UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU – IBB
Programa de Pós-Graduação em Ciências Biológicas (GENÉTICA)

# Estudo da variabilidade das regiões promotora e codificadora do gene *HLA-C* e assinaturas de seleção natural atuando nestes segmentos

**ANDRÉIA DA SILVA SOUZA**

**Botucatu - SP**

**2020**

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU – IBB

Programa de Pós-Graduação em Ciências Biológicas (GENÉTICA)

# Estudo da variabilidade das regiões promotora e codificadora do gene *HLA-C* e assinaturas de seleção natural atuando nestes segmentos

ANDRÉIA DA SILVA SOUZA

ORIENTADOR: PROF. DR. ERICK DA CRUZ CASTELLI

Tese apresentada ao Instituto de Biociências, Universidade Estadual Paulista "Júlio de Mesquita Filho", *Campus* de Botucatu, para obtenção do título de Doutora pelo Programa de Pós-Graduação em Ciências Biológicas (Genética).

**Botucatu – SP**

**2020**

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSANGELA APARECIDA LOBO-CRB 8/7500

Souza, Andréia da Silva.
Estudo da variabilidade das regiões promotora e codificadora do gene *HLA-C* e assinaturas de seleção natural atuando nestes segmentos / Andréia da Silva Souza. - Botucatu, 2020

Tese (doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Erick da Cruz Castelli
Capes: 20200005

1. Antígenos *HLA-C*. 2. Genes MHC Classe I. 3. Variação genética. 4. Seleção natural. 5. Benin. 6. Brasil.

Palavras-chave: Benin; Brasil; *HLA-C*; seleção natural; variabilidade.

Andréia da Silva Souza

# Estudo da variabilidade das regiões promotora e codificadora do gene *HLA-C* e assinaturas de seleção natural atuando nestes segmentos

**English Title:** *HLA-C* promoter and coding genetic diversity and evolutionary aspects

**Orientador:** Profº. Dr. Erick C. Castelli

Comissão examinadora

_____
Prof. Dr. Erick C. Castelli
Universidade Estadual Paulista (UNESP)

_____
Prof. Dr. Diogo Meyer
Universidade de São Paulo (USP)

_____
Prof. Dr. Celso Teixeira Mendes-Junior
Universidade de São Paulo (USP)

_____
Profa. Dra. Norma Lucena Cavalcanti Licinio da Silva
Fundação Oswaldo Cruz, Centro de Pesquisas Aggeu Magalhaes

_____
Profa. Dra. Camila Ferreira Bannwart
Universidade Estadual Paulista (UNESP)

Botucatu, 21 de fevereiro de 2020

## Dedicatória

*Dedico este trabalho...*

*Aos meus pais* ***Antônio*** *e* ***Lourdes*** *que são meu porto seguro e sempre apoiam minhas decisões. Obrigada por toda dedicação e amor.*

## Agradecimentos

- *Agradeço primeiramente a **Deus** por ter guiado os meus passos até aqui e pela sua constante presença em minha vida.*

- *Aos meus pais **Antônio** e **Lourdes,** ao meu irmão **Tony,** minha cunhada **Luana** e meu sobrinho **Anthony** pelo amor e apoio incondicional em todos os momentos.*

- *Às minhas tias **Rose** e **Lêda** por cuidar de mim, sempre me incentivar, aconselhar, apoiar minhas decisões e pela frase "Qualquer coisa, estamos aqui, viu!".*

- *A todos familiares e amigos que mesmo distantes estão sempre torcendo e ajudando em minha caminhada.*

- *Ao meu namorado **Marlon Jocimar,** por todo companheirismo, cuidado e zelo comigo e por me incentivar sempre. Obrigada por tornar essa jornada mais leve.*

- *À minha querida amiga **Thálitta Ayala,** por estar sempre ao meu lado e por ser a mão amiga a me ajudar sempre que precisei. Obrigada pelas conversas, pelos conselhos, por sempre me animar, pelo cuidado e pela amizade que nos sustentou em tantos momentos de dificuldade e que também nos trouxe tanta alegria e sorrisos durante essa jornada.*

- *À minha querida amiga **Letícia Pastore Mendonça** pela companhia, pelos conselhos, pelas boas conversas, risadas e pelos cafezinhos da tarde sempre cheios de histórias de vida, de família, de força e de superação que nos ajudou muitas vezes a aguentar a saudade de casa e as dificuldades do dia a dia. Muito obrigada.*

- *Às minhas amigas **Rosemary Cristina** e **Luciana Pizzani** pelo acolhimento, carinho e amizade que sempre encontro em vocês.*

- *Aos amigos especiais que fizeram parte dessa jornada **Iane de Oliveira Pires Porto**, **Michelle de Almeida Paz** e **Neilton Paulo Bezerra**. Aprendi a amar cada um com seu jeito único e foi maravilhoso contar com a amizade de vocês durante esse período.*

- *Aos amigos e colegas de trabalho do GemBio **Heloisa de Souza Andrade, Marília Rodrigues Silva Passos, Nayane dos Santos Brito Silva, Emiliana Weiss, Hecttor Sebastian Baptista, Arielle Lima da Rocha, Gabriela Pereira de Carvalho** e **Camila Ferreira Bannwart** que fizeram parte dessa jornada e trouxeram tantos aprendizados,*

*estiveram ao meu lado em momentos difíceis e momentos de muita descontração e boas risadas. Obrigada pelos conhecimentos e sonhos compartilhados, pela amizade e principalmente pelas sessões de terapia.*

- *Aos amigos de Botucatu que me acolheram aqui e se tornaram a família de Botucatu.*

- *Aos professores que fizeram parte da banca de qualificação deste trabalho, Dr. Ivan de Godoy Maia, Dra. Claudia Aparecida Rainho e Dr. Ramon Kaneno pelas contribuições científicas.*

- *Aos professores que compuseram a comissão examinadora desta tese, pelas correções e contribuições.*

- *Ao meu orientador Prof. Dr.* **Erick C. Castelli**, *por contribuir com o meu aprendizado e com minha formação acadêmica, pelos conselhos, pela oportunidade e por inspirar seus alunos com sua dedicação e profissionalismo. Muito obrigada, Professor!*

- *À Universidade Estadual Paulista Júlio Mesquita Filho, Campus de Botucatu, pela infraestrutura oferecida e pela oportunidade de construir minha formação científica nessa instituição.*

- *A todos os profissionais da UNESP com os quais tive contato durante esse período de formação.*

- *Ao programa de Pós-Graduação em Ciências Biológicas (Genética), pela oportunidade e credibilidade no desempenho do projeto.*

- *À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa de doutoramento.*

- *À Fundação de Amparo e Apoio a Pesquisa (FAPESP) pelo financiamento do projeto de pesquisa realizado pelo grupo, processo n° 2013/17084-2, do qual faz parte a pesquisa apresentada nesta tese.*

- *E a todos que fizeram parte dessa caminhada: muito obrigada!*

*“Quem acende uma luz é o primeiro a se beneficiar da claridade.”*

*G. K. Chesterton*

## SUMÁRIO

# Lista de Figuras

## Lista de Figuras – Capítulo I



## Lista de Figuras – Capítulo II



## Lista de Figuras - Material Suplementar

# Lista de Tabelas

## Lista de Tabelas - Capítulo I



## Lista de Tabelas - Capítulo II



## Lista de Tabelas - Material Suplementar

# Lista de Abreviaturas e Siglas

| | |
|---|---|
| **APC** | Célula Apresentadora de Antígeno (do inglês, *Antigen Presenting Cell*) |
| **ATP** | Adenosina Trifosfato (do inglês, *Adenosine triphosphate*) |
| **BiP** | Proteína Ligação de imunoglobulina (do inglês, *Binding Immunoglobulin Protein*) |
| **CTL** | Linfócito T Citotóxico |
| **DNA** | Ácido Desoxirribonucleico (do inglês, *Deoxyribonucleic acid*) |
| **ERAAP** | Aminopeptidases Associadas ao Processamento de antígeno do Retículo Endoplasmático |
| **HIV** | Vírus da Imunodeficiência Humana (do inglês, *Human Immunodeficiency Virus)* |
| **HLA** | Antígeno Leucocitário Humano (do inglês, *Human Leukocyte Antigen*) |
| **IFN** | Interferon |
| **IMGT** | International Immunogenetics Database |
| **KIR** | Receptor do tipo imunoglobulina (KIR, do inglês *Killer Cell Immunoglobulin Like Receptor*) |
| **KIRV** | Resíduo formado pelos aminoácidos Lisina (K), Isoleucina (I), Arginina (R) e Valina (V) |
| **Kb** | Kilobase ($10^3$ bases) |
| **LD** | Desequilíbrio de Ligação (*Linkage Disequilibrium*) |
| **Mb** | Megabase ($10^6$ bases) |
| **MHC** | Complexo Principal de Histocompatibilidade (do inglês, *Major Histocompatibility complex*) |
| **mRNA** | Ácido Ribonucleico mensageiro ou RNA mensageiro |
| **NK** | Célula Assassina Natural (do inglês, *Natural Killer*) |
| **NT** | Não Traduzida |
| **Pb** | Pares de bases |
| **PCR** | Reação em Cadeia da Polimerase (do inglês, *Polymerase Chain Reaction*) |
| **RE** | Retículo Endoplasmático |
| **SNP** | Polimorfismo de base única (do inglês, *Single Nucleotide Polymorphism*) |
| **TAP** | Transportador associado ao processamento de antígeno (do inglês, *Transporter Associated with Antigen Processing*) |
| **TCR** | Receptor de célula T |
| **TNF** | Fator de Necrose Tumoral |
| **UTR** | Região não traduzida (do inglês, *Untranslated region*) |

# RESUMO

O gene *HLA-C* está localizado dentro do Complexo Principal de Histocompatibilidade (MHC), a região mais variável do genoma humano. *HLA-C* codifica moléculas que participam principalmente do processo de apresentação de antígenos intracelulares aos linfócitos T citotóxicos e interage com receptores presentes nas células NK, modulando sua atividade. A variabilidade das moléculas HLA permite a apresentação de diferentes peptídeos por um mesmo indivíduo e aumenta o repertório de peptídeos apresentados por uma população. Além disso, *HLA-C* é expresso na interface materno-fetal pelo trofoblasto, onde possui uma importante função imunomodulatória por meio da interação com receptores KIR das células NK maternas. Devido a sua dupla função, alguns alelos de *HLA-C* ou combinações de alelos *HLA-C/KIR* têm sido associados a diversos contextos fisiológicos e patológicos. Contudo, a variabilidade desse gene tem sido explorada principalmente para os éxons, em especial os éxons 2 e 3, que codificam os domínios de ligação ao peptídeo, enquanto a variabilidade de outros segmentos gênicos importantes (como outros éxons, íntrons e regiões regulatórias) foram pouco estudadas. Este estudo avaliou a variabilidade genética e assinaturas de seleção natural ao longo do gene *HLA-C* em uma amostra de 418 indivíduos do Brasil e 108 indivíduos do Benin. Considerando as duas populações, detectamos 359 sítios de variação ao longo da região analisada. Os haplótipos promotores, codificadores e de 3'NT apresentaram uma correlação direta, com alelos que codificam proteínas semelhantes (grupos alélicos) associados a sequências específicas de promotores e de 3'NT. Observamos alta diversidade nucleotídica ao longo de todo o gene, com com destaque para o segmento que codifica o domínio transmembrana. Na região promotora foi encontrada a menor diversidade nucleotídica em uma região próxima à posição -700, um segmento monomórfico de 115pb que é compartilhado entre diferentes primatas. Para a 3'NT foi encontrada alta diversidade na segunda metade deste segmento, porém baixa diversidade no trecho inicial. Nas duas populações foram detectados valores positivos de *D de Tajima* e valores negativos de *F* de *Ewens-Watterson* para quase todos os segmentos do gene *HLA-C,* compatíveis com seleção balanceadora atuando em todo o gene. Além disso, ambas as populações apresentaram evidência de seleção positiva para o éxon 1, que é responsável por codificar o peptídeo de sinal. As frequências dos motivos HLA-C previamente associados à interação com KIR e regulação da expressão são semelhantes entre as duas populações, porém há mudanças na frequência de aminoácidos que influenciam o repertório de peptídeo ligantes, indicando que o repertório peptídico apresentado por essas populações pode ser diferente. Essas populações apresentam grandes diferenças na frequência de aminoácidos no segmento transmembrana e também no primeiro aminoácido do domínio alfa-1, mas não está claro se essas modificações alteram a função e a estabilidade do HLA-C. Estudos funcionais ainda são necessários para o melhor entendimento sobre a relação entre o padrão de variabilidade de *HLA-C* e sua expressão, capacidade de apresentação antigênica, estabilidade e perfil de ligação KIR/HLA-C.



**Palavras-chave:** *HLA-C*, variabilidade, seleção natural, NGS, Brasil, Benin.

## ABSTRACT

Human Leucocyte antigen-C (HLA-C) is a classical HLA class I molecule that binds and presents peptides to cytotoxic T lymphocytes in the cell surface. HLA-C has a dual function since it also interacts with KIR receptors expressed in NK and T cells, modulating their activity. The structure and diversity of the *HLA-C* regulatory regions, as well as the relationship among variants along the *HLA-C* locus, is poorly addressed, and no population-based study explored the complete *HLA-C* variability in different population samples. Here we first present a new molecular and bioinformatics method to evaluate the full HLA-C segment, including regulatory sequences. Then, we applied this method to survey the *HLA-C* diversity in two geographically distinct population samples, one admixed from Brazil (São Paulo State) and one less admixed from Benin. Our results indicate that the *HLA-C* promoter and 3'UTR are very polymorphic, with the presence of few but highly divergent haplotypes. However, both these regulatory regions present conserved segments that are shared among different primates. Nucleotide diversity was higher in other exonic segments rather than exons 2 and 3, and also higher in the second half of the 3'UTR region. We detected evidence of balancing selection on the entire *HLA-C* locus and positive selection in exon 1, for both populations. HLA-C motifs previously associated with KIR interaction and expression regulation are similar between both populations, but the frequency of amino acids influencing peptide-binding is different between samples. Each allele group is associated with specific regulatory sequences, supported by the high linkage disequilibrium along the entire *HLA-C* locus in both populations.



Key words: *HLA-C*, variability, natural selection, NGS, Brazil, Benin.

# Capítulo I - Revisão de Literatura

# INTRODUÇÃO

Os Antígenos Leucocitários Humanos (HLA, do inglês *Human Leucocytes Antigens)* são moléculas envolvidas no controle da reposta imunitária, principalmente no processo de apresentação de peptídeos aos linfócitos T (Shiina *et al.*, 2004; Shiina *et al.*, 2009). Os genes *HLA* estão localizados dentro do Complexo Principal de Histocompatibilidade (MHC, do inglês *Major Histocompatibility Complex*), no braço curto do cromossomo seis, em 6p21.3. Esse complexo possui aproximadamente 250 genes distribuídos em ~4Mb e está dividido didaticamente em três grupos ou classes, denominados classe I, II, ou III (Shiina *et al.*, 2004; Shiina *et al.*, 2009). Os genes *HLA* de classe I são responsáveis pela codificação de proteínas ligadas à membrana celular e possuem um papel essencial no desempenho da resposta imune inata e adaptativa, podendo interagir com receptores dos linfócitos T citotóxicos ou CD8+ e de células *Natural Killer* (NK). Essa classe de genes pode ainda ser subdividida em genes *HLA* de classe I clássicos ou Ia (*HLA-A*, *HLA-B* e *HLA-C*), associados principalmente a apresentação antigênica, e não clássicos ou Ib (*HLA-E*, *HLA-F* e *HLA-G*) que desempenham funções imunomodulatórias (Klein and Sato, 2000; Blais *et al.*, 2011; Donadi *et al.*, 2011). Os genes de classe I, principalmente os loci *HLA-A*, *HLA-B* e *HLA-C*, constituem os genes mais variáveis do genoma humano (Klein and Sato, 2000; Bernatchez and Landry, 2003; Abbas *et al.*, 2011). Atualmente, cerca de 18.441 diferentes alelos para estes genes estão depositados no banco de dados oficial para HLA (The IPD-IMGT/HLA Database) (Robinson *et al.*, 2015) (Tabela 1).

Tabela 1. Número de alelos identificados para os genes de classe I (The IPD-IMGT/HLA Database, versão 3.39.0 de 20-01-2020).

| **CLASSE I CLÁSSICOS** | | | **CLASSE I NÃO CLÁSSICOS** | | |
|---|---|---|---|---|---|
| Gene | Alelos | Proteínas | Gene | Alelos | Proteínas |
| *HLA-A* | 5.907 | 3.720 | *HLA-E* | 84 | 15 |
| *HLA-B* | 7.126 | 4.604 | *HLA-F* | 44 | 6 |
| ***HLA-C*** | **5.709** | **3.470** | *HLA-G* | 69 | 19 |
| Total | 18.742 | 11.794 | Total | 197 | 40 |

A região da classe II abrange aproximadamente 0,7 Mb de DNA e contém os genes que codificam as cadeias α e β dos HLA de classe II clássicos, *HLA-DP*, *HLA-DQ* e *HLA-DR*. Essas moléculas são expressas na superfície de células apresentadoras de antígenos (APC, do inglês *Antigen Presenting Cell*) e apresentam antígenos exógenos às células T auxiliares ou CD4+. A região da classe III, localizada entre as regiões de classe I e classe II, apresenta alta densidade

gênica, com aproximadamente um gene a cada 14.516 nucleotídeos, e contém genes do sistema complemento (*e.g., C2* e *C4*) genes de citocinas (*e.g., TNF*, *LTA* e *LTB*) e muitos genes sem relação direta conhecida com a função imunitária ou inflamação (Horton *et al.*, 2004; Shiina *et al.*, 2004; Shiina *et al.*, 2009).

**Estrutura e função das moléculas HLA de classe I**

Estruturalmente, as moléculas de classe I clássicas e não clássicas são semelhantes, sendo compostas por uma cadeia pesada α associada não covalentemente a uma cadeia leve de $\beta_2$-microglobulina (β2m), codificada por um gene localizado no cromossomo 15 (Abbas *et al.*, 2011). A cadeia α codificada pelos genes HLA de classe I é composta por cinco domínios: domínios α1 e α2, que constituem o sítio de ligação a peptídeos; domínio α3 semelhante à imunoglobulina, que interage com o co-receptor CD8 dos linfócitos; domínio transmembrana e domínio citoplasmático (Klein and Sato, 2000; Alberts *et al.*, 2015) (Figura 1).

No entanto, moléculas clássicas e não clássicas são diferencialmente expressas e possuem funções diferentes. As moléculas HLA de classe I clássicas (HLA-A, HLA-B e HLA-C) são expressas pela maioria das células somáticas e atuam principalmente no processo de apresentação de antígenos intracelulares aos linfócitos T citotóxicos ou TCD8+. As moléculas HLA de classe I não clássicas (HLA-E, HLA-F e HLA-G) possuem expressão celular mais restrita e desempenham principalmente um papel imunomodulatório, por meio de ligações com receptores específicos em células do sistema imune, tais como monócitos, células T, B e *NK* (*Natural Killer*) (Klein and Sato, 2000; Lepin *et al.*, 2000; Pietra *et al.*, 2009; Donadi *et al.*, 2011). No entanto, há evidências de que as moléculas não clássicas também participam da apresentação antigênica (D'souza *et al.*, 2019).

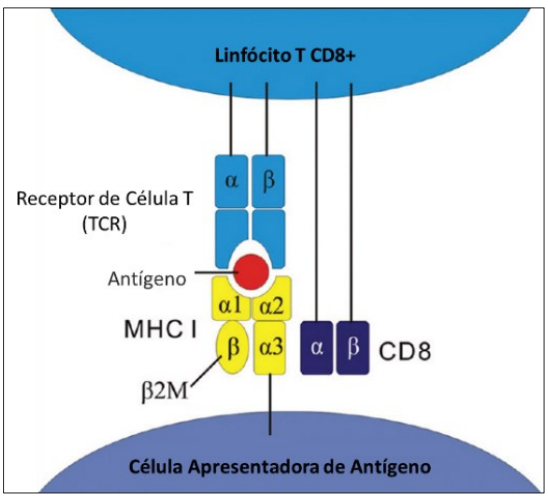



Figura 1. Representação esquemática da estrutura de uma molécula de MHC de classe I e apresentação de antígeno às células T CD8 (Adaptado de Li *et al.*, 2016).

As duas cadeias polipeptídicas das moléculas HLA de classe I (α e β2m), são sintetizadas separadamente e são associadas na superfície luminal do retículo endoplasmático (RE) pela ação de chaperonas, como a calreticulina, BiP (do inglês, *Binding immunoglobulin Protein*), p57 do retículo endoplasmático e calnexina (Yewdell *et al.*, 2003; Parcej and Tampé, 2010; Abbas *et al.*, 2011; Eggensperger and Tampé, 2015). Os peptídeos que serão associados às moléculas de HLA de classe I são gerados por um complexo multiprotéico denominado proteassoma, responsável pela degradação de proteínas citoplasmáticas, em um processo dependente de ubiquitina e ATP (ATP, do inglês, *Adenosine triphosphate*) (Yang *et al.*, 1996). Esses peptídeos são transportados ao lúmen do RE por meio do Transportador Associado ao Processamento de Antígeno (TAP), onde se associam as moléculas de HLA de classe I com o auxílio do complexo de carregamento de peptídeos (formado por TAP e chaperonas). A tapasina é uma chaperona fundamental para a ligação do peptídeo e interage diretamente com a molécula de MHC I (Hulpke *et al.*, 2012; Blees *et al.*, 2017) (Figura 2).

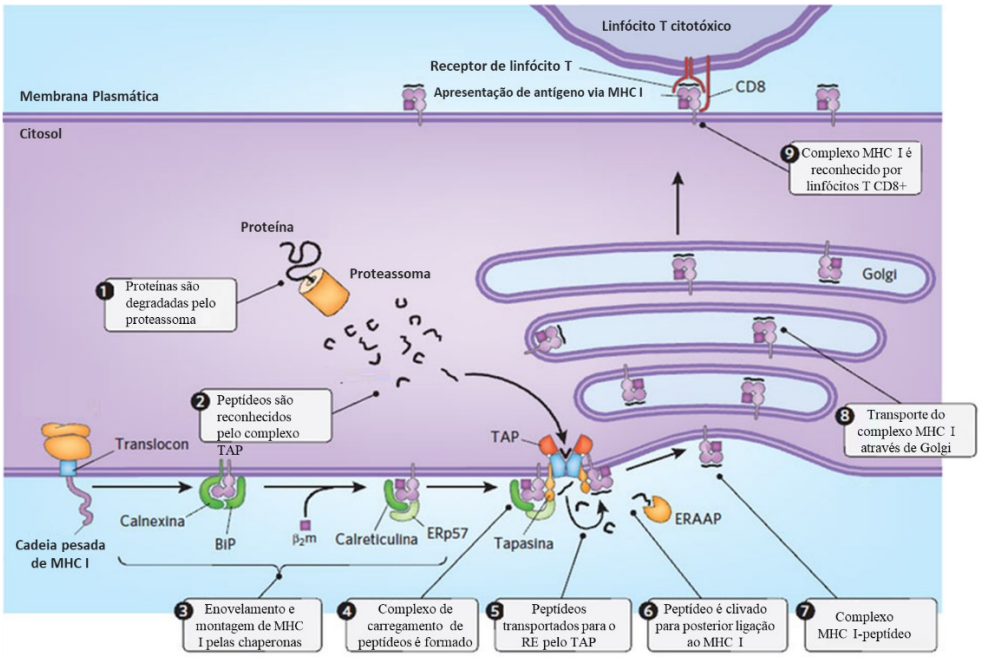



Figura 2. Processamento de antígeno pela via de moléculas MHC de classe I (Adaptado de Parcej e Tampé, 2010).

As moléculas HLA de classe I associam-se a peptídeos de aproximadamente 8 a 9 aminoácidos (AA), dando origem aos complexos HLA-peptídeo (Yang *et al.*, 1996). Peptídeos maiores precisam ser clivados pelas aminopeptidases associadas ao processamento de antígeno do RE (ERAAP) e assim adquirir o tamanho adequado para ligar a fenda de ligação a peptídeo da molécula HLA (Yewdell *et al.*, 2003; Parcej and Tampé, 2010; Blees *et al.*, 2017). Dada a ligação peptídica bem-sucedida, o complexo HLA-peptídeo se dissocia do complexo de

carregamento de peptídeos para ser exportado à superfície celular através da via secretora padrão. Na superfície celular, o complexo HLA-peptídeo será reconhecido pelos Receptores das Células TCD8+ (TCR, do inglês, *T-cell Receptor*) (Parcej and Tampé, 2010) (Figura 2).

Os peptídeos apresentados pelas moléculas HLA de classe I são, em sua maioria, citosólicos, oriundos de proteínas virais, de microrganismos intracelulares, da degradação de proteínas próprias ou ainda proteínas de células transformadas (Alberts *et al.*, 2015; Rock *et al.*, 2016). Uma vez que os receptores dos linfócitos TCD8+ tenham reconhecido peptídeos não-próprios via MHC, eles serão ativados, se diferenciarão em linfócitos efetores e podem então reconhecer qualquer célula alvo expressando os mesmos complexos HLA-peptídeo (Alberts *et al.*, 2015). Em conjunto as interações HLA/peptideo-TCR, HLA-CD8, moléculas de adesão e proteínas de sinalização intracelular, na interface entre as duas células (célula infectada e linfócito T), formam uma sinapse imunológica. A resposta citotóxica dos linfócitos T contra a célula infectada/alterada ocorre por meio da liberação de proteases no espaço sináptico ou por induzir a célula alvo à apoptose (Yang *et al.*, 1996; Alberts *et al.*, 2015).

O conjunto HLA-peptídeo interage com os TCRs, porém, os resíduos polimórficos no topo das alfas hélices das moléculas HLA são a base para especificidade de TCRs para uma forma alélica particular de MHC associada a um peptídeo antigênico (fenômeno chamado restrição de MHC) (Rock *et al.*, 2016). Isso ocorre porque as moléculas de HLA estão diretamente envolvidas na seleção do repertório de células T durante a fase de maturação no timo. O processo de seleção de linfócitos (timócitos) garante um repertório de células T funcionais, ou seja, capazes de interagir com moléculas de HLA do indivíduo. Apenas linfócitos portando TCRs que interagem com MHC-peptídeo próprio são selecionados (processo conhecido como seleção positiva) (Klein *et al.*, 2014). Além disso, o processo de maturação tímica fornece um repertório de linfócitos autotolerantes, impedindo que linfócitos portando TCRs com forte afinidade por MHC-peptídeo próprio sejam liberados na circulação periférica (seleção negativa) (Sebzda *et al.*, 1999; Wiegers *et al.*, 2011; Klein *et al.*, 2014). A falha nesse processo é considerada um dos principais mecanismos associados ao desenvolvimento de doenças autoimunes. Entretanto linfócitos com potencial autorreativo que escapam à seleção negativa e adentram a circulação periférica podem ainda ser inativados, suprimidos ou eliminados pela tolerância periférica para não desencadear autoimunidade (Ohashi, 2002; Abbas *et al.*, 2011; Wiegers *et al.*, 2011).

Assim, além da capacidade de apresentar antígenos, as moléculas HLA também são responsáveis pela seleção de células T aptas a distinguir adequadamente entre antígenos

próprios e não próprios. Deste modo, as moléculas HLA tem importante implicação em contextos fisiológicos e patológicos, relacionadas com autotolerância e autoimunidade, defesa contra patógenos, tumores e rejeição a transplantes alogênicos.

A habilidade das moléculas HLA de classe I clássicas de apresentar diversos peptídeos diferentes é devido à grande variabilidade existente nos domínios de ligação ao peptídeo (codificada pelos éxons 2 e 3). De fato, os resíduos polimórficos da fenda de ligação modificam a especificidade de ligação a peptídeos, porque modificam os bolsões (do inglês, *pockets*) que são preenchidos pelas cadeias laterais de aminoácidos dos peptídeos antigênicos (os chamados resíduos âncoras). Cada peptídeo antigênico tem poucos resíduos âncoras, que são aminoácidos específicos para ligação aos bolsões da molécula HLA. Os demais aminoácidos próximos aos resíduos âncoras, que interagem com a fenda, podem variar entre todas as possibilidades dos 20 aminoácidos conhecidos (Rock *et al.*, 2016) (Figura 3). Isso permite que o repertório de peptídeos apresentados por uma molécula de HLA seja bastante vasto.

