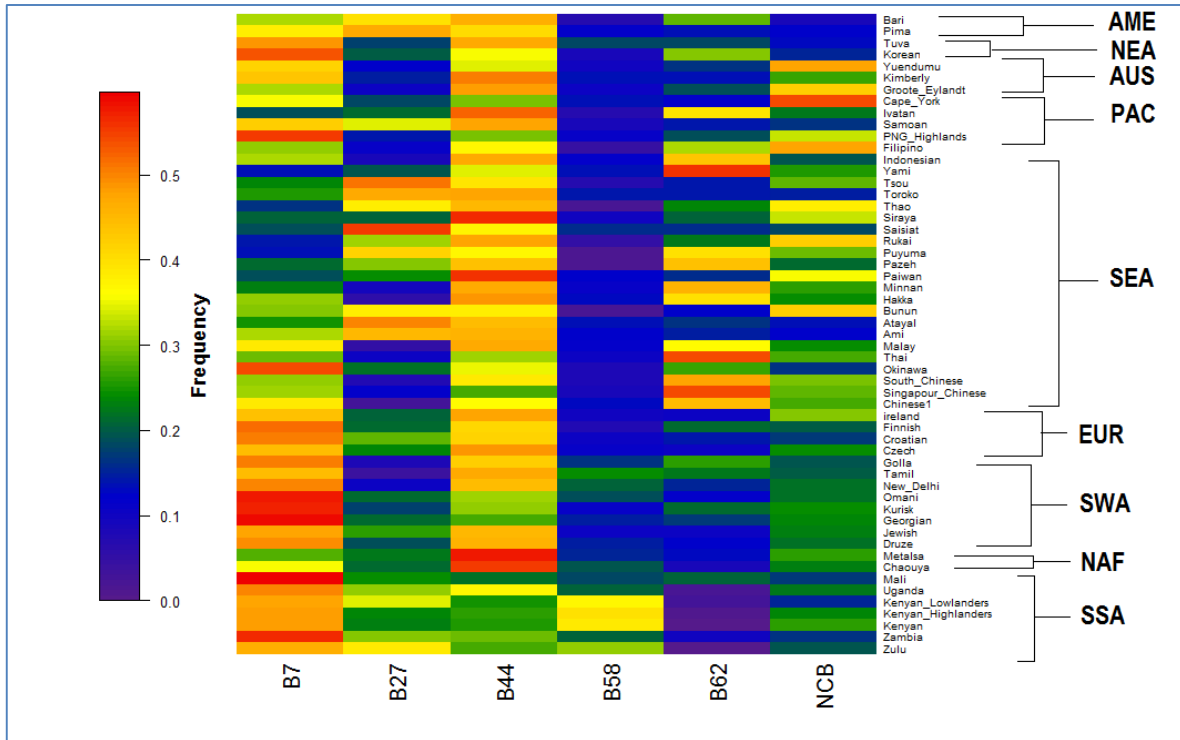
\*:  $p < 0.01$ ; \*\*:  $p < 0.001$ ; \*\*\*:  $p < 0.0001$



**Figure 1.1. Patterns of supertypes observed variation:** A. The distributions of frequencies for each of the four *HLA-A* and *-B* supertypes and the non-classified alleles; B. Number of observed individual alleles per population belonging to each supertypes; C. Number of populations showing just one allele for the referred supertype; D. Number of population lacking the referred supertype.



**Figure 1.2. *HLA-A* supertypes frequencies.** Color matrix summarizing the frequencies of the four *HLA-A* supertypes and the group of non-classified alleles. Population names are shown at the right side with the codes for populations regions: SSA, NAF, SWA, EUR, SEA, PAC, AUS, NEA and AME (Sub-Sahara Africa, North Africa, South West Asia, Europe, South East Asia, Pacific, Australia, North East Asia and Americas).



**Figure 1.3. *HLA-B* supertype frequencies.** The color matrix summarizes the frequencies of the four *HLA-B* supertypes and the group of non-classified alleles. Population's names are shown at the right side as the codes for populations regions: SSA, NAF, SWA, EUR, SEA, PAC, AUS, NEA and AME (Sub-Saharan Africa, North Africa, South West Asia, Europe, South East Asia, Pacific, Australia, North East Asia and Americas).

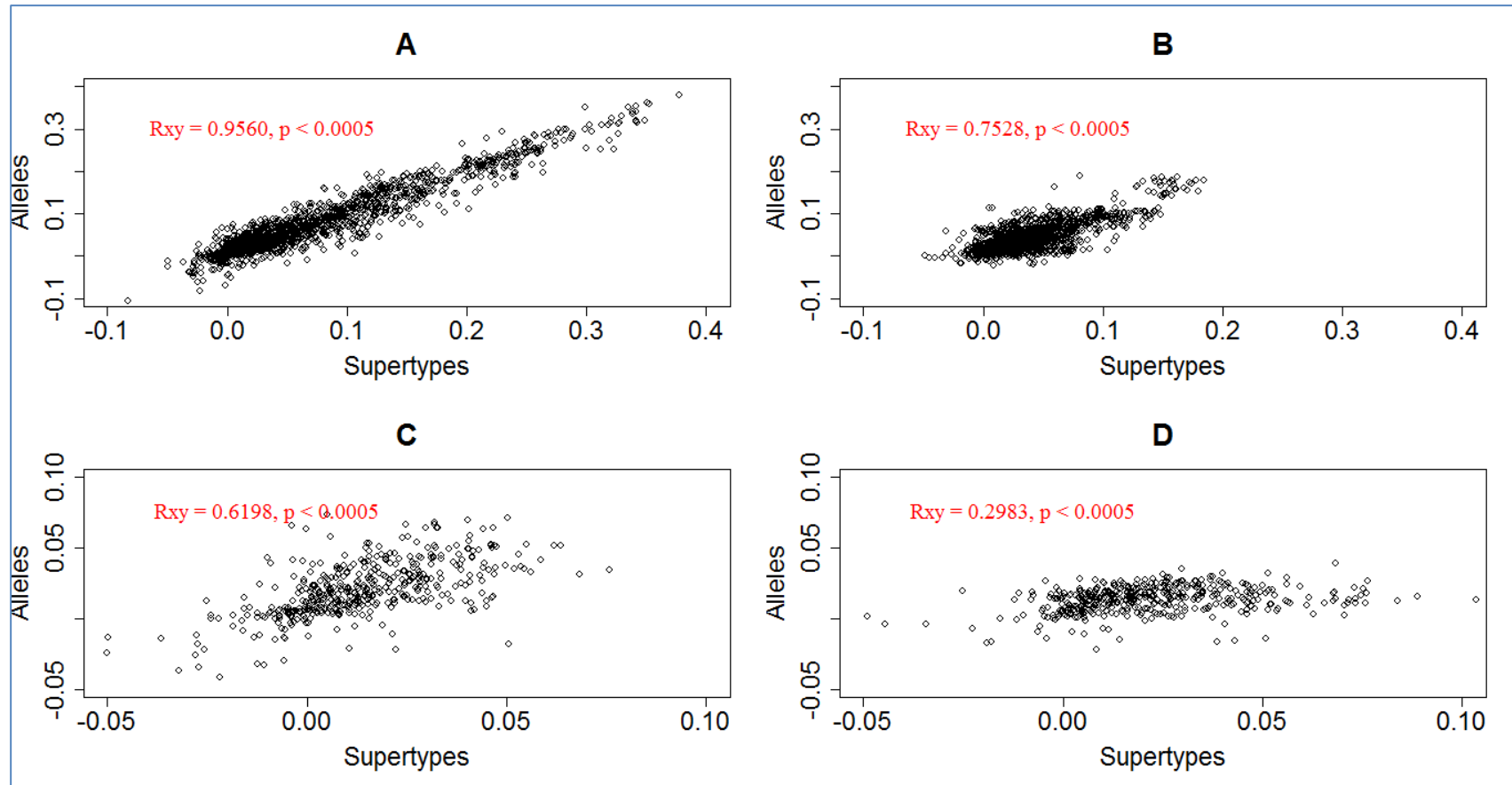
### 3.2. Heterozygosity, interpopulation differentiation and correlation with geography among alleles and supertypes.

The observed heterozygosity for the data treated at the allelic level of variation was always larger than that obtained for at the supertype level, using both complete and partial datasets. This is expected because alleles are nested within supertypes, so the heterozygosity of supertypes is constrained to be equal to or smaller than the allelic (table 1.3).

In order to define the degree to which genetic differentiation at the supertype and allelic levels were concordant, we estimated the correlation between these measures for genetic differentiation for all populations. Much of the differentiation among populations measured at the supertype level can be explained by the allele-level differentiation, especially for the *HLA-A* locus with a correlation index higher than 0.95 ( $p < 0.0005$ , figure 1.4A). A similar pattern was observed for the *HLA-B* locus ( $r_{xy}=0.75$ ,  $p < 0.0005$ , figure 1.4B). The removal of the Pacific, Australian, Taiwanese and American Populations provoked an overall drop of the  $F_{st}$  values and correlations. Despite this decrease, we still observed a strong correlation between the levels of differentiation obtained from *HLA-A* alleles and supertypes ( $r_{xy} = 0.62$ ,  $p < 0.0005$ , figure 1.4C). However, this pattern of correlation was reduced in the reduced dataset for the *HLA-B* locus, with a reduction of correlation index to 0.30 ( $p < 0.0005$ , figure 1.4D). The decrease in correlation after the removal of the Islanders and Americans was expected because these populations contribute with the large differentiation values. Furthermore, they also have few alleles per supertype, which probably increases the correlations between alleles and supertypes. When these populations were removed we observed a drastic reduction in the correlations in *HLA-B*, but not in *HLA-A*, because the later already presents in general a small number of alleles per supertype (figure 1.1B and 1.1C).

**Table 1.3.** Expected Heterozygosity (He) of alleles and supertypes.

<i>Loci</i>	Dataset	Average Allelic He	Average Supertype He
<b><i>HLA-A</i></b>	Complete	0.7761	0.6774
<b><i>HLA-A</i></b>	Reduced	0.8974	0.7504
<b><i>HLA-B</i></b>	Complete	0.8948	0.7577
<b><i>HLA-B</i></b>	Reduced	0.9429	0.7766



**Figure 1.4. Correlation between the  $G_{st}$  among all population pairs, obtained with the alleles and supertypes.** The correlation ( $R_{xy}$ ) and significance were obtained using a Mantel test. A. Correlation from *HLA-A* loci using the complete dataset; B. Correlation from *HLA-B* loci using the complete dataset; C. Correlation from *HLA-A* loci using the continental dataset; D. Correlation from *HLA-B* loci using the continental dataset.

### 3.3. Patterns of molecular variability for different pockets of *HLA* genes.

Our goal with the molecular analysis of *HLA-A* and *-B* peptide binding region (PBR) was to test the prediction that B and F pockets have the highest levels of variation, as a consequence of their role in peptide binding and consequently of a stronger intensity in balancing selection.

First, we compared the global levels of observed variation at *HLA-A* and *-B* PBR and observed significantly higher levels of nucleotide diversity ( $\pi$ ) in the latter compared with the former locus ( $p < 0.0000005$ , Mann-Whitney rank sum test). Moreover, these genes differed in the way the variation is distributed within the PBR (figure 1.5). The rank order of  $\pi$  levels in *HLA-A* and *-B* was  $pCDE \gg pB \gg pA > pF$  and  $pB \gg pF > pCDE \gg pA$ , respectively (where  $p$  is an abbreviation for “pocket”,  $\gg$  indicates a significant difference ( $p < 0.00001$ , Mann-Whitney rank sum test) and  $>$  indicates non-significant differences; figure 1.5). Among the *HLA-B* pockets,  $pB$  presented by far the highest variation inside the PBR, with half of the populations exhibiting  $\pi$  values ranging from 0.18 to 0.21 (figure 1.5). The remaining *HLA-B* pockets exhibited a relatively narrow  $\pi$  distribution, ranging from 0.10 to 0.12 in half of the populations (figure 1.5).

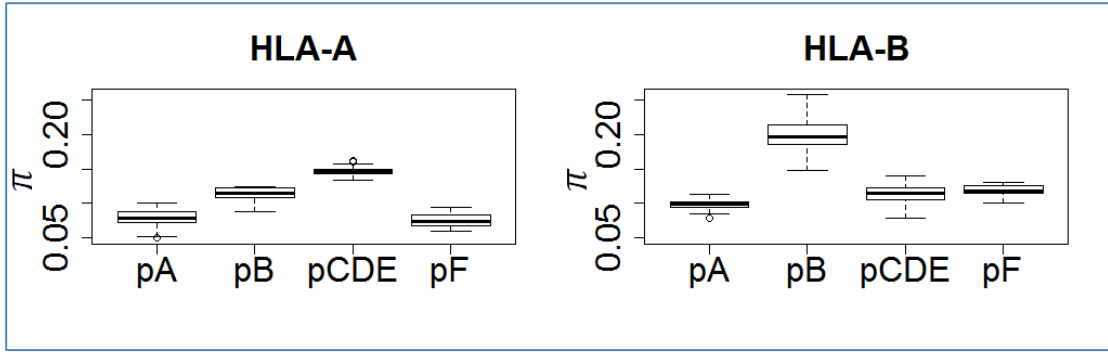
Among the *HLA-A* pockets, most of the variation was found at  $pCDE$  pockets, which compose the central region of the PBR, with half of the populations exhibiting  $\pi$  values ranging from 0.14 to 0.15 (figure 1.5). The  $pB$  was the second most variable with half of the populations presenting  $\pi$  values ranging from 0.11 to 0.12 (figure 1.5). The  $pA$  and  $pF$  showed the smallest levels of variation, with half of the populations exhibiting  $\pi$  values ranging from 0.07 to 0.09 (figure 1.5).

The hypothesis that the pockets involved in the supertype specification are the main targets of balancing selection was partially supported for *HLA-B*, since  $pB$  presented by far the highest levels of nucleotide diversity. On the other hand,  $pF$  exhibited relatively

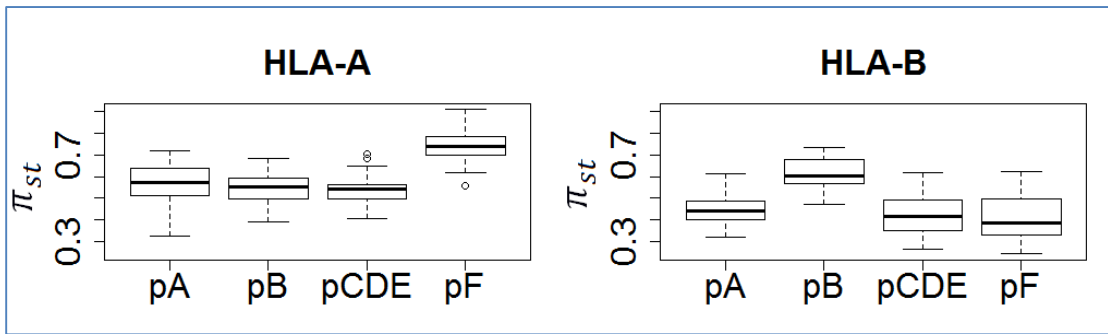
lower  $\pi$  levels in both loci, but it was not significantly different from pA and pCDE in *HLA-A* and *-B*, respectively, indicating that pF does not present a signature of balancing selection comparable to the remaining regions of the PBR. It is important to note that these results were obtained independently of the allele classification into supertypes, since the determination of the pockets codons were obtained from the classical study of Saper, Bjorkman, and Wiley (1991).

Finally, we analyzed how the nucleotide diversity is distributed between supertypes. Since the supertype categorization is based on the variations of pB and pF, it was expected that these pockets would present most of the differences between supertypes. Indeed, this prediction was confirmed in both loci, since pF and pB concentrate most of the differences between *HLA-A* and *-B* supertypes, respectively (figure 1.6). As pB presents the highest levels of variation in *HLA-B* and accounts for most of the differences between *HLA-B* supertypes, we conclude that variation between *HLA-B* supertypes accounts for most of the differences observed between *HLA-B* alleles. In other words, alleles classified into the same *HLA-B* supertype share more similarities than alleles assigned to different supertypes.

On the other hand, as most of the differences between *HLA-A* supertypes lie at pF, the pocket presenting the lowest  $\pi$  values for this gene, most of the variation between alleles are not accounted by the *HLA-A* supertypes (figure 1.6). In other words, *HLA-A* locus present more variation within supertypes than between them.



**Figure 1.5. Nucleotide diversity ( $\pi$ ) distribution among *HLA-A* and *-B* pockets:** each box represents the nucleotide diversity distribution among the populations of the complete dataset.



**Figure 1.6. Proportion of = nucleotide diversity between supertypes ( $\pi_{st}$ ) for *HLA-A* and *-B* pockets:** each box represents the proportion of nucleotide diversity distribution among between supertypes among the populations of the complete dataset.

### 3.3. Simulation results.

Alleles included in the same supertype have overlapping peptide binding specificities. The assignment of alleles to supertypes was randomized by permuting the supertype labels attributed to each allele motif allowing us to test the effect of supertype classification upon heterozygosity and differentiation. As the same patterns were obtained using the two different simulation approaches (see materials and methods), we will present the results generated with the second one, in which we did not impose constraints on the number of alleles associated to a specific supertype.

For *HLA-B*, we observed six out of 55 populations exhibiting significant lower expected heterozygosity ( $H_e$ ) than the predicted by the simulations ( $p > 0.95$ ; figure 1.7) and the same six populations also showed significant deviation at the reduced dataset. For *HLA-A*, we did not observe any significant results. Because the number of populations with individually significant  $p$ -values in either direction is low in our dataset, we investigated if the distribution of the  $p$ -values themselves are informative about a selective regime. To do this, we tested whether the entire distribution of  $p$ -values deviated from the expectation of equal numbers of values to either side of 0.5, by implementing an exact binomial test (figure 1.7). Doing this, we found that for *HLA-B* there is a significant skewing towards  $p$ -values greater than 0.5 indicating a tendency of lower observed  $H_e$  in comparison with the simulations. On the other hand, at the *HLA-A* locus, there was no clear tendency of higher or smaller simulated  $H_e$  values when compared with the observed dataset.

We did not obtain significant departures from the simulated  $G_{st}$  values with the complete dataset, for both *HLA-A* and *-B*. However, with the reduced dataset, the observed  $G_{st}$  was higher than 98% of the simulations for *HLA-B* (figure 1.8). For *HLA-A*, we did not observe significant departures from the simulation distribution even with the reduced dataset (figure 1.8).

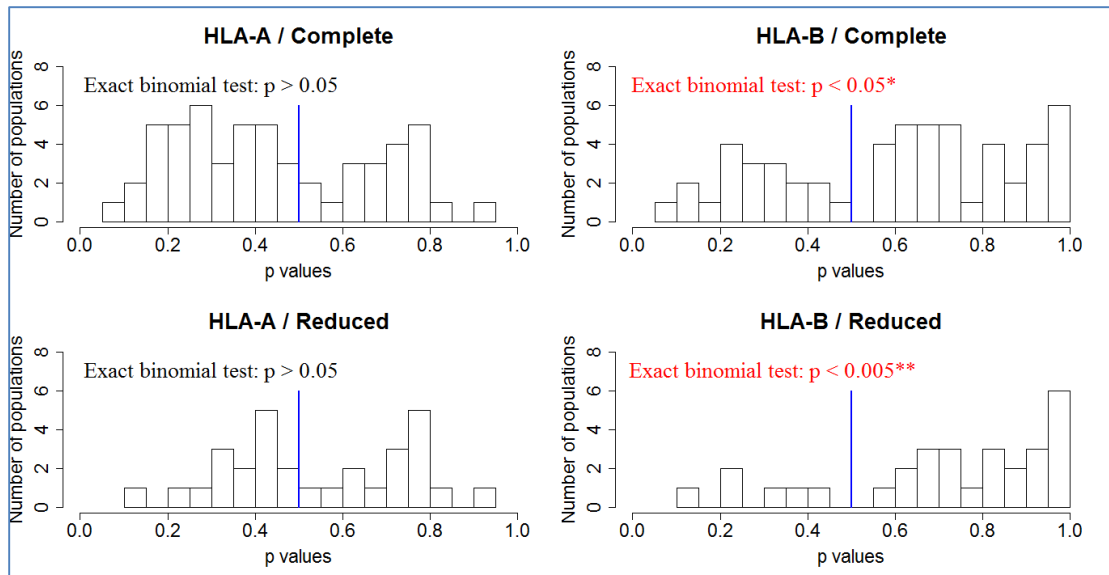
Although this result seems counterintuitive based on Sidney et al. (1996) who predicted an overall decrease in differentiation at the supertype level, it is in agreement with our description of the observed data. In our simulations, alleles were randomly assigned to groups, creating randomized supertypes with similar contents of common and rare alleles. The common alleles are expected to be assigned to different randomized supertypes in most of the simulations, just because they are less numerous than rare alleles, the same patterns we described for real *HLA-A* supertypes, which presented a low number of common alleles per populations (figure 1.1B and 1.1C). As we already

discussed, this pattern also explains the high correlation between the  $F_{st}$  measured at the allelic and supertype levels (figure 1.4). Finally, as we discussed for molecular data, there is less variation between *HLA-A* superotypes than within them, indicating that *-A* superotypes are composed by a heterogeneous set of alleles with few similarities at pF (figures 1.5 and 1.6), which facilitates the reproduction of the observed patterns of variation by the simulations.

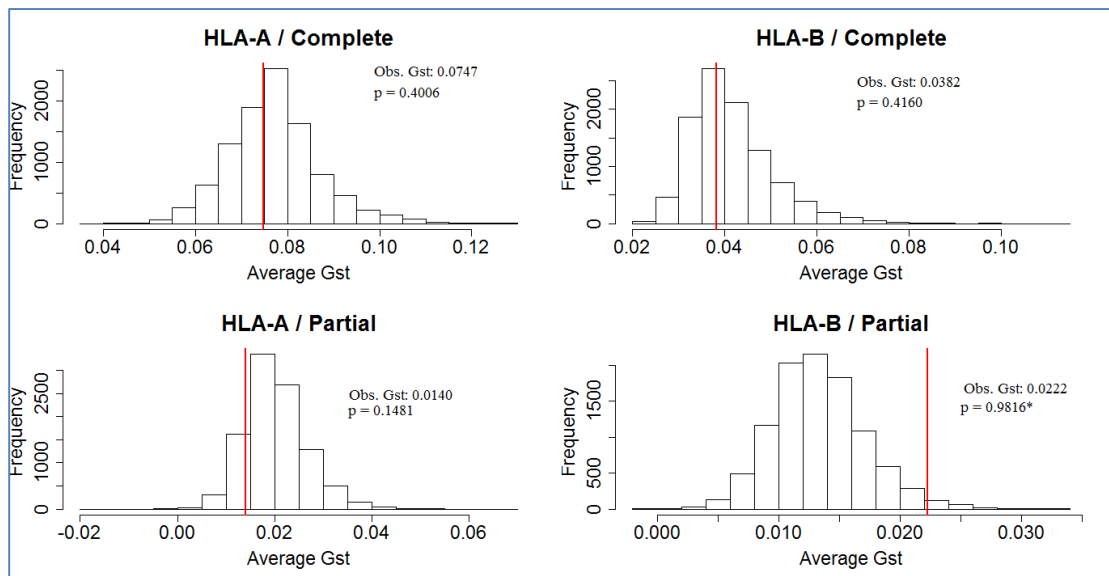
On the other hand, *HLA-B* superotypes seem to be composed by alleles that share more sequences similarities between each other, as demonstrated by the molecular analysis (figures 1.5 and 1.6). In a simple way, *HLA-B* superotypes are sets of alleles sharing B pocket resemblances. One can interpret these similarities directly in terms of the peptide presentation profiles, since *HLA-B* superotypes exhibit profound differences in the chemical properties of pB. Thus, our results of increased differentiation at the observed *HLA-B* superotypes may reflect the local adaptation of populations to different pathogen environments.

The simulations disrupted the patterns of allelic structure exhibited by *HLA-B* superotypes, creating randomized groups in the same way we described for *HLA-A*. The frequent allocation of common alleles into different randomized superotypes during the simulations provoked an increase of the  $H_e$  and decrease of population differentiation, when compared with the observed data at the *HLA-B* locus (figures 1.7 and 1.8).

The inclusion of the Islanders and American populations reduced this effect, because the patterns of variation for *HLA-B* at these populations resembles the observed for *HLA-A*, with a relatively low number of alleles belonging to different superotypes.



**Figure 1.7. Histograms of p-value distributions obtained with simulations involving the expected Heterozygosity ( $H_e$ ) for each of the analyzed populations:** each bar represents the number of populations that presented a p-value included into a 0.05 interval delimited by the X-axis. The p-value was defined as the proportion of simulated datasets with  $H_e$  larger than the observed  $H_e$ . Results obtained with the complete and reduced dataset are shown.



**Figure 1.8. Simulations results for  $G_{st}$ .** The red line represents the average observed  $G_{st}$ . We calculated the average  $G_{st}$  value for each simulated step and then determined the significance as the proportion of simulated values smaller than the observed. The results with the complete and reduced (partial dataset are shown).

#### 4. Conclusions

The supertype classification of *HLA-A* and *-B* alleles has been widely used in medical research, with the report of supertype variants contributing with the progression and/or resistance of series of pathogenic disease. This classification was proposed in the 90s as an attempt to find, as described by Sette and Sidney (1999): “the common denominators and similarities hidden within this very large degree of polymorphism”. These same authors also stated that, “the overall frequency of each of these superotypes is remarkably high and fairly conserved among very different ethnicities. Thus, there might be some advantage for human populations to present approximately five to ten main binding specificities, and that each one of these is maintained at relatively high frequency”.

Indeed, at least for *HLA-B* locus, superotypes seem to reflect the common denominator of an incredible amount of diversity, since most of the molecular variability we found at the PBR for this locus is accounted by differences between superotypes. We also demonstrated that the B pocket is probably the main target of the natural selection at this locus, presenting the highest levels of variation at *HLA-B*, and accounting for the main differences in the peptide presentation profiles from this gene.

However, going on the opposite direction from the classical expectations from loci evolving under balancing selection, our simulation results revealed that *HLA-B* supertype frequencies are not evenly distributed inside populations (lower  $H_e$  compared with the obtained from the random assigned groups of alleles). Moreover, the populations are more differentiated for *HLA-B* superotypes than the expected (higher  $G_{st}$  than the obtained from the random assigned groups of alleles). The differences on *HLA-B* supertype composition among populations can be interpreted in term of peptide recognition, since most of the differences between superotypes lies on B pocket. Thus, for *HLA-B* locus, our results

support the idea that populations present more differences in peptide presentation profiles than expected because of local adaptations to pathogens.

On the other hand, *HLA-A* supertypes appear to be composed by few representative alleles in terms of frequencies and most of the differences between alleles at this locus are not structured between supertypes, which supports our simulation results since the random assigned groups of alleles often reproduced the observed patterns of variation and differentiation of *HLA-A* supertypes. Since *HLA-A* alleles are more conserved at sites the sites involved in peptide engaging, it is reasonable that the *HLA-A* alleles present a more conserved profile of peptides across populations, contrasting with *HLA-B* as we previously discussed.

## Capítulo 2: Evolution of *HLA* genes in the American continent.

### 1. Introduction

*HLA* genes (human leucocyte antigens), located in the human Major Histocompatibility Complex (MHC) are among the most polymorphic loci in the human genome (The MHC sequencing consortium 1999). A large number of studies have been performed trying to disentangle the evolutionary mechanisms which drive such high levels of diversity (Albrechtsen, Moltke, and Nielsen 2010; Apanius et al. 1997; Bustamante et al. 2005; Hughes and Nei 1988; Knowles 2003; Meyer and Thomson 2001; Solberg et al. 2008). The role of the HLA molecules in the presentation of antigens to T lymphocytes during the first steps of the adaptive immune response (Doherty and Zinkernagel 1975) and their regulatory effect on natural killer (NK) cells via interaction with KIR receptor make the pathogen-driven selection one of the most accepted mechanisms for the maintenance of the high levels of diversity in these genes (Gillespie 1977).

There is a substantial body of work showing the association between *HLA* polymorphism and the resistance or susceptibility to infectious diseases (Carrington *et al.* 1999; Kaslow *et al.* 1996; Hraber, Kuiken, e Yusim, 2007; Campos-Lima *et al.* 1997), as well as correlations between levels of *HLA* allelic variation and regional pathogen diversity (Prugnolle et al. 2005; Qutob et al. 2011; Sanchez-Mazas, Lemaître, and Currat 2012). These studies corroborate the idea that pathogens probably played a central role in *HLA* evolution, leaving signs strong enough to surpass the dominant effect of human demographic history.

Among the evolutionary models of natural selection that can account for the increased diversity at *HLA* loci, strong support is found for those that invoke the co-evolution of pathogens and *HLA* (Borghans, Beltman, and De Boer 2004). Computer simulations have shown that these models predict that populations settling at different geographic locations would present distinct *HLA* allelic profiles providing adaptation to local immune challengers (Borghans, Beltman, and De Boer 2004).

In this context, studies involving Native American populations can be extremely valuable for understanding how the *HLA* genes have evolved recently on a regional scale. Multiple lines of evidence indicate that the occupation of the American continent occurred relatively recently, making it the last large continental region to be occupied by modern humans. During the north-south occupation, the ancestral Amerindian population was exposed to a striking variety of biomes, from the Nearctic region to the tropical landscapes from eastern South America, which probably was associated with a change in the pathogen profile faced by these populations. Based on the co-evolution hypothesis, we would expect to find genetic differences among Native American populations from different regions, which would indicate local adaptations.

In the early 90s, Belich et al (1992) and Watkins et al (1992), described a series of *HLA-B* alleles found only in South American populations and completely absent in all other world regions, contrasting with the general pattern obtained with other genetic markers like mitochondrial DNA, Y chromosome and even other *HLA* genes like *HLA-A* and *HLA-C*, in which the American variation is a subset of what is observed in Asia. Most of these new *HLA-B* alleles originated by gene conversion involving Asiatic alleles.

To explain these observations these authors proposed the hypothesis of allelic turnover, which stated that the ancestors of the native South American populations would have left the glacial environment from North America and started to face the higher levels of biodiversity of endemic pathogens during the colonization of the southern reaches of

the American continent. This process would have resulted in the selective pressures favoring new alleles, which were adaptive in the face of novel pathogens. The new alleles eventually replaced the old ones and the small effective population size associated with low levels of gene flow between populations would have contributed to the process of replacement of the old alleles by new ones. (Belich et al. 1992; Watkins et al. 1992; Cadavid and Watkins 1997; Parham et al. 1997).

Although, previous studies have described the frequencies of *HLA* alleles in the American population, none has explicitly tested the turnover hypothesis. Here, we investigate the turnover model by analyzing the correlation between the presence of alleles originated by gene conversion at high frequencies and deviations from neutral expectations. According to the turnover model, the rise of the new alleles would result from natural selection.

In order to examine the geographic distribution, allele frequencies and evidence for selection at *HLA* loci we investigated the variation of three *HLA* genes (-B, -C and -DRB1) using a large dataset consisting of samples from 32 Native American populations, increasing the number of populations from the eastern region of the South American continent (13 populations), which was systematically underrepresented in previous research on *HLA* variation. This region probably faced higher levels of genetic drift and lower levels of gene flow, when compared with other populations. To disentangle the effects of genetic drift and natural selection we used a dataset consisting of 61 microsatellite markers, spread throughout the genome in the same populations genotyped for the *HLA* loci. This data gave us insights about specific demographic events (e.g., recent bottlenecks resulting in loss of diversity) and allowed us to create the background scenario for the discussion of the patterns of *HLA* variation.

## **2- Methods and Materials**

### **2.1- Population sampling.**

We analyzed 32 populations, which were divided into five groups based on geographic and linguistic criteria, following Yang et al. (2010). The groups and populations are North America (NAM): Chipewyan, Cree and Ojibwa; Meso-America (MAM): Mixtec, Mixe, Zapotec, Quiche and Wayuu; Northwest South America (NWSAM): Cabecar, Guaymi, Kogi, Arhuaco, Zenu, Embera, Waunana; West South America (WSAM): Inga, Quechua, Aymara, Huilliche; East South America (ESAM): Arara do Iriri, Arara, Kayapo Krokaimoro, Kayapo Xicrim, Kaigang, Parakanã, Araweté, Urubu Kaapor, Asurini, Guarani Kaiowá, Ticuna Arara, Ticuna Tarapaca, Ache.

All populations belong to the Amerindian linguistic group except the Chipewyan, a Na Dene population, which was included in the North American group due to the high level of genetic similarity with Cree and Ojibwa Populations (Wang et al. 2007)(Reich et al. (2012).

## 2.2- Microsatellite dataset.

We used a subset of 61 out of the 678 autosomal microsatellites extracted from the dataset analyzed by Wang et al. (2007). The data from non-Brazilian samples was obtained directly from the dataset presented by Wang et al. (2007) whereas the Brazilians dataset was generated by the post graduate student Kelly Nunes during her Phd thesis (not published data). The final microsatellites dataset we used in the present study consisted in 32 populations, 22 of them obtained from the Wang et al. (2007) and the remaining 10 populations generated by our group (Supplementary Material)

## 2.3- *HLA* typing.

The region comprising the exons 2 and 3 from *HLA-B* and *-C* was amplified and alleles typing were carried for each individual. In-house PCR primers were used to amplify each locus (SumpkeS2). Applied Biosystems Big Dye terminator chemistry and sequencing primers were used to obtain the sequences of both strands of exons 2 and 3. The *HLA-DRB1* locus was amplified and sequenced using the AlleleSEQR class II kit (Abbott Molecular Inc, Des Plaines, IL). Reaction products were identified with Applied Biosystems 3730xl DNA analyzer (PE Applied Biosystems, Foster City, CA) and sequence interpretation used Assign software (Conexio Genomics, Applecross, Western Australia).

### ***HLA* nomenclature:**

The nomenclature of *HLA* alleles follows the rules of the World Health Organization (WHO) Nomenclature Committee (Robinson et al. 2011). An *HLA* allele corresponds to a unique sequence comprising at least exons 2 and 3 for class I and exon 2 for class II loci. The allele name carries the information of which locus the corresponding allele belongs followed by an “\*” and a unique number ranging from two to four sets of digits separated by colons (e.g. *B\*14:04*, *DRB1\*04:04:01* or *C\*07:01:01:01*). All alleles receive a name with

at least two sets of digits. The digits before the first colon describe the lineage, which often corresponds to the serological HLA antigen (e.g. *B\*51* or *DRB1\*16*). The next set of digits are used to list the unique protein-level alleles belonging to the same lineage (e.g. *B\*15:04* and *B\*15:08*). Alleles whose names differ on the first two sets of digits must differ by one or more non-synonymous mutations. Alleles that differ only by synonymous mutations are distinguished by the use of the third set of digits (e.g. *C\*03:04:02* and *C\*03:04:03*). Alleles that only differ by sequence polymorphisms in non-coding regions (introns and/or in the 5' or 3' untranslated regions) are distinguished by the use of the fourth set of digits (*DRB1\*15:01:01:01* and *DRB1\*15:01:01:02*).

The *HLA* alleles whose sequences differ outside the exons 2 and 3 for class I and exon 2 for class II loci are grouped and identified using an upper case 'G' after the third set of digits of the lowest numbered allele in the group (e.g. *B\*15:01:01G* and *C\*08:01:01G*).

As we discussed before, the dataset we used in this study was generated by sequencing of exons 2 and 3 for class I genes and exon 2 for *HLA-DRB1*. Therefore, we will present results at the highest possible typing resolution level. Eventually, we have found genotype ambiguities, when more than one pair of alleles is compatible with the observed sequences. These ambiguities were resolved with allele specific primers and in some cases applying the following frequency criteria: 1) an internal criterion based on the allelic composition of non-ambiguous genotypes we have found in the same populations in which the ambiguity was observed. In this case, the inferred genotype was composed by alleles we found in homozygosis and/or non-ambiguous combinations; 2) an external criterion using published datasets (Gonzalez-Galarza et al. 2011; Solberg et al. 2008), taking into account the alleles frequencies in populations at the same region than the one in which the ambiguities were observed.

## 2.4- Data Analysis

**Population variability:** We used the Arlequin software package (Excoffier and Lischer 2010) to estimate the allele frequencies from the genotypes using the Expectation-Maximization (EM) algorithm (Slatkin and Excoffier 1996) the number of alleles (k), and the observed and expected sample heterozygosity (Ho and He respectively). We also used Arlequin to perform the test for deviations from Hardy-Weinberg proportions using an exact test developed by Guo and Thompson (1992).

**Haplotype inference:** We inferred haplotypes using the EM algorithm as implemented in Arlequin. Haplotypes with frequency lower than the standard error were excluded from subsequent analyses.

**Allele Sharing between Populations:** We estimated the degree of allele sharing between all pairs of populations using the Prevost's Distance (Prevosti, Ocaña, and Alonso 1975; Wright 1978), defined as:

$$D = \frac{1}{2} \sum_{i=1}^k |p_i - q_i| , \quad (2)$$

where p and q are the frequencies of the ith allele on populations 1 and 2 and k is the number of observed alleles in these populations. A value of Prevost's Distance (D) equal to 1 is obtained from pairs of populations that do not share alleles while a D=0 is obtained when the same alleles are observed at the same frequencies in both populations. Thus, the Prevost's D value represents the direct proportion of differences on the observed allele frequencies between a pair of populations. We implemented this measure using the R package (S2).

**Tests of neutrality.** We tested for departure from neutrality-equilibrium using Tajima's-D (Tajima 1983) test implemented in Arlequin 3.5, which compares two estimators parameter  $\theta$  ( $4N\mu$ ),  $\pi$  (the average number of nucleotide differences between pairs of sequences) and  $S$  (the number of segregating sites in the sample, normalized by a value proportional to the sample size). Tajima (1989) demonstrated that under neutrality and equilibrium the expectation of the difference between these estimators (Tajima's D) is equal to zero. Positive Tajima's D values are expected when polymorphisms at intermediate frequencies are present in the samples, which is expected under the action of balancing selection or recent population bottlenecks. Tajima's D values are expected from samples containing a large proportion of lower frequencies polymorphisms, which is expected under directional and/or negative selections or population expansion.

### 3- Results And Discussion

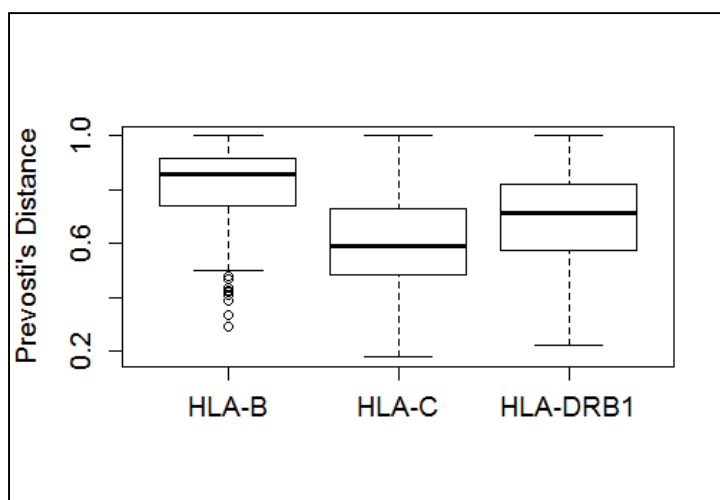
#### 3.1. Deviations from Hardy-Weinberg proportions.

We performed 96 Hardy-Weinberg tests (32 populations at 3 loci) and three, one and two deviations were detected at the 5% significance level at *HLA-B*, *-C* and *-DRB1* loci respectively (only one deviation at 1% level for both *HLA-B* and *-DRB1* loci) (table 1-S2). Deviations are due to an excess of homozygotes except for Kogi population. This overall result would indicate the presence of null alleles. However, six deviations spread for all loci is very close to that expected by chance alone (5 deviations) and none of them have remained significant after correction for multiple tests, suggesting the absence of null alleles and the efficiency of our methodology.

### 3.2. Allele frequency distributions of *HLA* loci.

#### 3.2.1. General patterns.

The *HLA-B* locus exhibited the highest differentiation, with 50% of the pairs of populations presenting a Prevost's Distance ranging from 0.74 to 0.92 (with a median value of 0.86; figure 2.1). On the other hand, *HLA-C* locus exhibited the smallest Prevost's D median (0.59) and a broader distribution, with 50% of the pairs of population presenting a Prevost's D ranging from 0.48 to 0.73 (figure 2.1). The *HLA-DRB1* locus showed an intermediate pattern, with 50% of the populations pairs presenting a Prevost's D ranging from 0.58 to 0.82 with a median of 0.70 (figure 2.1).