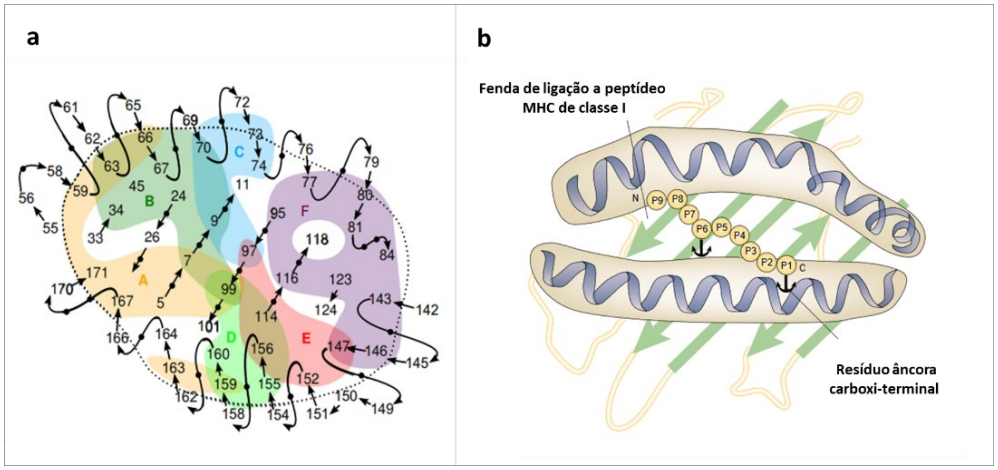



Figura 3. Visão geral esquemática da fenda de ligação a peptídeos. **a)** Painel apresentando os '*pockets*' de *A* a *F* na fenda de ligação a peptídeo de uma molécula de MHC de classe I (diferenciados por um mapa de cores), com as posições dos AA na molécula (Adaptado de Van Deutekom e Keşmir, 2015). **b)** Fenda de ligação do MHC de classe I com petídeo antigênico do '*pocket*', o peptídeo é um nonâmero com dois resíduos âncoras nas posições 1 e 6 (Adaptado de (Kloetzel, 2001).

De tal modo, a substituição de um único aminoácido na fenda de ligação ao peptídeo pode ter efeito no repertório de peptídeos compatíveis com uma determinada molécula HLA, porém depende da posição substituída e das propriedades físico-químicas dos aminoácidos envolvidos (Van Deutekom and Keşmir, 2015; Kaur *et al.*, 2017). De forma semelhante, a mudança de um resíduo âncora de um epítopo antigênico pode impedir sua ligação e apresentação por uma determinada molécula HLA (Rock *et al.*, 2016).

Van Deutekom e Keşmir (2015) usaram um método *in silico* para estudar o efeito de mutações únicas no repertório de peptídeos de várias molélulas HLA-A e HLA-B e mostraram

que as posições que mais alteram a ligação peptídica são as mais polimórficas. No entanto, as posições que raramente variam entre as moléculas HLA parecem não interferir no repertório de peptídeos apresentados. O mesmo estudo demonstrou que o efeito, no repertório de peptídeo ligantes, causado por uma substituição na fenda peptídica é significativamente maior em moléculas de HLA-B em comparação com as moléculas de HLA-A. Considerando que uma nova molécula HLA com um repertório de peptídeos alterado seja mantida na população, caso proporcione vantagens adaptativas, novas moléculas de HLA-B poderiam evoluir facilmente através de mutações pontuais e, portanto, alcançar uma maior diversidade a nível populacional (Van Deutekom and Keşmir, 2015).

Além da ampla gama de peptídeos que podem ser apresentados por cada molécula HLA de classe I individualmente, especialmente os genes clássicos que são muito polimórficos (ver Tabela 1), o repertório é ainda maior quando consideramos a redundância dessas moléculas na superfície celular. Cada indivíduo pode ter de três a seis moléculas HLA de classe I clássicas, considerando a co-expressão dessas moléculas e dependendo dos alelos herdados (Rock *et al.*, 2016).

A maior diversidade alélica dos genes HLA aumenta a capacidade de apresentação antigênica, favorecendo uma melhor adaptabilidade em um contexto de resposta a patógenos e combate a células tumorais. Todavia, essa elevada diversidade alélica pode representar um problema para a compatibilização genotípica entre dois indivíduos (doador e receptor) no contexto do transplante alogênico. Rejeições a enxertos já foram associados a pequenas diferenças (variações) na fenda ligadora de peptídeos das moléculas HLA dos doadores e receptores (Choo, 2007; Ayala García *et al.*, 2012; Tiercy, 2014).

Ainda, é válido ressaltar que apesar da elevada diversidade nucleotídica dos genes HLA estar associada a uma melhor capacidade de apresentação antigênica no contexto de resposta a patógenos e células tumorais, os vírus e as células tumorais podem evadir a resposta imunológica por meio de mecanismos que afetem o processo de apresentação antigênica (Seliger *et al.*, 2006).

Além da modificação de resíduos âncoras para evitar a ligação à molécula HLA, como já citado, os patógenos podem também: (a) escapar da degradação proteasomal, evitando assim a produção de epítopos imunogênicos (Seliger *et al.*, 2006); (b) codificar proteínas que inibem o transportador de peptídeos (TAP) (Van Hall *et al.*, 2007); (c) codificar proteínas que induzem a retenção de moléculas de MHC I no RE-Golgi (Ziegler *et al.*, 2000); (d) codificar proteínas que deslocam as moléculas MHC de classe I do RE ao citosol, onde são rapidamente degradadas

pelo proteassoma (Wiertz, Jones*, et al.*, 1996; Wiertz, Tortorella*, et al.*, 1996) e/ou (e) reduzir a expressão MHC classe I por induzir endocitose dessas moléculas, acúmulo em vesículas endossomais e degradação (Schwartz *et al.*, 1996).

A ausência ou redução da expressão das moléculas HLA de classe I em células tumorais também tem sido frequentemente observada em muitas neoplasias malignas humanas e representa um importante mecanismo de escape da resposta imune (Campoli *et al.*, 2002; Aptsiauri *et al.*, 2007). A diminuição da expressão de componentes da maquinaria de processamento de antígenos da via HLA classe I (*e.g.*, chaperonas e TAP) geralmente têm um impacto negativo no curso clínico dos tumores. Em um número substancial de neoplasias prostáticas, a maioria dos componentes da maquinaria de processamento de antígenos foi encontrada com expressão reduzida, com consequente redução ou ausência de HLA de classe I na superfície celular. Este fenômeno estaria relacionado com progressão e recorrência da doença (Seliger *et al.*, 2010).

Nesse sentido, as células NK são fundamentais na resistência do hospedeiro a infecções virais ou células tumorais, porque estão entre as primeiras células a detectar a liberação de citocinas pró-inflamatórias, bem como a redução da expressão de moléculas MHC de classe I na superfície de células infectadas (Jonjic *et al.*, 2008). O reconhecimento de moléculas HLA por receptores KIR (do inglês, *Killer cell Immunoglobulin like Receptor*) ativatórios e inibitórios promove um equilíbrio de sinais que regulam a atividade de células *NK* (Parham, 2004; Parham, 2005; Augusto and Petzl-Erler, 2015). Assim, a expressão reduzida de HLA na superfície celular pode induzir a ativação de células *NK* por '*missing self*' (ausência do próprio) (Parham, 2004; Parham, 2005; Jonjic *et al.*, 2008). Porém, os vírus adquiriram numerosos mecanismos destinados a subverter ou evadir a vigilância imune das células *NK* (Jonjic *et al.*, 2008; Béziat *et al.*, 2017). As células tumorais também podem aumentar a expressão de HLA classe I não clássicos como mecanismo de escape da resposta imune, visto que estas moléculas desempenham funções imunomodulatórias por meio de ligações a receptores específicos em várias células do sistema imunológico (Paul *et al.*, 1998; Ibrahim *et al.*, 2001; Wiendl *et al.*, 2002; Bossard *et al.*, 2012).

Este cenário evidencia uma constante co-evolução entre patógenos e sistema imune, que no caso das moléculas HLA precisa favorecer a constante mudança no repertório de peptídeos que podem ser apresentados, o escape à regulação da expressão mediada por proteínas dos patógenos e manutenção de certa conservação em domínios importantes para interação com co-receptores, receptores imunomodulatórios e moléculas da via de processamento de antígenos.

## *HLA-C* e suas características particulares

O gene *HLA-C,* alvo deste estudo, pertence à classe de genes mais polimórficos entre os genes *HLA*. Apesar disso, o estudo da variabilidade desse gene foi negligenciado por um longo tempo e *HLA-C* foi considerado pouco polimórfico comparado aos outros genes de classe I clássicos. Atualmente, conhecemos cerca de 5.709 alelos que codificam 3.470 proteínas HLA-C diferentes (Tabela 1), variabilidade que acompanha a de outros genes clássicos como *HLA-A* e *HLA-B*. Porém, algumas características específicas de HLA-C, como a restrição peptídica e menor expressão na superfície celular, ainda o diferencia dos seus correlatos (Blais *et al.*, 2011).

A menor expressão celular de HLA-C tem sido associada a vários fatores, tais como: (a) instabilidade do RNA mensageiro (Mccutcheon *et al.*, 1995); (b) retenção prolongada da molécula HLA-C no interior do retículo endoplasmático por associação ineficiente da cadeia α pesada de HLA-C com a β2m (Neefjes and Ploegh, 1988), por interações estáveis da molécula HLA-C com TAP ou tapasina ou ainda por causa da restrição do repertório de peptídeos (Neisig *et al.*, 1998; Sibilio *et al.*, 2008), que é consequência da maior conservação nos domínios de ligação de peptídeos de HLA-C (Zemmour and Parham, 1992; Neisig *et al.*, 1998; Turner *et al.*, 1998) (Figura 4) .

O motivo KIRV (resíduos nas posições 66, 67, 69 e 76) na α hélice de HLA-C é extremamente conservado na maioria de seus grupos alélicos, exceto para a algumas proteínas do grupo *HLA-C*07 e C*15*, que tem uma mudança de aminoácido na posição $^{Lys}66^{Asn}$ (IMGT database, 3.38.0) (Robinson *et al.*, 2015). Esse motivo conservado é característico apenas de *HLA-C,* não está presente em outros genes HLA de classe I (exceto *HLA-B*46*) e parece favorecer uma rigorosa seleção/restrição do repertório de peptídeos. Além disso, KIRV confere maior associação de HLA-C às chaperonas TAP e tapasina e acúmulo intracelular por deficiência em montagem ou instabilidade pós montagem da molécula (Sibilio *et al.*, 2008).

Um estudo recente com dois alelos de *HLA-C* (*HLA-C*05:01* e *HLA-C*07:02*) mostrou a influência dos domínios de ligação a peptídeo na expressão de superfície celular. *HLA-C*05*, mesmo com um promotor mais fraco que *HLA-C*07* (avaliado por expressão do gene repórter luciferase), apresentou uma expressão duas vezes maior que *HLA-C*07* na superfície celular. Ainda, a maior expressão de *C*05* em relação a *C*07* não foi consequente à regulação diferencial mediada pela porção 3'não traduzida (3'NT), uma vez que para descartar essa influência fora utilizada uma região do gene H-$2K^b$ (MHC de classe I murino) do íntron 3 até a porção 3'NT, tornando esse segmento idêntico para ambos os alelos. O mesmo estudo mostrou que a fenda peptídica de *HLA-C*05* é mais plana enquanto a de *HLA-C*07* é mais profunda e

estreita. Essa característica permitia *HLA-C*05* ligar um número de peptídeos três vezes maior que *HLA-C*07*, indicando que quanto maior a restrição a peptídeos, maior o acúmulo intracelular e menor a expressão na superfície celular (Kaur *et al.*, 2017). Estes achados fornecem informações sobre a complexidade da regulação da expressão de *HLA-C* que depende de fatores pré-transcricionais (importante influência da região promotora), pós-trancricionais (regulação diferencial por ligação de microRNAs a região 3'NT) e pós-traducionais (montagem, estabilidade e ligação a peptídeos), sendo a última, dependente da sequência do alelo codificador.

Contudo, outra característica de *HLA-C* é a imunomodulação, função característica de genes não-clássicos (Klein and Sato, 2000; Lepin *et al.*, 2000; Pietra *et al.*, 2009; Blais *et al.*, 2011; Donadi *et al.*, 2011). O gene *HLA-C* é o único entre os genes *HLA* de classe I clássicos expresso na interface materno-fetal, junto com os genes não clássicos *HLA-E,HLA-F* e *HLA-G* (Hackmon *et al.*, 2017). Vários estudos tem mostrado que combinações particulares de alelos *HLA-C* fetais e *KIR* inibitórios presentes nas células *NK* maternas são fundamentais para imunotolerância materno-fetal e sucesso da gravidez (Trowsdale and Moffett, 2008; Chazara *et al.*, 2011).

O *HLA-C* é considerado o mais importante entre os genes clássicos para a biologia das células *NK* (Parham, 2005). A variabilidade encontrada em algumas porções da molécula HLA-C, em especial nas α hélices, pode modificar a afinidade com receptores KIR (Boyington and Sun, 2002). A presença de um dimorfismo na posição 80 da sequência de aminoácidos na molécula HLA-C define dois grupos de ligantes a receptores KIRs: HLA-C1 (Asn80), ligante dos KIRs inibitórios KIR2DL2 e KIR2DL3, e HLA-C2 (Lys80), ligante do inibitório KIR2DL1 e do ativador KIR2DS1. O KIR ativador KIR2DS4 pode interagir com ambos os alótipos de HLA-C (Parham, 2005; Blais *et al.*, 2011; Parham and Moffett, 2013). A capacidade de interagir com vários KIRs inibitórios e ativadores pode ser um dos motivos por que determinados alelos de *HLA-C* ou combinações específicas de alelos de *HLA-C* e *KIR* têm sido associados a um grande número de doenças autoimunes, alérgicas, inflamatórias, cânceres, abortos recorrentes, pré-eclâmpsia (Kulkarni *et al.*, 2008; Kuśnierczyk, 2013), aloreatividade pós-transplantes e doença do enxerto *versus* hospedeiro (GVHD, do inglês *Graft versus Host disease*) (Parham, 2005).

A função imunomodulatória de HLA-C também aumentou as especulações sobre a importância dos níveis de expressão de HLA-C em infecções. A observação de que a proteína *Nef* do Vírus da Imunodeficiência Humana (HIV, do inglês, *Human Immunodeficiency Virus*)

reduz a expressão de HLA-A e HLA-B, mas não de HLA-C e HLA-E (Le Gall *et al.*, 1998; Cohen *et al.*, 1999; Williams *et al.*, 2002) (Figura 4) foi considerada um mecanismo de escape do vírus. A diminuição nos níveis de expressão das moléculas HLA-A e HLA-B permitiria ao vírus HIV escape da vigilância imunológica, neste caso, do reconhecimento pelas CTL, enquanto que a manutenção da expressão de HLA-C e HLA-E poderia favorecer a inibição de células *NK* (Williams *et al.*, 2002; Blais *et al.*, 2011).

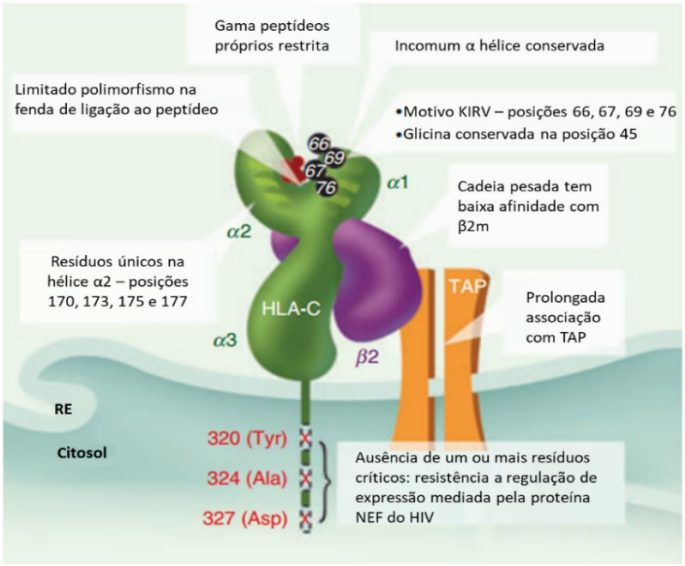



Figura 4. Representação esquemática das características distintas da molécula HLA-C (Adaptado de Blais; Dong and Rowland-Jones, 2011).

Contrariando a visão de que a expressão de HLA-C favoreceria a infecção por HIV, estudos de variabilidade genética em porções regulatórias de *HLA-C* mostrou que variantes favorecendo a expressão do gene estão associadas ao controle da infecção por HIV. Um polimorfismo de nucleotídeo único (SNP) a -35 Kilobases (Kb) de distância do gene *HLA-C* (rs9264942 C) está associada a maior expressão de HLA-C e a uma menor carga viral (Fellay *et al.*, 2007; Thomas *et al.*, 2009). Esse polimorfismo foi ainda encontrado em desequilíbrio de ligação com outra variante na região 3'NT (rs67384697), que também favorece a expressão de HLA-C, uma variante do tipo inserção/deleção, com a deleção afetando a ligação do microRNA miR-148a (Thomas *et al.*, 2009; Kulkarni *et al.*, 2011; Blais *et al.*, 2012; Chen *et al.*, 2012). Ainda, Apps e colaboradores (2013) mostraram que o aumento da expressão de HLA-C em pacientes infectados pelo HIV, independentemente do alótipo, pode aumentar a resposta citotóxica de linfócitos T, apontando um efeito protetor da alta expressão de HLA-C (Apps *et al.*, 2013).

A contribuição de determinados alelos de *HLA-C* no controle de doenças infecciosas, como nos casos de infecções pelo vírus da hepatite C (HCV, do inglês *Hepatitis C Virus*),

citomegalovírus (Kondo *et al.*, 2004; Lauer *et al.*, 2004; Bihl *et al.*, 2005), HIV (Bihl *et al.*, 2005; Makadzange *et al.*, 2010) e vírus Epstein–Barr (Ito *et al.*, 2007), também tem sido descrita. Estes achados parecem estar relacionados a uma apresentação antígeno-específica por determinadas variantes de *HLA-C*. Além disso, respostas eficazes contra infecções virais podem ser mediadas por interações HLA-C/KIR. O receptor ativatório KIR2DS2 de células *NK* é capaz de ligar a moléculas HLA-C*01:02 associadas a peptídeos virais derivados de regiões conservadas de RNA helicases da superfamília dos flavivírus (vírus da dengue, Zika e da febre amarela) e do HCV, levando a ativação de células *NK* (Naiyer *et al.*, 2017). É possível que todas essas respostas sejam impactadas positivamente pela maior expressão celular de HLA-C.

De modo oposto, a maior expressão de HLA-C em doenças autoimunes, como doença de Crohn (Apps *et al.*, 2013), ou maior expressão de um alótipo específico como HLA-C*06 em psoríase, parece ter um efeito deletério (Majorczyk *et al.*, 2014). Nesses casos, provavelmente, por aumento na expressão de alelos comprometidos com a apresentação antigênica e ativação de CTLs contra peptídeos próprios. A hipótese é reforçada pelo achado de efeitos epistáticos entre *HLA-C* e *ERAP1* (ou *ERAAP*) na psoríase. A susceptibilidade a psoríase é aumentada em indivíduos que possuem polimorfismos em *ERAP1* somente quando também apresentam o alelo de risco *HLA-C* (Strange *et al.*, 2010). A ERAP1 desempenha um papel importante no processamento de peptídeos de MHC classe I que serão apresentados a CTLs. De modo semelhante, também foi identificada relação entre os genes ERAP, KIR e HLA-C na suscetibilidade ao aborto espontâneo recorrente (Wilczyńska *et al.*, 2019). Isso porque a mudança no repertório de peptídeos gerados por ERAP pode alterar o complexo HLA-peptídeo e em consequência, alterar tanto a interação HLA-peptídeo-KIR como a interação HLA-peptídeo-TCR.

Em transplantes alogênicos de medula óssea com incompatibilidades permissivas para o gene *HLA-C*, entre doador e receptor, a aloreatividade pós-transplante pode ser impactada pelo nível de expressão da molécula HLA-C. Desta forma, como o nível de expressão do alótipo de HLA-C incompatível pode influenciar o aloreconhecimento por CTL, alelos *HLA-C* de baixa expressão podem representar melhores chances para transplantes com incompatibilidades permissivas (Tiercy, 2014). Assim, é possível que a complexidade dos efeitos de HLA-C, em transplantes e em diversas patogenias, exceda a especificidade de ligação a peptídeo e inclua os diferentes níveis de expressão do gene, que podem modular o potencial da resposta imune.

Entretanto, os mecanismos relacionados com a regulação da expressão não estão completamente elucidados, assim como a variabilidade inerente as regiões regulatórias

(promotora e 3'não-traduzida) do gene *HLA-C* e o impacto desta variabilidade na modulação da expressão gênica. Para a região 3'NT, os poucos estudos existentes apontam para uma influência de um polimorfismo na ligação do miR-148a-3p (Thomas *et al.*, 2009; Chen *et al.*, 2012; Apps *et al.*, 2013). Um outro estudo explorou a influência de variações na 3'NT relacionadas com os alelos *C*07:02* e *C*06:02* na regulação do *HLA-C*, em linhagens celulares homozigotas (Kulkarni *et al.*, 2011).

Em relação ao promotor do *HLA-C*, alguns estudos mostraram variantes que podem afetar a ligação de fatores de transcrição, como rs2395471 A/G, que está localizado em um motivo consenso de ligação do fator de transcrição Oct1 (Vince *et al.*, 2016) e duas variantes regulatórias que influenciam a resposta ao TNF-α (-196A>G, rs2524094) e IFN-γ (−166_−163delTCT, rs10657191) (Hundhausen *et al.*, 2012). Outro estudo explorando a variabilidade na região promotora de *HLA-C,* porém em linhagens celulares homozigóticas para HLA classe I, mostrou que alguns grupos de alelos *HLA-C* tais como *HLA-C *03*, *-C *07* e *-C *17* têm variações em um ou mais motivos do '*core*' promotor (Ramsuran *et al.*, 2017). Esses três grupos alélicos *HLA-C* também apresentam a variante na região 3'NT com afinidade para o miR-148a-3p e o alelo rs2395471G na região promotora, ambos associados a baixos níveis de expressão de HLA-C na superfície celular (Vince *et al.*, 2016; Ramsuran *et al.*, 2017).

Embora alguns estudos tenham demonstrado associações entre variantes regulatórias, alelos codificadores e perfis diferenciais de expressão, nenhum estudo explorou a variabilidade das regiões regulatórias e codificadora de *HLA-C* em uma amostra populacional. Geralmente, os segmentos regulatórios e íntrons de *HLA-C* são desconsiderados na maioria dos estudos abordando sua variabilidade. Ainda, o banco de dados IPD-IMGT/HLA, por exemplo, não contempla a variabilidade do *HLA-C* além do promotor proximal e 5'não traduzida (5'NT), e a maioria dos alelos já descritos para *HLA-C* carecem de sequências de íntrons, promotor e 3'NT.

Diante do exposto, é notória a importância dos níveis de expressão da molécula HLA-C para a realização de suas funções, apresentação antigênica e imunomodulação, e que a regulação da expressão é dependente da variabilidade tanto das regiões regulatórias quanto da região codificadora do gene. Portanto, a avaliação da variabilidade de todo o gene é de extrema importância para o entendimento da biologia e função deste gene paradoxal, que combina características de genes HLA clássicos e não clássicos.

**Seleção Natural atuando sobre os genes HLA**

Estudos de variabilidade genética permitem que o grau de polimorfismo do genoma seja usado como um indicador para identificação de regiões gênicas submetidas a diferentes regimes

de seleção natural (Satta *et al.*, 1998; Meyer *et al.*, 2017). Seleção natural atuando sobre uma região gênica pode, em nível populacional, elevar a frequência de alelos que aumentam o *fitness* de um indivíduo (seleção positiva ou balanceadora). Mas, diminuir a frequência ou eliminar variantes que reduzam o *fitness* do indivíduo (seleção negativa ou purificadora). Regiões gênicas sob seleção direcional (positiva ou negativa/purificadora) apresentam uma taxa reduzida de polimorfismos (Fu and Akey, 2013), enquanto aquelas submetidas à seleção balanceadora tendem a manter um grande número de variantes adaptativas na população, em frequências intermediárias, aumentando o grau de variabilidade e a heterozigose (Key *et al.*, 2014). Sem a ação da seleção natural, esses *loci* evoluem segundo a teoria da neutralidade, em que mutações que não afetam o *fitness* do organismo podem aumentar ou diminuir sua frequência em uma população por meio de deriva genética (Satta *et al.*, 1998).

Seleção balanceadora é um termo amplo que engloba vários regimes seletivos responsáveis pela manutenção de variantes genéticas adaptativas, aumentando a diversidade genética, entre eles: (a) Vantagem do heterozigoto (ou sobredominância), que ocorre quando o '*fitness*' do genótipo heterozigoto é maior que de ambos os alelos em homozigose, mantendo dois ou mais alelos indefinidamente na população; (b) a seleção variável ou flutuação (em tempo e espaço) mantém níveis de diversidade semelhantes ao esperado sobre vantagem do heterozigoto, porém, como o valor adaptativo dos genótipos podem variar ao longo do tempo e espaço, as frequências genotípicas também podem mudar; e (c) vantagem do alelo raro ou seleção dependente de frequência negativa, que pode ser explicada pela co-evolução antagônica entre patógenos e hospedeiros, uma vez que há forte seleção em patógenos para superar a resistência dos alelos MHC mais comuns, os alelos raros que podem ainda oferecer resistência ao patógeno tem uma vantagem seletiva e tendem a aumentar sua frequência (Meyer and Thomson, 2001; Spurgin and Richardson, 2010; Meyer *et al.*, 2017).

Vários autores apontaram para uma incontestável assinatura de seleção balanceadora em genes HLA de classe I clássicos, que seria a explicação para o alto grau de polimorfismos observados, excesso de variantes não-sinônimas e elevado desequilíbrio de ligação nesses genes (Hughes and Yeager, 1998; Meyer and Thomson, 2001; Garrigan and Hedrick, 2003; Sanchez-Mazas, 2007; Spurgin and Richardson, 2010; Meyer *et al.*, 2017). De fato, somente a ação de seleção natural poderia explicar a extrema diversidade genética de genes HLA de classe II e classe I clássicos, como também a alta conservação dos genes não-clássicos.

Seleção balanceadora atuando sobre determinado segmento de DNA pode se estender, por efeito carona, para regiões próximas até uma distância de pelo menos 10kb (Satta *et al.*,

1998). Porém, nos genes HLA, as assinaturas de seleção podem ser diferentes em segmentos muito próximos. No gene *HLA-G,* por exemplo, uma provável seleção purificadora atua sobre a região codificadora, limitando a variabilidade deste segmento (Mendes-Junior *et al.*, 2013). No entanto, as regiões regulatórias do gene *HLA-G* (promotora e *3' NT*) parecem estar sob ação de seleção balanceadora, elevando a heterozigose de haplótipos divergentes (Tan *et al.*, 2005; Castelli *et al.*, 2011; Castelli *et al.*, 2014; Sabbagh *et al.*, 2014; Gineau *et al.*, 2015).

Estudos sobre a variabilidade do gene *HLA-E* também apontam para perfis de seleção diferentes atuando em diferentes segmentos gênicos. O gene *HLA-E* apresenta um perfil de seleção balanceadora na região codificadora da fenda peptídica, especialmente em um ponto de variação não sinônima no éxon 3, mas o éxon 4 (que codifica α3) e regiões não-codificadoras apresentaram assinaturas de seleção purificadora. Adicionalmente, a região *3' NT* de *HLA-E* também é muito conservada (Veiga-Castelli *et al.*, 2012; Felicio *et al.*, 2014). De forma semelhante, alta conservação foi descrita para o gene *HLA-F* (Lima *et al.*, 2016).

Mesmo HLA-A que apresenta assinatura de seleção balanceadora em quase todo segmento gênico (região codificadora e regiões regulatórias), apresenta alta conservação e sinais de seleção purificadora no éxon 4, que codifica α3 (Lima *et al.*, 2019).

Em geral, a assinatura de seleção balanceadora observada para genes HLA de classe I clássicos foi detectada na região que codifica a fenda de ligação aos peptídeos (éxons 2 e 3), cuja manutenção da alta diversidade alélica aumentaria a capacidade de apresentar um maior repertório de peptídeos e consequentemente, o '*fitness*' na resistência contra patógenos (Meyer and Thomson, 2001; Sanchez-Mazas, 2007; Cagliani and Sironi, 2013; Dos Santos Francisco *et al.*, 2015; Bitarello *et al.*, 2016; Buhler *et al.*, 2016).

Ainda, levando em conta a vantagem do heterozigoto, a maior divergência entre os alelos (vantagem do alelo divergente, DAA) de um heterozigoto aumentaria a capacidade de apresentar um conjunto maior de peptídeos que aqueles heterozigotos com alelos mais próximos do ponto de vista de sequência (Buhler *et al.*, 2016). Porém, a grande proporção de homozigotos observada em um ou mais *loci* HLA em estudos populacionais, em pequenas e grandes populações, indica que mecanismos adicionais, além da vantagem do heterozigoto, estão envolvidos para garantir uma adequada proteção imune nessas populações (Buhler *et al.*, 2016).

Esses achados elevaram o interesse em abordagens que investiguem os padrões de variação genética de HLA observados em diferentes populações, em relação à funcionalidade das moléculas (especificidade de ligação ao peptídeo). Os resultados desse tipo de estudo confirmaram que os resíduos polimórficos não são distribuídos aleatoriamente dentro da fenda

de ligação a peptídeo, mas seguem um padrão de maior diversidade em sítios que definem o repertório de peptídeos (Dos Santos Francisco *et al.*, 2015; Van Deutekom and Keşmir, 2015; Bitarello *et al.*, 2016; Buhler *et al.*, 2016). Assim, Buhler e colaboradores (2016) propuseram o modelo de seleção assimétrica divergente dos genes HLA de classe I, que sugere que a falta de diversidade em um gene HLA, como observado em algumas populações, é contrabalanceada por propriedades complementares de ligação ao peptídeo das moléculas codificadas por outros genes. No entanto, Buhler observou que este modelo parece ser robusto para os genes *HLA-A* e *HLA-B*, porém *HLA-C* não aumenta significativamente o potencial de ligação de peptídeo dos HLA de classe I, sugerindo que este *locus* assume um papel mais importante em suas funções relacionadas a KIR (Buhler *et al.*, 2016). Além disso, é possível que a adaptação envolvendo genes HLA seja poligênica e que haplótipos (estendidos) carregando combinações de alelos HLA que apresentam maior número de peptídeos sejam favorecidos (Meyer *et al.*, 2017) . Em suporte a essa hipótese, já foi demostrado que alelos HLA em forte desequilíbrio de ligação, em média, tem menor sobreposição do repertório de peptídeo (Penman *et al.*, 2013).