**Figure 2.1. Prevost's Distance distributions between pairs of populations for *HLA-B*, *-C* and *-DRB1* loci.**

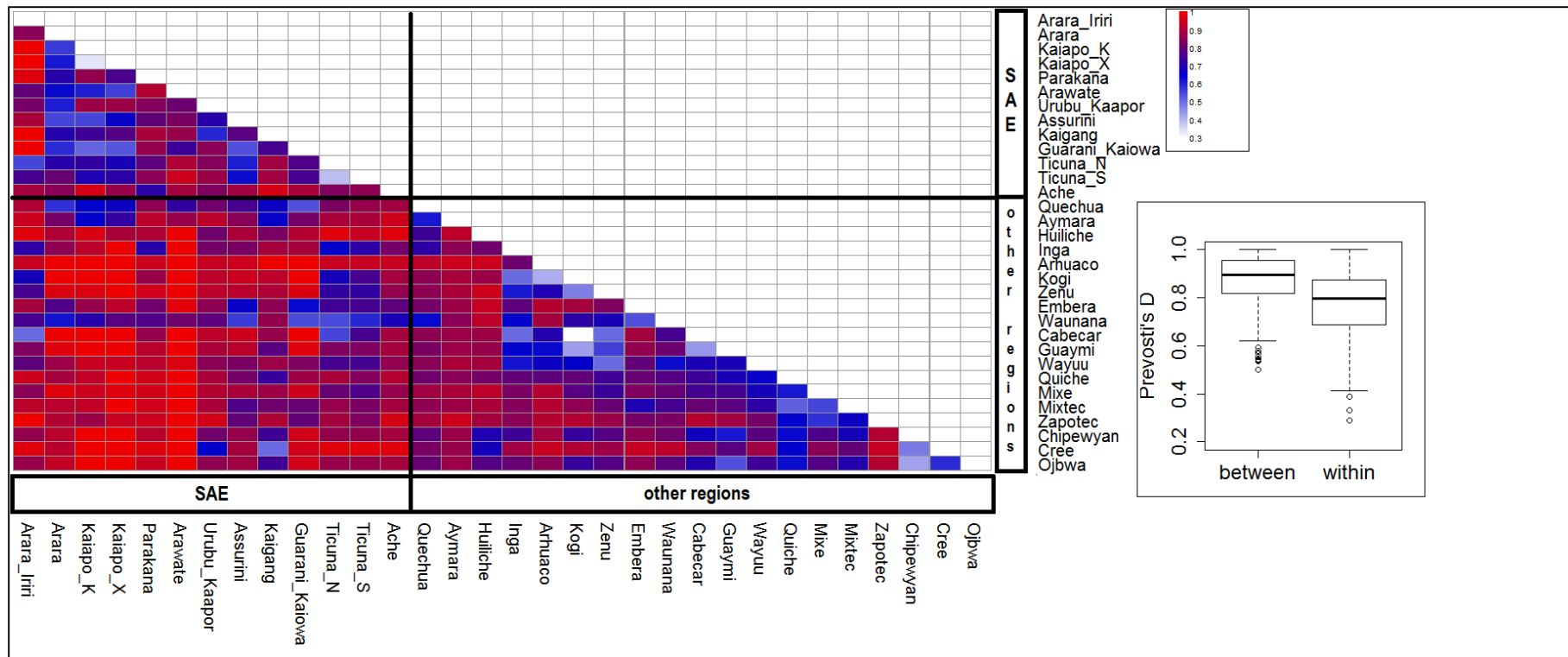
The *HLA-B* locus displayed the highest number of population pairs in which we did not observe sharing of alleles (Prevost's D = 1), with 44 pairs against four and six for *HLA-C* and *-DRB1* loci, respectively (figures 2.3 to 2.5). Thirty-nine out of these 44 pairs included a population from the eastern region of South America (SAE), making this region the highest contributor for the elevated Prevost's D values we obtained on the *HLA-B* and *-DRB1* analysis (figures 2.3 to 2.5).

The distinct *HLA-B* allelic component we observed in SAE were generated by gene conversion or in some cases recombination involving the *B\*15:01:01G*, *B\*27:05:02G*, *B\*35:01:01G*, *B\*37:01:01*, *B\*39:01:01G*, *B\*40:02:01G*, *B\*44:02:01G*, *B\*48:01:01G* and *B\*51:01:01G* alleles, as shown in Figure 1-S2. (hereinafter called parental alleles) which were found in North American populations (S2) and previously described in Northeast Asia (Gonzalez-Galarza et al. 2011; Robinson et al. 2011; Solberg et al. 2008). An alternative hypothesis to explain the emergence of the recombination/gene conversion alleles would require an incredible level of homoplasy to account the small stretches of sequences shared by them and the parental alleles (S2).

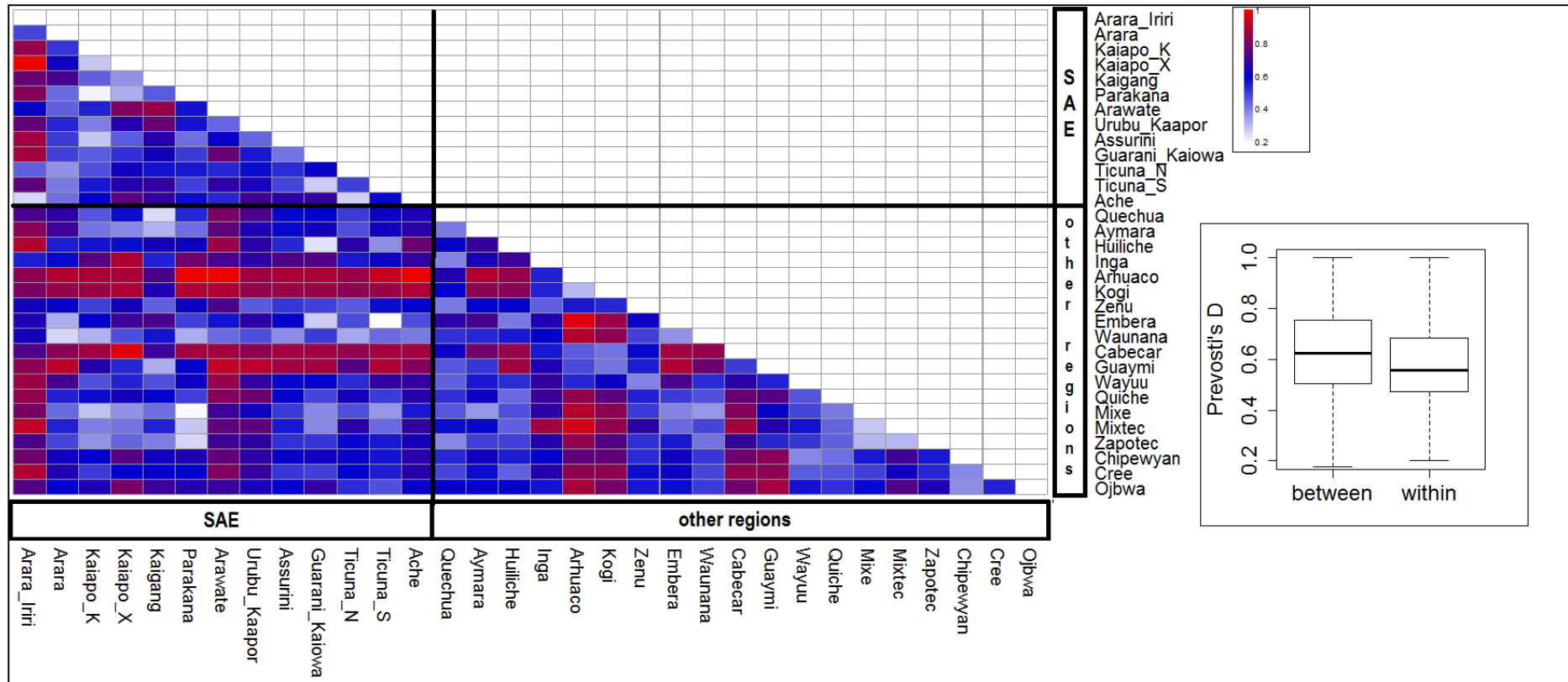
The *HLA-B* recombination/gene conversion alleles were practically absent in North American populations and start to appear in the Mesoamerican region (figure 2.6A and 2.6B). On the other hand, SAE populations exhibit the highest frequencies of these alleles except for Ticuna\_N, Ticuna\_S, Kaingang, Urubu Kaapor and Arara do Iriri, which presented cumulative frequencies of the recombination/gene conversion smaller than 0.50 (figure 2.6A, 2.6B and 2.6C).

We tested for a correlation between the frequencies of recombination/gene conversion alleles and the distance from the Bering Strait. A weak but significant correlation was obtained when all populations were considered ( $r^2=0.34$ ,  $p<0.001$ ) but it was lost after the removal of the North American populations ( $r^2=0.07$ ,  $p<0.15$ ) (figure 2.6C). This result indicates that the frequencies of the recombination/gene conversion *HLA-B* alleles are not gradually increasing from north to south on the American continent, but they show a small rise in frequency starting from the Mesoamerican (MAM) region, and then reach their maximum in the SAE populations (figure 2.6). Furthermore, according to the Prevost's D results, different sets of recombination/gene conversion alleles were found in distinct regions of the American continent (we will present a detailed description of the allelic content in the American populations in the next section).

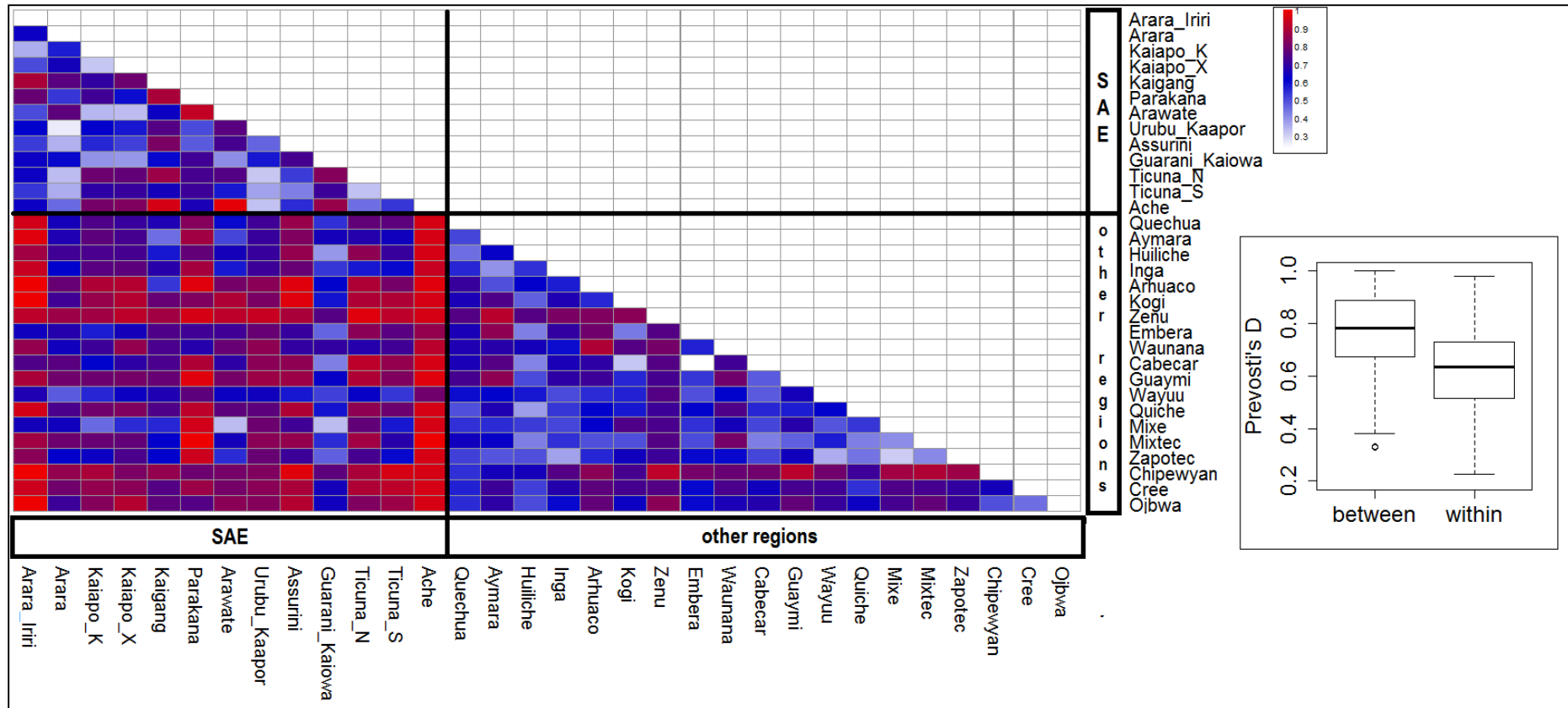
The differences in the allelic frequencies we observed for *HLA-DRB1* locus between the American regions could not be associated with the emergence of new alleles, except by *DRB1\*08:07* (which we will discuss in the next section). For both *HLA-C* and *-DRB1* loci, the same alleles were found along different American regions and except for *DRB1\*08:07*, all of them were previously described in Asian populations (see next section). Moreover, we could not distinguish any clear signal of recombination/gene conversion involving the observed alleles for both *HLA-C* and *-DRB1* loci.



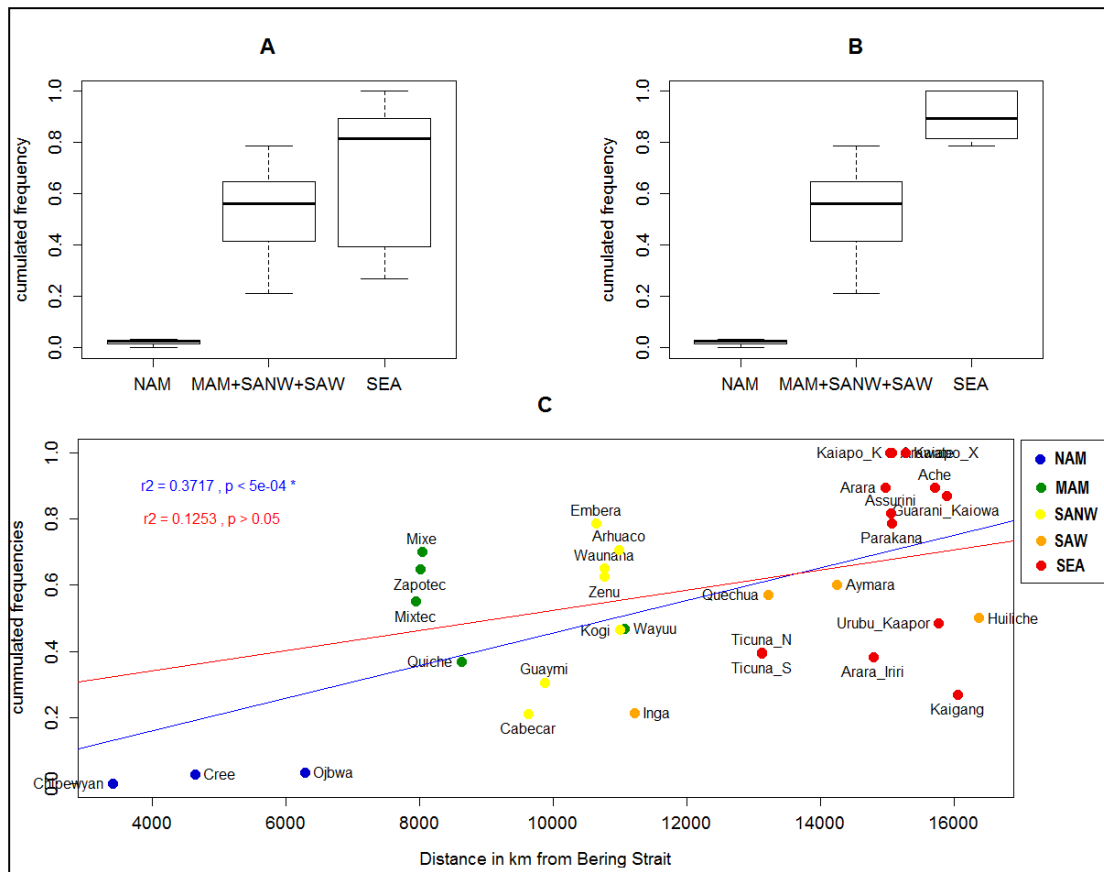
**Figure 2.2. Color grid representing the *HLA-B* pairwise Prevost's Distance.** The difference between the Prevost's D values obtained from the comparisons between and within groups was significant (Mann-Whitney test,  $p < 2.2 \times 10^{-16}$ ).



**Figure 2.3. Color grid representing the *HLA-C* pairwise Prevost's Distance:** in this figure, the populations are separated into two sets: 1) The eastern South American population (SAE) and the remaining populations. The difference between the Prevost's D values obtained from the comparisons between and within groups was significant (Mann-Whitney test,  $p > 0.001$ ).



**Figure 2.4. Color grid representing the *HLA-DRB1* pairwise Prevost's Distance:** in this figure, the populations are separated into two sets: 1) The eastern South American population (SAE) and the remaining populations. The difference between the Prevost's D values obtained from the comparisons between and within groups was significant (Mann-Whitney test,  $p < 2.2 \times 10^{-16}$ ).



**Figure 2.5. Cumulative frequencies distributions of the recombination/gene conversion alleles in the American continent:** North America (NAM), Mesoamerica (MAM), Northwest South America (SAWN), West South America (SAW) and East South America (SAE), respectively; **A.** Frequency distributions of the recombination/gene conversion alleles, using complete set of populations; **B.** Frequency distribution of the recombination/gene conversion alleles, excluding the Ticuna\_N, Ticuna\_S, Kaigang, Urubu Kaapor and Arara do Iriri populations (which presented higher frequencies of parental alleles, see text); **C.** Linear regression between the cumulative frequency of recombination/gene conversion alleles and the distance in km from Bering Strait. ( $r^2=0.12, p>0.05$ , tendency line shown in red), indicating an anchoring effect generated by these populations. We also can note the four red dots below the tendency lines that represent the Ticuna\_N, Ticuna\_S, Kaigang, Urubu Kaapor and Arara do Iriri populations (the dots representing the Ticuna populations overlap).

### 3.3. Correlation between diversity indexes and the geographic distance from the Bering Strait

We compared the patterns of variation obtained for autosomal microsatellites with those estimated for the *HLA* loci, attempting to disentangle the signs left by the

demographic history and natural selection. First, we tested for the presence of correlations between the expected heterozygosity ( $H_e$ ) with the distance from Bering Strait for both *HLA* and microsatellites loci. Wang et al. (2007), reported a significant negative correlation  $H_e$  at microsatellite loci and distance from Bering Strait ( $r^2 = 0.436$ ), which is in agreement with a main North-South colonization route for the American Continent. Here, we did not reproduce this result, although we also obtained a negative and significant correlation between the  $H_e$  and distance from Bering Strait, the intensity of the correlation was much smaller than the obtained by Wang et al. (2007) ( $r^2 = 0.133$ ,  $p < 0.05$ ; data not shown).

We can point out three reasons to explain the divergence between our result and the one obtained by Wang et al. (2007): 1) we are working with a different dataset, with the introduction of ten more populations of the eastern South American region; 2) we also are working with a subset of 61 out of the 678 microsatellites used by Wang et al. (2007); 3) and finally, we applied a much more simple approach to infer the distances between populations and the Bering Strait, using a great circle routes approach, not taking into account possible coastal routes or obligatory way points, like the Isthmus of Panama, which led us to underestimate the distances between populations and the Bering Strait. We tested the weight of the first and third hypothesis removing the additional eastern South American populations and inferring manually the distances between populations and the Bering Strait using the software google earth (4.3.7284.3916 beta).

The introduction of way points on the Pacific coast led to an increase of the correlation ( $r^2 = 0.1957$ ,  $p < 0.05$ ; figure 2.7A), although it was still not as large as the one obtained by Wang et al. (2007). On the other hand, the removal of eastern South American populations (SAE) led to the loss of the correlation ( $r^2 = 0.037$ ,  $p > 0.80$ ; figure 2.7A). The same effect was not achieved when only the outlier populations (Arara do Iriri and Ache) were removed ( $r^2 = 0.1891$ ,  $p < 0.05$ ; data not shown), indicating that the loss of significance caused by taking the SAE populations cannot be accounted by the outliers.

The negative correlation between  $H_e$  and the increase of the distance from Africa is a well-established result for humans, with correlations larger than 0.8 using different genetic markers (Prugnolle et al. 2005; Sanchez-Mazas, Lemaître, and Currat 2012; Wang et al. 2007). However, an important aspect of these studies, not previously discussed, is the fact that the American continent is the geographic region presenting the largest variance in the residuals of the correlation between genetic variation and distance from Africa.

According to our results, there is no correlation between microsatellite diversity and of distance from Bering Strait, unless SAE populations were included in the analysis (figure 2.7A). As already discussed, the removal of SAE populations led to loss of the correlation, probably due to an anchoring effect provoked by the significantly lower  $H_e$  we observed in these populations, when compared with the remaining American populations (figure 2.7B,  $p < 0.00001$ , Mann-Whitney rank sum test). The stable levels of  $H_e$  we observed for microsatellites along the American continent support the hypothesis of rapid colonization through the Pacific coast followed by lower levels of genetic drift outside SAE. This result is also supported by others genetic markers as Y chromosome and mitochondrial DNA.

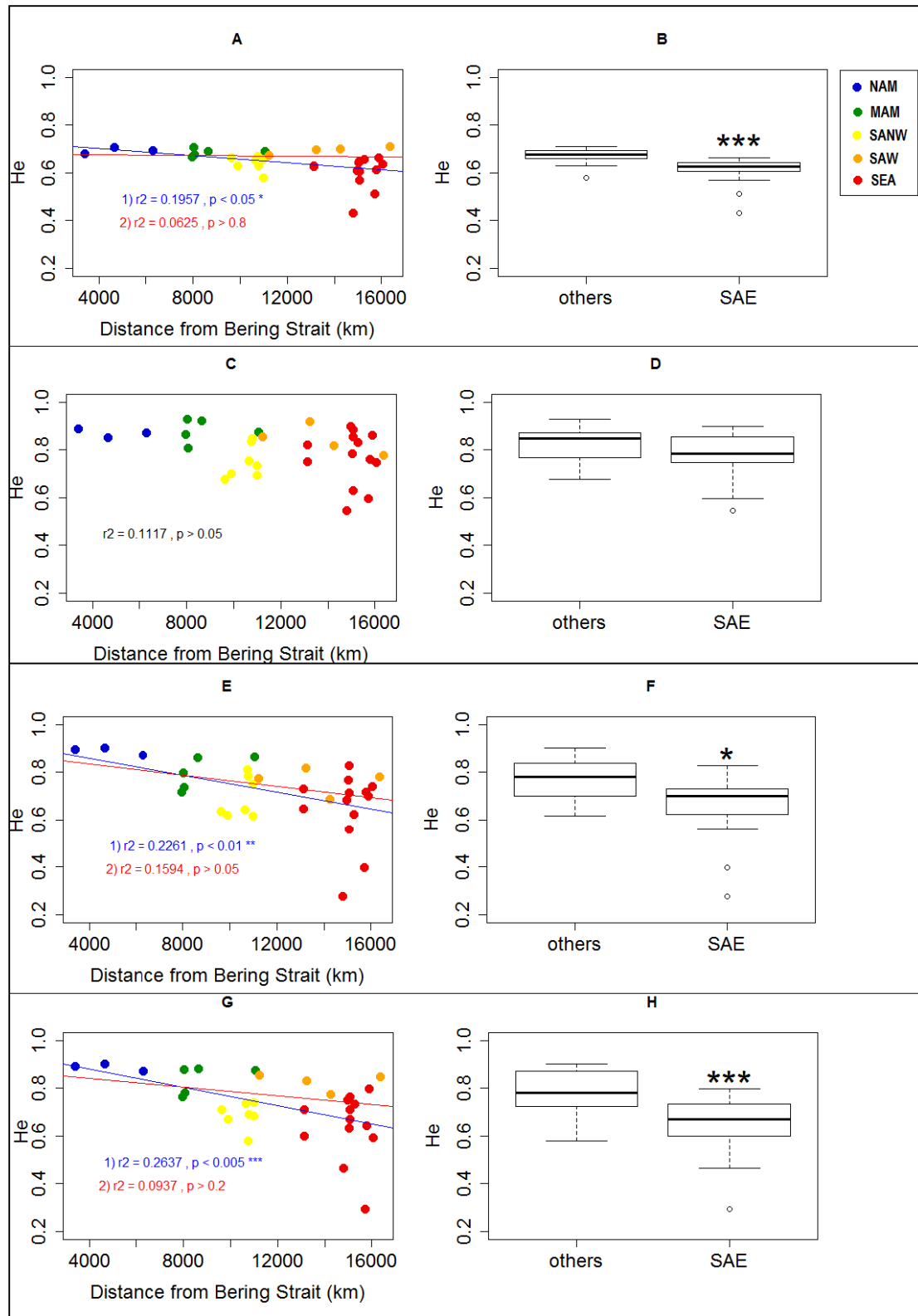
We applied this same approach on the *HLA* dataset, testing for correlation between  $H_e$  and the distance from Bering Strait. We observed the same patterns as for microsatellites at the *HLA-C* and *-DRB1* loci, with a modest but significant negative correlation between the loss of diversity and the increase of distance ( $r^2 = 0.2261$  ( $p < 0.01$ ) and  $r^2 = 0.2637$  ( $p < 0.005$ ) for *HLA-C* and *-DRB1* loci, respectively; figure 2.13E and 2.13G), which persisted even after the removal of the outliers Arara do Iriri and Ache populations (data not shown).

The removal of SAE populations also provoked the loss of the correlation with distance ( $r^2 = 0.1594$  ( $p > 0.05$ ) and  $r^2 = 0.0937$  ( $p > 0.20$ ) for *HLA-C* and *-DRB1* loci,

respectively; figure 2.13E and 2.13G) which can also be explained by the significant lower levels of  $H_e$  in SAE populations when compared with other American populations (figure 2.13F and 2.13H,  $p < 0.05$  and  $p < 0.01$ , for *HLA-C* and *-DRB1* loci, Mann-Whitney rank sum test). These results indicate that the genetic drift was probably the main force governing the *HLA-C* and *-DRB1* patterns of variation in the American continent, since they did not deviate from the general genome profile evidenced by the microsatellites.

We did not observe correlations between the levels of  $H_e$  of the *HLA-B* locus and the increase of the distance from Bering Strait ( $r^2 = 0.1117$ ,  $p > 0.05$ ; figure 2.7C) even before the removal of SAE populations. In agreement with this result, we also did not find significant differences in  $H_e$  between SAE and other Native American populations (figure 2.7D,  $p > 0.15$ , Mann-Whitney rank sum test). Differently from which was observed for *HLA-C*, *-DRB1* and microsatellites, the expected heterozygosity of *HLA-B* locus is not changing along the American continent, even in the Eastern South American region.

The observed patterns of variation for microsatellites, *HLA-C* and *-DRB1* loci are in agreement with which was described for other genetic markers and morphological data. All these evidences state that the populations at the Eastern South American region have probably suffered with more intense levels of genetic drift, which led to a significant decrease of the genetic diversity. This effect could be detected even in the high polymorphic *HLA* genes, except in *HLA-B*, which conserved the levels of diversity even under the general trend of loss of variability.



**Figure 2.6. Distribution of the expected Heterozygosity (He) of microsatellites and *HLA* loci along the American continent.** The correlations between He and distance from Bering Strait for the microsatellites, *HLA-B*, *HLA-C* and *HLA-DRB1* loci are shown in A, C, E and G, respectively. The correlation indexes ( $r^2$ ), p-values and vest-fit linear regression lines obtained with all populations are shown in blue while the ones obtained after the removal of eastern South American population (SAE) are shown in red. The

distribution of the  $H_e$  of SAE and the other American populations are shown in B, D, F.  
 \*for  $p < 0.05$ ; \*\*\* for  $p < 0.005$ .

### 3.4. Neutrality and Equilibrium Test

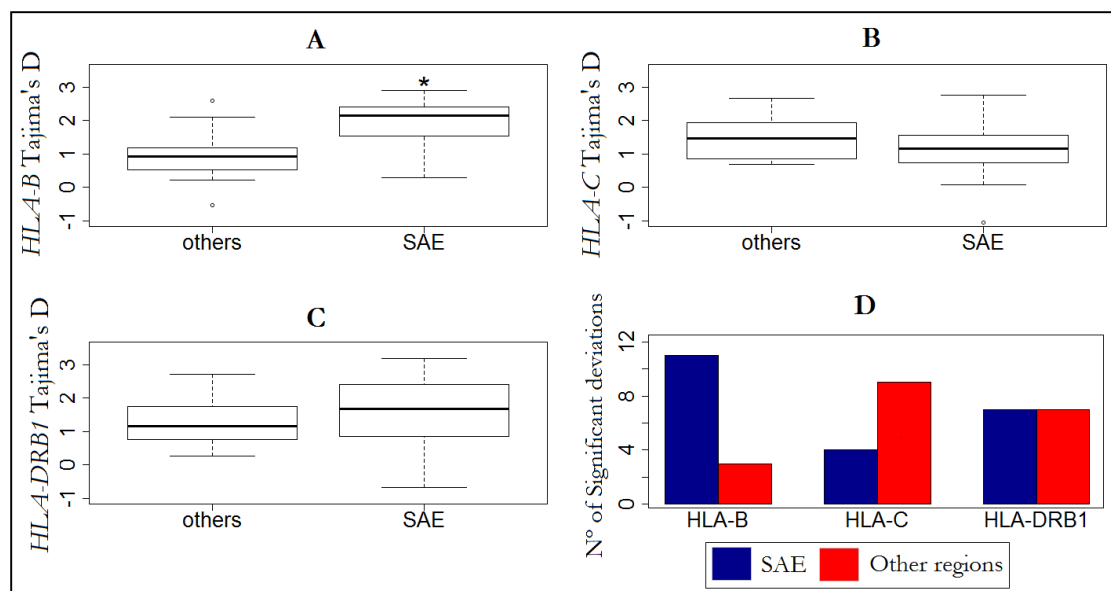
Deviation from the null hypothesis can result from long-term selection events, and the shared evidences for selection across populations can be attributed to events that occurred in ancestral populations. On the other hand, selection events that are not shared due to common ancestry are more likely to show lower patterns of sharing.

The analysis of Tajima's  $D$  revealed that all populations except for one for *HLA-B*, one for *HLA-C* and two for *HLA-DRB1*, exhibited  $D > 0$ , which is very unlikely for genes evolving under neutrality ( $p < 0.0000005$ , applying an exact binomial test to the entire set of populations). In fact, the excess of positive Tajima's  $D$  is even more unusual when we consider knowledge of the demographic history of human populations, and Native Americans in particular, which commonly show evidence of having experienced recent expansions, which drive the overall distribution of Tajima's  $D$  to negative values (Hammer et al. 2003; Jaruzelska et al. 1999).

This result indicates that the Native American populations we studied inherited a polymorphic array of *HLA* alleles from their Asian ancestors, and this set of divergent sequences had been evolving under balancing selection long before the divergence of the populations that occupied the Americas (Garrigan and Hedrick 2003). On the other hand, the highest and most significant positive Tajima's  $D$  values are not evenly distributed among the Native American populations. The *HLA-B* from Eastern South American populations (SAE) presented significant higher Tajima's  $D$  values when compared to populations from other regions (figure 2.8A,  $p < 0.001$ , Mann-Whitney rank sum test). The *HLA-C* and *-DRB1* loci did not present the same patterns, showing similar Tajima's  $D$  among regions (figures 2.8B and C,  $p > 0.30$  for both loci, Mann-Whitney rank sum

test). Moreover, taking into account only the significant results, the SAE populations concentrates the majority of the deviations observed for *HLA-B* locus, with 11 out of the total of 14 significant Tajima's D values (which was marginally significant with  $p = 0.06$ , applying an exact binomial test). This pattern was not seen at *HLA-C* and *-DRB1*, which presented four out of 13 and seven out of 14 significant deviations in SAE, respectively ( $p > 0.25$  for both loci, exact binomial test).

We have demonstrated that SAE populations exhibit gene conversion/recombination alleles in high frequencies (figure 2.6). These alleles are very divergent, differing in multiple nucleotide sites from the parental alleles involved in their origin (figure 1-S2). The presence of these divergent alleles in intermediate frequencies accounts for the elevated Tajima's D values we obtained for *HLA-B* locus in SAE populations. Therefore, the increased departures from neutral expectations we observed in SAE population at the *HLA-B* locus can be direct associated with the presence of the gene conversion / recombination alleles.



**Figure 2.8. Tajima's D** for Eastern South American (SAE) and remaining populations for *HLA-B* (A), *HLA-C* (B) and *HLA-DRB1* (C) loci. A significant difference was observed for *HLA-B* locus ( $p < 0.001$ , Mann-Whitney rank sum test). (D) Differences in the amount of significant Tajima's D values.

#### 4- Conclusions

The distinct *HLA-B* allelic profile of Central and South Native American populations was a remarkable discovery during the 90s. The relevance of this find lies in the contrast with other genetic markers, including other *HLA* genes, whose diversity in Central and South America seems to be a subset from which is found in North America and Northeast Asia. To explain such unusual pattern, Parham et al. (1997) proposed the allelic turnover hypothesis, which claims that the replacement of the ancestral alleles by the new ones, products of recombination and gene conversion, occurred under the selective pressures imposed by the endemic pathogens of the American tropical regions. The authors also postulated that genetic drift had an important role enhancing the efficiency of the replacement process.

In the present study, we revisited the turnover model and demonstrated that the Eastern South American (SAE) populations are those exhibiting the highest frequencies of gene conversion / recombination of *HLA-B* alleles (figure 2.6). These same populations also presented a significant increase of measures like Tajima's D and expected heterozygosity at the *HLA-B* locus, when compared with other populations from the American continent. These same patterns were not observed at the other *HLA* loci, reinforcing the interpretation that these results are due to the increased frequencies of gene conversion / recombination alleles.

## 5. Detailed Description of Allele Frequencies.

In this section we will provide a detailed description of the *HLA-B*, *-C* and *-DRB1* allelic distribution in the American continent. We found 91, 28 and 40 different alleles for these three loci respectively, taking into account the complete set of populations. The allele frequencies for each locus are summarized in figures 2.7 to 2.10. Starting with the *HLA-B* locus, about 74.1% of the allelic distribution in North American populations was comprised by the parental alleles (*B\*15:01:01G*, *B\*27:05:02G*, *B\*35:01:01G*, *B\*37:01:01*, *B\*39:01:01G*, *B\*40:02:01G*, *B\*44:02:01G*, *B\*48:01:01G* and *B\*51:01:01G*)(figure 2.7).

Some of the parental alleles were still found at high frequencies in the Mesoamerican region (*B\*35:01:01G*, *B\*40:02:01G* and *B\*51:01:01G*, at the accumulative frequency of 21%) but a second allelic component comprised by recombination/gene conversion alleles (*B\*35:12:01*, *B\*35:14:01*, *B\*35:17*, *B\*35:43:01G*, *B\*3549*, *B\*39:02:02* and *B\*39:05:01*, at the cumulative frequency of 47.8%) were observed (figure 2.7).

The populations located at the Isthmus of Panama and northwest of the South American Continent (the ones belonging to the Chibcha-Paezan linguistic group (SAWN)) presented two different sets of alleles. The more common alleles exhibited by the Chibcha populations (Guaymi, Cabecar, Arhuaco, Kogi and Zenu) were the parental *B\*15:01:01G*, *B\*35:01:01G* and *B\*40:02:01G* (at the cumulative frequency of 46.4%) added by the recombined *B\*35:43:01G* allele (30.4%) (Figure 2.7). On the other hand, the more frequent alleles at the Paezan populations (Embera and Waunana) were the parental *B\*40:02:01G* (19.1%) followed by *B\*15:04*, *B\*35:10*, *B\*40:04* and *B\*39:05:01*, the latter were probably generated by gene conversion and were found at the accumulated frequency of 60.3% (figure 2-3).

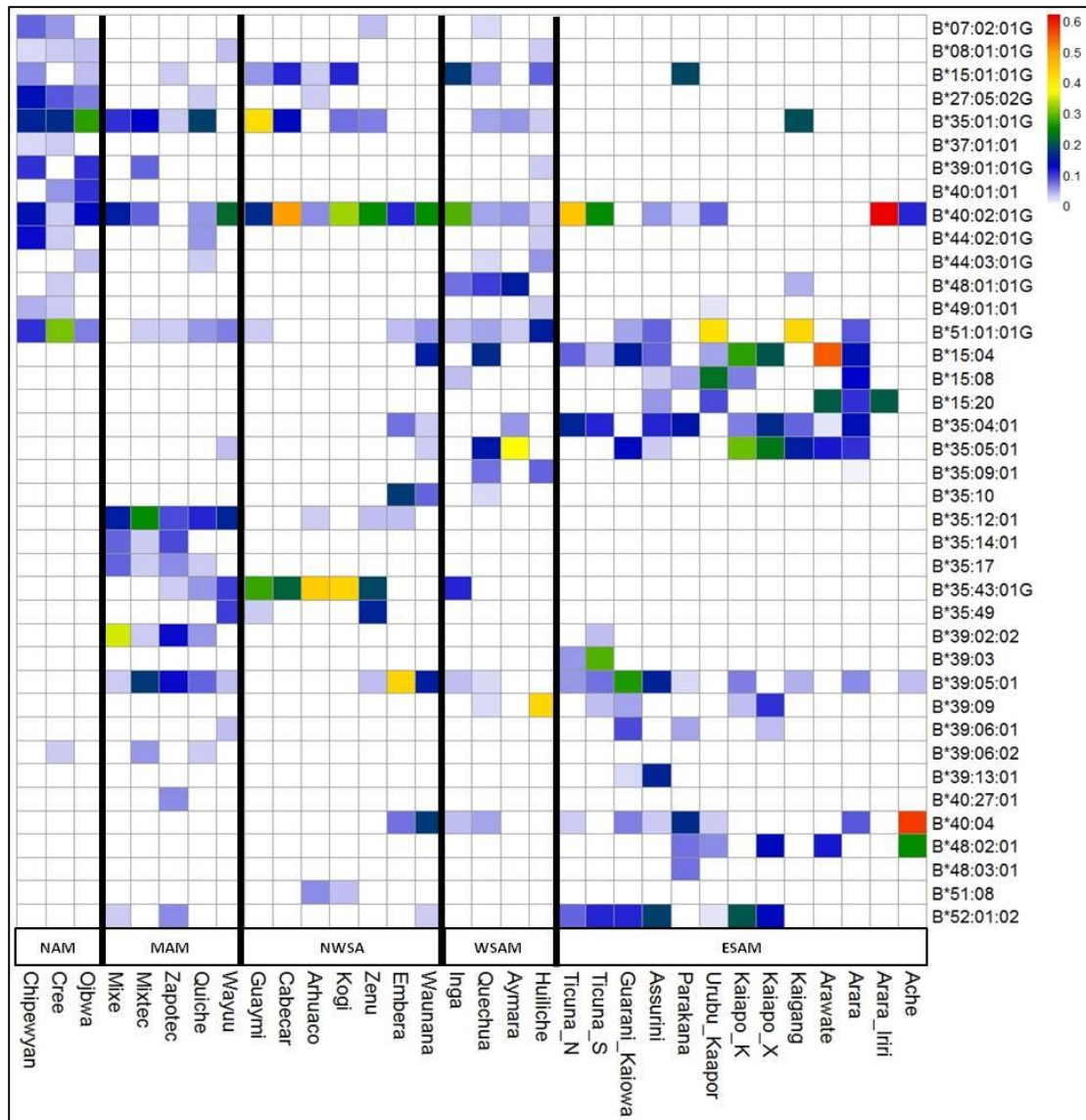
The West South America region (Andean region or SAW) collects a heterogeneous set of populations regarding the allelic content for the *HLA-B* locus. We found the

parental alleles *B\*15:01:01G*, *B\*35:01:01G*, *B\*40:02:01G*, *B\*48:01:01G* scattered along this geographic region but some populations missed one or another of these alleles (the Inga, Aymara and Huilliche populations lack *B\*35:05:01*, *B\*15:01:01G* and *B\*48:01:01G* alleles respectively)(figure 2-3). On the other hand, these populations differed from each other on the gene conversion/recombination allele content. The Inga population presented the *B\*35:43:01G* allele at 10.7% while the more common allele at the Aymara population was the *B\*35:05:01* (37.5%). At the Huilliche population, we observed *B\*39:09* (43.5%) and finally, for the Quechua population we observed *B\*15:04* and *B\*35:05:01* alleles at the frequencies of 16.7 and 14.3%, respectively (figure 2-3).

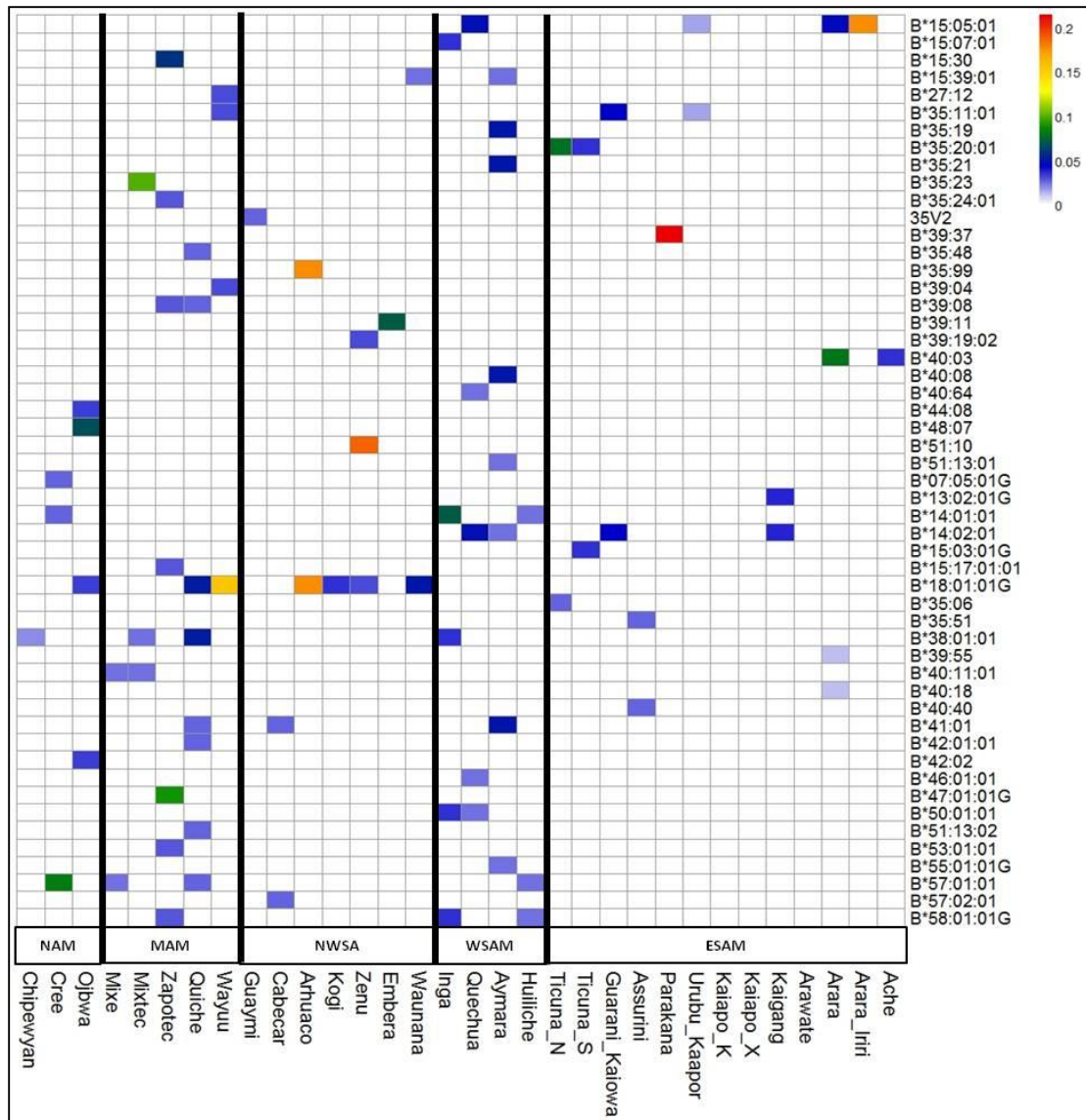
As we discussed before, we observed the alleles resulting from gene conversion/recombination at high frequencies in the eastern region of the South American (figure 2.2 and S2). Only two parental alleles were observed in more than one population at elevated frequencies: *B\*40:02:01G* (reaching its maximum frequency on the region, 61.8%, at the Arara do Iriri population, figure 2.2) and *B\*51:01:01G* (reaching its maximum frequency on the region, 40.9%, at the Urubu Kaapor population, figure 2.3). The *B\*15:01:01G* and *B\*35:01:01G* alleles were also observed, but only in the Parakana and Kaingang populations, at the frequencies of 19 and 19.2%, respectively (figure 2-3). The *B\*15:04*, *B\*15:08*, *B\*15:20*, *B\*35:04:01*, *B\*35:05:01*, *B\*39:05:01*, *B\*39:09*, *B\*40:04*, *B\*48:02:01* and *B\*52:01:02* (at the accumulated frequency of 64.8% in the region) comprise most of the variation of *HLA-B* at the eastern regions of the south American continent, and markedly contrast with the allelic composition we observed in North American Populations (figure 2-3).