Embora seja consenso que a alta diversidade do MHC é fundamental para defesa imune contra patógenos, é bem conhecida a associação entre diversos genes localizados nesse complexo e doenças humanas, entre elas, doenças autoimunes como diabetes tipo I, psoríase, doença de Crohn, doença celíaca, esclerose múltipla, lúpus, artrite reumatoide e espondilite anquilosante (Trowsdale and Knight, 2013; Matzaraki *et al.*, 2017).

O delicado equilíbrio entre variantes genéticas que favoreçam o controle de infecções, mas que podem aumentar a suscetibilidade à doença autoimune, pode ser considerado um importante paradigma em genética evolutiva (Chen *et al.*, 2012). De fato, à partir de uma perspectiva evolutiva, é intrigante a existência de condições auto-imunes, que reduzem as chances de sobrevivência e reprodução de um indivíduo, associadas a alelos HLA relativamente comuns (Meyer *et al.*, 2017). Contudo, isso pode ser resultado da seleção de alelos que conferem resistência a doenças infecciosas e ao mesmo tempo estão associadas a condições auto-imunes (Corona *et al.*, 2010; Abadie *et al.*, 2011; Sams and Hawks, 2014; Meyer *et al.*, 2017). Por exemplo, alelos *HLA-C* relacionados com psoríase e doença de Crohn também estão relacionados a proteção contra doenças infecciosas, como infeção pelo HIV (Blais *et al.*, 2012; Chen *et al.*, 2012; Apps *et al.*, 2013; Kulkarni *et al.*, 2013; Majorczyk *et al.*, 2014).

Algumas interações entre HLA e KIR (especialmente HLA-C/KIR) também têm sido associadas a diversas doenças autoimunes (Parham, 2005; Blais *et al.*, 2011; Augusto and Petzl-Erler, 2015). Há evidências de co-evolução entre os complexos KIR e HLA, e é possível que

seleção balanceadora esteja mantendo a diversidade excepcional do sistema KIR-HLA, aparentemente evoluindo sob a pressão proporcionada pelas exigências concorrentes da reprodução e da imunidade inata e adaptativa (Augusto and Petzl-Erler, 2015).

O gene *HLA-C* é altamente polimórfico e ao mesmo tempo mantém vários motivos conservados em sua fenda de ligação a peptídeo, que restringe seu repertório de peptídeos e regula a sua expressão na supeficie celular. Adicionalmente, a molécula HLA-C desempenha a função de apresentação antigênica a CTLs (como HLA-A e HLA-B), mas também tem um papel imunomodulatório importante semelhante aos genes não clássicos, ambas situações sendo impactadas pelos níveis de expressão do gene supracitado. Levando em consideração o papel crucial que o gene *HLA-C* desempenha na resposta imunitária, ressalta-se a importância dos estudos de variabilidade no intuito de esclarecer de que forma variações em sua sequência se relacionam com diferentes perfis de expressão. Além disso, estudos que avaliem os perfis de seleção natural em cada segmento de HLA-C é de fundamental importância para esclarecer de que maneira a variabilidade desse gene tem sido moldada evolutivamente. Desta forma, o presente estudo busca esclarecer se o perfil de seleção balanceadora se extende por toda região codificadora e segmentos regulatórios de *HLA-C*, ou se estes estão sob diferentes regimes de seleção natural, permitindo assim um melhor entendimento acerca da biologia do gene *HLA-C*.

**Testes de desvio de neutralidade e demografia das populações estudadas**

Para testar seleção natural atuando a nível de DNA em amostras populacionais, é possível utilizar testes estatísticos de desvio de neutralidade. Esses testes utilizam uma abordagem teórica baseada em simulações de coalescência para detectar desvios significativos de um modelo de neutralidade (Excoffier and Lischer, 2010; Gineau *et al.*, 2015).

Neste estudo as assinaturas de seleção presentes no gene *HLA-C* foram avaliadas por meio de três testes de desvio da neutralidade: (a) teste de Ewens-Watterson (Ewens, 1972; Watterson, 1978), que compara a homozigose observada na população em estudo com a homozigose esperada sob expectativas de neutralidade e em equilíbrio de Hardy-Weinberg, para um mesmo tamanho amostral (2n) e com o mesmo número de alelos únicos. O excesso de homozigose indica a possibilidade de uma seleção direcionada, enquanto um excesso de heterozigose pode indicar atuação de seleção balanceadora favorecendo genótipos heterozigotos; (b) teste *D* de Tajima (Tajima, 1989), que é calculado considerando a diferença entre as estimativas Θ (theta), que se baseia no número médio de diferenças observadas entre pares de sequências ($\pi$) e o número de sítios de segregação (S). A significância estatística do *D* de Tajima é testada através da geração de amostras aleatórias sob a hipótese de neutralidade

seletiva em uma população com tamanho constante e em equilíbrio, utilizando um número determinado de simulações (Excoffier and Lischer, 2010); e (c) Razão *dN/dS,* que compara as taxas de substituições não-sinônimas (*dN*) e sinônimas (*dS*) e permite inferir os regimes de seleção que estão atuando sobre a região codificadora do gene em estudo (Kumar *et al.*, 2016). Segmentos exônicos em que a razão *dN/dS* é superior a 1 podem estar sob seleção direcional positiva (Kumar *et al.*, 2016). Porém esse critério é considerado conservador, visto que apenas alguns códons podem estar sob seleção positiva (Bitarello *et al.*, 2016), enquanto a maior parte das mutações não-sinônimas são deletérias e são eliminadas da população sem contribuir com a evolução ou o polimorfismo (Kimura, 1991) (seleção purificadora). Assim, muitas vezes esse critério ($dN/dS > 1$) não é atendido quando analisamos genes ou segmentos exônicos inteiros, mesmo quando há presença de alguns códons sob seleção positiva, tornando necessária a análise de subconjuntos de códons. Razões *dN/dS* inferiores a 1 podem ser um indicativo de seleção direcional negativa (Kumar *et al.*, 2016).

O uso de mais de um teste avaliando as assinaturas de seleção evita resultados falso-positivos, uma vez que a variação das frequências alélicas pode sofrer influência de efeitos demográficos, tais como efeito fundador, migração, miscigenação, contração ou expansão populacional e deriva genética (Stajich and Hahn, 2005; Gineau *et al.*, 2015). Os valores positivos de *D* de Tajima, por exemplo, são indicativos de um excesso de alelos de frequência intermediária (seleção balanceadora) ou contração recente da população (gargalo), enquanto os valores *D* negativos de Tajima indicam um excesso de alelos raros, que pode ser resultado de expansão recente da população, varredura seletiva, seleção negativa fraca ou amostra proveniente de uma população miscigenada (Stajich and Hahn, 2005).

Além do uso de três testes distintos, nós também optamos por estudar populações de dois continentes diferentes. A população brasileira é sabidamente uma população miscigenada, formada por 5 séculos de mistura interétnica entre europeus, africanos e nativos americanos (Pena *et al.*, 2011), além das migrações do leste asiático ocorridas no último século, o que poderia levar a resultados tendenciosos nos testes de desvio de neutralidade devido à sua história demográfica. Por isso, nós inserimos nesse estudo uma população do Benin, país localizado no continente africano. Essa amostra é composta por indivíduos da etnia Toffin (etimologicamente “pessoas da água"). Os Toffin estão presentes em mais de 10 aldeias, representando pequenas comunidades autônomas, fugitivos das lutas encabeçadas pelos soldados franceses durante o período pré-colonial e localizados na mesma região há séculos (Sonon *et al.*, 2018). A população do Benin é menos miscigenada que a população brasileira,

mas também tem uma história demográfica que poderia influenciar uma análise de seleção natural de sua população quando estudada individualmente. Não obstante a história demográfica de cada população, temos duas populações com diferentes histórias demográficas, localizadas em continentes distintos e ambas apresentaram o mesmo perfil de assinaturas de seleção natural atuando no gene *HLA-C*.

# OBJETIVOS

## Objetivo Geral

Avaliar a variabilidade e os haplótipos encontrados na região promotora, codificadora e 3' não-traduzida do gene *HLA-C* em uma amostra brasileira e beninense e detectar assinaturas de seleção natural ao longo do gene.

## Objetivos específicos

- Consolidar uma metodologia para avaliação da variabilidade do gene *HLA-C* por sequenciamento massivo paralelo (ou sequenciamento de nova geração);
- Caracterizar os pontos de variação e haplótipos encontrados para o gene *HLA-C* em uma amostra do Brasil e do Benin;
- Caracterizar as amostras brasileiras utilizadas quanto a seus perfis de ancestralidade;
- Avaliar o perfil de seleção atuando ao longo de todos os segmentos do gene *HLA-C*.

# REFERÊNCIAS


ABADIE, V. et al. Integration of genetic and immunological insights into a model of celiac disease pathogenesis. **Annu Rev Immunol,** v. 29, p. 493-525, 2011. ISSN 0732-0582.

ABBAS, A. K.; LICHTMAN, A. H.; PILLAI, S. H. I. V. Moléculas do Complexo Principal de Histocompatibilidade e Apresentação do Antígeno aos Linfócitos T. In: (Ed.). **Imunologia Celular e Molecular**. 7ª. Rio de Janeiro: Elsevier, 2011. p.109-138.

ALBERTS, B. et al. The Innate and Adaptive Immune Systems: T cells and MHC proteins. In: (Ed.). **Molecular Biology of the Cell**. sixth edition. 711 Tird Avenue, New York, NY 10017, US: Garland Science, 2015. chap. 24, p.1297-1342.

APPS, R. et al. Influence of HLA-C expression level on HIV control. **Science,** v. 340, n. 6128, p. 87-91, Apr 05 2013. ISSN 0036-8075.

APTSIAURI, N. et al. Role of altered expression of HLA class I molecules in cancer progression. **Adv Exp Med Biol,** v. 601, p. 123-31, 2007. ISSN 0065-2598 (Print)0065-2598.

AUGUSTO, D. G.; PETZL-ERLER, M. L. KIR and HLA under pressure: evidences of coevolution across worldwide populations. **Hum Genet,** v. 134, n. 9, p. 929-40, Sep 2015. ISSN 0340-6717.

AYALA GARCÍA, M. A. et al. The Major Histocompatibility Complex in Transplantation. **Journal of Transplantation,** v. 2012, p. 842141-842141, 2012.

BERNATCHEZ, L.; LANDRY, C. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? **Journal of evolutionary biology,** v. 16, n. 3, p. 363-377, 2003.

BIHL, F. K. et al. Simultaneous assessment of cytotoxic T lymphocyte responses against multiple viral infections by combined usage of optimal epitope matrices, anti- CD3 mAb T-cell expansion and "RecycleSpot". **Journal of Translational Medicine,** v. 3, p. 20-20, 2005.

BITARELLO, B. D.; FRANCISCO RDOS, S.; MEYER, D. Heterogeneity of dN/dS Ratios at the Classical HLA Class I Genes over Divergence Time and Across the Allelic Phylogeny. **J Mol Evol,** v. 82, n. 1, p. 38-50, Jan 2016. ISSN 0022-2844.

BLAIS, M.-E.; DONG, T.; ROWLAND-JONES, S. HLA-C as a mediator of natural killer and T-cell activation: spectator or key player? **Immunology,** v. 133, n. 1, p. 1-7, 2011.

BLAIS, M. E. et al. High frequency of HIV mutations associated with HLA-C suggests enhanced HLA-C-restricted CTL selective pressure associated with an AIDS-protective polymorphism. **J Immunol,** v. 188, n. 9, p. 4663-70, May 01 2012. ISSN 0022-1767.

BLEES, A. et al. Structure of the human MHC-I peptide-loading complex. **Nature,** v. 551, n. 7681, p. 525-528, Nov 2017. ISSN 1476-4687.

BOSSARD, C. et al. HLA-E/beta2 microglobulin overexpression in colorectal cancer is associated with recruitment of inhibitory immune cells and tumor progression. **Int J Cancer,** v. 131, n. 4, p. 855-63, 2012 Aug 15 2012. ISSN 0020-7136.

BOYINGTON, J. C.; SUN, P. D. A structural perspective on MHC class I recognition by killer cell immunoglobulin-like receptors. **Molecular immunology,** v. 38, n. 14, p. 1007-1021, 2002.

BUHLER, S.; NUNES, J. M.; SANCHEZ-MAZAS, A. HLA class I molecular variation and peptide-binding properties suggest a model of joint divergent asymmetric selection. **Immunogenetics,** v. 68, n. 6-7, p. 401-16, Jul 2016. ISSN 0093-7711.

BÉZIAT, V. et al. Deciphering the killer-cell immunoglobulin-like receptor system at super-resolution for natural killer and T-cell biology. **Immunology,** v. 150, n. 3, p. 248-264, 03 2017. ISSN 1365-2567.

CAGLIANI, R.; SIRONI, M. Pathogen-driven selection in the human genome. **Int J Evol Biol,** v. 2013, p. 204240, 2013. ISSN 2090-8032 (Print)2090-052x.

CAMPOLI, M.; CHANG, C. C.; FERRONE, S. HLA class I antigen loss, tumor immune escape and immune selection. **Vaccine,** v. 20 Suppl 4, p. A40-5, 2002 Dec 19 2002. ISSN 0264-410X (Print)0264-410x.

CASTELLI, E. C. et al. A comprehensive study of polymorphic sites along the HLA-G gene: implication for gene regulation and evolution. **Molecular biology and evolution,** v. 28, n. 11, p. 3069-86, 2011.

CASTELLI, E. C. et al. Transcriptional and posttranscriptional regulations of the HLA-G gene. **J Immunol Res,** v. 2014, p. 734068, 2014. ISSN 2314-7156.

CHAZARA, O.; XIONG, S.; MOFFETT, A. Maternal KIR and fetal HLA-C: a fine balance. **J Leukoc Biol,** v. 90, n. 4, p. 703-16, Oct 2011. ISSN 0741-5400.

CHEN, H. et al. Psoriasis patients are enriched for genetic variants that protect against HIV-1 disease. **PLoS genetics,** v. 8, n. 2, p. e1002514-e1002514, 2012.

CHOO, S. Y. The HLA System: Genetics, Immunology, Clinical Testing, and Clinical Implications. **Yonsei Medical Journal,** v. 48, n. 1, p. 11-23, 2007.

COHEN, G. B. et al. The selective downregulation of class I major histocompatibility complex proteins by HIV-1 protects HIV-infected cells from NK cells. **Immunity,** v. 10, n. 6, p. 661-71, 1999 Jun 1999. ISSN 1074-7613 (Print)1074-7613.

CORONA, E.; DUDLEY, J. T.; BUTTE, A. J. Extreme evolutionary disparities seen in positive selection across seven complex diseases. **PLoS One,** v. 5, n. 8, p. e12236, Aug 17 2010. ISSN 1932-6203.

D'SOUZA, M. P. et al. Casting a wider net: Immunosurveillance by nonclassical MHC molecules. **PLoS Pathog,** v. 15, n. 2, p. e1007567, 02 2019. ISSN 1553-7374.

DONADI, E. A. et al. Implications of the polymorphism of HLA-G on its function, regulation, evolution and disease association. **Cell Mol Life Sci,** v. 68, n. 3, p. 369-95, Feb 2011. ISSN 1420-9071.

DOS SANTOS FRANCISCO, R. et al. HLA supertype variation across populations: new insights into the role of natural selection in the evolution of HLA-A and HLA-B polymorphisms. **Immunogenetics,** v. 67, n. 11-12, p. 651-63, Nov 2015. ISSN 0093-7711.

EGGENSPERGER, S.; TAMPÉ, R. The transporter associated with antigen processing: a key player in adaptive immunity. **Biol Chem,** v. 396, n. 9-10, p. 1059-72, Sep 2015. ISSN 1437-4315.

EWENS, W. J. The sampling theory of selectively neutral alleles. **Theoretical population biology,** v. 3, n. 1, p. 87-112, 1972.

EXCOFFIER, L.; LISCHER, H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. **Molecular ecology resources,** v. 10, n. 3, p. 564-567, 2010.

FELICIO, L. P. et al. Worldwide HLA-E nucleotide and haplotype variability reveals a conserved gene for coding and 3' untranslated regions. **Tissue antigens,** v. 83, n. 2, p. 82-93, 2014.

FELLAY, J. et al. A whole-genome association study of major determinants for host control of HIV-1. **Science,** v. 317, n. 5840, p. 944-7, Aug 17 2007. ISSN 0036-8075.

FU, W.; AKEY, J. M. Selection and adaptation in the human genome. **Annu Rev Genomics Hum Genet,** v. 14, p. 467-89, 2013. ISSN 1527-8204.

GARRIGAN, D.; HEDRICK, P. W. Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. **Evolution,** v. 57, n. 8, p. 1707-22, Aug 2003. ISSN 0014-3820 (Print)0014-3820.

GINEAU, L. et al. Balancing immunity and tolerance: genetic footprint of natural selection in the transcriptional regulatory region of HLA-G. **Genes Immun,** v. 16, n. 1, p. 57-70, 2015 Jan-Feb 2015. ISSN 1476-5470.

HACKMON, R. et al. Definitive class I human leukocyte antigen expression in gestational placentation: HLA-F, HLA-E, HLA-C, and HLA-G in extravillous trophoblast invasion on placentation, pregnancy, and parturition. **American Journal of Reproductive Immunology,** v. 77, n. 6, p. e12643-n/a, 2017. ISSN 1600-0897.

HORTON, R. et al. Gene map of the extended human MHC. **Nat Rev Genet,** v. 5, n. 12, p. 889-99, Dec 2004. ISSN 1471-0056 (Print)1471-0056.

HUGHES, A. L.; YEAGER, M. Natural selection at major histocompatibility complex loci of vertebrates. **Annu Rev Genet,** v. 32, p. 415-35, 1998. ISSN 0066-4197 (Print)0066-4197.

HULPKE, S.; BALDAUF, C.; TAMPÉ, R. Molecular architecture of the MHC I peptide-loading complex: one tapasin molecule is essential and sufficient for antigen processing. **FASEB J,** v. 26, n. 12, p. 5071-80, Dec 2012. ISSN 1530-6860.

HUNDHAUSEN, C. et al. Allele-specific cytokine responses at the HLA-C locus: implications for psoriasis. **J Invest Dermatol,** v. 132, n. 3 Pt 1, p. 635-41, Mar 2012. ISSN 0022-202x.

IBRAHIM, E. C. et al. Tumor-specific up-regulation of the nonclassical class I HLA-G antigen expression in renal carcinoma. **Cancer Res,** v. 61, n. 18, p. 6838-45, 2001 Sep 15 2001. ISSN 0008-5472 (Print)0008-5472.

ITO, Y. et al. Full-length EBNA1 mRNA-transduced dendritic cells stimulate cytotoxic T lymphocytes recognizing a novel HLA-Cw*0303- and -Cw*0304-restricted epitope on EBNA1-expressing cells. **J Gen Virol,** v. 88, n. Pt 3, p. 770-80, 2007 Mar 2007. ISSN 0022-1317 (Print)0022-1317.

JONJIC, S. et al. Immune evasion of natural killer cells by viruses. **Curr Opin Immunol,** v. 20, n. 1, p. 30-8, 2008 Feb 2008. ISSN 0952-7915 (Print)1879-0372 (Electronic).

KAUR, G. et al. Structural and regulatory diversity shape HLA-C protein expression levels. **Nat Commun,** v. 8, p. 15924, Jun 2017. ISSN 2041-1723.

KEY, F. M. et al. Advantageous diversity maintained by balancing selection in humans. **Curr Opin Genet Dev,** v. 29, p. 45-51, Dec 2014. ISSN 0959-437x.

KIMURA, M. The neutral theory of molecular evolution: A review of recent evidence. **The Japanese Journal of Genetics,** v. 66, n. 4, p. 367-386, 1991.

KLEIN, J.; SATO, A. The HLA system. First of two parts. **The New England journal of medicine,** v. 343, n. 10, p. 702-709, 2000.

KLEIN, L. et al. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). **Nat Rev Immunol,** v. 14, n. 6, p. 377-91, Jun 2014. ISSN 1474-1741.

KLOETZEL, P. M. Antigen processing by the proteasome. **Nat Rev Mol Cell Biol,** v. 2, n. 3, p. 179-87, Mar 2001. ISSN 1471-0072.

KONDO, E. et al. Identification of novel CTL epitopes of CMV-pp65 presented by a variety of HLA alleles. **Blood,** v. 103, n. 2, p. 630-8, 2004 Jan 15 2004. ISSN 0006-4971 (Print)0006-4971.

KULKARNI, S.; MARTIN, M. P.; CARRINGTON, M. The Ying and Yang of HLA and KIR in Human Disease. **Semin Immunol,** v. 20, n. 6, p. 343-52, 2008 Dec 2008. ISSN 1044-5323 (Print)1096-3618 (Electronic).

KULKARNI, S. et al. Genetic interplay between HLA-C and MIR148A in HIV control and Crohn disease. **Proceedings of the National Academy of Sciences of the United States of America,** v. 110, n. 51, p. 20705-20710, 2013.

KULKARNI, S. et al. Differential microRNA regulation of HLA-C expression and its association with HIV control. **Nature,** v. 472, n. 7344, p. 495-498, 2011.

KUMAR, S.; STECHER, G.; TAMURA, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. **Mol Biol Evol,** v. 33, n. 7, p. 1870-4, 07 2016. ISSN 1537-1719.

KUŚNIERCZYK, P. Killer Cell Immunoglobulin-Like Receptor Gene Associations with Autoimmune and Allergic Diseases, Recurrent Spontaneous Abortion, and Neoplasms. **Front Immunol,** v. 4, 2013 Jan 29 2013. ISSN 1664-3224 (Electronic).

LAUER, G. M. et al. High resolution analysis of cellular immune responses in resolved and persistent hepatitis C virus infection. **Gastroenterology,** v. 127, n. 3, p. 924-36, 2004 Sep 2004. ISSN 0016-5085 (Print)0016-5085.

LE GALL, S. et al. Nef interacts with the mu subunit of clathrin adaptor complexes and reveals a cryptic sorting signal in MHC I molecules. **Immunity,** v. 8, n. 4, p. 483-95, 1998 Apr 1998. ISSN 1074-7613 (Print)1074-7613.

LEPIN, E. J. et al. Functional characterization of HLA-F and binding of HLA-F tetramers to ILT2 and ILT4 receptors. **Eur J Immunol,** v. 30, n. 12, p. 3552-61, Dec 2000. ISSN 0014-2980.

LI, L.; DONG, M.; WANG, X. G. The Implication and Significance of Beta 2 Microglobulin: A Conservative Multifunctional Regulator. **Chin Med J (Engl),** v. 129, n. 4, p. 448-55, Feb 2016. ISSN 0366-6999.

LIMA, T. H. et al. HLA-F coding and regulatory segments variability determined by massively parallel sequencing procedures in a Brazilian population sample. **Hum Immunol,** v. 77, n. 10, p. 841-53, Oct 2016. ISSN 0198-8859.

LIMA, T. H. A. et al. HLA-A promoter, coding, and 3'UTR sequences in a Brazilian cohort, and their evolutionary aspects. **HLA,** v. 93, n. 2-3, p. 65-79, Feb 2019. ISSN 2059-2310.

MAJORCZYK, E. et al. A single nucleotide polymorphism -35 kb T>C (rs9264942) is strongly associated with psoriasis vulgaris depending on HLA-Cw(*)06. **Human immunology,** v. 75, n. 6, p. 504-507, 2014. ISSN 1879-1166 (Electronic)\r0198-8859 (Linking).

MAKADZANGE, A. T. et al. Characterization of an HLA-C-restricted CTL response in chronic HIV infection. **Eur J Immunol,** v. 40, n. 4, p. 1036-41, 2010 Apr 2010. ISSN 0014-2980.

MATZARAKI, V. et al. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. **Genome Biol,** v. 18, n. 1, p. 76, Apr 2017. ISSN 1474-760X.

MCCUTCHEON, J. A. et al. Low HLA-C expression at cell surfaces correlates with increased turnover of heavy chain mRNA. **The Journal of experimental medicine,** v. 181, n. 6, p. 2085-95, 1995.

MENDES-JUNIOR, C. T. et al. Genetic diversity of the HLA-G coding region in Amerindian populations from the Brazilian Amazon: a possible role of natural selection. **Genes Immun,** v. 14, n. 8, p. 518-26, Dec 2013. ISSN 1466-4879.

MEYER, D. et al. A genomic perspective on HLA evolution. **Immunogenetics**, Jul 2017. ISSN 1432-1211.

MEYER, D.; THOMSON, G. How selection shapes variation of the human major histocompatibility complex: a review. **Ann Hum Genet,** v. 65, n. Pt 1, p. 1-26, Jan 2001. ISSN 0003-4800 (Print)0003-4800.

NAIYER, M. M. et al. KIR2DS2 recognizes conserved peptides derived from viral helicases in the context of HLA-C. **Sci Immunol,** v. 2, n. 15, 2017 Sep 15 2017. ISSN 2470-9468.

NEEFJES, J. J.; PLOEGH, H. L. Allele and locus-specific differences in cell surface expression and the association of HLA class I heavy chain with beta 2-microglobulin: differential effects of inhibition of glycosylation on class I subunit association. **European journal of immunology,** v. 18, n. 5, p. 801-810, 1988.

NEISIG, A.; MELIEF, C. J.; NEEFJES, J. Reduced cell surface expression of HLA-C molecules correlates with restricted peptide binding and stable TAP interaction. **Journal of immunology (Baltimore, Md. : 1950),** v. 160, n. 1, p. 171-179, 1998.

OHASHI, P. S. T-cell signalling and autoimmunity: molecular mechanisms of disease. **Nat Rev Immunol,** v. 2, n. 6, p. 427-38, Jun 2002. ISSN 1474-1733.

PARCEJ, D.; TAMPÉ, R. ABC proteins in antigen translocation and viral inhibition. **Nat Chem Biol,** v. 6, n. 8, p. 572-80, Aug 2010. ISSN 1552-4469.

PARHAM, P. Killer cell immunoglobulin-like receptor diversity: balancing signals in the natural killer cell response. **Immunol Lett,** v. 92, n. 1-2, p. 11-3, Mar 29 2004. ISSN 0165-2478 (Print)0165-2478.

PARHAM, P. MHC class I molecules and kirs in human history, health and survival. **Nat Rev Immunol,** v. 5, n. 3, p. 201-214, 2005.

PARHAM, P.; MOFFETT, A. Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. **Nat Rev Immunol,** v. 13, n. 2, p. 133-44, Feb 2013. ISSN 1474-1733.

PAUL, P. et al. HLA-G expression in melanoma: a way for tumor cells to escape from immunosurveillance. **Proc Natl Acad Sci U S A,** v. 95, n. 8, p. 4510-5, Apr 1998. ISSN 0027-8424.

PENA, S. D. et al. The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. **PLoS One,** v. 6, n. 2, p. e17063, 2011. ISSN 1932-6203.

PENMAN, B. S. et al. Pathogen selection drives nonoverlapping associations between HLA loci. **Proc Natl Acad Sci U S A,** v. 110, n. 48, p. 19645-50, Nov 26 2013. ISSN 0027-8424.

PIETRA, G. et al. HLA-E and HLA-E-bound peptides: recognition by subsets of NK and T cells. **Curr Pharm Des,** v. 15, n. 28, p. 3336-44, 2009. ISSN 1873-4286.

RAMSURAN, V. et al. Sequence and Phylogenetic Analysis of the Untranslated Promoter Regions for HLA Class I Genes. **J Immunol,** v. 198, n. 6, p. 2320-2329, Mar 15 2017. ISSN 0022-1767.

ROBINSON, J. et al. The IPD and IMGT/HLA database: allele variant databases. **Nucleic Acids Res,** v. 43, n. Database issue, p. D423-31, Jan 2015. ISSN 1362-4962.

ROCK, K. L.; REITS, E.; NEEFJES, J. Present Yourself! By MHC Class I and MHC Class II Molecules. **Trends Immunol,** v. 37, n. 11, p. 724-737, Nov 2016. ISSN 1471-4981.

SABBAGH, A. et al. Worldwide genetic variation at the 3' untranslated region of the HLA-G gene: balancing selection influencing genetic diversity. **Genes Immun,** v. 15, n. 2, p. 95-106, Mar 2014. ISSN 1476-5470.

SAMS, A.; HAWKS, J. Celiac disease as a model for the evolution of multifactorial disease in humans. **Hum Biol,** v. 86, n. 1, p. 19-36, Winter 2014. ISSN 0018-7143.

SANCHEZ-MAZAS, A. An apportionment of human HLA diversity. **Tissue Antigens,** v. 69 Suppl 1, p. 198-202, Apr 2007. ISSN 0001-2815 (Print)0001-2815.

SATTA, Y.; LI, Y. J.; TAKAHATA, N. The neutral theory and natural selection in the HLA region. **Frontiers in bioscience : a journal and virtual library,** v. 3, p. d459-67, 1998.