There is a vast list of other alleles for *HLA-B* which we did not discuss (figure 2-4). Most of these alleles were found at low frequencies and/or exclusive to only one population. Some of these rare alleles were also the result of recombination and gene conversion and are listed on figure S2. Other alleles, for instance the *B\*14:01:01*,

*B\*14:02:01*, *B\*18:01:01G* and *B\*50:01:01*, which were observed at relatively high frequencies in our sample (figure 2-4) were probably introduced on the Amerindian population by recent gene flow with Africans and Europeans. The global distribution of these alleles is in agreement with this explanation, since they were described at high frequencies in Africans and Europeans and at very low frequencies in other populations (Gonzalez-Galarza et al. 2011; Solberg et al. 2008).



**Figure 2.9. Color grid representing the *HLA-B* allelic frequencies along the American Continent (part 1):** the populations are represented as columns and are grouped by geographic region, where NAM, MAM, NWSA, WSAM and ESAM represent North America, Meso-America, Northwest South America, West South America and East South America, respectively. Alleles were represented in rows.



**Figure 2.10. Color grid representing the *HLA-B* allelic frequencies along the American Continent (part 2):** the populations were represented at the columns and grouped by geographic region, where NAM, MAM, NWSA, WSAM and ESAM mean North America, Meso-America, South America Northwest, West South America and East South America, respectively. Alleles were represented at the rows.

Differently from what we observed at *HLA-B* locus, the *HLA-C* alleles we observed at the southern region of the American continent proved to be a subset from those we found in North American populations. The *C\*01:02:01G*, *C\*02:02:02G*, *C\*03:04:01G*, *C\*04:01:01G*, *C\*04:04:01*, *C\*05:01:01G*, *C\*06:02:01G*, *C\*07:01:01G*, *C\*07:02:01G*, *C\*12:03:01G* and *C\*15:02:01G* alleles together comprised 88.8% of the variation in North American native populations. From these alleles, *C\*02:02:02G*,

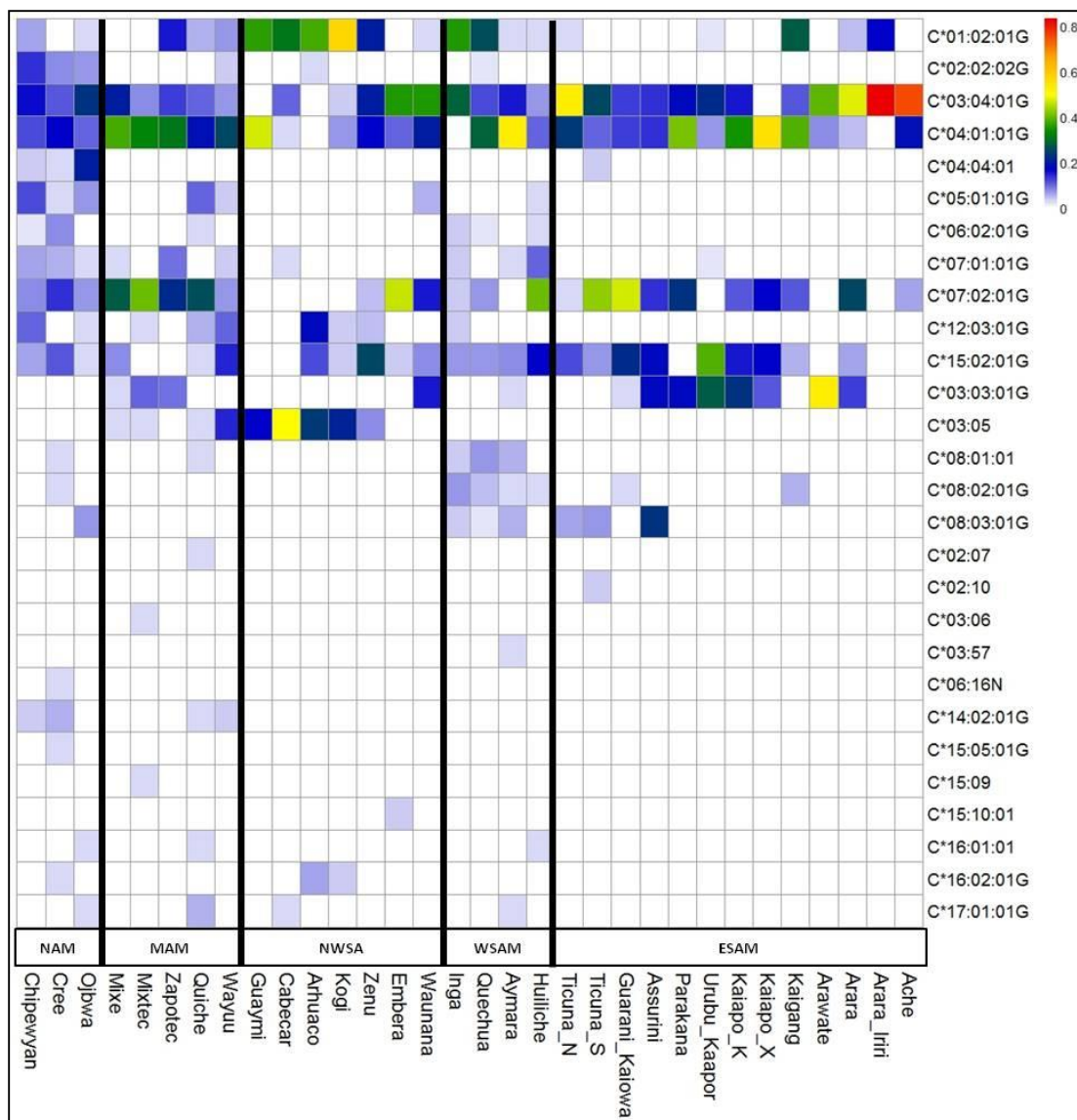
*C\*04:04:01*, *C\*05:01:01G*, *C\*06:02:01G*, *C\*07:01:01G* and *C\*12:03:01G* were absent or found at lower frequencies outside North America, but the remaining alleles were found along all the American continent (figure 2.5).

More than half of the allelic frequency (55.1%) at Meso-American populations was contributed by *C\*04:01:01G* and *C\*07:02:01G* alleles. Furthermore, the *C\*01:02:01G*, *C\*03:04:01G*, *C\*07:01:01G*, *C\*12:03:01G* and *C\*15:02:01G* alleles made up together an additional 27.5% of the distribution (figure 2.5). As we observed for *HLA-B* locus, the Chibcha and Paezan populations also differed on the allelic composition for *HLA-C*. The more common allele we found at the Chibcha populations was *C\*01:02:01G* (at the cumulative frequency of 36.9%) followed by *C\*03:05* (at the accumulated frequency of 25.6%) (Figure 2.5). The *C\*03:05* allele was previously described at Northeast Asia (Gonzalez-Galarza et al. 2011; Solberg et al. 2008) and its origin cannot be associated to a single gene conversion or recombination event involving others *HLA-C* alleles we observed in this study (Robinson 2003). The more common alleles at Paezan populations were *C\*03:04:01G*, *C\*07:02:01G* and *C\*04:01:01G* at the accumulated frequency of 35.3, 27.9 and 16.2%, respectively (figure 2-5).

At the west region of the South American continent, the *C\*01:02:01G*, *C\*03:04:01G*, *C\*04:01:01G*, *C\*07:02:01G* and *C\*15:02:01G* alleles made up 77.9% of the frequency distribution (figure 2-5). We also observed *C\*08* alleles at relatively higher frequencies (*C\*08:01:01*, *C\*08:02:01G* and *C\*08:03:01G*, at the accumulated frequency of 11.4%). We also observed *C\*08* alleles in North American populations but at lower frequencies (figure 2-5) and these alleles were previously described in Asiatic, African and European populations (Gonzalez-Galarza et al. 2011; Solberg et al. 2008).

At the eastern regions of the South American continent we found the *C\*03:04:01G*, *C\*04:01:01G*, *C\*07:02:01G* and *C\*15:02:01G* alleles at the accumulated frequency of 78% (figure 2-5). Apart from these, we found the *C\*03:03:01G* allele at

relatively high frequency (15.2%) (Figure 2-5). This allele was found in other American regions but at lower frequencies, and its sequence differs from *C\*03:04:01G* in only one nucleotide at the last position of the exon 2 (Robinson 2003).



**Figure 2.11. Color grid representing the *HLA-C* allelic frequencies along the American Continent:** the populations were represented at the columns and grouped by geographic region, where NAM, MAM, NWSA, WSAM and ESAM mean North America, Meso-America, South America Northwest, West South America and East South America, respectively. Alleles were represented at the rows.

As the same way we observed for *HLA-C*, the North American populations collect several of the frequent alleles we observed at the American continent (except by *DRB1\*04:11:01*, *DRB1\*08:07* and *DRB1\*14:06:01*)(figure 2.6). On the other hand, *HLA-DRB1* locus presented a higher difference on the allelic frequencies between populations than the observed for *HLA-C* (figure 2.1) getting close to the pattern we found for *HLA-B* locus.

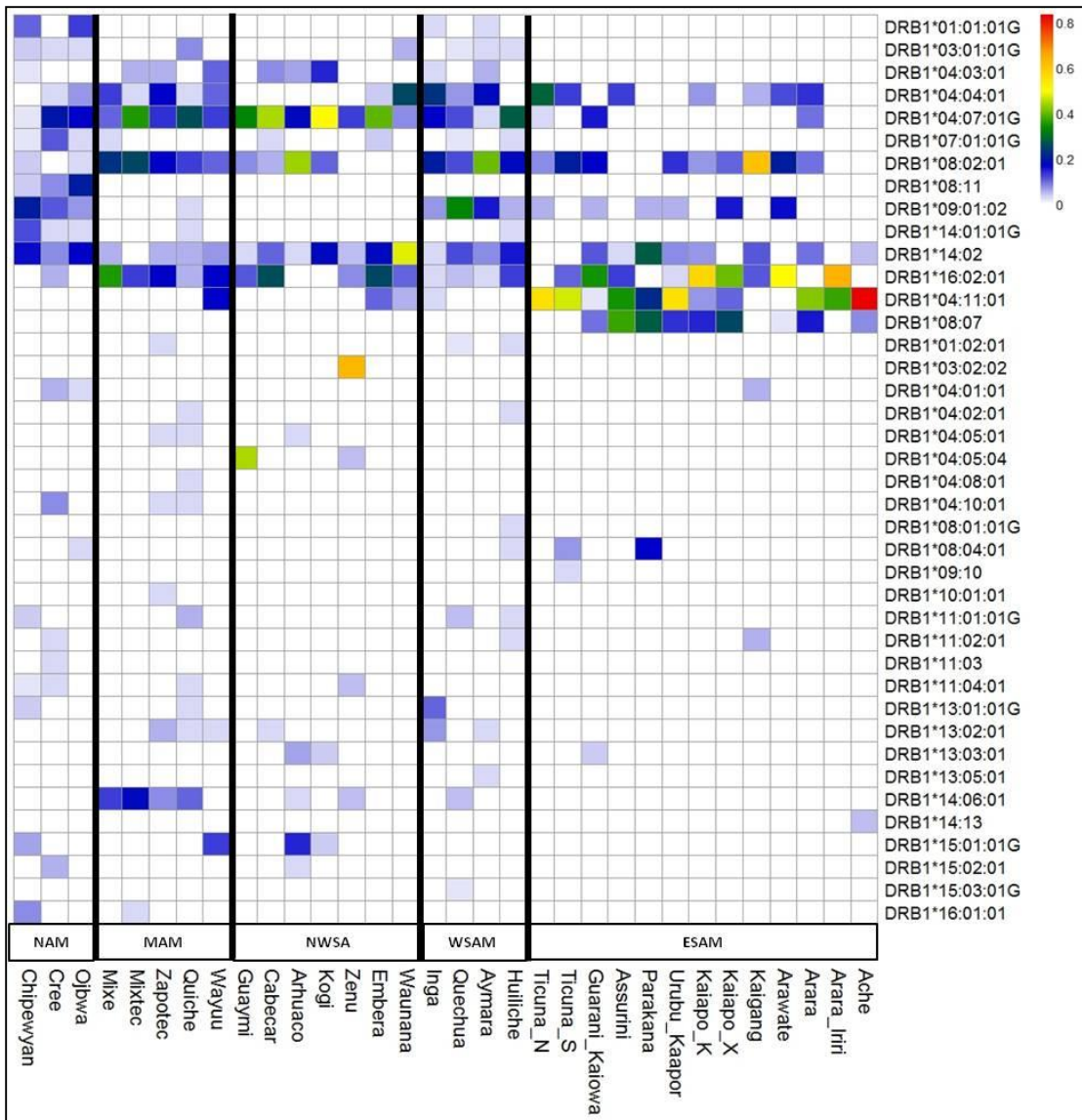
The more frequent alleles we observed in North America were *DRB1\*04:07:01G*, *DRB1\*07:01:01G*, *DRB1\*08:11*, *DRB1\*09:01:02*, *DRB1\*14:01:01G* and *DRB1\*14:02*, at the accumulated frequency of 60.3% (figure 2.6). Heading south to the Meso-American region, we have found a second set of allele at elevated frequency. The *DRB1\*04:04:01*, *DRB1\*04:07:01G*, *DRB1\*08:02:01*, *DRB1\*14:02*, *DRB1\*14:06:01* and *DRB1\*16:02:01* were found at the accumulated frequency of 78.8% and except by *DRB1\*14:06:01*, all of them have been observed in North American populations (figure 2.6). The *DRB1\*14:06:01* allele were previously described outside the Americas presenting elevated frequencies at Northeast Asia (Gonzalez-Galarza et al. 2011; Robinson 2003; Solberg et al. 2008). We also observed the *DRB1\*04:11:01* and *DRB1\*15:01:01G* allele at relatively high frequencies (16.7 and 13.3%, respectively) in the Wayuu population. *DRB1\*04:11:01* were found at high frequencies in South American populations as we will discuss later and *DRB1\*15:01:01G* presents elevated frequency in Europeans and Asiatic populations, but is very rare in other Amerindian populations (Gonzalez-Galarza et al. 2011; Solberg et al. 2008).

The Chibcha-Paez populations (NWSAM) shared less alleles at *HLA-DRB1* when compared with the other loci. Two alleles were found in all population: *DRB1\*04:07:01G* and *DRB1\*14:02*, with frequencies ranging from 7.5 to 50% and 2.8 to 47.5%, respectively (figure 2.6). The *DRB1\*04:03:01*, *DRB1\*08:02:01* and *DRB1\*16:02:01* were also observed, but not at all populations (figure 2.6). The *DRB1\*03:02:02*,

*DRB1\*04:05:02* and *DRB1\*15:01:01G* alleles were observed in the Zenu, Guaymi and Arhuaco populations and are very common at non-Amerindian populations (Gonzalez-Galarza et al. 2011; Solberg et al. 2008), suggesting a recent introgression into the continent by gene flow.

West South American populations were very similar to North American population regarding the allelic content (figure 2.6). We observed the *DRB1\*04:04:01*, *DRB1\*04:07:01G*, *DRB1\*08:02:01*, *DRB1\*09:01:02*, *DRB1\*14:02* and *DRB1\*16:02:01* at the accumulated frequency of 73.7% (figure 2.6) and all of them were observed in North American populations. The *DRB1\*09:01:02* alongside *DRB1\*01:01:01G* and *DRB1\*03:01:01G* which were very rare or absent at Meso-American and Chibcha-Paezan population, have appeared again at the west South American populations.

At the populations from East South America we have found the *DRB1\*04:04:01*, *DRB1\*04:07:01G*, *DRB1\*08:02:01*, *DRB1\*09:01:02*, *DRB1\*14:02* and *DRB1\*16:02:01* alleles at the accumulated frequency 55.2%. These alleles were also found at the North American populations (figure 2.6). The *DRB1\*04:11:01* and *DRB1\*08:07* alleles, which we did not observe in North American population, reached high frequencies in eastern South American population (at the accumulated frequencies of 30.3 and 11.4%, respectively). The patterns of frequency distribution from these two alleles resemble the one we described for *HLA-B*, but the origin of these two *HLA-DRB1* alleles cannot be associated to single recombination/gene conversion/mutation events involving alleles found in North American or other region of the continent (S2).



**Figure 2.12.** Color grid representing the *HLA-DRB1* allelic frequencies along the American Continent: the populations were represented at the columns and grouped by geographic region, where NAM, MAM, NWSA, WSAM and ESAM mean North America, Meso-America, South America Northwest, West South America and East South America, respectively. Alleles were represented in rows.

The *HLA-C* and *-DRB1* allelic distributions were very similar and distinct from which we observed for *HLA-B* locus: most of the alleles we have observed for the formers loci was found on the North American populations (figure 2.5 and 2.6) while at least other two allelic components beside the one observed in North America could be observed on the *HLA-B* locus, one at the Meso-American populations and other at the east South America (figure 2.2).

### 5.1. Main Haplotypes Frequency Distribution.

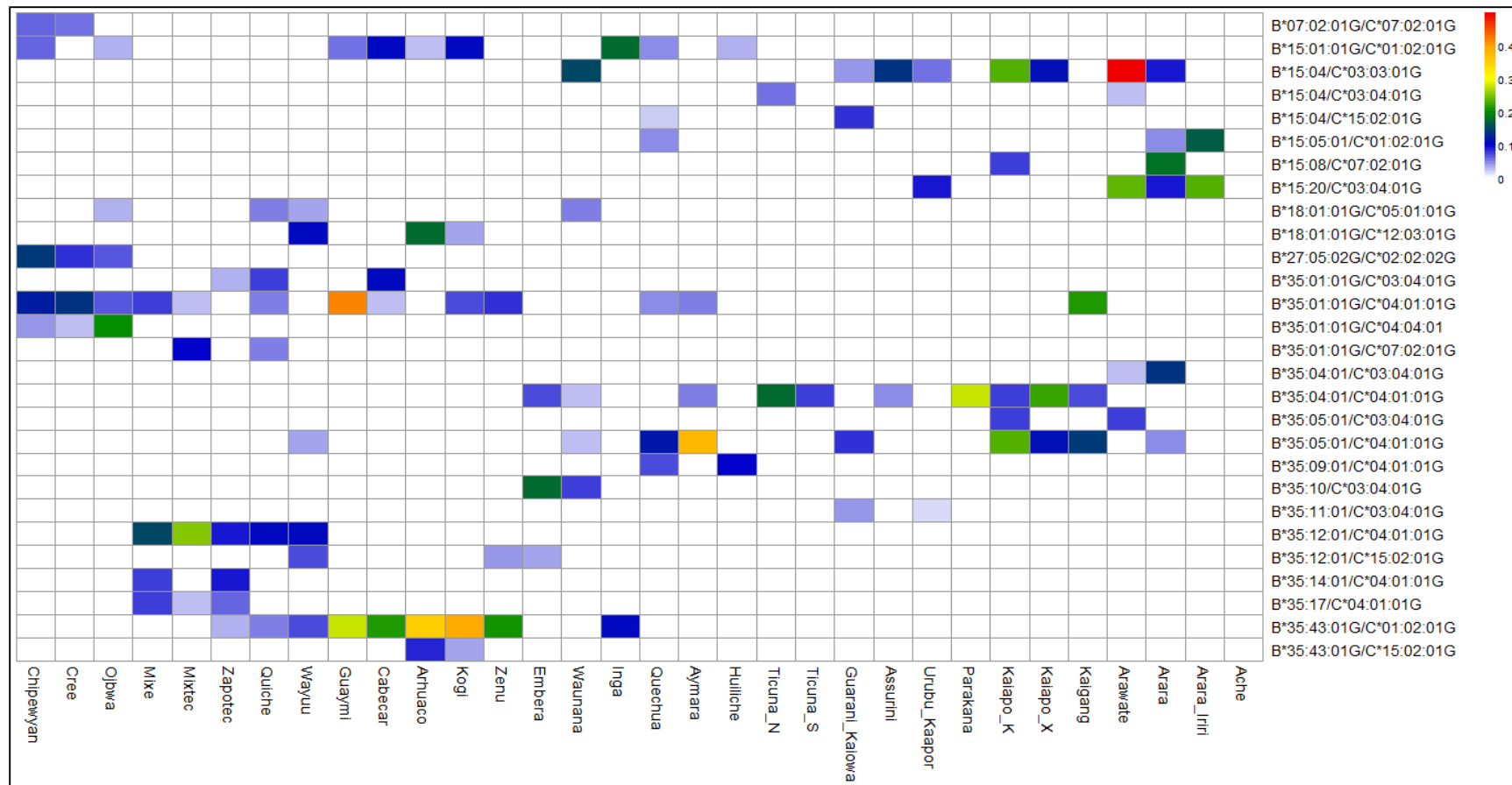
We observed an incredible high number of haplotypes involving *HLA-B*, *-C* and *-DRB1* alleles (180), even after the removal of the estimations with poor accuracy. Even taking only the haplotypes involving the *HLA-B* and *-C* loci, which are physically linked, 123 different haplotypes were observed (the entire haplotype list is provided in Supplementary Material 2 (S2)).

Multiple *HLA-DRB1* alleles were observed in combination with individuals *HLA-B/-C* haplotype (figure 2.7) which is expected due to the distance between *HLA-DRB1* and the other genes (1330 megabases (MB)). Despite this pattern, we observed some haplotypes spread over vast geographical regions at high frequencies. We selected the haplotypes that were observed at least in three populations and presented them on table 2.2. Six  $\pi$  of the ten haplotypes that fulfilled this criterion contained a *HLA-B* allele which probably originated by gene conversion (or recombination in the case of *B\*35:43:01G*). The other four haplotypes included the *B\*40:02:01G* allele, two of them with *C\*03:04:01* and the other two with *C\*03:05* (table 2.2). The *DRB1\*04:11:01* and *DRB1\*16:02:01* were observed in four haplotype, two of them being the most frequent in this study (table 2.2).

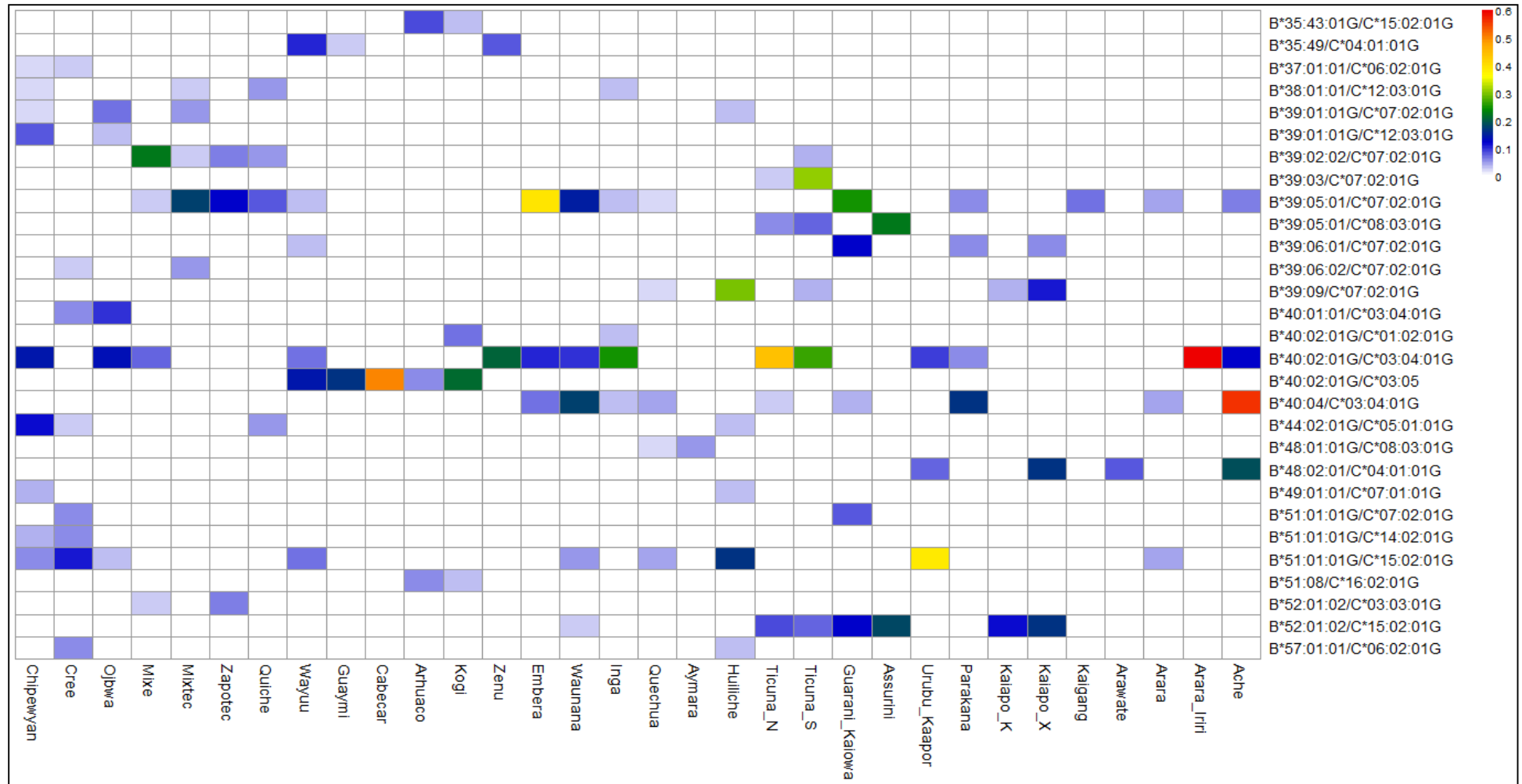
Considering only the *HLA-B* and *-C* loci, a general pattern can be drawn: the *HLA-B* gene conversion/recombination alleles were often observed in association with the same *-C* alleles then their parental alleles, except by *B\*15:04*, which was associated with *C\*03:03:01G* while *B\*15:01:01G* was linked to *C\*01:02:01G* (figure 2.7).

**Table 1.2.** Main *HLA-C/-B/-DRB1* haplotypes frequencies.

<i>HLA-C/-B/-DRB1</i> Haplotype	Populations	Average Frequency	min - max
<b><i>C*03:03:01G/B*15:04/DRB1*16:02:01</i></b>	Assurini, Kaiapo_K and Arawete	24.02%	9.09 - 43.75%
<b><i>C*03:04:01G/B*40:04/DRB1*04:11:01</i></b>	Ticuna_N, Arara and Ache	21.58%	2.94 - 56.25%
<b><i>C*01:02:01G/B*35:43:01G/DRB1*04:07:01G</i></b>	Quiche, Wayuu, Guaymi, Cabecar, Arhuaco, Kogi and Inga	17.17%	5.26 - 39.29%
<b><i>C*03:04:01G/B*40:02:01G/DRB1*04:11:01</i></b>	Wayuu, Ticuna_N and Ticuna_S	13.98%	7.14 - 26.47%
<b><i>C*07:02:01G/B*39:05:01/DRB1*04:07:01G</i></b>	Mixe, Mixtec, Zapotec, Embera and Guarani Kaiowá	13.42%	2.5 - 28.57%
<b><i>C*04:01:01/B*35:12:01/DRB1*08:02:01</i></b>	Mixe, Mixtec, Zapotec and Quiche	11.19%	7.89 - 15%
<b><i>C*03:05/B*40:02:01G/DRB1*04:07:01G</i></b>	Mixtec, Guaymi, Cabecar and Zenu	10.02%	2.5 - 23.68%
<b><i>C*03:04:01G/B*40:02:01G/DRB1*04:01:01</i></b>	Waunana, Embera, Ticuna_N and Ticuna_S	9.68%	3.57 - 17.65%
<b><i>C*03:05/B*40:02:01G/DRB1*04:03:01</i></b>	Cabecar, Arhuaco and Kogi	8.48%	5.26 - 14.29%
<b><i>C*15:02:01G/B*52:01:02/DRB1*16:02:01</i></b>	Waunana, Ticuna_S and Kaiapo_K	6.75%	2.5 - 8.33%



**Figure 2.13.** Color grid representing the *HLA-B/-C* haplotypes along the American Continent (part 1): the populations were represented at the columns. Alleles were represented at the rows.



**Figure 2.12. Color grid representing the *HLA-B/-C* haplotypes along the American Continent (part 2):** the populations were represented at the columns. Alleles were represented in rows.

## Considerações Finais

Os genes *HLA* apresentam os mais altos níveis de variação do genoma e uma vasta gama de evidências apontam a seleção natural dirigida por patógenos como o principal fator evolutivo contribuindo para a manutenção dessa variação. Partindo desse pressuposto, nossa motivação consistiu na identificação dos principais alvos da atuação da seleção natural nos genes *HLA*. No primeiro capítulo, nosso objetivo foi a caracterização das regiões que apresentam as mais fortes evidências da atuação de seleção dentro dos genes *HLA-A* e *-B*. No segundo capítulo, nos perguntamos quais dentre os genes *HLA-B*, *-C* e *-DRB1* evoluíram mais rapidamente durante a ocupação relativamente recente do continente Americano.

As diferenças nos regimes seletivos atuantes nas distintas porções das moléculas HLA foram inicialmente exploradas por Hughes e Nei (1988) que demonstraram que os códons dos aminoácidos que compõem a fenda de ligação dos peptídeos apresentam uma maior taxa de variação não sinônima quando comparados aos demais códons dos genes *HLA*. Entretanto, esse estudo não levou em consideração as diferenças entre as regiões da fenda de ligação dos peptídeos. Como descrevemos na introdução do primeiro manuscrito, há evidências que apontam os bolsões B e F como os mais críticos na restrição do repertório de peptídeos apresentados pelas moléculas HLA de classe I. Além disso, Sidney e colaboradores (1996), Sette e Sidney (1999) e Sidney e colaboradores (2008) demonstraram que os alelos dos genes *HLA-A* e *-B* poderiam ser classificados em quatro e cinco supertipos, respectivamente, baseando-se no grau de compartilhamento do repertório de apresentação de peptídeos, uma propriedade definida basicamente pela variação nos bolsões B e F. Baseando-se nessas características, nos perguntamos se os bolsões B e F apresentariam níveis mais elevados de variação e se os supertipos estariam

pouco estruturados entre as populações humanas, sinais condizentes com a atuação da seleção balanceadora.

Essas hipóteses foram parcialmente corroboradas na análise do gene *HLA-B*. O sítio de ligação de peptídeos desse gene apresentou níveis significativamente mais elevados de variação quando comparado à mesma região em *HLA-A*. Também demonstramos que o bolsão B concentra boa parte da variação observada em *HLA-B* e a maioria das diferenças entre os supertipos desse loco. Por outro lado, o bolsão F apresentou os níveis mais baixos de variação em ambos os *loci* e boa parte das diferenças entre os supertipos de *HLA-A*.

Esses resultados indicam uma provável restrição evolutiva na variação do bolsão F que pode ter sido imposta pelos padrões de geração e transporte dos peptídeos até a luz do retículo endoplasmático, realizados pelo proteossomo e proteínas TAP, respectivamente. Essa interpretação é corroborada pelo resultado de que boa parte dos peptídeos extraídos de moléculas HLA de classe I apresentam extremidades carboxila hidrofóbicas e essa limitação pode ser a força por trás dos baixos níveis de diversidade no bolsão F. Por outro lado, a restrição da variação no bolsão F provavelmente tornou vantajosa a maximização da variação no bolsão B, pois indivíduos contendo alelos que codificam bolsões B distintos teriam uma capacidade mais ampla de ligação de peptídeos e consequentemente, de desencadear uma resposta imunológica.

Os elevados níveis de variação no bolsão B de *HLA-B* corroboram essa ideia, indicando que esse bolsão provavelmente é o alvo da atuação da seleção balanceadora nesse loco. Em *HLA-A*, por outro lado, o bolsão B não é a região com os mais altos níveis de variação indicando que outras regiões da fenda podem contribuir para a ancoragem dos peptídeos na molécula, ou que o repertório de apresentação de peptídeos pelas moléculas HLA-A seja mais restrito e/ou conservado quando comparado a HLA-B. Os resultados com as simulações envolvendo os supertipos corroboram as interpretações obtidas a partir

das análises moleculares, uma vez que as simulações não foram capazes de reproduzir os padrões observados de variação e diferenciação populacional que obtivemos com os supertipos de *HLA-B*. Esses resultados permitem a interpretação de que as especificidades de ligação de peptídeo das moléculas *HLA-B* encontram-se mais estruturadas entre as populações do que seria esperado pelo acaso o que pode ser um reflexo de adaptações locais contra patógenos.

Os nossos resultados permitem a inferência de que as moléculas *HLA-B* provavelmente possuem os mais amplos repertórios de apresentação de peptídeos. Essa maior diversidade na capacidade de apresentação seria uma consequência dos níveis mais elevados de variação no bolsão B associados à estruturação dessa variação entre os supertipos, fazendo dessas moléculas a principal linha de defesa contra novos agentes patogênicos ou mutações de escape. Essa interpretação está completamente de acordo os resultados que apresentamos no segundo capítulo dessa tese.

O gene *HLA-B* apresenta vários alelos originados por conversão gênica e/ou recombinação, presentes em altas frequências nas regiões tropicais das Américas, especialmente no leste da América do Sul, sendo praticamente ausentes nas populações nativas da América do Norte. Esse perfil alélico distinto foi descoberto no início da década de 90 e a relevância deste achado deve-se ao contraste com outros marcadores genéticos, incluindo outros genes de *HLA*, cuja diversidade parece ser um subconjunto do encontrado em outras regiões das Américas e nordeste da Ásia. Para explicar esse padrão incomum, Parham e colaboradores (1997) propuseram a hipótese *turnover* de alelos que aponta a seleção imposta pelos patógenos endêmicos das regiões tropicais das Américas como a principal força que contribuiu para a substituição dos alelos ancestrais e aumento da frequência dos novos alelos produtos de recombinação e conversão de gênica. Apesar de intuitiva, essa hipótese nunca foi formalmente testada, não havendo estudos

populacionais que tenham comparado as frequências dos novos alelos de *HLA-B* com desvios do esperado segundo o modelo de neutralidade e equilíbrio populacional.

No segundo capítulo dessa tese nós revisitamos esse modelo e demonstramos que as populações do leste América do Sul (SAE) apresentam as maiores frequências dos alelos de *HLA-B* gerados por conversão gênica/recombinação. Também demonstramos que boa parte da variação observada nos *loci* *HLA-C* e *-DRB1* foi constituída por um subconjunto do observado nas demais regiões do continente Americano.

Corroborando com os dados presentes na literatura, demonstramos que as populações SAE apresentam níveis de variação obtidos a partir de 61 microssatélites espalhados pelo genoma e dos genes *HLA-C* e *-DRB1*, significativamente inferiores ao observado nas demais regiões do continente. Esse resultado está de acordo com a interpretação de que as populações SAE experimentaram elevadas taxas de deriva genética, causando uma diminuição global dos níveis de variação do genoma. Entretanto, contrastando com esse padrão geral, demonstramos que os níveis de diversidade do gene *HLA-B* não sofreram esse mesmo decaimento, indicando que a seleção natural provavelmente contribuiu na manutenção da diversidade nesse loco no leste da América do Sul.

Esse resultado indicou uma provável relação entre a presença dos novos alelos de *HLA-B* e a atuação da seleção natural. Essa reação foi confirmada com a aplicação do teste D de Tajima, uma vez que as populações do leste da América do Sul apresentaram para o gene *HLA-B* os mais elevados e significativos valores para essa estatística. Por outro lado, não observamos esse mesmo padrão nos genes *HLA-C* e *-DRB1*. Valores significativos e positivos de D de Tajima são obtidos de populações que apresentam alelos divergentes em frequências intermediárias. Como já discutimos, as populações SAE apresentam as mais altas frequências dos alelos de *HLA-B* originados por eventos de conversão gênica/recombinação. Esses alelos são muito divergentes, diferindo um dos

outros em múltiplas posições e a sua presença em frequências intermediárias explica os valores elevados de Tajima D que obtivemos. O aumento da frequência dos alelos gerados por conversão gênica/recombinação em detrimento de alelos gerados por mutações de ponto foi provavelmente a consequência da combinação da maior velocidade geração de novos alelos associada à maior probabilidade de criação de novas especificidades peptídicas. Dentre os genes que analisamos, *HLA-B* foi o que teve o maior número de alelos que foram aumentados de frequência desde que o leste da América do Sul foi ocupado.

As conclusões obtidas a partir de ambas as abordagens do presente trabalho apontam o gene *HLA-B* como o principal alvo da seleção natural, uma vez que esse loco concentra as maiores evidências de atuação de seleção natural recente quando comparado aos demais genes *HLA* analisados. Nós também demonstramos com as análises intragênicas que o bolsão B do sítio de ligação de peptídeos de *HLA-B* concentra por boa parte das diferenças observadas entre as populações, implicando em diferenças nos perfis de apresentação de peptídeos entre essas mesmas populações, o que pode ser interpretado como um indicativo de adaptações locais aos conjuntos de patógenos presentes em distintas regiões geográficas.

## Referencias Bibliograficas

- Albrechtsen, Anders, Ida Moltke, and Rasmus Nielsen. 2010. "Natural selection and the distribution of identity-by-descent in the human genome." *Genetics* 186(1): 295–308.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2940294&tool=pmcentrez&rendertype=abstract> (August 7, 2013).
- Apanius, V, D Penn, P R Slev, L R Ruff, and W K Potts. 1997. "The nature of selection on the major histocompatibility complex." *Critical reviews in immunology* 17(2): 179–224.  
<http://www.ncbi.nlm.nih.gov/pubmed/9094452> (August 19, 2012).
- Bamshad, Michael, and Stephen P Wooding. 2003. "Signatures of natural selection in the human genome." *Nature reviews. Genetics* 4(2): 99–111.  
<http://www.ncbi.nlm.nih.gov/pubmed/12560807> (September 20, 2013).
- Borghans, José A M, Joost B Beltman, and Rob J De Boer. 2004. "MHC polymorphism under host-pathogen coevolution." *Immunogenetics* 55(11): 732–9. <http://www.ncbi.nlm.nih.gov/pubmed/14722687> (August 14, 2013).
- Buhler, Stéphane, and Alicia Sanchez-Mazas. 2011. "HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events." *PloS one* 6(2): e14643.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3051395&tool=pmcentrez&rendertype=abstract> (September 12, 2011).
- Bustamante, Carlos D, Adi Fladel-Alon, Scott Williamson, Rasmus Nielsen, Melissa Todd Hubisz, Stephen Glanowski, David M Tanenbaum, Thomas J White, John J Sninsky, Ryan D Hernandez, Daniel Civello, Mark D Adams, Michele Cargill, and Andrew G Clark. 2005. "Natural selection on protein-coding genes in the human genome." *Nature* 437(7062): 1153–7. <http://www.ncbi.nlm.nih.gov/pubmed/16237444> (July 18, 2011).
- Campos-Lima, P O, V Levitsky, M P Imreh, R Gavioli, and M G Masucci. 1997. "Epitope-dependent selection of highly restricted or diverse T cell receptor repertoires in response to persistent infection by Epstein-Barr virus." *The Journal of experimental medicine* 186(1): 83–9.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2198955&tool=pmcentrez&rendertype=abstract> (August 19, 2012).