SCHWARTZ, O. et al. Endocytosis of major histocompatibility complex class I molecules is induced by the HIV-1 Nef protein. **Nat Med,** v. 2, n. 3, p. 338-42, 1996 Mar 1996. ISSN 1078-8956 (Print)1078-8956.

SEBZDA, E. et al. Selection of the T cell repertoire. **Annu Rev Immunol,** v. 17, p. 829-74, 1999. ISSN 0732-0582.

SELIGER, B.; RITZ, U.; FERRONE, S. Molecular mechanisms of HLA class I antigen abnormalities following viral infection and transformation. **Int J Cancer,** v. 118, n. 1, p. 129-38, 2006 Jan 1 2006. ISSN 0020-7136 (Print)0020-7136.

SELIGER, B. et al. Association of HLA class I antigen abnormalities with disease progression and early recurrence in prostate cancer. **Cancer Immunol Immunother,** v. 59, n. 4, p. 529-40, 2010 Apr 2010. ISSN 0340-7004 (Print)1432-0851 (Electronic).

SHIINA, T. et al. The HLA genomic loci map: expression, interaction, diversity and disease. **Journal of Human Genetics,** v. 54, n. 1, p. 15-39, 2009.

SHIINA, T.; INOKO, H.; KULSKI, J. K. An update of the HLA genomic region, locus information and disease associations: 2004. **Tissue Antigens,** v. 64, n. 6, p. 631-649, 2004.

SIBILIO, L. et al. A single bottleneck in HLA-C assembly. **J Biol Chem,** v. 283, n. 3, p. 1267-74, Jan 2008. ISSN 0021-9258.

SONON, P. et al. HLA-G, -E and -F regulatory and coding region variability and haplotypes in the Beninese Toffin population sample. **Mol Immunol,** v. 104, p. 108-127, 12 2018. ISSN 1872-9142.

SPURGIN, L. G.; RICHARDSON, D. S. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. **Proc Biol Sci,** v. 277, n. 1684, p. 979-88, Apr 07 2010. ISSN 0962-8452.

STAJICH, J. E.; HAHN, M. W. Disentangling the effects of demography and selection in human history. **Mol Biol Evol,** v. 22, n. 1, p. 63-73, Jan 2005. ISSN 0737-4038.

STRANGE, A. et al. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. **Nat Genet,** v. 42, n. 11, p. 985-90, 2010 Nov 2010. ISSN 1061-4036.

TAJIMA, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. **Genetics,** v. 123, n. 3, p. 585-595, 1989.

TAN, Z.; SHON, A. M.; OBER, C. Evidence of balancing selection at the HLA-G promoter region. **Hum Mol Genet,** v. 14, n. 23, p. 3619-28, Dec 2005. ISSN 0964-6906.

THOMAS, R. et al. HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. **Nat Genet,** v. 41, n. 12, p. 1290-1294, 2009.

TIERCY, J.-M. HLA-C Incompatibilities in Allogeneic Unrelated Hematopoietic Stem Cell Transplantation. **Frontiers in immunology,** v. 5, p. 216-216, 2014.

TROWSDALE, J.; KNIGHT, J. C. Major histocompatibility complex genomics and human disease. **Annu Rev Genomics Hum Genet,** v. 14, p. 301-23, 2013. ISSN 1527-8204.

TROWSDALE, J.; MOFFETT, A. NK receptor interactions with MHC class I molecules in pregnancy. **Seminars in Immunology,** v. 20, n. 6, p. 317-320, 2008/12/01/ 2008. ISSN 1044-5323.

TURNER, S. et al. Sequence-based typing provides a new look at HLA-C diversity. **J Immunol,** v. 161, n. 3, p. 1406-13, 1998 Aug 1 1998. ISSN 0022-1767 (Print)0022-1767.

VAN DEUTEKOM, H. W. M.; KEŞMIR, C. Zooming into the binding groove of HLA molecules: which positions and which substitutions change peptide binding most? **Immunogenetics,** v. 67, n. 8, p. 425-436, 2015.

VAN HALL, T. et al. The Varicellovirus-Encoded TAP Inhibitor UL49.5 Regulates the Presentation of CTL Epitopes by Qa-1b1. 2007-01-15 2007.

VEIGA-CASTELLI, L. C. et al. Non-classical HLA-E gene variability in Brazilians: a nearly invariable locus surrounded by the most variable genes in the human genome. **Tissue antigens,** v. 79, n. 1, p. 15-24, 2012.

VINCE, N. et al. HLA-C Level Is Regulated by a Polymorphic Oct1 Binding Site in the HLA-C Promoter Region. **Am J Hum Genet,** v. 99, n. 6, p. 1353-1358, Dec 01 2016. ISSN 0002-9297.

WATTERSON, G. A. The homozygosity test of neutrality. **Genetics,** v. 88, n. 2, p. 405-417, 1978.

WIEGERS, G. J. et al. Shaping the T-cell repertoire: a matter of life and death. **Immunol Cell Biol,** v. 89, n. 1, p. 33-9, Jan 2011. ISSN 1440-1711.

WIENDL, H. et al. A functional role of HLA-G expression in human gliomas: an alternative strategy of immune escape. **J Immunol,** v. 168, n. 9, p. 4772-80, May 2002. ISSN 0022-1767.

WIERTZ, E. J. et al. The human cytomegalovirus US11 gene product dislocates MHC class I heavy chains from the endoplasmic reticulum to the cytosol. **Cell,** v. 84, n. 5, p. 769-79, 1996 Mar 8 1996. ISSN 0092-8674 (Print)0092-8674.

WIERTZ, E. J. et al. Sec61-mediated transfer of a membrane protein from the endoplasmic reticulum to the proteasome for destruction. **Nature,** v. 384, n. 6608, p. 432-8, 1996 Dec 5 1996. ISSN 0028-0836 (Print)0028-0836.

WILCZYŃSKA, K. et al. ERAP, KIR and HLA-C gene interaction in susceptibility to recurrent spontaneous abortion in the Polish population. **Hum Immunol,** v. 80, n. 5, p. 344-348, May 2019. ISSN 1879-1166.

WILLIAMS, M. et al. Direct Binding of Human Immunodeficiency Virus Type 1 Nef to the Major Histocompatibility Complex Class I (MHC-I) Cytoplasmic Tail Disrupts MHC-I Trafficking. **Journal of Virology,** v. 76, n. 23, p. 12173-12184, 2002.

YANG, Y.; SEMPÉ, P.; PETERSON, P. A. Molecular mechanisms of class I major histocompatibility complex antigen processing and presentation. **Immunologic research,** v. 15, n. 3, p. 208-33, 1996.

YEWDELL, J. W.; REITS, E.; NEEFJES, J. Making sense of mass destruction: quantitating MHC class I antigen presentation. **Nat Rev Immunol,** v. 3, n. 12, p. 952-61, Dec 2003. ISSN 1474-1733.

ZEMMOUR, J.; PARHAM, P. Distinctive polymorphism at the HLA-C locus: implications for the expression of HLA-C. **The Journal of experimental medicine,** v. 176, n. 4, p. 937-50, 1992.

ZIEGLER, H. et al. The luminal part of the murine cytomegalovirus glycoprotein gp40 catalyzes the retention of MHC class I molecules. In: (Ed.). **EMBO J**, v.19, 2000. p.870-81. ISBN 0261-4189 (Print)1460-2075 (Electronic).

# Capítulo II - Artigo

## Article

# *HLA-C* genetic diversity in two geographically distinct samples from Brazil and Benin and evolutionary insights

Andreia S. Souza[1,2], Michelle A. Paz [1,3], Paulin Sonon[4], Thálitta H. A. Lima[1,2], Iane O. P. Porto[1,3], Luciana C. Veiga-Castelli [5], Maria Luiza G. Oliveira[5], Eduardo A. Donadi[6], Diogo Meyer[7], Audrey Sabbagh[8], Celso T. Mendes-Junior[9], David Courtin [8], Erick C. Castelli[1,2,3]

[1] São Paulo State University (UNESP), Molecular Genetics and Bioinformatics Laboratory - Experimental Research Unity, School of Medicine, Botucatu, State of São Paulo, Brazil. [2] São Paulo State University (UNESP), Genetics Program, Institute of Biosciences of Botucatu, Botucatu, State of São Paulo, Brazil. [3]São Paulo State University (UNESP), Pathology Program, School of Medicine, Botucatu, State of São Paulo, Brazil. [4] Laboratório de Biologia Molecular,Universidade de São Paulo, Programa de Imunologia Básica e Aplicada (IBA), Faculdade de Medicina de Ribeirão Preto (FMRP-USP), Estado de São Paulo, SP, Brazil. [5] Department of Genetics, School of Medicine of Ribeirão Preto, University of São Paulo (USP), Ribeirão Preto, State of São Paulo, Brazil. [6] Department of Medicine, School of Medicine of Ribeirão Preto, University of São Paulo (USP), Ribeirão Preto, State of São Paulo, Brazil. [7] University of São Paulo, Department of Genetics and Evolutionary Biology, São Paulo, Brazil. [8] UMR 216 MERIT, IRD, Université Paris Descartes, Faculté de Pharmacie, Sorbonne Paris Cité, Paris, France. [9] Departamento de Química, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, São Paulo, Brasil

**Contact**
Erick C. Castelli
Departamento de Patologia, Faculdade de Medicina de Botucatu, Unesp – Botucatu, SP
CEP: 18618970, Brazil, Phone: +55 14 3880-1696, E-mail address: erick.castelli@unesp.br
ORCID 0000-0003-2142-7196

## Abstract

Human Leucocyte antigen-C (HLA-C) is a classical HLA class I molecule that binds and presents peptides to cytotoxic T lymphocytes in the cell surface. HLA-C has a dual function since it also interacts with KIR receptors expressed in NK and T cells, modulating their activity. The structure and diversity of the *HLA-C* regulatory regions, as well as the relationship among variants along the *HLA-C* locus, is poorly addressed, and no population-based study explored the complete *HLA-C* variability in different population samples. Here we present a molecular and bioinformatics method to evaluate the entire *HLA-C* diversity, including regulatory sequences. Then, we applied this method to survey the *HLA-C* diversity in two geographically distinct population samples, one admixed from Brazil and one less admixed from Benin. The *HLA-C* promoter and 3'UTR were very polymorphic with the presence of few but highly divergent haplotypes, and also presenting conserved segments that are shared among different primates. Nucleotide diversity was higher in other exonic segments rather than exons 2 and 3, and also higher in the second half of the 3'UTR region. We detected evidence of balancing selection on the entire *HLA-C* locus and positive selection in exon 1, for both populations. HLA-C motifs previously associated with KIR interaction and expression regulation are similar between both populations, but the frequency of amino acids influencing peptide-binding is different between samples. Each allele group is associated with specific regulatory sequences, supported by the high linkage disequilibrium along the entire *HLA-C* locus in both populations.

Key words: *HLA-C*, variability, natural selection, NGS, Brazilian population, Beninese population.

# 1. Introduction

Human Leucocyte Antigen C (*HLA-C*) gene encodes an important molecule for antigen presentation and Natural Killer (NK) cell modulation. *HLA-C*, together with other classical HLA class I genes, *HLA-A* and *HLA-B,* are among the most polymorphic human genes. This variability is mainly associated with their function because HLA-C binds different intracellular peptides and present them at the cell surface to T CD8+ cells. The HLA-peptide complex on the cell surface allows the recognition of self and non-self antigens by T lymphocytes, triggering an immune response against cells presenting foreign or abnormal peptides, such as virally infected or tumor cells (Rock, Reits and Neefjes, 2016). Classical HLA class I molecules are also able to interact with activating and inhibitory Killer-cell immunoglobulin-like receptors (KIR) of NK cells, modulating its activity (Parham, 2004; Yawata *et al.*, 2006; Parham and Moffett, 2013; Penman *et al.*, 2016). However, *HLA-C* shows some peculiar features that make it an unusual classical HLA gene.

*HLA-C* has a polymorphic nature, with about 5,653 alleles reported in the IPD-IMGT/HLA database, version 3.38, following other classical *HLA* class I genes like *HLA-B* (7,053 alleles) and *HLA-A* (5,735 alleles). However, *HLA-C* presents lower cell surface expression levels when compared with *HLA-A* and *HLA-B* and, thus, it has been less associated with restricted CTL responses (Blais, Dong and Rowland-Jones, 2011). Nonetheless, higher HLA-C expression levels might influence many clinical outcomes, such as HIV infection (Apps *et al.*, 2013; Thomas *et al.*, 2009; Kulkarni *et al.*, 2011; Fellay *et al.*, 2007), unrelated hematopoietic cell transplantation (Petersdorf *et al.*, 2014; Tiercy, 2014), and Crohn's disease (Kulkarni *et al.*, 2013), among others, suggesting that the molecule expression levels might influence the HLA-C-restricted CTL response.

*HLA-C* regulatory variants play a role in HLA-C expression levels. For instance, a single-nucleotide polymorphism (SNP) located 35 kb upstream *HLA-C* (rs9264942) and a variant at the 3' untranslated region (3'UTR, rs67384697) were associated with increased HLA-C expression levels and decreased HIV viral load (Kulkarni *et al.*, 2011; Thomas *et al.*, 2009; Apps *et al.*, 2013; Fellay *et al.*, 2007). The 3'UTR variant affects miR-148a binding and thus HLA-C expression. Moreover, it was also associated with a deleterious effect in psoriasis and Crohn's diseases (Kulkarni *et al.*, 2013; Chen *et al.*, 2012). The influence on binding of transcription factors has previously been described for some *HLA-C* promoter variants. It has been suggested that the variant rs2395471, located about 800 bp upstream the transcription start site may influence the binding of Oct1 (Vince *et al.*, 2016), while variants rs2524094 and

rs10657191 could affect TNF-α and IFN-γ responses, respectively (Hundhausen *et al.*, 2012). Some *HLA-C* allele groups, such as HLA-C*03, -C*07, and -C*17, present variations in one or more transcription factor binding sites (TFBSs) at the core promoter (Ramsuran *et al.*, 2017). These alleles also carry rs2395471*G (negatively influencing Oct1 binding) and an intact miR-148a binding site, both associated with lower HLA-C cell surface expression levels (Ramsuran *et al.*, 2017; Vince *et al.*, 2016).

*HLA-C* is the only classical HLA class I gene expressed at the maternal-fetal interface. Along with the non-classical genes *HLA-G* and *HLA-E,* it plays a pivotal immunomodulatory role for placentation and pregnancy success (Chazara, Xiong and Moffett, 2011; Hackmon *et al.*, 2017). HLA-C is the most important among the HLA class I molecules for NK cell activity regulation, since all HLA-C proteins bind to activating and inhibitory KIR ligands (Sharkey *et al.*, 2008). Considering that both *KIR* and *HLA-C* are highly polymorphic, pregnancy success may depend on the combination of fetal *HLA-C* and maternal *KIRs* (Chazara, Xiong and Moffett, 2011; Moffett-King, 2002; Hiby *et al.*, 2004). Specific *KIR/HLA-C* combinations have also been associated with susceptibility to autoimmune and inflammatory diseases, outcomes in a series of infectious diseases (reviewed by (Kuśnierczyk, 2013; Kulkarni, Martin and Carrington, 2008; Parham, 2005)), and alloreactivity in hematopoietic stem cell transplantation (reviewed by (Parham, 2005)). Given the HLA-C role for immune defense and reproduction, selective pressures may have shaped its genetic variability and polymorphism frequencies (Parham *et al.*, 2012; Augusto and Petzl-Erler, 2015).

The high polymorphism levels across classical HLA *loci* are related to the presentation of a wide range of peptides within a population (Meyer and Thomson, 2001). *HLA-C* follows *HLA-A* and *HLA-B* polymorphic nature, and they are all related to antigen presentation. Nonetheless, HLA-C is also associated with the modulation of NK cells activity, such as performed by the highly conserved non-classical HLA genes, *HLA-G* and *HLA-E.* Thus, it is not clear (i) how HLA-C presents both features considering its polymorphic nature and expression pattern, and (ii) how natural selection has shaped this gene variability since the coding region of classical and non-classical HLA genes present different selection profiles. Both *HLA-G* and *HLA-E* coding region are under the influence of purifying selection (Castelli *et al.*, 2017; Ramalho *et al.*, 2017) while balancing selection acting on the segments encoding the peptide-binding groove of classical HLA molecules is well documented (Sanchez-Mazas, 2007; Lima *et al.*, 2019; Meyer *et al.*, 2006; Hedrick and Thomson, 1983; Hedrick, Whittam and Parham, 1991; Garrigan and Hedrick, 2003).

The structure and diversity of the HLA-C regulatory regions, as well as the relationship among variants along the HLA-C locus (including the promoter and 3' UTR) have been poorly addressed, and no population-based study explored the complete *HLA-C* variability. Furthermore, most of the known *HLA-C* alleles that were submitted to the IPD-IMGT/HLA database lack the proximal promoter, 3'UTR, and intron sequences. Here we propose a second-generation sequencing and bioinformatic approach to evaluate the complete *HLA-C* variability encompassing at least 1,500bp of the promoter segment, the coding region and the entire 3'UTR in a highly admixed Brazilian population sample and less admixed Beninese population sample. The bioinformatics workflow optimizes sequence mapping at *HLA-C* and infers haplotypes combining the phase information obtained directly at the sequencing data and probabilistic models. We observed that the regulatory segments present few promoter and 3'UTR haplotypes, but highly divergent among them, and that each coding allele group is associated with similar regulatory sequences. Our results suggest that most of the *HLA-C* segments may be under balancing selection, but a different scenario arises when the leader-peptide, encoded by exon 1, the promoter, and initial 3'UTR segment is taken into account.

## 2. Materials and Methods

### 2.1. Brazilian samples

We evaluated *HLA-C* variability in 418 unrelated samples from the state of São Paulo, Southeastern Brazil. Each individual signed an informed consent term before blood withdrawal. This study protocol was reviewed and approved by the Human Research Ethics Committee of the School of Medicine (Unesp/Brazil) - Protocol #24157413.7.0000.5411. To assess ancestry composition of the population sample, we randomly selected 205 samples to evaluate Ancestry Informative Markers (AIMs). Ancestry composition of the population sample was estimated at 75.5% European, 16% African, and 8.5% Amerindian/Asian.

### 2.2. Beninese samples

We evaluated *HLA-C* variability of 108 unrelated individuals of the Toffin ethnic group from Sô-Ava, an area located in Southern region of Benin, 12 km North of Cotonou, the economic capital of Benin. Informed consent was obtained from all individual participants included in the study before blood collection. The study was approved by the Ethics Committee of the ''Faculté des Sciences de la Santé (FSS)'' of Cotonou, Benin and registered at No.12/ 03/2012/CEIFSS/UAC.

### 2.3. *HLA-C* gene amplification and sequencing libraries

Genomic DNA was extracted by a salting-out procedure or using GeneJET Genomic DNA Purification kit (Thermo Fisher Scientific, Waltham, MA), according to manufacturer's instructions. DNA samples were quantified with Qubit dsDNA Broad Range Assays (Thermo Fisher Scientific Inc., Waltham, MA, USA) and normalized to 50 ng/µL.

*HLA-C* was amplified as a single amplicon of approximately 5,671 nucleotides (not considering primers sequences), comprehending the segment between nucleotides 31,268,667 and 31,274,338 considering the chromosome 6 sequence from human genome assembly hg38. Polymerase Chain Reaction (PCR) was carried out with primers HCPR.F1 (5'-TGAAGAACTGAACAGCAACTA-3') and HCUT.R1 (5'-GTCTGAGGGATAAGGGGCA-3') in a final volume of 50 µL, containing 0.30 µM of each primer, 0.20 mM of each dNTP (Invitrogen, Carlsbad, CA, EUA), 1.25 units of DNA polymerase (PrimeSTAR GXL, TaKaRa Bio Company), 1X the PCR buffer solution supplied with the DNA polymerase and 50 ng of genomic DNA. Cycling conditions were set as follows: 30 cycles of 98°C for 10 seconds, 60°C for 15 seconds and 68°C for 6 minutes. Amplicons were evaluated on 1% agarose gel stained with GelRed® (Biotium™, Hayward, USA), purified using Illustra ExoProStar (GE Healthcare), quantified using Qubit dsDNA High-Sensitivity Assays (ThermoFisher Scientific, USA), and normalized to 0.2 ng/µL.

In order to prepare sequencing libraries, we used the Nextera XT Library Preparation Kit and Nextera XT Index Kit (both from Illumina, Inc.). Libraries were quantified by qPCR using Kapa (Kapa Biosystems, Wilmington, USA). The fragmentation pattern was assessed with High-Sensitivity DNA Bioanalyzer chips (Agilent Technologies, CA, USA), and we normalized samples based on the quantification and the fragmentation pattern. Sequencing was performed using MiSeq Reagent Kit (V2, 500 cycles, 2 x 250bp) in a MiSeq Platform, as recommended by Illumina Inc.

### 2.4. Raw data processing, mapping, and genotyping

Given the polymorphic nature of HLA genes and the high sequence similarity among them due to their paralogous origins, which may bias read mapping (Brandt *et al.*, 2015), and because all other class I genes were sequenced together with *HLA-C* (data not presented here), we used *hla-mapper dna* to optimize read mapping at the *HLA-C locus* (Castelli *et al.*, 2018), with error threshold set to 0.05, minimum read size set 70, tolerance set to 0.04, and also using the flag --multiple_hits_MQ0.

After the mapping procedure, we observed the underrepresentation of reads at intron 2. This low sequence depth is not related to mapping bias since further investigations using other library preparation kits circumvented this issue. Thus, this low sequence depth is related to a fragmentation bias of Nextera XT kit, which affects especially CG-rich regions such as intron 2. This was already reported for the *HLA-A* locus (Lima *et al.*, 2019).

We used the Genome Analysis Toolkit (*GATK*, version 4.1) *HaplotypeCaller* in the GVCF mode to infer genotypes, using hg38 as a reference (McKenna *et al.*, 2010). The multi-sample G.VCF file was created with *GATK CombineGVCFs* and the multi-sample VCF file with genotype likelihoods [created with *GATK GenotypeGVCFs*] was annotated considering the dbSNP version 150.

We processed the multi-sample VCF in order to introduce missing alleles on genotypes with low likelihoods or on unbalanced genotypes, using *vcfx version 2.0,* available at www.castelli-lab.net/apps/vcfx. The missing alleles were introduced on genotypes presenting a likelihood lower than 99.999% [using *vcfx checkpl*], and on unbalanced genotypes [using *vcfx checkad,* with default parameters]. After, we refined variants using *vcfx evidence*, with an additional step in which we manually checked the variants that were excluded. Since *vcfx* introduced a large number of missing alleles at intron 2 because of the low sequencing depth and because of the large number of highly unbalanced genotypes in this segment, we opted to exclude the entire intron 2 from the analysis. However, it should be noted that this issue only affects studies using Nextera, and we haven't observed this issue in WGS or using other fragmentation strategies.

### 2.5. Phasing and *HLA-C* allele calling

We removed singletons using *vcfx singleton* before phasing. Then, we used *GATK ReadBackedPhasing* to detect the phase among closely located sites, i.e., occurring at the same DNA fragment or read (McKenna *et al.*, 2010). *ReadBackedPhasing* (RBP) considered a haplotype quality threshold of 2,000. Nonetheless, this algorithm ignores indels and multi-allelic *loci*. Then we used *phasex* to perform the haplotyping analysis (available upon request). This software considers the phasing sets detected by RBP and creates several VCF replicates, each of them considering a random phase set for each sample. Then, *phasex* uses Beagle 4.1 to phase each replicate (Browning and Browning, 2007), and the results are compared. For *HLA-C* we used 100 replicates. When a sample presents the same pair of haplotypes in at least 95/100 of the replicates, this pair is fixed and passed forward as entirely phased to the next round of replicates. This strategy goes on until no new sample is fixed (for *HLA-C*, it was 5 iterations), and, after the final iteration, all samples presenting the same haplotype in at least 70% of the

replicates is considered phased and this haplotype passed forward for further analysis. Considering all heterozygous sites, 77.51% were directly phased using RBP. The combination *phasex/Beagle 4.1* phased the remaining 22.49%, most of them referring to indels or multiallelic *loci*, and also imputed the 0.37% of missing alleles observed after the *vcfx* treatment. Since we removed intron 2 as discussed earlier, the association between the segment upstream and downstream intron 2 was entirely evaluated using probabilistic models by Beagle 4.1 (Browning and Browning, 2007).

After haplotyping, whenever possible, we manually re-introduced singletons. Then, we used *vcfx fasta* to create a complete genomic sequence and *vcfx transcript* to create the CDS sequences – two for each individual. Since *HLA-C* is encoded at the GRCh38 chromosome 6 reverse strand, sequences were reversed and complemented using *emboss revseq* (Rice, Longden and Bleasby, 2000).

We designed Perl scripts coupled with a local BLAST server with a database containing all known HLA alleles described by the IPD-IMGT/HLA database (version 3.36.0) to identify the closest known *HLA-C* coding allele for each different sequence we have detected, and the mutations observed when compared to it (Robinson *et al.*, 2015). We inferred the encoded proteins with *emboss transeq* (Rice, Longden and Bleasby, 2000). Promoters and 3' UTR haplotypes were named according to sequence similarities and their phylogenetic relationship.

### 2.6. Other analyses

We used the Arlequin 3.5 to assess haplotype frequencies, departures of Hardy-Weinberg equilibrium (HWE) expectations, nucleotide diversity, gene diversity, population differentiation parameters, $F_{ST}$ between populations, and Tajima's *D* for each *HLA-C* exon, intron, regulatory segment, and *HLA-C* as a whole. The significance of the *D* statistic was tested by generating random samples under the hypothesis of selective neutrality, using 5,000 simulations (Excoffier and Lischer, 2010). The input files for Arlequin were created using *vcfx arlequin*. The nucleotide diversity and Tajima's *D* plots were calculated by using *VariScan* (Vilella *et al.*, 2005) in a sliding window approach of 150bp and a step size of 3. We used *Pypop* to perform the Ewens-Watterson test (Lancaster *et al.*, 2007). The $d_N/d_S$ ratio test, which evaluates the ratio of synonymous and nonsynonymous nucleotide substitution, was calculated using *MEGA* version 7.0.20 (Kumar *et al.*, 2016) for each *HLA-C* exon. The frequency of each promoter, coding, and 3'UTR sequence was compared between samples using the Fisher exact test, using Bonferroni correction ($P_c$) and considering the number of different sequences observed in each case.

Linkage Disequilibrium (LD) across the *HLA-C locus* was assessed using *Haploview* 4.2 (Barrett *et al.*, 2005), considering only variable sites with a minor allele frequency (MAF) higher than 1%. Haplotype blocks were inferred by using Gabriel's method. The PED and MAP file for Haploview were generated using *vcfx haploview*. As a quality control, we used *Optitype* (Szolek *et al.*, 2014) to get *HLA-C* coding alleles, comparing the results with the ones obtained with our workflow.

## 3. Results

We evaluated *HLA-C* variability encompassing at least 1,500 nucleotides from the upstream promoter to the end of *HLA-C* last exon, exception made to intron 2, in two different population samples. There were 359 variable sites across *HLA-C*. The spreadsheet file available as supplemental material (Table S1) lists all these variants as seem in the VCF file when using HaplotypeCaller, together with their chromosome positions, SNPid, the reference allele frequencies, and other information regarding each site. Most of them are already reported in the IPD-IMGT/HLA database. Genotype frequencies were consistent with the Hardy-Weinberg equilibrium expectations for more than 96% of the variants, and the remaining are randomly distributed across *HLA-C* in both population samples.

The *HLA-C* upstream segment (approximately 1500 bases upstream HLA-C, which includes the distal promoter, the proximal promoter, and the 5'UTR) presented 86 variants arranged into 33 different sequences (Table 1 and Supplemental alignment 1). The *HLA-C* CDS (coding sequence, all exons) presented 47 different sequences (Table 2) encoding 40 HLA-C protein molecules (Table 3). The *HLA-C* 3'UTR segment is encoded in the last exon, and this segment presented 40 variable sites arranged into 25 different sequences (Table 4 and Supplemental alignment 2).

### 3.1. *HLA-C* genetic diversity

*HLA-C* promoters were named following sequence similarities and their phylogenetic relationships (Table 1 and supplemental alignment 1). There were 13 different promoter groups, with many shared variants within each group. In both samples, the most frequent promoter sequence was *P01:01,* with a frequency of 16.7% among Brazilians and 27.3% among Beninese (Table 1). This promoter is linked with alleles from the *HLA-C*04* group Promoter frequencies differ between populations, as observed for *P01:01* ($P_c$ = 0.0198), *P04:02* ($P_c$ = 0.0003), *P05:01* ($P_c$ = 0.0066), and *P10:01* ($P_c$ = 0.0003). Each *HLA-C* promoter group is associated with a

specific *HLA-C* coding allele group, supporting the high Linkage Disequilibrium (LD) observed across the gene (Figure S1).

The *HLA-C* coding segment (all exons and introns, except intron 2) presented 82 different sequences, named as described in Methods (Table S2). Among them, 67 (81.71%) are identical to a known IPD-IMGT/HLA alleles, and the 15 remaining sequences (new alleles, grouped as unknown sequences at Table S2) have a summed frequency of 2.09% considering both populations pooled together. Some of these new alleles were detected more than once. For instance, there were six copies of C*18:01:01 with a different nucleotide at the 3'UTR, and because of that all C*18:01:01 are placed under unknown sequences in Table S2.