- Carrington, M, G W Nelson, M P Martin, T Kissner, D Vlahov, J J Goedert, R Kaslow, S Buchbinder, K Hoots, and S J O'Brien. 1999. "HLA and HIV-1: heterozygote advantage and B\*35-Cw\*04 disadvantage." *Science (New York, N.Y.)* 283(5408): 1748–52. <http://www.ncbi.nlm.nih.gov/pubmed/10073943> (August 19, 2012).
- Chakraborty, Sajib, Taibur Rahman, Rajib Chakravorty, Alison Kuchta, Atai Rabby, and Munsu Sahiuzzaman. 2013. "HLA supertypes contribute in HIV type 1 cytotoxic T lymphocyte epitope clustering in Nef and Gag proteins." *AIDS research and human retroviruses* 29(2): 270–8. <http://www.ncbi.nlm.nih.gov/pubmed/23061377> (September 13, 2013).
- Cordery, Damien V, Allison Martin, Janaki Amin, Anthony D Kelleher, Sean Emery, and David A Cooper. 2012. "The influence of HLA supertype on thymidine analogue associated with low peripheral fat in HIV." *AIDS (London, England)* 26(18): 2337–44. <http://www.ncbi.nlm.nih.gov/pubmed/23032422> (September 13, 2013).
- Doherty, P C, and R M Zinkernagel. 1975. "A biological role for the major histocompatibility antigens." *Lancet* 1(7922): 1406–9. <http://www.ncbi.nlm.nih.gov/pubmed/49564> (August 19, 2012).
- Ewens, W J. 1972. "The sampling theory of selectively neutral alleles." *Theoretical population biology* 3(1): 87–112. <http://www.ncbi.nlm.nih.gov/pubmed/4667078> (August 26, 2013).
- Excoffier, Laurent, and Heidi E L Lischer. 2010. "Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows." *Molecular ecology resources* 10(3): 564–7. <http://www.ncbi.nlm.nih.gov/pubmed/21565059> (July 8, 2011).
- Garrigan, Daniel, and Philip W Hedrick. 2003. "Perspective: detecting adaptive molecular polymorphism: lessons from the MHC." *Evolution; international journal of organic evolution* 57(8): 1707–22. <http://www.ncbi.nlm.nih.gov/pubmed/14503614>.
- Gilchuk, Pavlo, Charles T Spencer, Stephanie B Conant, Timothy Hill, Jennifer J Gray, Xinnan Niu, Mu Zheng, John J Erickson, Kelli L Boyd, K Jill McAfee, Carla Oseroff, Sine R Hadrup, Jack R Bennink, William Hildebrand, Kathryn M Edwards, James E Crowe, John V Williams, Søren Buus, Alessandro Sette, Ton N M Schumacher, Andrew J Link, and Sebastian Joyce. 2013. "Discovering naturally processed antigenic

- determinants that confer protective T cell immunity.” *The Journal of clinical investigation* 123(5): 1976–87.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3635741&tool=pmcentrez&rendertype=abstract> (August 8, 2013).
- Gillespie, J H. 1977. “Sampling theory for alleles in a random environment.” *Nature* 266(5601): 443–5.  
<http://www.ncbi.nlm.nih.gov/pubmed/859613> (August 19, 2012).
- Gonzalez-Galarza, Faviel F, Stephen Christmas, Derek Middleton, and Andrew R Jones. 2011. “Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations.” *Nucleic acids research* 39(Database issue): D913–9.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013710&tool=pmcentrez&rendertype=abstract> (August 7, 2013).
- Guo, S W, and E A Thompson. 1992. “Performing the exact test of Hardy-Weinberg proportion for multiple alleles.” *Biometrics* 48(2): 361–72.  
<http://www.ncbi.nlm.nih.gov/pubmed/1637966> (August 19, 2012).
- Hammer, Michael F, Felisa Blackmer, Dan Garrigan, Michael W Nachman, and Jason A Wilder. 2003. “Human population structure and its effects on sampling Y chromosome sequence variation.” *Genetics* 164(4): 1495–509.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1462677&tool=pmcentrez&rendertype=abstract> (September 18, 2013).
- Hill, A V, A Jepson, M Plebanski, and S C Gilbert. 1997. “Genetic analysis of host-parasite coevolution in human malaria.” *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 352(1359): 1317–25.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1692024&tool=pmcentrez&rendertype=abstract> (August 3, 2012).
- Hraber, Peter, Carla Kuiken, and Karina Yusim. 2007. “Evidence for human leukocyte antigen heterozygote advantage against hepatitis C virus infection.” *Hepatology (Baltimore, Md.)* 46(6): 1713–21.  
<http://www.ncbi.nlm.nih.gov/pubmed/17935228> (August 19, 2012).
- Hughes, A L, and M Nei. 1988. “Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection.” *Nature* 335(6186): 167–70.  
<http://www.ncbi.nlm.nih.gov/pubmed/3412472> (September 26, 2011).
- Jaruzelska, J, E Zietkiewicz, M Batzer, D E Cole, J P Moisan, R Scozzari, S Tavaré, and D Labuda. 1999. “Spatial and temporal distribution of the

neutral polymorphisms in the last ZFX intron: analysis of the haplotype structure and genealogy.” *Genetics* 152(3): 1091–101.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1460666&tool=pmcentrez&rendertype=abstract> (September 18, 2013).

Jost, Lou. 2008. “G ST and its relatives do not measure differentiation.”

*Molecular Ecology* 17(18): 4015–4026.

<http://doi.wiley.com/10.1111/j.1365-294X.2008.03887.x> (June 16, 2011).

Karlsson, Ingrid, Lea Brandt, Lasse Vinner, Ingrid Kromann, Lars Vibe Andreasen, Peter Andersen, Jan Gerstoft, Gitte Kronborg, and Anders Fomsgaard. 2013. “Adjuvanted HLA-supertype restricted subdominant peptides induce new T-cell immunity during untreated HIV-1-infection.” *Clinical immunology (Orlando, Fla.)* 146(2): 120–30.

<http://www.ncbi.nlm.nih.gov/pubmed/23314272> (September 13, 2013).

Karlsson, Ingrid, Henrik Kløverpris, Kristoffer Jarlov Jensen, Anette Stryhn, Søren Buus, Annika Karlsson, Lasse Vinner, Philip Goulder, and Anders Fomsgaard. 2012. “Identification of conserved subdominant HIV Type 1 CD8(+) T Cell epitopes restricted within common HLA Supertypes for therapeutic HIV Type 1 vaccines.” *AIDS research and human retroviruses* 28(11): 1434–43.

<http://www.ncbi.nlm.nih.gov/pubmed/22747336> (September 13, 2013).

Kaslow, R A, M Carrington, R Apple, L Park, A Muñoz, A J Saah, J J Goedert, C Winkler, S J O’Brien, C Rinaldo, R Detels, W Blattner, J Phair, H Erlich, and D L Mann. 1996. “Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection.” *Nature medicine* 2(4): 405–11.

<http://www.ncbi.nlm.nih.gov/pubmed/8597949> (August 19, 2012).

Knowles, L. L. 2003. “The burgeoning field of statistical phylogeography.”

*Journal of Evolutionary Biology* 17(1): 1–10.

<http://doi.wiley.com/10.1046/j.1420-9101.2003.00644.x> (July 22, 2011).

Kuniholm, M H, K Anastos, A Kovacs, X Gao, D Marti, A Sette, R M Greenblatt, M Peters, M H Cohen, H Minkoff, S J Gange, C L Thio, M A Young, X Xue, M Carrington, and H D Strickler. 2013. “Relation of HLA class I and II supertypes with spontaneous clearance of hepatitis C virus.” *Genes and immunity* 14(5): 330–5.

<http://www.ncbi.nlm.nih.gov/pubmed/23636221> (September 13, 2013).

Lazaryan, Aleksandr, Elena Lobashevsky, Joseph Mulenga, Etienne Karita, Susan Allen, Jianming Tang, and Richard A Kaslow. 2006. "Human leukocyte antigen B58 supertype and human immunodeficiency virus type 1 infection in native Africans." *Journal of virology* 80(12): 6056–60. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1472610&tool=pmcentrez&rendertype=abstract> (September 13, 2013).

Meyer, D, and G Thomson. 2001. "How selection shapes variation of the human major histocompatibility complex: a review." *Annals of human genetics* 65(Pt 1): 1–26. <http://www.ncbi.nlm.nih.gov/pubmed/11415519> (December 16, 2011).

Naugler, Christopher, and Robert Liwski. 2008. "An evolutionary approach to major histocompatibility diversity based on allele supertypes." *Medical hypotheses* 70(5): 933–7. <http://www.ncbi.nlm.nih.gov/pubmed/18063318> (October 8, 2011).

Nei, M, and R K Chesser. 1983. "Estimation of fixation indices and gene diversities." *Annals of human genetics* 47(Pt 3): 253–9. <http://www.ncbi.nlm.nih.gov/pubmed/6614868> (September 13, 2013).

Nei, Masatoshi. 1987. "Molecular Evolutionary Genetics."

Ovsyannikova, Inna G, Robert A Vierkant, V Shane Pankratz, Megan M O'Byrne, Robert M Jacobson, and Gregory A Poland. 2009. "HLA haplotype and supertype associations with cellular immune responses and cytokine production in healthy children after rubella vaccine." *Vaccine* 27(25-26): 3349–58. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2693336&tool=pmcentrez&rendertype=abstract> (September 13, 2013).

Parham, P, K L Arnett, E J Adams, A M Little, K Tees, L D Barber, S G Marsh, T Ohta, T Markow, and M L Petzl-Erler. 1997. "Episodic evolution and turnover of HLA-B in the indigenous human populations of the Americas." *Tissue antigens* 50(3): 219–32. <http://www.ncbi.nlm.nih.gov/pubmed/9331945> (December 16, 2011).

Parham, Peter. 2005. "MHC class I molecules and KIRs in human history, health and survival." *Nature reviews. Immunology* 5(3): 201–14. <http://www.ncbi.nlm.nih.gov/pubmed/15719024> (July 15, 2012).

- Piertney, S B, and M K Oliver. 2006. "The evolutionary ecology of the major histocompatibility complex." *Heredity* 96(1): 7–21.  
<http://www.ncbi.nlm.nih.gov/pubmed/16094301> (June 20, 2011).
- Prevosti, A., J. Ocaña, and G. Alonso. 1975. "Distances between populations of *Drosophila subobscura*, based on chromosome arrangement frequencies." *Theoretical and Applied Genetics* 45(6): 231–241.  
<http://link.springer.com/10.1007/BF00831894> (August 16, 2013).
- Prugnolle, Franck, Andrea Manica, Marie Charpentier, Jean François Guégan, and François Balloux. 2007. "UKPMC Funders Group Pathogen-driven selection and worldwide HLA class I diversity." *Genetics* 15(11): 1022–1027.
- Prugnolle, Franck, Andrea Manica, Marie Charpentier, Jean François Guégan, Vanina Guernier, and François Balloux. 2005. "Pathogen-driven selection and worldwide HLA class I diversity." *Current biology : CB* 15(11): 1022–7. <http://www.ncbi.nlm.nih.gov/pubmed/15936272> (July 30, 2012).
- Qutob, Nouar, Francois Balloux, Towfique Raj, Hua Liu, Sophie Marion de Procé, John Trowsdale, and Andrea Manica. 2011. "Signatures of historical demography and pathogen richness on MHC class I genes." *Immunogenetics*. <http://www.ncbi.nlm.nih.gov/pubmed/21947542> (November 30, 2011).
- Reich, David, Nick Patterson, Desmond Campbell, Arti Tandon, Stéphane Mazieres, Nicolas Ray, et al. 2012. "Reconstructing Native American population history." *Nature*: 1–6.  
<http://www.nature.com/doifinder/10.1038/nature11258> (July 12, 2012).
- Robinson, James. 2003. "IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex." *Nucleic Acids Research* 31(1): 311–314.  
<http://nar.oxfordjournals.org/content/31/1/311.short> (July 6, 2011).
- Robinson, James, Kavita Mistry, Hamish McWilliam, Rodrigo Lopez, Peter Parham, and Steven G E Marsh. 2011. "The IMGT/HLA database." *Nucleic acids research* 39(Database issue): D1171–6.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013815&tool=pmcentrez&rendertype=abstract> (September 1, 2013).
- Sanchez-Mazas, Alicia, Jean-François Lemaître, and Mathias Currat. 2012. "Distinct evolutionary strategies of human leucocyte antigen loci in

- pathogen-rich environments.” *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 367(1590): 830–9.  
<http://www.ncbi.nlm.nih.gov/pubmed/22312050> (August 19, 2012).
- Saper, M A, P J Bjorkman, and D C Wiley. 1991. “Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 Å resolution.” *Journal of molecular biology* 219(2): 277–319.  
<http://www.ncbi.nlm.nih.gov/pubmed/2038058> (December 16, 2011).
- Sette, a, and J Sidney. 1999. “Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism.” *Immunogenetics* 50(3-4): 201–12.  
<http://www.ncbi.nlm.nih.gov/pubmed/10602880>.
- Shehzadi, Abida, Shahid Ur Rehman, and Tayyab Husnain. 2012. “Selection of epitope-based vaccine targets of HCV genotype 1 of Asian origin: a systematic in silico approach.” *Bioinformation* 8(20): 957–62.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3524940&tool=pmcentrez&rendertype=abstract> (September 13, 2013).
- Sidney, John, Howard M Grey, Ralph T Kubo, and Alessandro Sette. 1996. “Practical, biochemical and evolutionary implications of the discovery of HLA class I supermotifs.” *Immunology today* 17(6): 261–266.
- Sidney, John, Bjoern Peters, Nicole Frahm, Christian Brander, and Alessandro Sette. 2008. “HLA class I supertypes: a revised and updated classification.” *BMC immunology* 9: 1.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2245908&tool=pmcentrez&rendertype=abstract> (July 17, 2011).
- Slatkin, M, and L Excoffier. 1996. “Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm.” *Heredity* 76 ( Pt 4): 377–83. <http://www.ncbi.nlm.nih.gov/pubmed/8626222> (August 26, 2013).
- Solberg, Owen D, Steven J Mack, Alex K Lancaster, Richard M Single, Yingssu Tsai, Alicia Sanchez-Mazas, and Glenys Thomson. 2008. “Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies.” *Human immunology* 69(7): 443–64.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2632948&tool=pmcentrez&rendertype=abstract> (June 13, 2011).
- Tajima, F. 1983. “Evolutionary relationship of DNA sequences in finite populations.” *Genetics* 105(2): 437–60.

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1202167&tool=pmcentrez&rendertype=abstract>.

Takahata, N, Y Satta, and J Klein. 1992. "Polymorphism and balancing selection at major histocompatibility complex loci." *Genetics* 130(4): 925–38.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1204941&tool=pmcentrez&rendertype=abstract> (September 13, 2012).

The MHC sequencing consortium. 1999. "Complete sequence and gene map of a human major histocompatibility complex." 401(October): 921–923.

Trachtenberg, Elizabeth, Bette Korber, Cristina Sollars, Thomas B Kepler, Peter T Hraber, Elizabeth Hayes, Robert Funkhouser, Michael Fugate, James Theiler, Yen S Hsu, Kevin Kunstman, Samuel Wu, John Phair, Henry Erlich, and Steven Wolinsky. 2003. "Advantage of rare HLA supertype in HIV disease progression." *Nature medicine* 9(7): 928–35.  
<http://www.ncbi.nlm.nih.gov/pubmed/12819779> (September 13, 2013).

Wang, Sijia, Cecil M Lewis, Mattias Jakobsson, Sohini Ramachandran, Nicolas Ray, Gabriel Bedoya, Winston Rojas, Maria V Parra, Julio A Molina, Carla Gallo, Guido Mazzotti, Giovanni Poletti, Kim Hill, Ana M Hurtado, Damian Labuda, William Klitz, Ramiro Barrantes, Maria Cátira Bortolini, Francisco M Salzano, Maria Luiza Petzl-Erler, Luiza T Tsuneto, Elena Llop, Francisco Rothhammer, Laurent Excoffier, Marcus W Feldman, Noah A Rosenberg, and Andrés Ruiz-Linares. 2007. "Genetic variation and population structure in native Americans." *PLoS genetics* 3(11): e185.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2082466&tool=pmcentrez&rendertype=abstract> (August 7, 2013).

Watterson, G A. 1978. "The homozygosity test of neutrality." *Genetics* 88(2): 405–17.  
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1213809&tool=pmcentrez&rendertype=abstract> (December 16, 2011).

Wright, Sewall. 1978. *Evolution and the Genetics of Populations*. Chicago: University of Chicago.

Xavier Eurico de Alencar, Liciana, Ulisses de Mendonça Braga-Neto, Eduardo José Moura do Nascimento, Marli Tenório Cordeiro, Ana Maria Silva, Carlos Alexandre Antunes de Brito, Maria da Paz Carvalho

da Silva, Laura Helena Vega Gonzales Gil, Silvia Maria Lucena Montenegro, and Ernesto Torres de Azevedo Marques. 2013. "HLA-B\*44 Is Associated with Dengue Severity Caused by DENV-3 in a Brazilian Population." *Journal of tropical medicine* 2013: 648475. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3684019&tool=pmcentrez&rendertype=abstract> (September 13, 2013).

**Material Suplementar 1**  
**(S1)**

Table 1-S1. Samples sizes and geographic distribution.

Populations	Regions	N	
		<i>HLA-A</i>	<i>HLA-B</i>
Zulu	SSA	186	201
Zambia	SSA	43	44
Kenyan	SSA	143	143
Kenyan_Highlanders	SSA	241	240
Kenyan_Lowlanders	SSA	265	265
Uganda	SSA	163	161
Mali	SSA	138	138
Chaouya	NAF	67	68
Metalsa	NAF	72	68
Druze	SWA	100	100
Jewish	SWA	117	109
Georgian	SWA	105	107
Kurisk	SWA	30	29
Omani	SWA	119	120
New_Delhi	SWA	66	66
Tamil	SWA	50	49
Golla	SWA	88	104
Czech	EUR	105	106
Croatian	EUR	150	150
Finnish	EUR	90	90
Ireland	EUR	999	1000
Chinese1	SEA	149	149
Singapour_Chinese	SEA	86	86
South_Chinese	SEA	282	281

Okinawa	SEA	105	104
Thai	SEA	98	99
Malay	SEA	124	101
Ami	SEA	98	98
Atayal	SEA	106	106
Bunun	SEA	101	101
Hakka	SEA	55	55
Minnan	SEA	102	102
Paiwan	SEA	51	51
Pazeh	SEA	55	55
Puyuma	SEA	50	50
Rukai	SEA	50	50
Saisiat	SEA	51	51
Siraya	SEA	51	51
Thao	SEA	30	30
Toroko	SEA	55	55
Tsou	SEA	51	51
Yami	SEA	50	50
Indonesian	SEA	50	49
Filipino	SEA	94	94
PNG_Highlands	PAC	92	75
Samoan	PAC	50	50
Ivatan	PAC	50	50
Cape_York	AUS	103	100
Groote_Eylandt	AUS	75	75
Kimberly	AUS	36	38
Yuendumu	AUS	191	193

<b>Korean</b>	NEA	191	200
<b>Tuva</b>	NEA	188	180
<b>Pima</b>	AME	86	89
<b>Bari</b>	AME	92	82

**Table 2-S1. Supertypes frequencies**

Populations	Region	HLA-A				HLA-B						
		A1	A2	A3	A24	NCA	B7	B27	B44	B58	B62	NCB
<b>Zulu</b>	SSA	0.201613	0.204301	0.206989	0.123656	0.263441	0.283582	0.233831	0.169154	0.189055	0.004975	0.119403
<b>Zambia</b>	SSA	0.337209	0.22093	0.151163	0.081395	0.209302	0.409091	0.193182	0.181818	0.113636	0.022727	0.079545
<b>Kenyan</b>	SSA	0.20979	0.276224	0.248252	0.090909	0.174825	0.297203	0.143357	0.157343	0.237762	0.003497	0.160839
<b>Kenyan_Highlanders</b>	SSA	0.215768	0.396266	0.172199	0.080913	0.134855	0.302083	0.145833	0.158333	0.247917	0	0.145833
<b>Kenyan_Lowlanders</b>	SSA	0.209434	0.266038	0.298113	0.09434	0.132075	0.30566	0.216981	0.15283	0.232075	0.003774	0.088679
<b>Uganda</b>	SSA	0.196319	0.303681	0.312883	0.116564	0.070552	0.273292	0.18323	0.21118	0.136646	0.049689	0.145963
<b>Mali</b>	SSA	0.068841	0.246377	0.264493	0.235507	0.184783	0.597826	0.126812	0.101449	0.054348	0.083333	0.036232
<b>Chaouya</b>	NAF	0.268657	0.268657	0.19403	0.11194	0.156716	0.235294	0.117647	0.397059	0.102941	0.014706	0.132353
<b>Metalsa</b>	NAF	0.291667	0.222222	0.284722	0.145833	0.055556	0.176471	0.117647	0.507353	0.036765	0.007353	0.154412
<b>Druze</b>	SWA	0.185	0.25	0.245	0.21	0.11	0.4	0.08	0.365	0.035	0.01	0.11
<b>Jewish</b>	SWA	0.307692	0.247863	0.25641	0.123932	0.064103	0.275229	0.16055	0.261468	0.077982	0.077982	0.146789
<b>Georgian</b>	SWA	0.147619	0.342857	0.3	0.195238	0.014286	0.490654	0.107477	0.168224	0.037383	0.065421	0.130841
<b>Kurisk</b>	SWA	0.266667	0.133333	0.316667	0.183333	0.1	0.482759	0.068966	0.206897	0	0.103448	0.137931
<b>Omani</b>	SWA	0.340336	0.252101	0.302521	0.088235	0.016807	0.404167	0.120833	0.2	0.104167	0.045833	0.125
<b>New_Delhi</b>	SWA	0.143939	0.19697	0.507576	0.136364	0.015152	0.356061	0.030303	0.310606	0.113636	0.068182	0.121212
<b>Tamil</b>	SWA	0.2	0.15	0.45	0.16	0.04	0.255102	0.05102	0.265306	0.153061	0.142857	0.132653

<b>Golla</b>	SWA	0.181818	0.193182	0.443182	0.153409	0.028409	0.307692	0.052885	0.259615	0.100962	0.158654	0.120192
<b>Czech</b>	EUR	0.22381	0.328571	0.290476	0.12381	0.033333	0.292453	0.141509	0.320755	0.051887	0.04717	0.146226
<b>Croatian</b>	EUR	0.253333	0.283333	0.253333	0.183333	0.026667	0.353333	0.176667	0.283333	0.036667	0.063333	0.086667
<b>Finnish</b>	EUR	0.138889	0.344444	0.394444	0.1	0.022222	0.355556	0.122222	0.272222	0.016667	0.122222	0.111111
<b>Ireland</b>	EUR	0.284785	0.295295	0.286286	0.074074	0.05956	0.2905	0.1195	0.317	0.0415	0.042	0.1895
<b>Chinese1</b>	SEA	0.030201	0.342282	0.422819	0.16443	0.040268	0.218121	0.057047	0.208054	0.104027	0.244966	0.167785
<b>Singapour_Chinese</b>	SEA	0.034884	0.343023	0.389535	0.215116	0.017442	0.19186	0.081395	0.168605	0.063953	0.319767	0.174419
<b>South_Chinese</b>	SEA	0.021277	0.29078	0.45922	0.179078	0.049645	0.181495	0.085409	0.213523	0.088968	0.252669	0.177936
<b>Okinawa</b>	SEA	0.214286	0.247619	0.190476	0.342857	0.004762	0.403846	0.120192	0.235577	0	0.163462	0.076923
<b>Thai</b>	SEA	0.010204	0.316327	0.44898	0.127551	0.096939	0.176768	0.090909	0.186869	0.090909	0.287879	0.166667
<b>Malay</b>	SEA	0.056452	0.181452	0.318548	0.221774	0.221774	0.237624	0.029703	0.292079	0.069307	0.222772	0.148515
<b>Ami</b>	SEA	0	0.040816	0.112245	0.627551	0.219388	0.22449	0.372449	0.377551	0	0.02551	0
<b>Atayal</b>	SEA	0.080189	0.169811	0.132075	0.617925	0	0.136792	0.448113	0.382076	0	0.033019	0
<b>Bunun</b>	SEA	0.193069	0.113861	0.108911	0.584158	0	0.188119	0.237624	0.237624	0	0.069307	0.267327
<b>Hakka</b>	SEA	0.063636	0.2	0.536364	0.145455	0.054545	0.181818	0.081818	0.254546	0.109091	0.218182	0.154545
<b>Minnan</b>	SEA	0.044118	0.29902	0.455882	0.186275	0.014706	0.147059	0.078431	0.264706	0.088235	0.259804	0.161765
<b>Paiwan</b>	SEA	0.039216	0.068627	0.019608	0.862745	0.009804	0.078431	0.137255	0.470588	0.009804	0.04902	0.254902
<b>Pazeh</b>	SEA	0.027273	0.2	0.436364	0.336364	0	0.136364	0.181818	0.254546	0.036364	0.254545	0.136364
<b>Puyuma</b>	SEA	0.03	0.2	0.09	0.64	0.04	0.08	0.26	0.23	0	0.25	0.18
<b>Rukai</b>	SEA	0.14	0.06	0.04	0.76	0	0.07	0.2	0.32	0	0.13	0.28
<b>Saisiat</b>	SEA	0.04902	0.137255	0.245098	0.568627	0	0.04902	0.588235	0.323529	0	0	0.039216
<b>Siraya</b>	SEA	0.019608	0.186275	0.313725	0.470588	0.009804	0.127451	0.127451	0.352941	0.058824	0.127451	0.205882
<b>Thao</b>	SEA	0.016667	0.116667	0.266667	0.6	0	0.116667	0.216667	0.25	0.05	0.15	0.216667
<b>Toroko</b>	SEA	0.218182	0.245455	0.090909	0.445455	0	0.145455	0.418182	0.427273	0	0	0.009091
<b>Tsou</b>	SEA	0.04902	0.029412	0.137255	0.784314	0	0.137255	0.362745	0.264706	0	0.058824	0.176471
<b>Yami</b>	SEA	0	0.02	0.39	0.54	0.05	0	0.07	0.26	0	0.52	0.15
<b>Indonesian</b>	SEA	0.11	0.18	0.31	0.21	0.19	0.204082	0.030612	0.316327	0.05102	0.285714	0.112245

<b>Filipino</b>	SEA	0.031915	0.106383	0.340426	0.244681	0.276596	0.180851	0.101064	0.207447	0.074468	0.18617	0.25
<b>PNG_Highlands</b>	PAC	0	0.005435	0.108696	0.782609	0.103261	0.46	0.033333	0.193333	0	0.08	0.233333
<b>Samoan</b>	PAC	0.06	0.26	0.22	0.35	0.11	0.3	0.23	0.35	0	0.05	0.07
<b>Ivatan</b>	PAC	0	0.34	0.12	0.32	0.22	0.1	0.12	0.38	0	0.27	0.13
<b>Cape_York</b>	AUS	0.072816	0.174757	0.237864	0.223301	0.291262	0.255	0.07	0.195	0.025	0	0.455
<b>Groote_Eylandt</b>	AUS	0.033333	0.106667	0.246667	0.293333	0.32	0.213333	0.006667	0.373333	0.006667	0.086667	0.313333
<b>Kimberly</b>	AUS	0	0.111111	0.111111	0.097222	0.680556	0.368421	0.013158	0.460526	0	0	0.157895
<b>Yuendumu</b>	AUS	0.010471	0.112565	0.10733	0.329843	0.439791	0.305699	0.020725	0.240933	0	0.064767	0.367876
<b>Korean</b>	NEA	0.094241	0.32199	0.301047	0.227749	0.054974	0.3175	0.125	0.215	0.06	0.185	0.0975
<b>Tuva</b>	NEA	0.140957	0.255319	0.340426	0.25	0.013298	0.313889	0.102778	0.3	0.108333	0.105556	0.069444
<b>Pima</b>	AME	0.011628	0.511628	0.104651	0.360465	0.011628	0.286517	0.382022	0.308989	0	0.016854	0.005618
<b>Bari</b>	AME	0.005435	0.407609	0	0.586957	0	0.207317	0.27439	0.329268	0	0.176829	0.012195

# **Material Suplementar 2**

## **(S2)**

**Table.1-S2.** Sample size and Hardy Weinberg Results

Population	HLA-B				HLA-C				HLA-DRB1			
	N	Ho	He	p value	N	Ho	He	p value	N	Ho	He	p value
Arara do Iriri	17	0.58824	0.5615	0.53409	15	0.2	0.28736	0.3255	15	0.46667	0.48046	1
Arara	43	0.90698	0.90752	0.05803	24	0.66667	0.69947	0.63685	32	0.71875	0.76339	0.04898
Kayapo Krokaimoro	15	0.6	0.81149	0.16983	13	0.92308	0.79692	1	14	0.78571	0.65608	0.83807
Kayapo Xicrim	15	0.8	0.85977	0.35033	9	0.66667	0.66013	0.51475	10	0.8	0.77368	0.70502
Kaigang	13	0.69231	0.77846	0.35637	9	0.66667	0.78431	0.65168	9	0.55556	0.62745	0.16931
Parakanã	21	0.85714	0.87456	0.29109	11	1	0.74892	0.3923	9	1	0.81046	1
Araweté	31	0.64516	0.63987	0.1612	19	0.42105	0.5761	0.32851	25	0.72	0.68327	0.28017
Urubu Kaapor	33	0.66667	0.7711	0.04183	27	0.59259	0.72956	0.0787	18	0.61111	0.66032	0.71206
Asurini	19	0.89474	0.90754	0.64608	11	0.72727	0.8658	0.72504	19	0.84211	0.73115	0.80232
Guarani Kaiowá	23	0.91304	0.87923	0.79707	16	0.75	0.72177	0.49221	26	0.92308	0.81448	0.4781
Ticuna Tarapaca	19	0.52632	0.76956	0.00585	17	0.58824	0.66488	0.59683	19	0.63158	0.61735	0.22038
Ticuna Arara	14	0.71429	0.85185	0.53674	14	0.5	0.75661	0.07315	15	0.73333	0.73333	0.54713
Ache	14	0.71429	0.61905	0.92019	8	0.5	0.425	1	12	0.33333	0.30797	1
Quechua	21	0.90476	0.93961	0.46605	21	0.7619	0.83624	0.11356	21	0.85714	0.8525	0.30404
Aymara	20	0.85	0.83718	0.89564	20	0.65	0.70513	0.21925	20	0.7	0.79487	0.34019
Huilliche	20	0.7	0.79744	0.0793	15	0.73333	0.8092	0.49978	20	0.8	0.86923	0.60764
Inga	14	0.85714	0.88624	0.36519	14	0.71429	0.80159	0.51534	15	0.86667	0.88276	0.6023
Arhuaco	17	0.82353	0.75579	0.56484	17	0.82353	0.77184	0.52942	17	0.82353	0.76471	0.08495
Kogi	14	0.85714	0.71958	0.02552	14	0.78571	0.64021	0.30864	14	0.92857	0.70899	0.05985
Zenu	16	0.875	0.86089	0.96299	12	0.91667	0.84783	0.72766	12	0.58333	0.60507	0.49621
Embera	14	0.78571	0.78307	0.61628	14	0.64286	0.66667	0.13631	14	0.85714	0.76455	0.51877
Waunana	20	0.85	0.87051	0.22489	20	0.9	0.80385	0.3255	20	0.65	0.70897	0.00483

<b>Cabecar</b>	19	0.73684	0.69417	0.95656	19	0.63158	0.65434	0.96628	19	0.78947	0.72831	0.98609
<b>Guaymi</b>	18	0.94444	0.73651	0.22566	18	0.77778	0.63651	0.70566	18	0.88889	0.69048	0.50665
<b>Wayuu</b>	16	0.9375	0.90323	0.9402	14	0.78571	0.89683	0.60618	15	1	0.90345	0.98649
<b>Quiche</b>	19	0.94737	0.94737	0.88759	19	0.89474	0.88478	0.04635	19	0.94737	0.90469	0.95519
<b>Mixe</b>	20	0.9	0.82949	0.74194	20	0.75	0.75513	1	20	0.8	0.80256	0.22438
<b>Mixtec</b>	20	0.95	0.8859	0.83949	20	0.75	0.73462	0.95553	20	0.75	0.78462	0.68665
<b>Zapotec</b>	17	0.94118	0.95544	0.77936	16	0.875	0.82258	0.76997	18	1	0.90317	0.63683
<b>Chipewyan</b>	25	0.88	0.90612	0.26628	25	0.92	0.91347	0.60298	25	0.96	0.90857	0.31021
<b>Cree</b>	18	0.83333	0.87619	0.6194	18	0.88889	0.92857	0.82419	18	0.88889	0.92857	0.21939
<b>Ojibwa</b>	15	0.93333	0.90115	0.85763	15	0.93333	0.90115	0.60412	15	1	0.90115	0.99548



**Table 2-S2.** Ewens-Watterson test results.

pop name	HLA-B				HLA-C				HLA-DRB1			
	2n	Fo	Fe	P-value	2n	Fo	Fe	P-value	2n	Fo	Fe	P-value
Arara do Iriri	34	0.45502	0.60138	0.1979	30	0.72222	0.75615	0.4381	30	0.53556	0.75586	0.1583
Arara	86	<b><u>0.10303</u></b>	<b><u>0.19715</u></b>	<b><u>0.0015</u></b>	48	0.3151	0.37297	0.3789	64	0.24854	0.39367	0.0856
Kaiapo_K	30	0.21556	0.28249	0.2031	26	<u>0.23373</u>	<u>0.38174</u>	<u>0.0206</u>	28	0.36735	0.32441	0.7402
Kaiapo_X	60	<b><u>0.16889</u></b>	<b><u>0.33974</u></b>	<b><u>0.0021</u></b>	18	0.37654	0.42959	0.4039	20	0.265	0.35601	0.1638
Kaigang	26	0.25148	0.23235	0.721	18	0.25926	0.284	0.472	18	0.40741	0.28332	0.9376
Parakanã	42	0.14626	0.21113	0.081	22	<u>0.28512</u>	<u>0.45271</u>	<u>0.0397</u>	18	<u>0.23457</u>	<u>0.34543</u>	<u>0.0485</u>
Arawete	62	0.37045	0.45519	0.337	38	0.43906	0.61239	0.153	50	0.3304	0.43819	0.2403
Urubu Kaapor	66	0.24059	0.21963	0.7059	54	0.28395	0.38343	0.2242	36	0.35802	0.34612	0.6301
Assurini	38	0.11634	0.14652	0.1729	22	<b><u>0.17355</u></b>	<b><u>0.30312</u></b>	<b><u>0.0004</u></b>	38	0.28809	0.41672	0.1283
Guarani Kaiowa	46	0.13989	0.19573	0.1133	32	0.30078	0.33782	0.444	62	0.20118	0.28568	0.1466
Ticuna_N	38	0.25069	0.23126	0.7058	34	0.35467	0.34273	0.6374	38	0.39889	0.41392	0.5447
Ticuna_S	28	0.17857	0.18049	0.5865	28	0.27041	0.27674	0.5698	30	0.29111	0.33182	0.4152
Ache	28	0.40306	0.39078	0.6426	16	0.60156	0.54115	0.742	24	<u>0.70486</u>	<u>0.45743</u>	<u>0.9706</u>
Quechua	42	0.08277	0.09101	0.3723	42	0.18367	0.21143	0.4026	42	0.1678	0.16958	0.5957
Aymara	40	0.18375	0.13619	0.9174	40	<u>0.3125</u>	<u>0.18578</u>	<u>0.9697</u>	40	0.225	0.18535	0.8263
Huilliche	40	<u>0.2225</u>	<u>0.13647</u>	<u>0.9726</u>	30	0.21778	0.18617	0.8138	40	0.1525	0.13617	0.7653
Inga	28	0.14541	0.1274	0.8228	28	0.22704	0.18113	0.8589	30	0.14667	0.1652	0.4012
Arhuaco	34	0.26644	0.25461	0.6575	34	0.25087	0.34239	0.1726	34	0.25779	0.22213	0.7814
Kogi	28	0.30612	0.32407	0.5152	28	0.38265	0.27701	0.898	28	0.31633	0.32619	0.5643
Zenu	32	0.16602	0.19112	0.3769	24	0.1875	0.2614	0.0844	24	<u>0.42014</u>	<u>0.26216</u>	<u>0.9688</u>

Embera	28	0.2449	0.23797	0.6595	28	0.35714	0.38882	0.4933	28	0.26276	0.32569	0.2778
Waunana	40	0.15125	0.18643	0.2641	40	0.21625	0.30742	0.1422	40	0.30875	0.35678	0.4084
Cabecar	38	0.3241	0.35307	0.4744	38	0.36288	0.35373	0.6242	38	0.29086	0.30375	0.5452
Guaymi	35	0.29959	0.34541	0.4033	36	0.38117	0.60842	0.0561	36	0.3287	0.41358	0.2973
Wayuu	32	0.125	0.13584	0.4678	28	0.1352	0.16011	0.2767	30	<b><u>0.12667</u></b>	<b><u>0.21213</u></b>	<b><u>0.0018</u></b>
Quiche	38	0.07756	0.07945	0.5684	38	0.1385	0.1329	0.6785	38	0.11911	0.1016	0.8482
Mixe	40	0.19125	0.20801	0.4787	40	0.26375	0.3078	0.3987	40	0.2175	0.30669	0.1423
Mixtec	40	0.13625	0.15039	0.4355	40	0.28375	0.26676	0.6844	40	0.235	0.30807	0.2255
Zapotec	34	0.07266	0.08744	0.1258	32	<u>0.20312</u>	<u>0.3378</u>	<u>0.0221</u>	36	0.12191	0.15937	0.1357
Chipewyan	50	<b><u>0.112</u></b>	<b><u>0.18091</u></b>	<b><u>0.014</u></b>	50	<b><u>0.1048</u></b>	<b><u>0.182</u></b>	<b><u>0.0027</u></b>	50	0.1096	0.13604	0.2194
Cree	36	0.14815	0.11773	0.893	36	0.09722	0.11804	0.2039	36	0.09722	0.118	0.2001
Ojibwa	30	0.12889	0.13164	0.5691	30	0.12889	0.13155	0.5887	30	0.12889	0.1463	0.3719

Table 3-S2. Tajima's D test results.

pop_name	HLA-B			HLA-C			HLA-DRB1		
	2n	Tajima's D	P-value	2n	Tajima's D	P-value	2n	Tajima's D	P-value
Arara do Iriri	34	<b><u>2.901102</u></b>	<b><u>0.9994</u></b>	30	0.484689	0.738	30	<b>3.191907</b>	<b>1</b>
Arara	86	<b><u>2.255566</u></b>	<b><u>0.9909</u></b>	48	1.165555	0.9068	64	<b>3.10497</b>	<b>0.9996</b>
Kaiapo_K	30	<u>1.498306</u>	<u>0.9553</u>	26	<b><u>2.365598</u></b>	<b><u>0.9972</u></b>	28	1.678131	0.9664
Kaiapo_X	60	<b><u>2.44449</u></b>	<b><u>0.9934</u></b>	18	1.344724	0.9401	20	1.690865	0.9747
Kaigang	26	0.29117	0.6746	18	1.027998	0.89	18	0.120702	0.5962
Parakanã	42	<b><u>2.283072</u></b>	<b><u>0.9917</u></b>	22	<b><u>2.759251</u></b>	<b><u>0.9994</u></b>	18	1.049376	0.8976
Arawete	62	<u>1.864597</u>	<u>0.9739</u>	38	-1.06374	0.1476	50	<b>2.412669</b>	<b>0.9963</b>
Urubu Kaapor	66	<b><u>2.395515</u></b>	<b><u>0.9933</u></b>	54	0.735853	0.8151	36	0.919014	0.8659
Assurini	38	<b><u>2.484834</u></b>	<b><u>0.9977</u></b>	22	<u>1.564884</u>	<u>0.9635</u>	38	<b><u>2.646943</u></b>	<b><u>0.9985</u></b>
Guarani Kaiowa	46	<u>2.82927</u>	<u>0.9877</u>	32	<u>1.547453</u>	<u>0.9561</u>	62	<u>1.87893</u>	<u>0.977</u>
Ticuna_N	38	<u>1.702371</u>	<u>0.9685</u>	34	1.042047	0.8855	38	-0.67065	0.2766
Ticuna_S	28	<u>1.527817</u>	<u>0.9556</u>	28	1.318537	0.9297	30	0.845631	0.8489
Ache	28	1.100747	0.8988	16	0.078052	0.5833	24	-0.50356	0.3441
Quechua	42	0.850133	0.8542	42	1.459493	0.9477	42	<u>1.769513</u>	<u>0.9754</u>
Aymara	40	0.520051	0.7685	40	0.705586	0.8136	40	1.347757	0.9359
Huilliche	40	1.077851	0.8976	30	0.729486	0.8216	40	0.820287	0.8403
Inga	28	0.431261	0.7208	28	1.039216	0.8865	30	0.829592	0.8497
Arhuaco	34	-0.53976	0.3265	34	<u>1.969984</u>	<u>0.9834</u>	34	<b>2.266925</b>	<b>0.9928</b>
Kogi	28	0.526885	0.7631	28	0.79871	0.831	28	0.290425	0.6741
Zenu	32	1.245103	0.9203	24	<u>1.557685</u>	<u>0.9576</u>	24	0.261279	0.6683
Embera	28	1.035601	0.8905	28	<u>1.935543</u>	<u>0.9841</u>	28	0.856211	0.8469
Waunana	40	<b><u>2.08943</u></b>	<b><u>0.9892</u></b>	40	<u>1.573751</u>	<u>0.9594</u>	40	1.197821	0.9122
Cabecar	38	0.549131	0.7701	38	0.688681	0.8057	38	0.64597	0.7951
Guaymi	35	0.21451	0.6463	36	<b><u>2.664837</u></b>	<b><u>0.9971</u></b>	36	0.599609	0.7754

Wayuu	32	1.058406	0.8935	28	1.294624	0.9312	30	<u>2.12534</u>	<u>0.9899</u>
Quiche	38	0.702254	0.8126	38	0.777405	0.8324	38	0.686991	0.8095
Mixe	40	0.51594	0.7612	40	<b><u>2.559057</u></b>	<b><u>0.9967</u></b>	40	<u>1.734986</u>	<u>0.9721</u>
Mixtec	40	0.912205	0.86	40	<u>1.952746</u>	<u>0.9804</u>	40	<b>2.71159</b>	<b>0.9982</b>
Zapotec	34	1.115692	0.9045	32	<b>2.627072</b>	<b>0.9979</b>	36	1.166305	0.9121
Chipewyan	50	<b><u>2.586672</u></b>	<b><u>0.9973</u></b>	50	1.736498	0.9696	50	<u>2.89717</u>	<u>0.9894</u>
Cree	36	<u>1.7149</u>	<u>0.9725</u>	36	1.317669	0.9296	36	<u>1.700495</u>	<u>0.9639</u>
Ojibwa	30	1.340873	0.9351	30	0.91977	0.8654	30	1.080667	0.8937