The most frequent CDS sequences among Brazilians are *C*04:01:01* (16.7%), *C*06:02:01* (9.09%), and *C*07:01:01* (8.97%), while for Beninese they are *C*04:01:01* (27.3%), *C*16:01:01* (14.3%), and *C*17:01:01* (13%) (Table 2). As observed for the promoter segment, allele frequencies vary between both population samples, as observed for CDS alleles *C*04:01:01* ($P_c$ = 0.0282), *C*06:02:01* ($P_c$ = 0.0047), *C*16:01:01* ($P_c$ = 0.00047), and *C*17:01:01* ($P_c$ = 0.0004). Likewise, protein frequencies shift between both population samples (Table 3), with C*04:01 being the most frequent in both samples, although twice as frequent in Benin. Only two CDS sequences are not identical to any sequence described in the IPD-IMGT/HLA database, but they encode previously described proteins.

The haplotypes within the 3'UTR segment are arranged in four distinct groups, with many mutational steps apart (Supplemental alignment 2). The most frequent 3'UTR haplotype in both populations was *U01:05,* in LD with *HLA-C*04* alleles, and others. As observed for the promoter region, each 3'UTR lineage (U01, U02, U03, and U04) is in LD with specific *HLA-C* allele groups (Table 4).

When we combine the promoter, CDS, and 3'UTRs sequences as extended haplotypes, there are 74 different *HLA-C* sequences (Table 5), and each HLA-C allele groups is related to specific promoter and 3'UTR sequences, supporting the high LD across the gene (Figure S1) and the single segregation block observed when all samples are pooled together.

*HLA-C* extended haplotypes were under the Hardy-Weinberg equilibrium (HWE) expectations in both populations, Beninese ($P$=0.66080±0.01259) and Brazilian ($P$=0.49396±0.01388). The population differentiation exact test based on haplotype frequencies revealed significant differences between these population samples ($P$<0.00001), supported by the frequency shifts when considering any *HLA-C* segment (promoter, CDS, and 3'UTR). Population differentiation measured by $F_{ST}$ presented a low level of genetic differentiation

between the studied populations ($F_{ST}$=0.0173), however, this $F_{ST}$ value was statistically significant ($P$<0.0001).

### 3.2. *HLA-C* evolutive aspects

Exception for exon 6, all *HLA-C* segments presented high nucleotide diversity (π), positive Tajima's D, and negative normalized F values (Table 6). Both populations presented significant positive Tajima's for exon 2, exon 3, exon 4, exon 5, the 3'UTR, and the entire CDS. Likewise, they presented significant negative normalized F (Ewens-Watterson) for exon 1, exon 4, and exon 5. Using a 150bp sliding window with a step size of 3 to calculate nucleotide diversity and Tajima's D, both populations presented the same pattern. Because of that, we opted to plot data only for the largest sample (Brazilians). Nucleotide diversity variates across the promoter segment, with a highly conserved segment between position -500 and -800, and with negative Tajima's D (Figure 1). Nucleotide diversity was high throughout HLA-C CDS, except for two segments. The first one presenting low nucleotide diversity is in the second half of exon 2 around nucleotide position 265, and it coincides with the highest Tajima's D observed across the *HLA-C* CDS (D=3.71 for Brazilians and D=2.98 for Beninese). This segment encodes residues between positions 41 and 90 (alpha 1 domain) of the HLA-C mature protein and it is important for both antigen presentation and KIR binding. In this segment, there were 6 highly frequent amino acid exchanges, including the dimorphism C1/C2 at position 80 important for KIR interaction (Table S3). The second is in the middle of exon 4 around nucleotide 740, with intermediate positive values for Tajima's D (Figure 1). The highest nucleotide diversity coincides with exon 5 around nucleotide 940, but with intermediate values for Tajima's D. Likewise, there are frequent non-synonymous mutations in this segment, including the insertion of 6 amino acids related to the C*17 alleles (amino acid 301 to 306, Table S3). The first half of the 3'UTR segment is very conserved in both samples, but the second half presented nucleotide diversity higher than the observed for exons 2 and 3 (Figure 1). The high nucleotide diversity in exons 2, 3, and 5 is also demonstrated when we considered each segment separately in both samples (Table 6).

The $d_N/d_S$ ratio test indicated an excess of non-synonymous changes at exon 1, which is consistent with positive selection in both populations (Table 7). This is also supported by the positive and significant normalized F and Tajima's for exon 1. All variable sites detected in exon 1 are frequent non-synonymous mutations (Table S3), and at least two of them present frequency shifts between samples (amino acid 7 and 20 in the leader peptide). Furthermore, we

found the same profile in all populations from the 1000Genomes project, and also in the sequences available at the IPD-IMGT/HLA database (data not shown).

## 4. Discussion

Here we present a bioinformatics approach to evaluate *HLA-C* variability when using second-generation sequencing, providing accurate genotypes and haplotypes from the upstream promoter up to the complete 3'UTR segment. This methodology relies entirely on publicly available software. The performance of our method is supported by the following perceptions. The majority of variants and sequences here detected have already been described by the IPD-IMGT/HLA database, but sequences from this database do not influence phasing whatsoever. We found at least one allele of each main *HLA-C* allele group. For the Brazilian sample, the CDS (and genomic) allele frequencies are compatible with the ones reported in previous studies using different typing techniques (Rodrigues *et al.*, 2015; González-Galarza *et al.*, 2015) (www.allelefrequencies.net). This comparison is not possible for Beninese sample since this is the first survey addressing their *HLA-C* genetic diversity. Most 3'UTR haplotypes here detected are compatible with partial or full 3'UTR sequences reported by IPD-IMGT/HLA. Moreover, 90% of the coding alleles reported here are identical to the ones called by the Optitype software (Szolek *et al.*, 2014), and the majority of the differences between methods include alleles that differ in segments not covered by the Optitype database. Unfortunately, the Optitype database available for download is outdated. Allele imputation did not exceed 0.37%, and the GATK *ReadBackedPhasing* phased directly 77.51% of the heterozygous sites.

As observed for other HLA genes such as *HLA-A* and *HLA-G* (Lima *et al.*, 2019; Castelli *et al.*, 2017), we observed a strong LD across the entire *HLA-C locus*, with a single segregation block (Figure S1) and many variants in complete LD. This LD profile supports the close relationship among coding and regulatory haplotypes. Since each allele group is related to specific regulatory sequences, different transcription factors and microRNAs may modulate their expression levels.

Since *HLA-C* has been presented as an unusual classical HLA class I *locus* because of its expression pattern, i.e. together with classical HLA presenting peptides in somatic tissues and together with nonclassical HLA in placenta (Blais, Dong and Rowland-Jones, 2011), we evaluated natural selection signatures across the *HLA-C locus* to better understand HLA-C biology and function, and also how natural selection has shaped its variability and haplotype structure. In this regard, the presence of divergent haplotypes is compatible with balancing selection, as observed across the entire *HLA-C locus* with few exceptions. Positive Tajima's *D*

values are indicative of an excess of alleles with an intermediate frequency (balancing selection) or recent population contraction (bottleneck). Negative Tajima's *D* values indicate an excess of rare alleles, which could be resulted from recent population expansion, selective sweep, weak negative selection, or sample coming from an admixed population (Stajich and Hahn, 2005).

Our results indicate an excess of heterozygosis and evidence of balancing selection not only at exons 2 and 3 as expected for an antigen presentation gene, but at others *HLA-C* segments such as part of the promoter segment, exon 4 and 5, and the second half of the *HLA-C* 3'UTR, and also in some intronic segments. It is not clear whether these results reflect a hitchhiking effect due to the high LD throughout *HLA-C* or not. One may claim that our results may be biased because of the Brazilian population demographic history. However, we detected the same pattern among Beninese. Moreover, the Ewens-Watterson test for neutrality also indicates a low homozygosity rate and supports the evidence of balancing selection. Balancing selection was described for classical HLA coding regions at exons 2 and 3, enhancing antigen presentation capabilities, and hardly any demographic or genetic factors would explain the high degree of polymorphism and excess of nonsynonymous variants at these genes (reviewed by (Meyer *et al.*, 2018)).

Here, we detected a high degree of polymorphism in all *HLA-C* segments and a remarkable excess of nonsynonymous variants in exons 1, 2, 3, and 5, with evidence of positive selection in exon 1 (Table 7). This is intriguing since exon 1 encodes the HLA-C leader peptide and it is primarily involved in targeting HLA-C to the cellular membrane. This pattern was not observed for HLA-A in the same Brazilian sample (Lima *et al.*, 2019). Both samples (Brazil and Benin) presented positive selection for the leader peptide. The leader peptide plays other roles besides addressing the molecule to the secretory pathway. After being cleaved, they may act as hormones, neurotransmitters, or even self-antigens (Hegde, 2002). Besides, leader peptides from HLA class I molecules can bind and stabilize HLA-E molecules, inducing its cell surface expression and modulating NK cell activity (Lemberg *et al.*, 2001). Interestingly, the two most frequent *HLA-C* exon 1 sequences encode the leader peptides VMAPRALLL (from 20.4% in Benin to 23.5% in Brazil) and VMAPRTLIL (from 58.4% in Benin to 64.0% in Brazil), which binds more efficiently to HLA-E*01:01 and HLA-E*01:03 molecules, respectively (Celik *et al.*, 2016). These HLA-E molecules are the two most frequently detected worldwide, including this same Brazilian and Beninese sample (Ramalho *et al.*, 2017; Sonon *et al.*, 2018). There is evidence of purifying selection acting on the *HLA-E locus*, but the dimorphism (Arg107Gly) that differentiates the two major HLA-E molecules and causes an

alteration of the HLA-E peptide repertoire (Celik *et al.*, 2016) seems to be maintained by balancing selection (Felicio *et al.*, 2014; Veiga-Castelli *et al.*, 2012). Veiga-Castelli *et al* (2012) also performed the synonymous and non-synonymous nucleotide substitution test for *HLA-E* and, although non-significant for exon 3, this segment presented different behaviors in comparison with other *HLA-E* segments, suggesting possible positive selection (Veiga-Castelli *et al.*, 2012).

HLA-E and HLA-C are co-expressed at the maternal-fetal interface, along with HLA-G and HLA-F, and they are important to pregnancy success because of their immunomodulatory role (Hackmon *et al.*, 2017). Celik *et al* (2016) showed that HLA-E*01:03 is the most protective HLA-E molecule against NK cell lysis and that the highest HLA-E-mediated protection against NK cell lysis is provided by HLA-E*01:03$^{VMAPRTLFL}$ complexes (a peptide from HLA-G) followed by HLA-E*01:03$^{VMAPRTLIL}$ (a peptide from HLA-C). However, the stabilization of HLA-E*01:03$^{VMAPRTLFL}$ complexes is least efficient. Also, the VMAPRTLIL peptide can bind to HLA-E*01:01, even less efficiently, but providing better protection against NK lysis than E*01:01 $^{VMAPRALLL}$ complexes. Thus, there is an evident biological benefit in the interaction between HLA-E molecules and VMAPRTLIL leader peptide from HLA-C at the pregnancy context, which would justify, at least in part, the increased frequency of this peptide.

We detected a high nucleotide diversity and evidence of balancing selection for exons 4 (α3 domain) and 5 (transmembrane domain) in both population samples. Since the α3 domain interacts with the CD8 co-receptor to facilitate TCR signaling, we expected conservation of this *HLA-C* segment as it was observed for *HLA-A* (Lima *et al.*, 2019). In spite of that, the HLA-C α3/CD8 interaction may be preserved since the amino acids located within residues 222-245 that are important to HLA/CD8 interaction are preserved (Salter *et al.*, 1990; Wesley *et al.*, 1993; Salter *et al.*, 1989), with only one rare variant (Glu229Gln) within this region (Table S3) that is associated with allele HLA-C*15:13. Many variants are surrounding this segment and presenting high heterozygosis, but it is uncertain whether these variants influence α3/CD8 interaction. The segment encoding the transmembrane domain presented the highest nucleotide diversity among all *HLA-C* segments since it presents frequent non-synonymous mutations and a frequent inframe indel.

HLA-C presents many conversed motifs in the α1 and α2 domains that are exclusive to *HLA-C*. This gene also presents a reduced diversity at the antigen recognition site and therefore a reduced set of self-peptides able to bind HLA-C in comparison with HLA-A and HLA-B (reviewed by (Blais, Dong and Rowland-Jones, 2011)). Moreover, the motifs responsible for

KIR interaction are also conserved. Because of that, nucleotide diversity shifts across exons 2 and 3 as shown in Figure 1. We detected this same phenomenon at other *HLA-C* segments, including the promoter and 3'UTR.

The promoter region presents a conserved segment around position -700, with a monomorphic region of 115 nucleotides from position 6:31,272,722 to 6:31,272,836 (hg38). Although Ramsuran *et al* (2017) had previously detected greater conservation in this segment studying homozygous cell lines (Ramsuran *et al.*, 2017), here we detected the same conservation in 526 individuals from two different populations. This segment is also conserved among different primates. Notwithstanding, some frequent variations at the promoter region (such as rs2395471 A/G) appears to be crucial for gene regulation, as have been shown by the different mRNA expression levels in cells that present promoter haplotypes that differ by few mutations (Ramsuran *et al.*, 2017; Vince *et al.*, 2016).

Nucleotide diversity and Tajima's D were low at the beginning of the 3'UTR segment but very high at the second half (Figure 1), in both population samples. Usually, the beginning of the 3'UTR segment is important for miRNA binding and post-transcription regulation (Gaidatzis *et al.*, 2007; Grimson *et al.*, 2007; Majoros and Ohler, 2007). This conservation contrasts with the excess of heterozygosis detected across HLA-C. The first half of the 3'UTR might have been maintained because of its critical role for *HLA-C* post-transcriptional regulation. This region might be under purifying selection. MicroRNA binding analysis indicate that the second half of the 3'UTR is not an important target for miRNAs with few exceptions, such as miR-148a (data not shown). It is well established that variants modifying the miR-148a binding site influence HLA-C expression levels (Kulkarni *et al.*, 2011; Kulkarni *et al.*, 2013). Thus, this region might be under a relaxed purifying selection, and it coincides with the presence of many frequent variants.

In general, we may find both conserved and highly polymorphic segments throughout *HLA-C*. While variants that may affect gene regulation such as the one within the miR-148a binding site or other variants within the promoter are maintained at high heterozygosis in both populations, there are highly conserved segments in the regulatory regions. This may contribute to a complex mechanism of transcriptional and post-transcriptional regulation of gene expression in tissues with different microenvironments. Thus, we suggest that HLA-C expression is finely regulated due to its immunomodulatory function, as have been proposed for the non-classical *HLA-G locus* (Castelli *et al.*, 2014a; Castelli *et al.*, 2014b). Besides, the interaction between HLA-C and T and NK cells receptors, as also with the CD8 co-receptor, is

apparently conserved considering the variability here detected in a population-based study in two samples from different continents.

*HLA-C* diversity is high in both populations considering the entire locus and each segment separately. Some haplotypes are more frequent in one population than the other, but they are usually present in both. For instance, HLA-C*16 and HLA-C*17, and their associated regulatory sequences, are frequent in Benin and not in Brazil, while the opposite is observed for HLA-C*05 and HLA-C*06. Nevertheless, they are all represented in both populations. HLA-C*04 is twice as frequent in Benin than in Brazil. Because of that, population differentiation measured by *FST* presented a low level of genetic differentiation between the studied populations (*FST*=0.0173), since both populations present the majority of the alleles, but this *FST* value was statistically significant ($P<0.0001$) due to the frequency shifts. Interestingly, when the amino acid sequence is taken into account (Table S3), no major frequency shifts are observed in important motifs for the HLA-C function and expression regulation. For instance, motif KYRV (amino acids 66,67,69, and 76), which is associated with lower cell surface HLA-C expression (Sibilio *et al.*, 2008), presents a frequency of 84% in Brazil and 87% in Benin. The dimorphism C1/C2 (amino acid 80) important for KIR interaction is also common in both populations (C2 frequency of 45.7% in Brazil and 54.2% in Benin). The Guanine deletion influencing miR-148a binding (and HIV outcome) presents a frequency of 41.15% in Brazil and 32.41% in Benin, and it is included in the U4 3'UTR haplotypes (See supplemental alignment 2 and Table 4). Variant rs2395471 at the promoter segment, in which allele Adenine influences the binding of Oct1 and is associated with higher HLA-C expression (Vince *et al.*, 2016) presents the same frequency in both populations (Table S1). The same for rs10657191 that influences the response to IFN-γ (Hundhausen *et al.*, 2012). The only exception in this scenario would be variant rs2524094, which may influence the response to TNF-α (Hundhausen *et al.*, 2012) since the variant associated with a functional TNF response is much more frequent in Benin than Brazil.

The frequency of amino acids influencing peptide binding and thus the HLA-C peptide repertoire (Kaur *et al.*, 2017) also shifts between samples, as observed for amino acid 9, 114, 116, and 156 (Table S3). Because of that, the peptide repertoire presented by these populations might be different. These populations present major amino acid frequency shifts in the transmembrane segment and also in the first alpha-1 amino acid, but it is not clear whether these modifications (which includes a large in-frame indel) modifies HLA-C function and stability.

Hence, in conclusion, here we present a molecular and bioinformatic approach to evaluate the entire *HLA-C* variability using NGS and freely available software. Then, we applied this method do survey HLA-C diversity in two geographically distinct population samples, from Brazil and Benin. The *HLA-C* promoter was very polymorphic, with the presence of few haplotypes presenting many mutational steps apart, but also presenting a monomorphic segment around position -700 that is shared among different primates. Nucleotide diversity was very high at exons 2 and 3 and higher in other exonic segments and the second half of the 3'UTR region. We detected evidence of balancing selection on the entire *HLA-C* locus (exception made to the promoter region) and positive selection in exon 1, for both populations. The frequencies of HLA-C motifs previously associated with KIR interaction and expression regulation are similar between both populations, while we detected major shifts in the frequency of amino acids that influence the peptide binding repertoire and the transmembrane region. Linkage disequilibrium along the HLA-C locus is high, with many variants in complete LD. Because of that, each allele group is associated with specific regulatory sequences, and the same pattern is observed in both populations.

**Conflicts of interest**

The authors declare that they have no conflict of interest.

**Acknowledgements**

This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP/Brazil (Grant# 2013-17084-2). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Finance Code 001.

**Table 1. List of *HLA-C* promoter sequences in two population samples from Brazil and Benin, their frequencies and the HLA-C molecules associated with them.**

| *HLA-C* promoters [a] | Brazilian Frequency (*2n*=836) | Benin Frequency (*2n*=216) | Associated HLA-C allele groups | Associated HLA-C molecules |
|---|---|---|---|---|
| P01:01 | 0.167 | 0.273 | C*04 | C*04:01, C*04:07, C*04:09N |
| P01:02 | 0.002 | - | C*04 | C*04:01 |
| P01:03 | 0.001 | - | C*04 | C*04:01 |
| P02:01 | 0.002 | - | C*01 | C*01:02 |
| P02:02 | 0.022 | 0.014 | C*01 | C*01:02 |
| P03:01 | 0.004 | - | C*14 | C*14:02 |
| P03:02 | 0.029 | - | C*14 | C*14:02, C*14:03 |
| P04:01 | 0.001 | - | C*06 | C*06:02 |
| P04:02 | 0.140 | 0.028 | C*06, C*12 | C*06:02, C*12:02, C*12:03 |
| P04:03 | 0.001 | - | C*06 | C*06:02 |
| P04:04 | 0.007 | - | C*06 | C*06:02 |
| P04:05 | 0.002 | - | C*12 | C*12:03 |
| P04:06 | 0.001 | - | C*12 | C*12:02 |
| P04:07 | 0.004 | - | C*12 | C*12:02 |
| P05:01 | 0.062 | 0.144 | C*16 | C*16:01, C*16:02, C*16:04 |
| P06:01 | 0.037 | 0.005 | C*15 | C*15:02, C*15:05, C*15:08, C*15:09, C*15:13 |
| P07:01 | 0.001 | - | C*08 | C*08:02 |
| P07:02 | 0.001 | - | C*08 | C*08:01 |
| P07:03 | 0.093 | 0.069 | C*05, C*08 | C*05:01, C*08:02, C*08:04 |
| P07:04 | 0.001 | - | C*08 | C*08:03 |
| P08:01 | 0.047 | 0.028 | C*02 | C*02:02, C*02:10, C*02:14 |
| P08:02 | 0.012 | 0.046 | C*02 | C*02:10 |
| P08:03 | - | 0.005 | C*02 | C*02:10 |
| P09:01 | 0.097 | 0.056 | C*03 | C*03:02, C*03:03, C*03:04 |
| P10:01 | 0.030 | 0.130 | C*17 | C*17:01, C*17:03, C*17:38 |
| P11:01 | 0.011 | - | C*07 | C*07:04 |
| P11:02 | 0.001 | - | C*07 | C*07:04 |
| P12:01 | 0.016 | 0.032 | C*18 | C*18:01, C*18:02 |
| P13:01 | 0.011 | - | C*07 | C*07:01 |
| P13:02 | 0.108 | 0.125 | C*07 | C*07:01, C*07:06, C*07:18, C*07:35 |
| P13:03 | 0.005 | - | C*07 | C*07:01 |
| **P13:04** | **0.083** | **0.046** | **C*07** | **C*07:02** |
| P13:05 | 0.001 | - | C*07 | C*07:02 |

[a] Please refer to the supplemental alignment 1 for the promoter sequences.

**Table 2. List of *HLA-C* CDS sequences detected in two population samples from Brazil and Benin, and their frequencies.**

| *HLA-C* CDS [a] | Brazilian Frequency (*2n*=836) | Beninese Frequency (*2n*=216) | *HLA-C* CDS [a] | Brazilian Frequency (*2n*=836) | Beninese Frequency (*2n*=216) |
|---|---|---|---|---|---|
| C*01:02:01 | 0.0239 | 0.0139 | C*08:01:01 | 0.0012 | - |
| C*02:02:02 | 0.0383 | 0.0139 | C*08:02:01 | 0.0514 | 0.0417 |
| C*02:10:01 | 0.0191 | 0.0648 | C*08:03:01 | 0.0012 | - |
| C*02:14:02 | 0.0012 | - | C*08:04:01 | 0.0012 | 0.0231 |
| C*03:02:02 | 0.0084 | 0.0324 | C*12:02:02 | 0.0084 | - |
| C*03:03:01 | 0.0431 | - | C*12:03:01 | 0.0562 | 0.0093 |
| C*03:03:04 | - | 0.0046 | C*14:02:01 | 0.0287 | - |
| C*03:04:01 | 0.0335 | - | C*14:03:01 | 0.0036 | - |
| C*03:04:02 | 0.0096 | 0.0185 | C*15:02:01 | 0.0287 | - |
| C*03:04:58 | 0.0012 | - | C*15:05:02 | 0.0024 | 0.0046 |
| C*04:01:01 | 0.1675 | 0.2731 | C*15:08:01 | 0.0012 | - |
| C*04:07:01 | 0.0012 | - | C*15:09 | 0.0012 | - |
| C*04:09N | 0.0024 | - | C*15:13:01 | 0.0036 | - |
| C*05:01:01 | 0.0419 | 0.0046 | C*16:01:01 | 0.0431 | 0.1435 |
| C*06:02:01 | 0.0909 | 0.0139 | C*16:02:01 | 0.0120 | - |
| C*06:02:03 | - | 0.0046 | C*16:04:01 | 0.0072 | - |
| C*07:01:01 | 0.0897 | 0.0602 | C*17:01:01 | 0.0239 | 0.1296 |
| C*07:01:02 | 0.0156 | - | C*17:03:01 | 0.0048 | - |
| C*07:01:09 | 0.0012 | - | C*17:38 | 0.0012 | - |
| C*07:02:01 | 0.0837 | 0.0463 | C*18:01:01 | 0.0084 | - |
| C*07:04:01 | 0.0120 | - | C*18:02:01 | 0.0072 | 0.0324 |
| C*07:06:01 | - | 0.0231 | unknown1 [b] | 0.0012 | - |
| C*07:18:01 | 0.0167 | 0.0324 | unknown2 [b] | 0.0012 | - |
| C*07:35 | - | 0.0093 | | | |

[a] *HLA-C* coding alleles according to the IPD-IMGT/HLA database version 3.36.0. The symbol (-) represents the absence of this allele.

[b] There were two rare new CDS sequences, but encoding known HLA-C protein molecules.

**Table 3. List of HLA-C encoded proteins detected in two population samples from Brazil and Benin, and their frequencies.**

| HLA-C Protein [a] | Brazilian Frequency (*2n*=836) | Beninese Frequency (*2n*=216) | HLA-C Protein [a] | Brazilian Frequency (*2n*=836) | Beninese Frequency (*2n*=216) |
|---|---|---|---|---|---|
| C*01:02 | 0.0239 | 0.0139 | C*08:02 | 0.0514 | 0.0417 |
| C*02:02 | 0.0383 | 0.0139 | C*08:03 | 0.0012 | - |
| C*02:10 | 0.0191 | 0.0648 | C*08:04 | 0.0012 | 0.0231 |
| C*02:14 | 0.0012 | - | C*12:02 | 0.0096 | - |
| C*03:02 | 0.0084 | 0.0324 | C*12:03 | 0.0562 | 0.0093 |
| C*03:03 | 0.0431 | 0.0046 | C*14:02 | 0.0287 | - |
| C*03:04 | 0.0443 | 0.0185 | C*14:03 | 0.0036 | - |
| C*03:96 | 0.0012 | - | C*15:02 | 0.0287 | - |
| C*04:01 | 0.1675 | 0.2731 | C*15:05 | 0.0024 | 0.0046 |
| C*04:07 | 0.0012 | - | C*15:08 | 0.0012 | - |
| C*04:09N | 0.0024 | - | C*15:09 | 0.0012 | - |
| C*05:01 | 0.0419 | 0.0046 | C*15:13 | 0.0036 | - |
| C*06:02 | 0.0909 | 0.0185 | C*16:01 | 0.0431 | 0.1435 |
| C*07:01 | 0.1065 | 0.0602 | C*16:02 | 0.0120 | - |
| C*07:02 | 0.0837 | 0.0463 | C*16:04 | 0.0072 | - |
| C*07:04 | 0.0120 | - | C*17:01 | 0.0239 | 0.1296 |
| C*07:06 | - | 0.0231 | C*17:03 | 0.0048 | - |
| C*07:18 | 0.0167 | 0.0324 | C*17:38 | 0.0012 | - |
| C*07:35 | - | 0.0093 | C*18:01 | 0.0084 | - |
| C*08:01 | 0.0012 | - | C*18:02 | 0.0072 | 0.0324 |

[a] *HLA-C* proteins according to the IPD-IMGT/HLA database version 3.36.0. The symbol (-) represents the absence of this sequence.

**Table 4. List of *HLA-C* 3'UTR sequences detected in two population samples from Brazil and Benin, and their frequencies.**

| *HLA-C* 3'UTR sequence [a] | Brazilian Frequency (*2n*=836) | Beninese Frequency (*2n*=216) | Associated HLA-C molecules |
|---|---|---|---|
| U01:01 | 0.075 | 0.056 | C*03:02, C*03:03, C*03:04 |
| U01:02 | 0.019 | - | C*03:04 |
| U01:03 | 0.002 | - | C*03:04 |
| U01:04 | 0.032 | - | C*14:02, C*14:03 |
| U01:05 | 0.208 | 0.315 | C*01:02, C*04:01, C*04:07, C*04:09N, C*18:01, C*18:02 |
| U01:06 | 0.001 | - | C*04:01 |
| U01:07 | 0.001 | - | C*18:01 |
| U01:08 | - | 0.005 | C*04:01 |
| U02:01 | 0.029 | 0.125 | C*17:01, C*17:03, C*17:38 |
| U02:02 | 0.001 | - | C*17:03 |
| U02:03 | - | 0.005 | C*17:01 |
| U03:01 | 0.106 | 0.069 | C*07:01, C*07:35 |
| U03:02 | 0.012 | - | C*07:04 |
| U03:03 | 0.081 | 0.046 | C*07:02 |
| U03:04 | 0.002 | - | C*07:02 |
| U03:05 | 0.017 | 0.056 | C*07:06, C*07:18 |
| U04:01 | 0.193 | 0.032 | C*06:02, C*12:02, C*12:03, C*15:02, C*15:05, C*15:08, C*15:09, C*15:13 |
| U04:02 | 0.001 | - | C*06:02 |
| U04:03 | 0.061 | 0.144 | C*16:01, C*16:02, C*16:04 |
| U04:04 | 0.001 | - | C*16:01 |
| U04:05 | 0.097 | 0.069 | C*05:01, C*08:01, C*08:02, C*08:03, C*08:04 |
| U04:06 | 0.055 | 0.074 | C*02:02, C*02:10, C*02:14 |
| U04:07 | 0.002 | - | C*02:02 |
| U04:08 | 0.001 | - | C*02:02 |
| U04:09 | - | 0.005 | C*02:10 |

[a] Please refer to the supplemental alignment 2 for the 3'UTR sequences.

**Table 5. List of *HLA-C* extended haplotypes detected in two population samples from Brazil and Benin, and their frequencies.**

| HLA-C haplotypes | | | Brazilian Frequency | Beninese Frequency |
|---|---|---|---|---|
| Promoter [a] | CDS[b] | UTR[c] | (2n=836) | (2n=216) |
| P01:01 | C*04:01:01 | U01:05 | 0.1627 | 0.2685 |
| P01:01 | C*04:01:01 | U01:06 | 0.0012 | - |
| P01:01 | C*04:01:01 | U01:08 | - | 0.0046 |
| P01:01 | C*04:07:01 | U01:05 | 0.0012 | - |
| P01:01 | C*04:09N | U01:05 | 0.0024 | - |
| P01:02 | C*04:01:01 | U01:05 | 0.0024 | - |
| P01:03 | C*04:01:01 | U01:05 | 0.0012 | - |
| P02:01 | C*01:02:01 | U01:05 | 0.0024 | - |
| P02:02 | C*01:02:01 | U01:05 | 0.0215 | 0.0139 |
| P03:01 | C*14:02:01 | U01:04 | 0.0036 | - |
| P03:02 | C*14:02:01 | U01:04 | 0.0251 | - |
| P03:02 | C*14:03:01 | U01:04 | 0.0036 | - |
| P04:01 | C*06:02:01 | U04:02 | 0.0012 | - |
| P04:02 | C*06:02:01 | U04:01 | 0.0813 | 0.0139 |
| P04:02 | C*06:02:03 | U04:01 | - | 0.0046 |
| P04:02 | C*12:02:02 | U04:01 | 0.0036 | - |
| P04:02 | unknown2 | U04:01 | 0.0012 | - |
| P04:02 | C*12:03:01 | U04:01 | 0.0538 | 0.0093 |
| P04:03 | C*06:02:01 | U04:01 | 0.0012 | - |
| P04:04 | C*06:02:01 | U04:01 | 0.0072 | - |
| P04:05 | C*12:03:01 | U04:01 | 0.0024 | - |
| P04:06 | C*12:02:02 | U04:01 | 0.0012 | - |
| P04:07 | C*12:02:02 | U04:01 | 0.0036 | - |
| P05:01 | C*16:01:01 | U04:03 | 0.0419 | 0.1435 |
| P05:01 | C*16:01:01 | U04:04 | 0.0012 | - |
| P05:01 | C*16:02:01 | U04:03 | 0.0120 | - |
| P05:01 | C*16:04:01 | U04:03 | 0.0072 | - |
| P06:01 | C*15:02:01 | U04:01 | 0.0287 | - |
| P06:01 | C*15:05:02 | U04:01 | 0.0024 | 0.0046 |
| P06:01 | C*15:08:01 | U04:01 | 0.0012 | - |
| P06:01 | C*15:09 | U04:01 | 0.0012 | - |
| P06:01 | C*15:13:01 | U04:01 | 0.0036 | - |
| P07:01 | C*08:02:01 | U04:05 | 0.0012 | - |
| P07:02 | C*08:01:01 | U04:05 | 0.0012 | - |
| P07:03 | C*05:01:01 | U04:05 | 0.0419 | 0.0046 |
| P07:03 | C*08:02:01 | U04:05 | 0.0502 | 0.0417 |
| P07:03 | C*08:04:01 | U04:05 | 0.0012 | 0.0231 |
| P07:04 | C*08:03:01 | U04:05 | 0.0012 | - |
| P08:01 | C*02:02:02 | U04:06 | 0.0347 | 0.0139 |
| P08:01 | C*02:02:02 | U04:07 | 0.0024 | - |
| P08:01 | C*02:02:02 | U04:08 | 0.0012 | - |
| P08:01 | C*02:10:01 | U04:06 | 0.0072 | 0.0139 |
| P08:01 | C*02:14:02 | U04:06 | 0.0012 | - |
| P08:02 | C*02:10:01 | U04:06 | 0.0120 | 0.0463 |
| P08:03 | C*02:10:01 | U04:09 | - | 0.0046 |
| P09:01 | C*03:02:02 | U01:01 | 0.0084 | 0.0324 |
| P09:01 | C*03:03:01 | U01:01 | 0.0431 | - |
| P09:01 | unknown1 | U01:01 | 0.0012 | - |
| P09:01 | C*03:03:04 | U01:01 | - | 0.0046 |
| P09:01 | C*03:04:02 | U01:01 | 0.0096 | 0.0185 |
| P09:01 | C*03:04:01 | U01:01 | 0.0120 | - |
| P09:01 | C*03:04:01 | U01:02 | 0.0191 | - |
| P09:01 | C*03:04:01 | U01:03 | 0.0024 | - |
| P09:01 | C*03:04:58 | U01:01 | 0.0012 | - |
| P10:01 | C*17:01:01 | U02:01 | 0.0239 | 0.1250 |
| P10:01 | C*17:01:01 | U02:03 | - | 0.0046 |
| P10:01 | C*17:03:01 | U02:01 | 0.0036 | - |
| P10:01 | C*17:03:01 | U02:02 | 0.0012 | - |
| P10:01 | C*17:38 | U02:01 | 0.0012 | - |
| P11:01 | C*07:04:01 | U03:02 | 0.0108 | - |
| P11:02 | C*07:04:01 | U03:02 | 0.0012 | - |
| P12:01 | C*18:01:01 | U01:05 | 0.0072 | - |
| P12:01 | C*18:01:01 | U01:07 | 0.0012 | - |
| P12:01 | C*18:02:01 | U01:05 | 0.0072 | 0.0324 |
| P13:01 | C*07:01:02 | U03:01 | 0.0108 | - |
| P13:02 | C*07:01:01 | U03:01 | 0.0897 | 0.0602 |
| P13:02 | C*07:01:09 | U03:01 | 0.0012 | - |
| P13:02 | C*07:06:01 | U03:05 | - | 0.0231 |
| P13:02 | C*07:18:01 | U03:05 | 0.0167 | 0.0324 |
| P13:02 | C*07:35 | U03:01 | - | 0.0093 |

| | | | | |
|---|---|---|---|---|
| P13:03 | C*07:01:02 | U03:01 | 0.0048 | - |
| P13:04 | C*07:02:01 | U03:04 | 0.0024 | - |
| P13:04 | C*07:02:01 | U03:03 | 0.0801 | 0.0463 |
| P13:05 | C*07:02:01 | U03:03 | 0.0012 | - |

[a] *HLA-C* promoter sequences. Please refer to the supplemental alignment 1 for the promoter sequences and Table 1 for their frequencies.
[b] *HLA-C* CDS sequences. Please refer to Table 2 for their frequencies.
[c] *HLA-C* 3'UTR sequences. Please refer to the supplemental alignment 2 for the promoter sequences and Table 4 for their frequencies.

**Table 6. Nucleotide diversity and Neutrality tests across the *HLA-C* locus.**

| Region | Length | Brazilian Samples | | | Beninese Samples | | |
|---|---|---|---|---|---|---|---|
| | | Nucleotide diversity | Tajima's D[1] | Ewens- Watterson[2] | Nucleotide diversity | Tajima's D[1] | Ewens- Watterson[2] |
| Promoter and 5'UTR | 1525 | 0.0132±0.0064 | 1.1752, P=0.8950 | Fo= 0.0945, Fe= 0.1316, F= -0.8032, P= 0.1778 | 0.0137±0.0067 | 1.0439, P=0.8876 | Fo= 0.1427, Fe= 0.2325, F= -1.0615, P= 0.0713 |
| Exon 1 | 73 | 0.0240±0.0154 | 1.5110, P=0.9242 | **Fo=0.2750, Fe=0.5434, F= -1.4445, P= 0.0296** | 0.0306±0.0187 | **1.8870, P=0.9608** | **Fo= 0.2345, Fe= 0.4755, F= -1.4739, P=0.0135** |
| Intron 1 | 130 | 0.0193±0.0115 | 1.1091, P=0.8890 | **Fo= 0.1304, Fe= 0.3594, F= -1.6580, P= 0.0001*** ** | 0.0186±0.0112 | 1.0223, P=0.8610 | Fo= 0.1960, Fe= 0.3482, F= -1.1953, P= 0.0505 |
| Exon 2 | 270 | 0.0259±0.0135 | **3.1783, P=0.9970** | **Fo= 0.0860, Fe= 0.2138, F= -1.5571, P= 0.0001*** ** | 0.0252±0.0132 | **2.2558, P=0.9836** | Fo= 0.1457, Fe= 0.2325, F= -1.0256, P= 0.0856 |
| Intron 2 | 250 | - | - | - | - | - | - |
| Exon 3 | 276 | 0.0288±0.0148 | **1.9489, P=0.9678** | **Fo= 0.0921, Fe= 0.1637, F= -1.1845, P= 0.0285** | 0.0292±0.0150 | **1.6789, P=0.9542** | Fo= 0.1557, Fe= 0.2025, F= -0.6480, P= 0.2742 |
| Intron 3 | 587 | 0.0145±0.0074 | 1.6654, P=0.9488 | **Fo= 0.1061, Fe= 0.2353 , F= -1.4116, P= 0.0037**** | 0.0158±0.0080 | 1.4061, P=0.9326 | Fo= 0.1480, Fe= 0.2325, F= -0.9987, P= 0.0963 |
| Exon 4 | 276 | 0.0212±0.0112 | **2.0722, P=0.9746** | **Fo= 0.1613, Fe= 0.3142, F= -1.2411, P= 0.0314** | 0.0229±0.0120 | **2.1107, P=0.9784** | **Fo= 0.1593, Fe= 0.2921, F= -1.2379, P= 0.0308** |
| Intron 4 | 124 | 0.0456±0.0242 | 1.1960, P=0.8914 | **Fo= 0.2184, Fe= 0.4175, F= -1.2697, P= 0.0423** | 0.0498±0.0263 | 0.7793, P=0.8248 | Fo= 0.2023, Fe= 0.3482, F= -1.1458, P= 0.0648 |
| Exon 5 | 120 | 0.0411±0.0218 | **1.8281, P=0.9614** | **Fo= 0.1518, Fe= 0.3873, F= -1.5870, P= 0.0008**** | 0.0685±0.0349 | **1.8032, P=0.9664** | **Fo= 0.1569, Fe= 0.3482, F= -1.5019, P= 0.0032**** |
| Intron 5 | 440 | 0.0169±0.0088 | **2.2718, P=0.9836** | **Fo= 0.1527, Fe= 0.2956, F= -1.2378, P= 0.0282** | 0.0158±0.0082 | 1.3146, P=0.9094 | Fo= 0.1831, Fe= 0.3188, F= -1.1532, P= 0.0579 |
| Exon 6 | 33 | 0.0011±0.0034 | **-1.2183, P=0.0440** | Fo= 0.9624, Fe= 0.6725, F= 1.4607, P= 0.9329 | 0.0032±0.0059 | -0.3723, P=0.2864 | Fo= 0.8951, Fe= 0.8330, F= 0.3713, P= 0.4691 |
| Intron 6 | 107 | 0.0132±0.0089 | 0.8581, P=0.8302 | Fo= 0.4147, Fe= 0.4964, F= -0.4589, P= 0.3918 | 0.0139±0.0093 | 0.9534, P=0.8502 | Fo= 0.2960, Fe= 0.4755, F= -1.0982, P= 0.1091 |
| Exon 7 | 48 | 0.0144±0.0121 | 1.7087, P=0.9350 | Fo= 0.6524, Fe= 0.6725, F= -0.1012, P= 0.4884 | 0.0119±0.0107 | 0.9407, P=0.8502 | Fo= 0.7161, Fe= 0.8330, F= -0.6997, P= 0.2624 |
| Intron 7 | 164 | 0.0164±0.0097 | 1.4725, P=0.9322 | Fo= 0.2281, Fe= 0.3873, F= -1.0730, P= 0.0966 | 0.0152±0.0091 | 0.9087, P=0.8418 | Fo= 0.3312, Fe= 0.3834, F= -0.3754, P= 0.4331 |
| Exon8/3'UTR | 425 | 0.0232±0.0118 | **1.8501, P=0.9626** | Fo= 0.1229, Fe= 0.1942,F= -0.9666, P= 0.1032 | 0.0253±0.0128 | **1.7389, P=0.9590** | Fo= 0.1599, Fe= 0.2498, F= -0.9781, P= 0.1101 |
| CDS | 1101 | 0.0259±0.0125 | **2.4623, P=0.9886** | Fo= 0.0697, Fe= 0.0946,F= -0.8383, P= 0.1626 | 0.0300±0.0145 | **2.2109, P=0.9882** | Fo= 0.1291, Fe= 0.1431, F= -0.2942, P= 0.4689 |

Significant P-values are marked in boldface. [1] Tajima's D values computed by Arlequin 3.5 software. *D* statistic tested by generating random samples under the hypothesis of selective neutrality and population equilibrium, and the ***P*** was obtained as the proportion of simulated *D* statistics less than or equal to the D observed (No. of simulation = 5.000). [2] Ewens-Watterson's F normalized values computed by Pypop software (numReplicates=20.000). [3] Intron 2 was not evaluated in this study. ** significant at the 1% level and *** significant at the 0.1% level.

**Table 7. dN/dS ratio test of the *HLA-C* exons and CDS.**

| Region | Size | Codons | Brazilian sample | | | Beninese sample | | |
|---|---|---|---|---|---|---|---|---|
| | | | *Ha = Neutrality* *(dn ≠ ds)* | *Ha* = Positive Selection (dn > ds) | *Ha* = Purifying Selection (dn < ds) | *Ha = Neutrality* *(dn ≠ ds)* | *Ha* = Positive Selection (dn > ds) | *Ha* = Purifying Selection (dn < ds) |
| Exon 1 | 72 | 24 | **dn–ds = 2.031. P= 0.044** | **dn–ds = 2.034. P= 0.022** | ds–dn = -2.051. P= 1.000 | **dn–ds = 2.387. P= 0.019** | **dn–ds = 2.378 . P= 0.009** | ds–dn = -2.363. P= 1.000 |
| Exon 2 | 270 | 90 | dn–ds = 0.382. P= 0.703 | dn–ds = 0.381. P= 0.352 | ds–dn = -0.376. P= 1.000 | dn–ds = 0.537. P= 0.592 | dn–ds = 0.554. P= 0.290 | ds–dn = -0.541. P= 1.000 |
| Exon 3 | 276 | 92 | dn–ds = 0.113. P= 0.910 | dn–ds = 0.118. P= 0.453 | ds–dn = -0.115. P= 1.000 | dn–ds = -0.028. P= 0.978 | dn–ds = -0.028. P= 1.000 | ds–dn = 0.028. P= 0.489 |
| Exon 4 | 276 | 92 | dn–ds = -1.403. P=0.163 | dn–ds = -1.368. P= 1.000 | ds–dn = 1.402. P= 0.082 | dn–ds = -1.135. P= 0.259 | dn–ds = -1.140. P= 1.000 | ds–dn = 1.159. P= 0.124 |
| Exon 5 | 120\|\|138 | 40\|\|46 | dn–ds = -0.021. P= 0.983 | dn–ds = -0.022. P=1.000 | ds–dn = 0.022. P= 0.491 | dn–ds = 0.340. P= 0.735 | dn–ds = 0.339. P= 0.368 | ds–dn = -0.339. P= 1.000 |
| Exon 6 | 33 | 11 | dn–ds = 0.649. P= 0.518 | dn–ds = 0.631. P= 0.265 | ds–dn = -0.625. P= 1.000 | dn–ds = 1.025. P= 0.307 | dn–ds = 1.037. P= 0.151 | ds–dn = -1.039. P= 1.000 |
| Exon 7 | 48 | 16 | dn–ds = 1.505. P= 0.135 | dn–ds = 1.472. P=0.072 | ds–dn = -1.506. P= 1.000 | dn–ds = 1.508. P= 0.134 | dn–ds = 1.504. P= 0.068 | ds–dn = -1.489. P= 1.000 |
| CDS | 1101\|\|1119 | 367\|\|373 | dn–ds = 0.016. P= 0.987 | dn–ds = 0.016. P= 0.494 | ds–dn = -0.016. P= 1.000 | dn–ds = 0.409. P= 0.683 | dn–ds = 0.409. P= 0.342 | ds–dn = -0.410. P= 1.000 |

dN/dS ratio test was performed using all sequences found for each segment of the *HLA-C* locus, using bootstrap method (5,000 replicates). Significant P-values are marked in boldface.

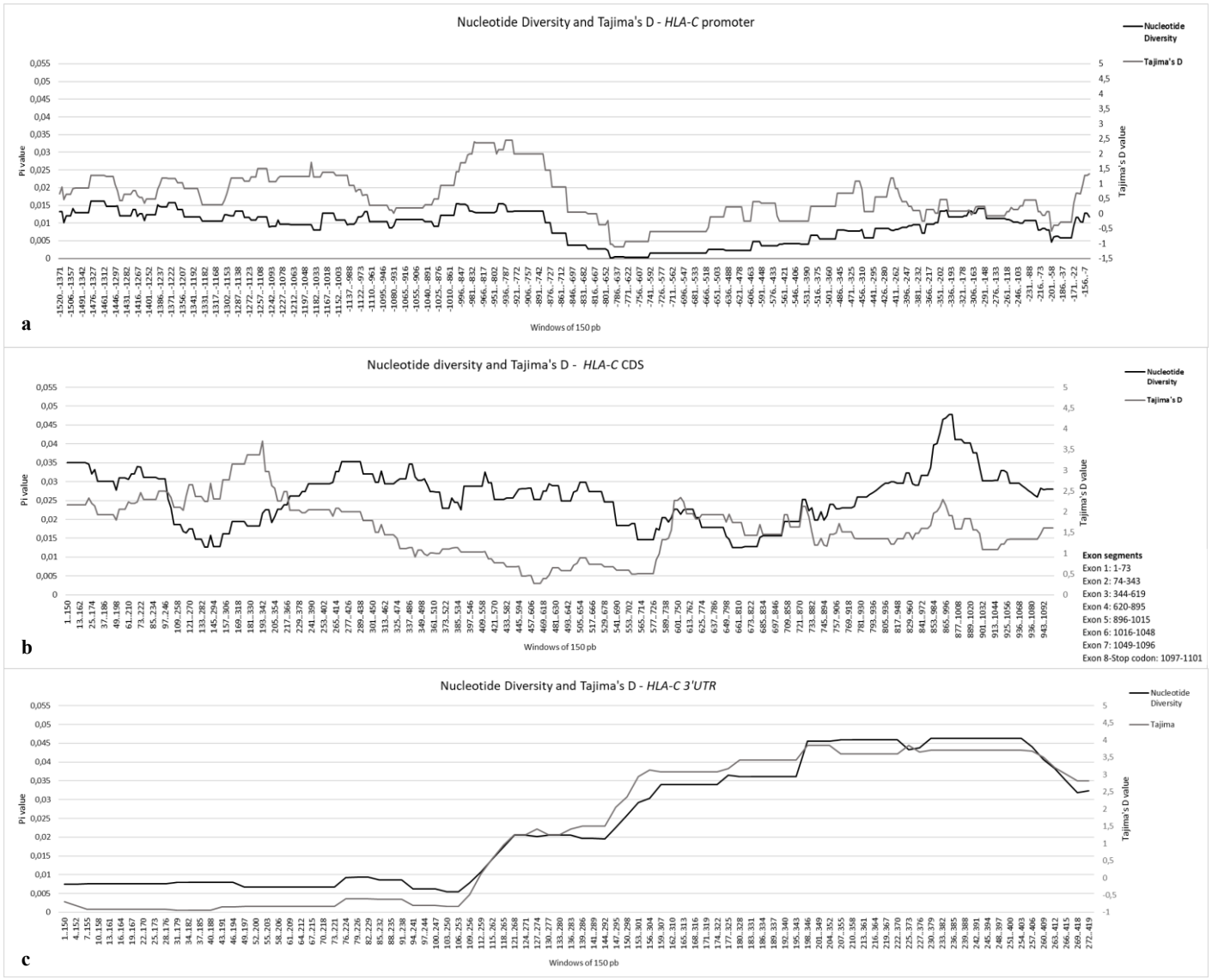
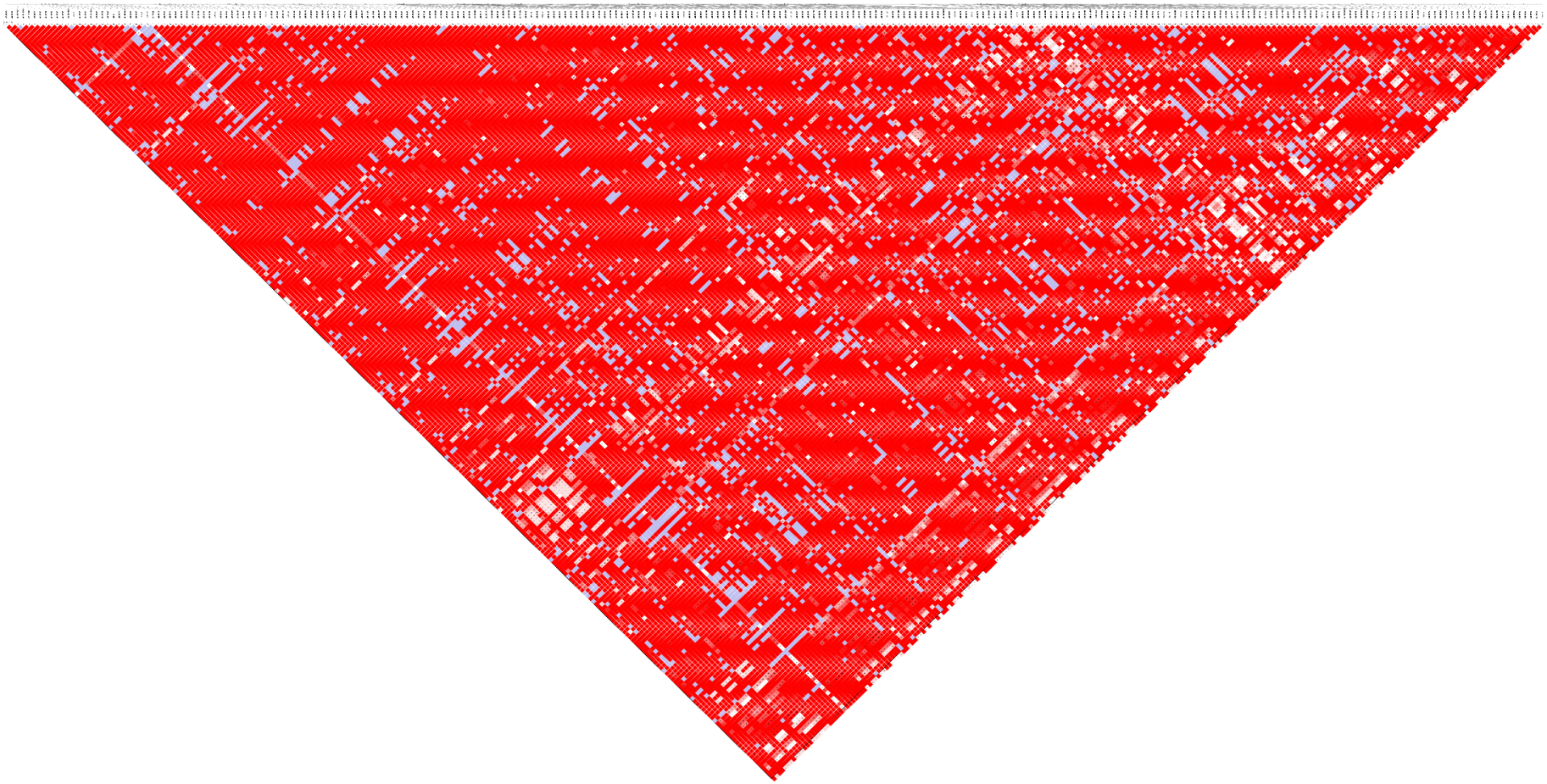



**Figure 1. Nucleotide diversity and Tajima's D at the *HLA-C* promoter, CDS and 3'UTR.** (**a**) Nucleotide diversity and Tajima's D of *HLA-C* 5' upstream regulatory segment examined in all aligned sequences using sliding window of 150bp with a step size of 3. The x-axis represents the windows of 150pb comprehending the *HLA-C* segments and the nucleotide position relative to the alignment using as reference the position -1 from IMGT database (genomic data); (**b**) Nucleotide diversity and Tajima's D of *HLA-C* exons. The position 1(from window 1..150) in the x-axis corresponds to first base from exon 1, the subsequent positions was given by alignment position (see caption of Exon segments); and (**c**) Nucleotide diversity of *HLA-C 3'UTR*.

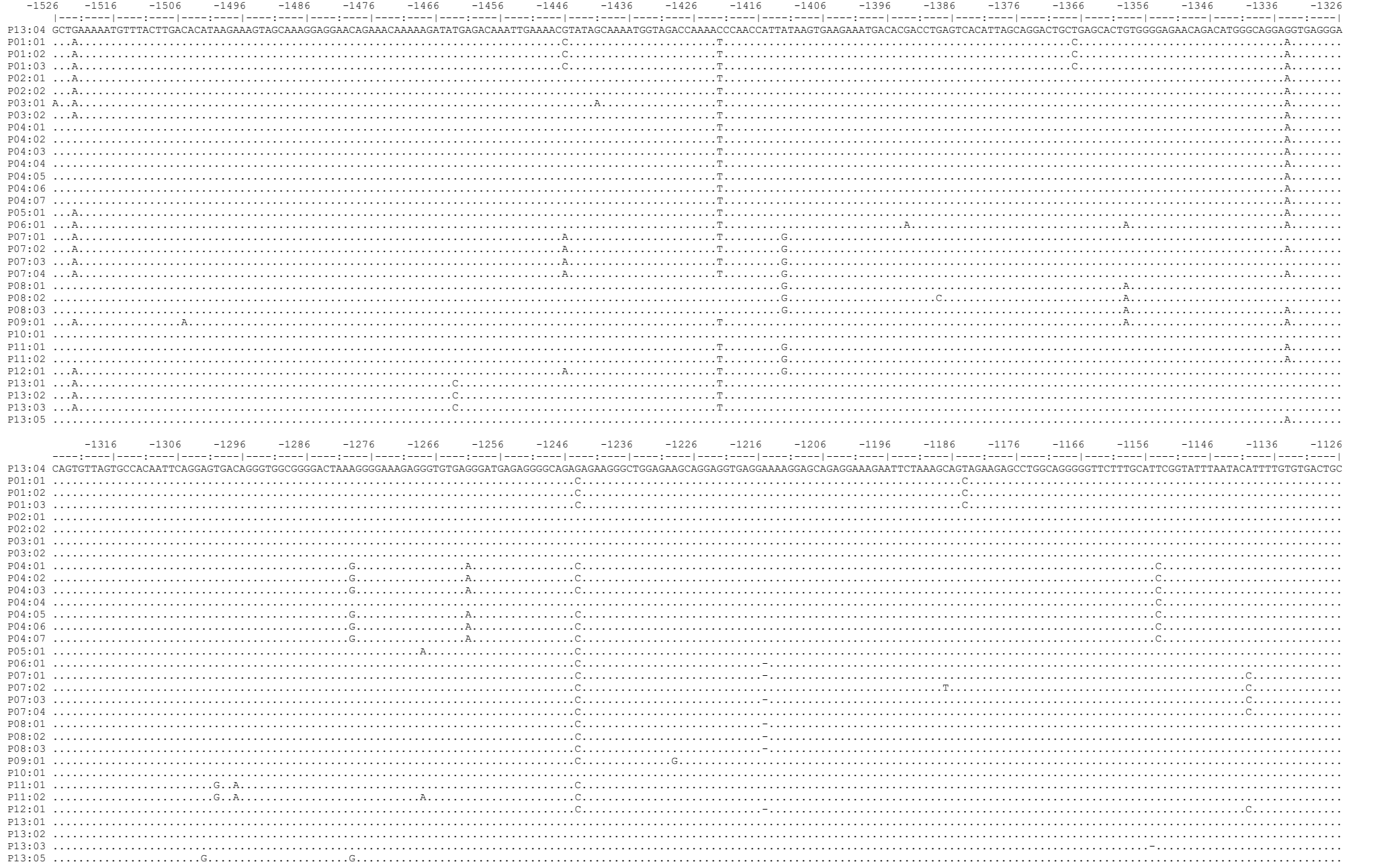
# *Supplemental Material*

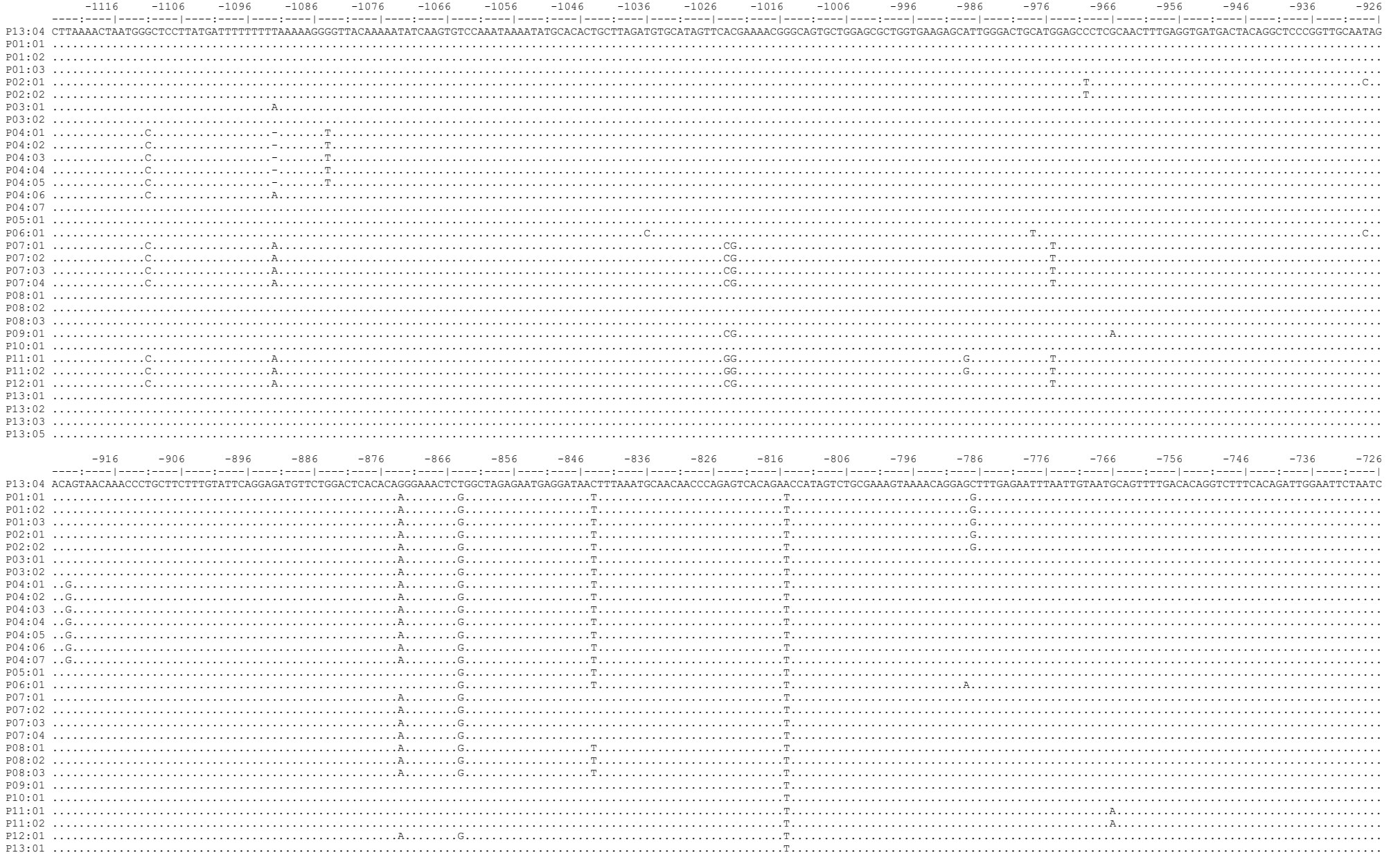


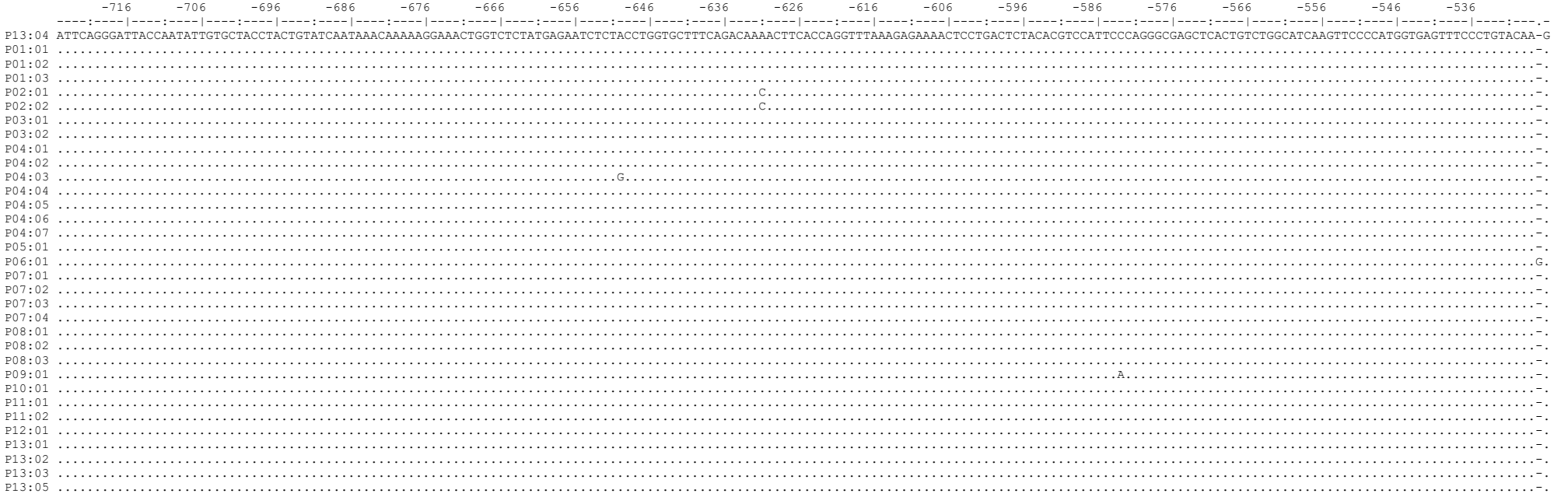
**Supplemental Figure 1. *Linkage Disequilibrium* (LD) between pair of single nucleotide polymorphisms (SNPs) of the *HLA-C* locus, considering genomic positions from 3'UTR to Promoter (31268757- 31273596, from hg38).** LD plot generated by Haploview 4.2, considering variable sites with a minor allele a minor allele frequency (MAF) ≥ 1% and the fraction of strong LD in informative comparison set to 0.9. Areas in red color indicate strong LD (LOD score ≥ 2, D' score =1); areas in light red or pink indicate moderate LD (LOD ≥ 2, D'<1); areas in blue indicate weak LD (LOD<2, D'=1); and areas in very weak LD LD (LOD<2, D'<1); LOD: Log of the odds – measure of confidence in the value of D'.

**Supplemental alignment 1. *HLA-C* promoter sequences in two population samples (Brazil and Benin).**

P13:04 is identical to the hg38 reference sequence (inverted and complemented), and it is used as reference for this alignment. The first translated ATG is the three last nucleotides. Please refer to Table 1 for the associated *HLA-C* coding alleles and the frequency of each of these *HLA-C* promoters.

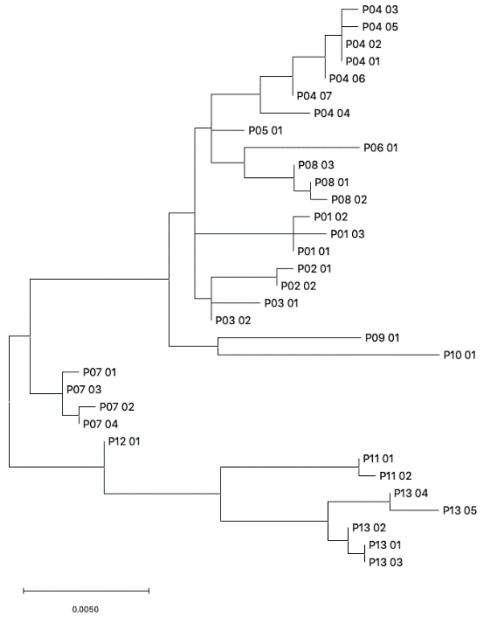



-1116 -1106 -1096 -1086 -1076 -1066 -1056 -1046 -1036 -1026 -1016 -1006 -996 -986 -976 -966 -956 -946 -936 -926
P13:04 CTTAAAACTAATGGGCTCCTTATGATTTTTTTTTAAAAAGGGGTTACAAAAATATCAAGTGTCCAAATAAAATATGCACACTGCTTAGATGTGCATAGTTCACGAAAACGGGCAGTGCTGGAGCGCTGGTGAAGAGCATTGGGACTGCATGGAGCCCTCGCAACTTTGAGGTGATGACTACAGGCTCCCGGTTGCAATAG
P01:01
P01:02
P01:03
P02:01
P02:02
P03:01
P03:02
P04:01
P04:02
P04:03
P04:04
P04:05
P04:06
P04:07
P05:01
P06:01
P07:01
P07:02
P07:03
P07:04
P08:01
P08:02
P08:03
P09:01
P10:01
P11:01
P11:02
P12:01
P13:01
P13:02
P13:03
P13:05
-916 -906 -896 -886 -876 -866 -856 -846 -836 -826 -816 -806 -796 -786 -776 -766 -756 -746 -736 -726
P13:04 ACAGTAACAAACCCTGCTTCTTTGTATTCAGGAGATGTTCTGGACTCACACAGGGAAACTCTGGCTAGAGAATGAGGATAACTTTAAATGCAACAACCCAGAGTCACAGAACCATAGTCTGCGAAAGTAAAACAGGAGCTTTGAGAATTTAATTGTAATGCAGTTTTGACACAGGTCTTTCACAGATTGGAATTCTAATC
P01:01
P01:02
P01:03
P02:01
P02:02
P03:01
P03:02
P04:01
P04:02
P04:03
P04:04
P04:05
P04:06
P04:07
P05:01
P06:01
P07:01
P07:02
P07:03
P07:04
P08:01
P08:02
P08:03
P09:01
P10:01
P11:01
P11:02
P12:01
P13:01
P13:02
P13:03
P13:05



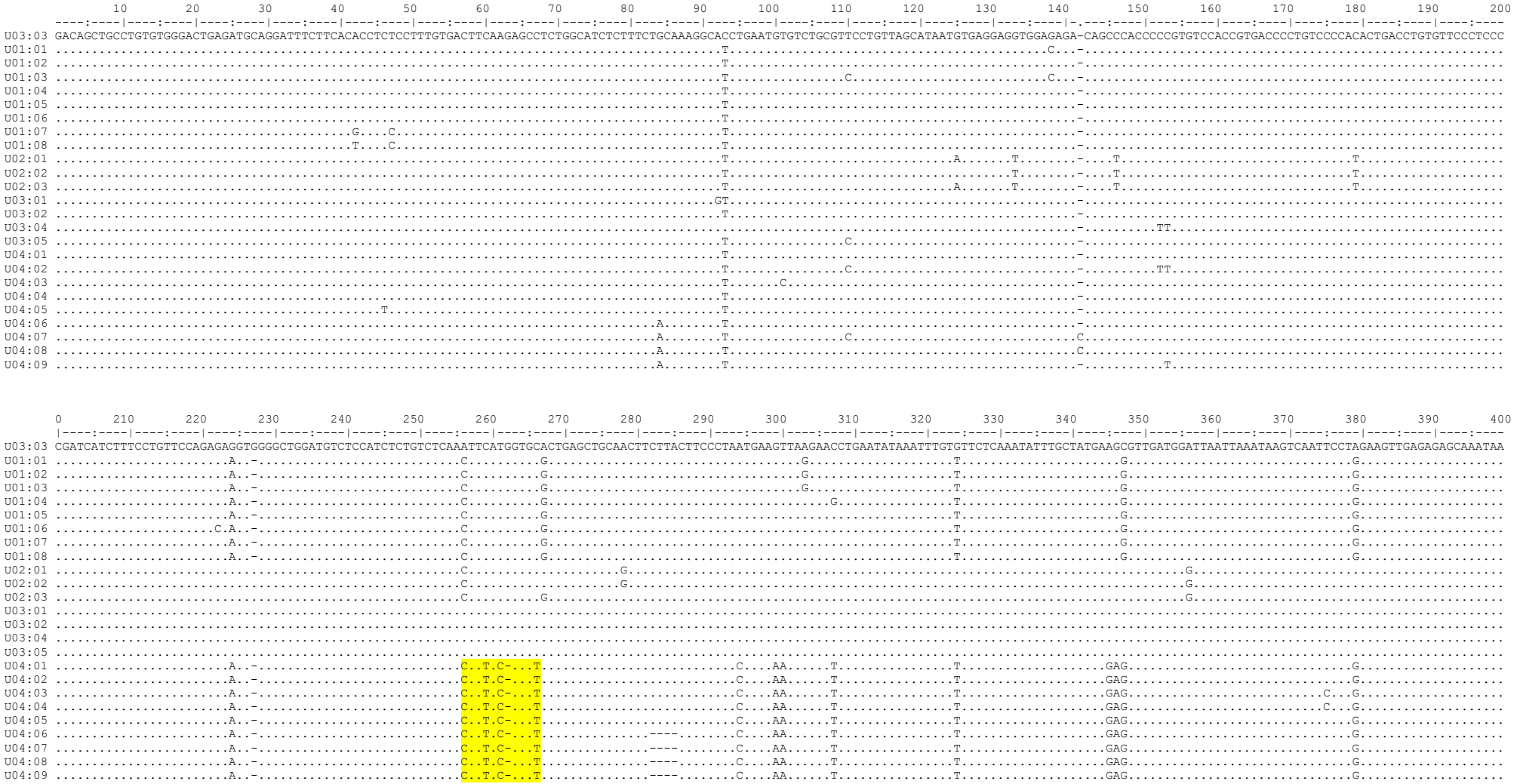
```
    -326      -316      -306      -296      -286      -276      -266      -256      -246      -236      -226      -216      -206      -196      -186      -176      -166      -156      -146      -136
       |----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|
P13:04 ACTCAGGCACACAGTGTGACAAAGATGCTTGGTGTAGGAGAAGAGGGATCAGGACGAAGTCCCAGGTCCCGGGCGGGGCTCTCAGGGTCTCAGGCTCCAAGGGCCGTGTCTGCATTGGGGAGGCGCCGCGTTGGGGATTCTCCACTCCC------CTGAGTTTCACTTCT---CCCAACCTGCGTCGGGTCCTTCTTCCT
P01:01 .....................................................A.....................................................C.....................------.................TCT.....................
P01:02 .....................................................A.....................................................C.....................------.................TCT.....................
P01:03 ...........................C.........A.....................................................C.....................------.................TCT.....................
P02:01 ..........................................................T..................C.....................A.....------.................TCT.....................
P02:02 ..........................................................T..................C.....................A.....------.................TCT.....................
P03:01 ..........................................................................C.....................A.....------.................TCT.....................
P03:02 ..........................................................................C.....................A.....------.................TCT.....................
P04:01 ..........................................................................C.....................A.....------.................TCT.....................
P04:02 ..........................................................................C.....................A.....------.................TCT.....................
P04:03 ..........................................................................C.....................A.....------.................TCT.....................
P04:04 ..........................................................................C.....................A.....------.................TCT.....................
P04:05 ..........................................................................C.....................A.....------.................TCT..T..................
P04:06 ..........................................................................C.....................A.....------.................TCT.....................
P04:07 ..........................................................................C.....................A.....------.................TCT.....................
P05:01 ..........................................................................C.....................A.....------.................TCT.....................
P06:01 ..........................A...............................................C.....................A.....------.................TCT..........A..........
P07:01 ..........................................................................C.....................A.....------.................TCTT....................
P07:02 ..........................................................................C.....................A.....------.................TCTT....................
P07:03 ..........................................................................C.....................A.....------.................TCTT....................
P07:04 ..........................................................................C.....................A.....------.................TCTT....................
P08:01 ..........................................................................C.....................A.....------.................TCT..........A..........
P08:02 ..........................................................................C.....................A.....------.................TCT..........A..........
P08:03 ..........................................................................C.....................A.....------.................TCT..........A..........
P09:01 .....................................................A.....................................C.....................A.....------.................TCT.....................
P10:01 ...........................................................T.......A....C.....................................A.............C........ACTCCC.............TCT.....................
P11:01 ..........................................................................C.....................A.....------.................---.....................
P11:02 ..........................................................................C.....................A.....------.................---.....................
P12:01 ..........................................................................................................................------.................---.....................
P13:01 ...................................................................A......................................................------.................---.....................
P13:02 ..........................................................................................................................------.................---.....................
P13:03 ...................................................................A......................................................------.................---.....................
P13:05 ..........................................................................................................................------.................---.....................
```

```
       -126      -116      -106       -96       -86       -76       -66       -56       -46       -36       -26       -16        -6   -1
       ----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:---
P13:04 GAATACTCATGACGCGTCCCCAATTCCCACTCCCATTGGGTGTCGGGTTCTAGAGAAGCCAATCAGCGTCTCCGCAGTCCCGGTTCTAAAGTCCCCAGTCACCCACCCGGACTCACATTCTCCCCAGAGGCCGAG|ATG|
P01:01 ..........................................................................................................G.....C.....
P01:02 ..........................................................................................................G.....C.....
P01:03 ..........................................................................................................G.....C.....
P02:01 ..........................................................................................................G.....C.....
P02:02 ..........................................................................................................G.....C.....
P03:01 ..........................................................................................................G.....C.....
P03:02 ..........................................................................................................G.....C.....
P04:01 ..........................................................................................................GG....C.....
P04:02 ..........................................................................................................GG....C.....
P04:03 ..........................................................................................................GG....C.....
P04:04 ..........................................................................................................GG....C.....
P04:05 ..........................................................................................................GG....C.....
P04:06 ..........................................................................................................GG....C.....
P04:07 ..........................................................................................................GG....C.....
P05:01 ..........................................................................................................GG....C.....
P06:01 ..........................................................................................................GG....C.....
P07:01 ..............................T...........................................................GG....C.....
P07:02 ..........................................................................................................GG....C.....
P07:03 ..........................................................................................................GG....C.....
P07:04 ..........................................................................................................GG....C.....
P08:01 ..........................................................................................................GG....C.....
P08:02 ..........................................................................................................GG....C.....
P08:03 ..........................................................................................................GG....C.....
P09:01 .....................................................A.....................G.....G.....C.....
P10:01 ...................................................A.....G....-------.....G.....C.....
P11:01 ..........................................................................................................
P11:02 ..........................................................................................................
P12:01 ..........................................................................................................
P13:01 ..........................................................................................................
P13:02 ..........................................................................................................
P13:03 ..........................................................................................................
P13:05 ..........................................................................................................
```



**Supplemental alignment 2. *HLA-C* 3′UTR sequences in two population samples (Brazil and Benin).**

**U03:03 is identical to the hg38 reference sequence (inverted and complemented), and it is used as reference for this alignment. The first base corresponds to the first nucleotide after the stop codon. Please refer to Table 4 for the associated *HLA-C* coding alleles and the frequency of each of these HLA-C 3′UTR sequences. The yellow region indicates the region failing to bind to *miRNA148a* according to (Kulkarni et al., 2011)**



```
       0        410        420
       |----:----|----:----|
U03:03 AGACCTGAGAACCTTCCAGAA
U01:01 ............T........
U01:02 ............T........
U01:03 ............T........
U01:04 .....................
U01:05 .....................
U01:06 .....................
U01:07 .....................
U01:08 .....................
U02:01 .....................
U02:02 .....................
U02:03 .....................
U03:01 .....................
U03:02 .....................
U03:04 .....................
U03:05 .....................
U04:01 .....................
U04:02 .....................
U04:03 .....................
U04:04 .....................
U04:05 .....................
U04:06 .....................
U04:07 .....................
U04:08 .....................
U04:09 .....................
```

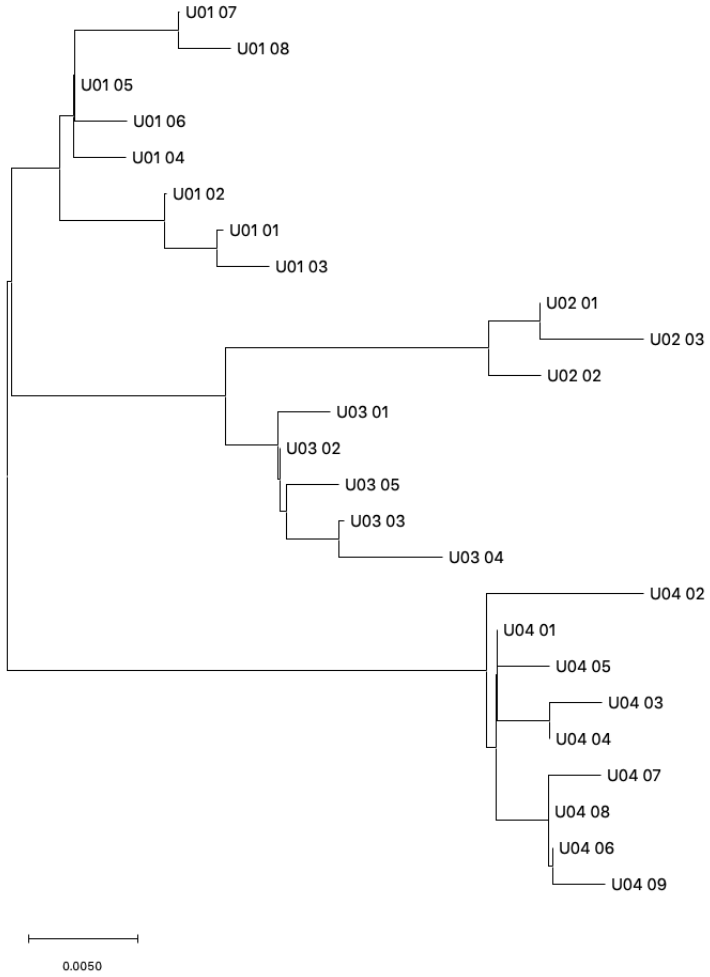


**Table S1. List of all variants observed across *HLA-C* and the reference allele frequency in Brazil and Benin, as genotyped by the GATK HaplotypeCaller.**

| Chr6/hg38 Position | SNPid | HLA-C segment | Reference allele | Alternative alleles | Brazilian Sample Reference allele frequency (2n=836) | Beninese Sample Reference allele frequency (2n=216) | Notes | IPD-IMGT/HLA Positions | IMGT variants |
|---|---|---|---|---|---|---|---|---|---|
| 31268757 | rs35075694 | 3'UTR | G | A | 0,9031 | 0,9444 | present on IMGT | 3306 | C>T |
| 31268790 | rs1049281 | 3'UTR | T | C | 0,2488 | 0,3009 | present on IMGT | 3273 | G>A |
| 31268794 | rs114027487 | 3'UTR | A | G | 0,9378 | 0,8565 | present on IMGT | 3269 | T>C |
| 31268813 | rs115510686 | 3'UTR | T | C | 0,9701 | 0,8704 | present on IMGT | 3250 | A>G |
| 31268821 | rs67157575 | 3'UTR | CG | C | 0,5885 | 0,6759 | present on IMGT - variable sites from 3239 to 3241 | 3241 | G>C |
| 31268822 | rs3189472 | 3'UTR | G | C,* | 0,2488 | 0,3009 | | 3240 | G>A |
| 31268824 | rs9281298 | 3'UTR | T | TC | 0,5885 | 0,6759 | | 3239 | A>G |
| 31268845 | rs1130538 | 3'UTR | C | A | 0,2488 | 0,3009 | present on IMGT | 3218 | T>G |
| 31268862 | rs1071643 | 3'UTR | G | C,A | 0,5562 | 0,6759 | present on IMGT | 3201 | C>T>G |
| 31268866 | rs1130552 | 3'UTR | T | C | 0,9031 | 0,9444 | present on IMGT | 3197 | A>G |
| 31268869 | rs1130554 | 3'UTR | A | T | 0,5885 | 0,6759 | present on IMGT | 3194 | T>A |
| 31268870 | rs1130558 | 3'UTR | C | T | 0,5885 | 0,6759 | present on IMGT | 3193 | G>A |
| 31268875 | rs1130559 | 3'UTR | T | G | 0,5885 | 0,6759 | present on IMGT | 3188 | A>C |
| 31268880 | rs369987786 | 3'UTR | GAAGT | G | 0,9414 | 0,9213 | present on IMGT | 3179 | delACTT |
| 31268891 | rs1130576 | 3'UTR | T | C | 0,9701 | 0,8750 | present on IMGT | 3172 | A>G |
| 31268902 | rs1130580 | 3'UTR | T | C | 0,6603 | 0,6204 | present on IMGT - variable sites from 3153 to 3161 | 3161 | G>A |
| 31268903 | rs1130586 | 3'UTR | G | A | 0,5885 | 0,6759 | | 3160* | **C>T** |
| 31268906 | rs879642168 | 3'UTR | CCAT | C | 0,5885 | 0,6759 | | 3155*; 3157* | **T>C; delG** |
| 31268910 | rs3207555 | 3'UTR | G | GTA | 0,5885 | 0,6759 | | 3153* | **C>T** |
| 31268913 | rs1130592 | 3'UTR | T | G | 0,2189 | 0,1713 | present on IMGT | 3150 | C>A |
| 31268938 | rs9279068 | 3'UTR | GC | G | 0,2488 | 0,3009 | present on IMGT | 3125 | insG |
| 31268945 | rs1094 | 3'UTR | C | T | 0,2488 | 0,3009 | present on IMGT | 3119 | G>A |
| 31268947 | rs529358989 | 3'UTR | C | G | 0,9988 | 1,0000 | alternative allele is not present on IMGT | 3117 | - |
| 31268990 | rs1049579 | 3'UTR | G | A | 0,9701 | 0,8704 | present on IMGT | 3074 | C>T |
| 31269016 | . | 3'UTR | G | A | 0,9964 | 0,9954 | present on IMGT | 3048 | C>T |
| 31269017 | . | 3'UTR | G | A | 0,9964 | 1,0000 | alternative allele is not present on IMGT | 3047 | - |
| 31269023 | rs116229144 | 3'UTR | G | A | 0,9701 | 0,8704 | present on IMGT | 3041 | C>T |
| 31269026 | rs201629107 | 3'UTR | T | TG | 0,9964 | 1,0000 | present on IMGT | 3038 | insC |
| 31269031 | rs1049650 | 3'UTR | C | G | 0,9222 | 0,9444 | present on IMGT | 3033 | G>C |
| 31269036 | rs1049663 | 3'UTR | C | A | 0,9701 | 0,8704 | present on IMGT | 3028 | G>T |
| 31269044 | rs1049668 | 3'UTR | C | T | 0,9713 | 0,8704 | present on IMGT | 3020 | G>A |
| 31269059 | rs41289069 | 3'UTR | A | G | 0,9773 | 0,9444 | present on IMGT | 3005 | T>C |
| 31269068 | rs3176007 | 3'UTR | A | G | 0,9390 | 0,8565 | present on IMGT | 2996 | T>C |
| 31269076 | rs1065711 | 3'UTR | G | A | 0,0837 | 0,0463 | present on IMGT | 2988 | T>C |
| 31269077 | rs1049709 | 3'UTR | T | C | 0,8935 | 0,9306 | present on IMGT | 2987 | A>G |
| 31269085 | rs1049724 | 3'UTR | C | T | 0,9414 | 0,9213 | present on IMGT | 2979 | G>A |
| 31269122 | rs193100294 | 3'UTR | A | G | 0,9988 | 0,9954 | present on IMGT | 2942 | T>C |
| 31269123 | rs1049853 | 3'UTR | G | A | 0,9031 | 0,9306 | present on IMGT | 2941 | C>T |
| 31269127 | rs183137172 | 3'UTR | T | A,C | 0,9988 | 0,9954 | second alternative is not present on IMGT | 2937 | A>T |
| 31269182 | rs192019825 | Intron 7 | C | A | 0,9928 | 1,0000 | present on IMGT | 2882 | G>T |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 31269221 | rs2001181 | Intron 7 | C | T | 0,0837 | 0,0463 | present on IMGT | 2843 | A>G |
| 31269225 | rs17885436 | Intron 7 | G | C | 0,9031 | 0,9444 | present on IMGT | 2839 | C>G |
| 31269237 | rs3998381 | Intron 7 | G | A | 0,2189 | 0,1713 | present on IMGT | 2827 | T>C |
| 31269239 | rs41544314 | Intron 7 | C | T | 0,9701 | 0,8704 | present on IMGT | 2825 | G>A |
| 31269240 | rs2394960 | Intron 7 | C | T | 0,2488 | 0,3009 | present on IMGT | 2824 | A>G |
| 31269250 | rs2394961 | Intron 7 | C | T | 0,2488 | 0,3009 | present on IMGT | 2814 | A>G |
| 31269263 | rs41555414 | Intron 7 | A | T | 0,9617 | 0,9861 | present on IMGT | 2801 | T>A |
| 31269271 | rs9264594 | Intron 7 | A | G | 0,2189 | 0,1713 | present on IMGT | 2793 | C>T |
| 31269284 | rs2894204 | Intron 7 | C | T | 0,2811 | 0,3009 | present on IMGT | 2780 | A>G |
| 31269310 | rs17885557 | Intron 7 | C | A | 0,8433 | 0,9722 | present on IMGT | 2754 | G>T |
| 31269323 | rs41548117 | Intron 7 | TC | T | 0,9414 | 0,9213 | present on IMGT | 2740 | delG |
| 31269338 | rs41544716 | Exon 7 | CT | C | 0,9976 | 1,0000 | present on IMGT | 2725 | delA |
| 31269347 | rs1130838 | Exon 7 | T | C | 0,2201 | 0,1713 | present on IMGT | 2717 | G>A |
| 31269385 | rs35708511 | Exon 7 | C | G | 0,2189 | 0,1713 | present on IMGT | 2679 | C>G |
| 31269423 | rs281860589 | Intron 6 | G | A | 0,9988 | 1,0000 | present on IMGT | 2641 | C>T |
| 31269453 | rs9264596 | Intron 6 | A | G | 0,2189 | 0,1713 | present on IMGT | 2611 | C>T |
| 31269456 | rs9264597 | Intron 6 | A | G | 0,2488 | 0,3009 | present on IMGT | 2608 | C>T |
| 31269460 | rs17881458 | Intron 6 | G | A | 0,9665 | 0,9954 | present on IMGT | 2604 | C>T |
| 31269477 | rs41562921 | Intron 6 | C | T | 0,9414 | 0,9213 | present on IMGT | 2587 | G>A |
| 31269478 | rs2523609 | Intron 6 | G | A | 0,2775 | 0,2500 | present on IMGT | 2586 | T>C |
| 31269480 | rs17879162 | Intron 6 | C | T | 0,9390 | 0,8565 | present on IMGT | 2584 | G>A |
| 31269498 | rs41559915 | Exon 6 | G | A | 0,9833 | 0,9444 | present on IMGT | 2566 | C>T |
| 31269509 | rs999710500 | Exon 6 | G | A | 0,9988 | 1,0000 | present on IMGT | 2555 | C>T |
| 31269518 | . | Exon 6 | T | C | 0,9988 | 1,0000 | alternative allele is not present on IMGT | 2546 | - |
| 31269546 | rs68094471 | Intron 5 | A | G | 0,2189 | 0,1713 | present on IMGT | 2518 | C>T |
| 31269556 | rs66772001 | Intron 5 | T | A | 0,2189 | 0,1713 | present on IMGT | 2508 | T>A |
| 31269565 | rs72558163 | Intron 5 | G | T | 0,9031 | 0,9444 | present on IMGT | 2499 | C>A |
| 31269576 | rs67827555 | Intron 5 | T | C | 0,2189 | 0,1713 | present on IMGT | 2488 | G>A |
| 31269577 | rs9264601 | Intron 5 | G | A | 0,9414 | 0,9213 | present on IMGT | 2487 | C>T |
| 31269628 | rs56010430 | Intron 5 | C | T | 0,2189 | 0,1713 | present on IMGT | 2436 | A>G |
| 31269660 | rs66459704 | Intron 5 | T | C | 0,2189 | 0,1713 | present on IMGT | 2404 | G>A |
| 31269661 | rs9264603 | Intron 5 | G | A | 0,9378 | 0,8565 | present on IMGT | 2403 | C>T |
| 31269672 | rs68037221 | Intron 5 | T | C | 0,2069 | 0,1713 | present on IMGT | 2392 | G>A |
| 31269680 | rs72558161 | Intron 5 | T | C | 0,9677 | 1,0000 | present on IMGT | 2384 | A>G |
| 31269684 | rs72502556 | Intron 5 | G | A | 0,7572 | 0,6806 | present on IMGT | 2380 | T>C |
| 31269686 | rs66620546 | Intron 5 | C | G | 0,2189 | 0,1713 | present on IMGT | 2378 | C>G |
| 31269736 | rs72558159 | Intron 5 | T | C | 0,9701 | 0,8704 | present on IMGT | 2328 | A>G |
| 31269749 | rs75141084 | Intron 5 | C | CA | 0,2488 | 0,3009 | present on IMGT | 2314 | delT |
| 31269786 | rs3819287 | Intron 5 | C | T | 0,2189 | 0,1713 | present on IMGT | 2277 | A>G |
| 31269794 | rs72558157 | Intron 5 | A | G | 0,9701 | 0,8704 | present on IMGT | 2269 | T>C |
| 31269798 | rs9279069 | Intron 5 | TCA | T | 0,2189 | 0,1713 | present on IMGT | 2264 | insTG |
| 31269806 | rs373826500 | Intron 5 | GCT | G | 0,9845 | 1,0000 | present on IMGT | 2257 | delAG |
| 31269815 | rs9264606 | Intron 5 | A | C | 0,2189 | 0,1713 | present on IMGT | 2250 | G>T |
| 31269828 | rs9264607 | Intron 5 | G | A | 0,7464 | 0,9028 | present on IMGT | 2237 | C>T |
| 31269859 | rs41541318 | Intron 5 | T | G,C | 0,9342 | 0,8704 | present on IMGT | 2206 | A>G>C |
| 31269883 | rs9264608 | Intron 5 | A | G | 0,2189 | 0,1713 | present on IMGT | 2182 | C>T |
| 31269887 | rs9264609 | Intron 5 | C | T | 0,2189 | 0,1713 | present on IMGT | 2178 | A>G |
| 31269900 | rs750145033 | Intron 5 | T | C | 1,0000 | 0,9954 | present on IMGT | 2165 | A>G |
| 31269926 | rs17886880 | Intron 5 | G | A | 0,6615 | 0,6250 | present on IMGT | 2139 | T>C |
| 31269946 | rs9264610 | Intron 5 | G | A | 0,2189 | 0,1667 | present on IMGT | 2119 | T>C |
| 31269950 | rs2074497 | Intron 5 | T | C | 0,2189 | 0,1667 | present on IMGT | 2115 | G>A |
| 31269955 | rs184081117 | Intron 5 | C | A | 0,9964 | 0,9954 | present on IMGT | 2110 | G>T |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 31269983 | rs1143551 | Exon 5 | C | T | 0,9701 | 0,8704 | present on IMGT | 2082 | G>A |
| 31269984 | rs41560617 | Exon 5 | A | G | 0,9701 | 0,8704 | present on IMGT | 2081 | T>C |
| 31269985 | rs41540416 | Exon 5 | C | T | 0,9701 | 0,8704 | present on IMGT | 2080 | G>A |
| 31269988 | rs2308650 | Exon 5 | C | A | 0,9378 | 0,8565 | present on IMGT | 2077 | G>T |
| 31269989 | rs41542414 | Exon 5 | A | T | 1,0000 | 0,9769 | present on IMGT | 2076 | T>A |
| 31269990 | rs1130935 | Exon 5 | T | C | 0,2189 | 0,1713 | present on IMGT | 2075 | G>A |
| 31269992 | rs1050105 | Exon 5 | G | A | 0,2488 | 0,3009 | present on IMGT | 2073 | T>C |
| 31269994 | rs1050106 | Exon 5 | G | A | 0,2189 | 0,1713 | present on IMGT | 2071 | T>C |
| 31269996 | rs1130947 | Exon 5 | T | C | 0,2189 | 0,1713 | present on IMGT | 2069 | G>A |
| 31269997 | rs41540512 | Exon 5 | G | C | 0,2488 | 0,3009 | present on IMGT | 2068 | G>C |
| 31269999 | rs1050118 | Exon 5 | C | T | 0,7093 | 0,8981 | present on IMGT | 2066 | G>A |
| 31270002 | rs146911342 | Exon 5 | C | T | 0,8289 | 0,7269 | present on IMGT | 2063 | G>A |
| 31270009 | rs41556617 | Exon 5 | A | T | 0,2189 | 0,1713 | present on IMGT | 2056 | A>T |
| 31270025 | rs1050147 | Exon 5 | A | G | 0,2273 | 0,1713 | present on IMGT | 2040 | C>T |
| 31270037 | rs148706212 | Exon 5 | A | AGGACAGCCAGGACAGCTG | 0,9701 | 0,8704 | present on IMGT | 2028-2029;2033 | TG>CA; insTCCTGGCTGTCCTGGCTG |
| 31270044 | rs41545712 | Exon 5 | C | A | 0,9701 | 0,8704 | present on IMGT | 2021 | G>T |
| 31270056 | rs1050180 | Exon 5 | T | C,A | 0,2189 | 0,1713 | present on IMGT | 2009 | G>A>T |
| 31270058 | rs1065600 | Exon 5 | A | T | 0,9701 | 0,8704 | present on IMGT | 2007 | T>A |
| 31270069 | rs11757919 | Exon 5 | G | A | 0,9845 | 1,0000 | present on IMGT | 1996 | C>T |
| 31270081 | rs34794906 | Exon 5 | T | C | 0,6304 | 0,4954 | present on IMGT | 1984 | G>A |
| 31270085 | rs9264621 | Exon 5 | T | C | 0,9031 | 0,9306 | present on IMGT | 1980 | A>G |
| 31270094 | rs9461679 | Intron 4 | G | A | 0,9833 | 0,9444 | present on IMGT | 1971 | C>T |
| 31270099 | rs9264622 | Intron 4 | A | G | 0,2189 | 0,1713 | present on IMGT | 1966 | C>T |
| 31270116 | rs72558154 | Intron 4 | G | T | 0,9701 | 0,8750 | present on IMGT | variants between 1923and 1954 | - |
| 31270118 | rs765198307 | Intron 4 | CCCCCAAG | C | 0,9904 | 0,9815 | present on IMGT | | |
| 31270119 | . | Intron 4 | CCCCAAG | ACCCAAG,* | 0,9605 | 0,8519 | present on IMGT | | |
| 31270120 | rs2234776 | Intron 4 | C | T,* | 0,7476 | 0,6620 | present on IMGT | | |
| 31270128 | rs752693431 | Intron 4 | TCAGCCCTGACCCG | T | 0,9904 | 0,9815 | present on IMGT | | |
| 31270140 | rs17884362 | Intron 4 | C | T,* | 0,9031 | 0,9444 | present on IMGT | | |
| 31270141 | rs116744327 | Intron 4 | G | A,* | 0,8971 | 0,9028 | present on IMGT | | |
| 31270154 | rs35424484 | Intron 4 | CAGA | C | 0,2584 | 0,3194 | present on IMGT | 1910 | insTCT |
| 31270161 | rs41542114 | Intron 4 | A | G | 0,9904 | 0,9815 | present on IMGT | 1907 | T>C |
| 31270164 | rs41560915 | Intron 4 | T | C | 0,9904 | 0,9815 | present on IMGT | 1904 | A>G |
| 31270173 | rs9264623 | Intron 4 | T | C | 0,2189 | 0,1713 | present on IMGT | 1895 | G>A |
| 31270174 | rs767135148 | Intron 4 | C | T | 0,9988 | 1,0000 | alternative allele is not present on IMGT | 1894 | - |
| 31270178 | rs9264624 | Intron 4 | T | G | 0,2189 | 0,1713 | present on IMGT | 1890 | C>A |
| 31270181 | rs41545214 | Intron 4 | G | C | 0,9904 | 0,9815 | present on IMGT | 1887 | C>G |
| 31270183 | rs9264625 | Intron 4 | G | C | 0,2285 | 0,1898 | present on IMGT | 1885 | G>C |
| 31270185 | rs9264626 | Intron 4 | T | C | 0,2285 | 0,1898 | present on IMGT | 1883 | G>A |
| 31270186 | rs9264627 | Intron 4 | G | A | 0,2189 | 0,1713 | present on IMGT | 1882 | T>C |
| 31270188 | rs17886232 | Intron 4 | G | C | 0,6699 | 0,6435 | present on IMGT | 1880 | G>C |
| 31270194 | rs9264628 | Intron 4 | C | T | 0,2189 | 0,1713 | present on IMGT | 1874 | A>G |
| 31270198 | rs9264629 | Intron 4 | T | C | 0,2189 | 0,1713 | present on IMGT | 1870 | G>A |
| 31270210 | rs41556321 | Exon 4 | C | T | 0,7835 | 0,5648 | present on IMGT | 1858 | G>A |
| 31270214 | rs2308628 | Exon 4 | G | T | 0,2189 | 0,1713 | present on IMGT | 1854 | A>C |
| 31270224 | rs41546713 | Exon 4 | A | C | 0,9701 | 0,8704 | present on IMGT | 1844 | T>G |
| 31270225 | rs41558713 | Exon 4 | G | A | 0,9701 | 0,8704 | present on IMGT | 1843 | C>T |
| 31270232 | rs1131014 | Exon 4 | T | C | 0,4725 | 0,2685 | present on IMGT | 1836 | G>A |
| 31270233 | rs1131015 | Exon 4 | T | G | 0,2488 | 0,3009 | present on IMGT | 1835 | C>A |
| 31270244 | rs61759945 | Exon 4 | G | A | 0,9988 | 1,0000 | present on IMGT | 1824 | C>T |
| 31270250 | rs41540117 | Exon 4 | C | A | 0,8289 | 0,7269 | present on IMGT | 1818 | G>T |
| 31270252 | rs2308622 | Exon 4 | T | C | 0,2189 | 0,1713 | present on IMGT | 1816 | G>A |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 31270271 | rs41542914 | Exon 4 | C | T | 0,9701 | 0,8704 | present on IMGT | 1797 | G>A |
| 31270276 | rs707908 | Exon 4 | G | C | 0,2488 | 0,3009 | present on IMGT | 1792 | G>C |
| 31270291 | rs1050276 | Exon 4 | C | T | 0,9761 | 0,9861 | present on IMGT | 1777 | A>G |
| 31270348 | rs41547622 | Exon 4 | C | G | 0,9964 | 1,0000 | present on IMGT | 1720 | G>C |
| 31270349 | rs2308618 | Exon 4 | G | A | 0,9031 | 0,9444 | present on IMGT | 1719 | C>T |
| 31270358 | rs1050317 | Exon 4 | G | A | 0,2189 | 0,1713 | present on IMGT | 1710 | T>C |
| 31270361 | rs1050320 | Exon 4 | C | T | 0,2189 | 0,1713 | present on IMGT | 1707 | A>G |
| 31270370 | rs1050326 | Exon 4 | C | G,A | 0,5885 | 0,6759 | second alternative is not present on IMGT | 1698 | G>C |
| 31270378 | rs1050328 | Exon 4 | G | A | 0,6603 | 0,6250 | present on IMGT | 1690 | T>C |
| 31270402 | rs41562012 | Exon 4 | C | T | 0,9617 | 0,9861 | present on IMGT | 1666 | G>A |
| 31270453 | rs1050716 | Exon 4 | G | C | 0,2189 | 0,1713 | present on IMGT | 1615 | G>C |
| 31270455 | rs1050343 | Exon 4 | G | A | 0,9378 | 0,8565 | present on IMGT | 1613 | C>T |
| 31270457 | rs1050344 | Exon 4 | G | A | 0,2189 | 0,1713 | present on IMGT | 1611 | T>C |
| 31270482 | rs1131096 | Exon 4 | G | T,C | 0,2189 | 0,1713 | present on IMGT | 1586 | A>C>G |
| 31270490 | rs41561715 | Intron 3 | C | T | 0,8134 | 0,6944 | present on IMGT | 1578 | G>A |
| 31270520 | rs9264636 | Intron 3 | C | T | 0,2189 | 0,1713 | present on IMGT | 1548 | A>G |
| 31270531 | rs9264637 | Intron 3 | G | C | 0,2189 | 0,1713 | present on IMGT | 1537 | G>C |
| 31270541 | rs9264638 | Intron 3 | G | A | 0,2488 | 0,3009 | present on IMGT | 1527 | T>C |
| 31270545 | rs9264639 | Intron 3 | C | T | 0,9031 | 0,9306 | present on IMGT | 1523 | G>A |
| 31270546 | rs140505137 | Intron 3 | G | A | 0,9916 | 1,0000 | present on IMGT | 1522 | C>T |
| 31270557 | rs41540318 | Intron 3 | C | T | 0,9414 | 0,9213 | present on IMGT | 1511 | G>A |
| 31270568 | rs17883395 | Intron 3 | T | C | 0,9031 | 0,9444 | present on IMGT | 1500 | A>G |
| 31270587 | rs41544416 | Intron 3 | G | A | 0,9701 | 0,8704 | present on IMGT | 1481 | C>T |
| 31270595 | rs9264640 | Intron 3 | C | G | 0,2189 | 0,1713 | present on IMGT | 1473 | C>G |
| 31270600 | rs9264641 | Intron 3 | G | C | 0,0514 | 0,0000 | present on IMGT | 1468 | G>C |
| 31270629 | rs41543713 | Intron 3 | C | G | 0,9701 | 0,8704 | present on IMGT | 1439 | G>C |
| 31270651 | rs41540812 | Intron 3 | A | T | 0,9701 | 0,8704 | present on IMGT | 1417 | T>A |
| 31270713 | rs9264642 | Intron 3 | T | C | 0,2811 | 0,3148 | present on IMGT | 1355 | G>A |
| 31270715 | rs9264643 | Intron 3 | C | A | 0,2189 | 0,1713 | present on IMGT | 1353 | T>G |
| 31270718 | rs17884390 | Intron 3 | G | A | 0,9378 | 0,8565 | present on IMGT | 1350 | C>T |
| 31270730 | rs41544520 | Intron 3 | C | T | 0,9617 | 0,9630 | present on IMGT | 1338 | G>A |
| 31270745 | rs9264644 | Intron 3 | C | T | 0,7464 | 0,9028 | present on IMGT | 1323 | G>A |
| 31270747 | rs9264645 | Intron 3 | G | A | 0,2488 | 0,3009 | present on IMGT | 1321 | T>C |
| 31270758 | rs9264646 | Intron 3 | T | C | 0,2488 | 0,3009 | present on IMGT | 1310 | G>A |
| 31270759 | rs17884702 | Intron 3 | G | A | 0,9378 | 0,8565 | present on IMGT | 1309 | C>T |
| 31270761 | rs41562714 | Intron 3 | G | C | 0,8289 | 0,7269 | present on IMGT | 1307 | C>G |
| 31270788 | rs114343571 | Intron 3 | G | A | 1,0000 | 0,9769 | present on IMGT | 1280 | C>T |
| 31270812 | rs41549515 | Intron 3 | T | C | 0,9701 | 0,8704 | present on IMGT | 1256 | A>G |
| 31270836 | rs9264647 | Intron 3 | A | G | 0,2189 | 0,1713 | present on IMGT | 1232 | C>T |
| 31270855 | rs17886924 | Intron 3 | C | T | 0,9031 | 0,9444 | present on IMGT | 1213 | G>A |
| 31270884 | rs9264648 | Intron 3 | G | A | 0,2488 | 0,3009 | present on IMGT | 1184 | T>C |
| 31270885 | rs9264649 | Intron 3 | A | G | 0,2189 | 0,1713 | present on IMGT | 1183 | C>T |
| 31270918 | rs41547523 | Intron 3 | G | C | 0,9701 | 0,8704 | present on IMGT | 1150 | C>G |
| 31270931 | rs9264650 | Intron 3 | G | A | 0,9187 | 0,9352 | present on IMGT | 1137 | C>T |
| 31270939 | rs41550819 | Intron 3 | A | G | 0,9701 | 0,8704 | present on IMGT | 1129 | T>C |
| 31270951 | rs9264651 | Intron 3 | G | C | 0,2225 | 0,1713 | present on IMGT | 1117 | G>C |
| 31270982 | rs17884428 | Intron 3 | C | T | 0,3158 | 0,2407 | present on IMGT | 1086 | A>G |
| 31270985 | rs9264652 | Intron 3 | T | C | 0,2189 | 0,1713 | present on IMGT | 1083 | G>A |
| 31271016 | rs41561913 | Intron 3 | G | A | 0,9629 | 0,9954 | present on IMGT | 1052 | C>T |
| 31271024 | rs41544614 | Intron 3 | C | G | 0,9761 | 0,9861 | present on IMGT | 1044 | C>G |
| 31271037 | rs9264653 | Intron 3 | T | C | 0,2488 | 0,3009 | present on IMGT | 1031 | G>A |
| 31271038 | rs41544212 | Intron 3 | A | G | 0,7572 | 0,6806 | present on IMGT | 1030 | C>T |
| 31271045 | rs41561812 | Intron 3 | G | A | 0,9629 | 0,9954 | present on IMGT | 1023 | C>T |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 31271074 | rs2308604 | Exon 3 | T | C | 0,2069 | 0,1713 | present on IMGT | 994 | G>A |
| 31271091 | rs1131103 | Exon 3 | C | T | 0,8911 | 0,9306 | present on IMGT | 977 | G>A |
| 31271097 | rs41552417 | Exon 3 | C | T,G | 0,9976 | 1,0000 | present on IMGT | 971 | G>A>C |
| 31271098 | rs1131104 | Exon 3 | G | A | 0,9031 | 0,9444 | present on IMGT | 970 | C>T |
| 31271103 | rs1050357 | Exon 3 | C | T | 0,9031 | 0,9444 | present on IMGT | 965 | G>A |
| 31271112 | rs2308598 | Exon 3 | T | C | 0,9701 | 0,8704 | present on IMGT | 956 | A>G |
| 31271132 | rs1050686 | Exon 3 | G | T,A | 0,8146 | 0,7361 | present on IMGT | 936 | C>A>T |
| 31271133 | rs1050685 | Exon 3 | T | C,G | 0,8146 | 0,7361 | present on IMGT | 935 | A>G>C |
| 31271151 | rs766416978 | Exon 3 | TCA | T | 0,9880 | 1,0000 | present on IMGT | variants between 914 and 916 | - |
| 31271153 | . | Exon 3 | AG | CG,TG,* | 0,3732 | 0,3843 | present on IMGT | | |
| 31271154 | rs697743 | Exon 3 | G | A,GTC | 0,7656 | 0,8935 | present on IMGT | | |
| 31271165 | rs2308590 | Exon 3 | G | T | 0,2835 | 0,3148 | present on IMGT | 903 | A>C |
| 31271166 | rs41552817 | Exon 3 | C | T | 0,9976 | 1,0000 | present on IMGT | 902 | G>A |
| 31271180 | rs1050366 | Exon 3 | A | C | 0,2488 | 0,3009 | present on IMGT | 888 | G>T |
| 31271193 | rs142570222 | Exon 3 | T | A | 0,9701 | 0,8704 | present on IMGT | 875 | A>T |
| 31271206 | rs2308585 | Exon 3 | G | C,T | 0,2428 | 0,1852 | present on IMGT | 862 | C>G>A |
| 31271207 | rs2308584 | Exon 3 | G | T | 0,9055 | 0,9306 | present on IMGT | 861 | C>A |
| 31271215 | rs41550715 | Exon 3 | G | C,A | 0,9605 | 0,8704 | present on IMGT | 853 | C>T>G |
| 31271218 | rs41553316 | Exon 3 | G | A | 0,9438 | 0,9907 | present on IMGT | 850 | C>T |
| 31271233 | rs1050371 | Exon 3 | G | A | 0,6842 | 0,6389 | present on IMGT | 835 | C>T |
| 31271239 | rs1050373 | Exon 3 | G | A | 0,9031 | 0,9306 | present on IMGT | 829 | C>T |
| 31271272 | rs1065406 | Exon 3 | G | T | 0,9665 | 1,0000 | present on IMGT | 796 | C>A |
| 31271273 | rs713032 | Exon 3 | G | A,T | 0,5263 | 0,4537 | present on IMGT | 795 | A>C>T |
| 31271280 | rs2308575 | Exon 3 | C | T | 0,6866 | 0,4954 | present on IMGT | 788 | G>A |
| 31271282 | rs41555122 | Exon 3 | T | C | 1,0000 | 0,9907 | present on IMGT | 786 | A>G |
| 31271283 | rs2308574 | Exon 3 | A | G | 0,9629 | 0,9954 | present on IMGT | 785 | T>C |
| 31271305 | rs1050384 | Exon 3 | G | C | 0,8134 | 0,6944 | present on IMGT | 763 | C>G |
| 31271313 | rs34592426 | Exon 3 | G | C | 0,9031 | 0,9444 | present on IMGT | 755 | C>G |
| 31271323 | rs1131114 | Exon 3 | A | G | 0,9222 | 0,8981 | present on IMGT | 745 | T>C |
| 31271324 | rs1131115 | Exon 3 | G | A,T,C | 0,0837 | 0,0463 | present on IMGT | 744 | G>A>T>C |
| 31271331 | rs1131118 | Exon 3 | T | A | 0,7345 | 0,8148 | present on IMGT | 737 | T>A |
| 31271337 | rs1071649 | Exon 3 | G | T,A | 0,8325 | 0,8426 | present on IMGT | 731 | C>A>T |
| 31271339 | rs1131119 | Exon 3 | G | A | 0,8660 | 0,9398 | present on IMGT | 729 | C>T |
| 31271347 | rs41543218 | Exon 3 | C | A | 0,9701 | 0,8704 | present on IMGT | 721 | G>T |
| Intron 2 | - | Intron 2 | - | - | - | - | not included in this study | - | - |
| 31271599 | rs1131122 | Exon 2 | C | T | 0,9557 | 0,9954 | present on IMGT | 473 | G>A |
| 31271601 | rs1131123 | Exon 2 | T | G | 0,4964 | 0,4954 | present on IMGT | 471 | C>A |
| 31271630 | rs17408553 | Exon 2 | G | T | 0,5431 | 0,4583 | present on IMGT | 442 | C>A |
| 31271640 | rs2308557 | Exon 2 | C | T | 0,5431 | 0,4583 | present on IMGT | 432 | G>A |
| 31271653 | rs41543814 | Exon 2 | C | T | 0,5921 | 0,6343 | present on IMGT | 419 | A>G |
| 31271672 | rs28626310 | Exon 2 | C | G | 0,8397 | 0,8704 | present on IMGT | 400 | G>C |
| 31271724 | rs1050409 | Exon 2 | G | T | 0,8301 | 0,7269 | present on IMGT | 348 | C>A |
| 31271729 | rs1050414 | Exon 2 | C | G | 0,8816 | 0,9491 | present on IMGT | 343 | G>C |
| 31271741 | rs1050420 | Exon 2 | C | T | 0,5048 | 0,4213 | present on IMGT | 331 | G>A |
| 31271766 | rs1050428 | Exon 2 | C | T | 0,9031 | 0,9306 | present on IMGT | 306 | G>A |
| 31271800 | rs707911 | Exon 2 | A | C | 0,3493 | 0,2361 | present on IMGT | 272 | T>G |
| 31271808 | rs1050437 | Exon 2 | C | T | 0,8038 | 0,8611 | present on IMGT | 264 | G>A |
| 31271816 | rs41542719 | Exon 2 | T | C | 0,6998 | 0,6713 | present on IMGT | 256 | A>G |
| 31271824 | rs151341100 | Exon 2 | C | T | 0,9414 | 0,9213 | present on IMGT | 248 | G>A |
| 31271825 | rs281860336 | Exon 2 | G | A | 0,9761 | 0,9861 | present on IMGT | 247 | T>C |
| 31271830 | rs41542423 | Exon 2 | G | A | 0,8289 | 0,7269 | present on IMGT | 242 | C>T |
| 31271837 | rs1050444 | Exon 2 | G | A | 0,8074 | 0,8611 | present on IMGT | 235 | C>T |
| 31271839 | rs1050445 | Exon 2 | C | A | 0,7727 | 0,7130 | present on IMGT | 233 | T>G |
| 31271840 | rs1050446 | Exon 2 | G | T | 0,7727 | 0,7130 | present on IMGT | 232 | A>C |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 31271844 | rs1071650 | Exon 2 | T | G,A | 0,7727 | 0,7130 | present on IMGT | 228 | T>A>C |
| 31271845 | rs9264668 | Exon 2 | C | A | 0,3266 | 0,2222 | present on IMGT | 227 | T>G |
| 31271853 | rs1131151 | Exon 2 | C | T | 0,9761 | 0,9861 | present on IMGT | 219 | A>G |
| 31271875 | rs41556116 | Intron 1 | G | A | 0,7967 | 0,7269 | present on IMGT | 197 | C>T |
| 31271876 | rs41560121 | Intron 1 | C | A | 0,9414 | 0,9213 | present on IMGT | 196 | G>T |
| 31271883 | rs41559317 | Intron 1 | G | T | 0,7967 | 0,7269 | present on IMGT | 189 | C>A |
| 31271904 | rs9264669 | Intron 1 | A | T | 0,2978 | 0,2176 | present on IMGT | 168 | A>T |
| 31271920 | rs281860317 | Intron 1 | C | T | 0,9880 | 1,0000 | present on IMGT | 152 | G>A |
| 31271945 | rs9264670 | Intron 1 | A | C | 0,1112 | 0,0787 | present on IMGT | 127 | G>T |
| 31271950 | rs29029490 | Intron 1 | C | A | 0,9761 | 0,9861 | present on IMGT | 122 | T>G |
| 31271956 | rs2074494 | Intron 1 | C | T | 0,9761 | 0,9861 | present on IMGT | 116 | A>G |
| 31271965 | rs41553018 | Intron 1 | G | C | 0,8289 | 0,7269 | present on IMGT | 107 | C>G |
| 31271975 | rs9264671 | Intron 1 | C | T,A | 0,2345 | 0,2037 | present on IMGT | 97 | A>T>G |
| 31271984 | rs41550615 | Intron 1 | C | T | 0,9677 | 1,0000 | present on IMGT | 88 | G>A |
| 31271989 | rs9264672 | Intron 1 | C | T | 0,9031 | 0,9306 | present on IMGT | 83 | G>A |
| 31271999 | rs2074493 | Exon 1 | A | C | 0,7022 | 0,5417 | present on IMGT | 73 | T>G |
| 31272002 | rs41553415 | Exon 1 | C | T | 0,9952 | 1,0000 | present on IMGT | 70 | G>A |
| 31272013 | rs41549413 | Exon 1 | G | A | 0,9701 | 0,8704 | present on IMGT | 59 | C>T |
| 31272025 | rs1050451 | Exon 1 | C | G | 0,2345 | 0,2037 | present on IMGT | 47 | C>G |
| 31272044 | rs2308527 | Exon 1 | G | T | 0,3600 | 0,4167 | present on IMGT | 28 | A>C |
| 31272050 | rs2308525 | Exon 1 | C | T | 0,2644 | 0,3333 | present on IMGT | 22 | A>G |
| 31272052 | rs41548123 | Exon 1 | C | T | 0,9701 | 0,8704 | present on IMGT | 20 | G>A |
| 31272067 | rs772484756 | Exon 1 | C | G | 1,0000 | 0,9907 | present on IMGT | 5 | G>C |
| 31272078 | rs7767581 | 5'Upstream | C | G | 0,2345 | 0,2037 | present on IMGT | -7 | C>G |
| 31272091 | rs9264674 | 5'Upstream | G | C | 0,2345 | 0,2037 | present on IMGT | -20 | G>C |
| 31272092 | rs2074492 | 5'Upstream | T | C | 0,5885 | 0,6759 | present on IMGT | -21 | A>G |
| 31272106 | rs200353551 | 5'Upstream | TGACTGGG | T | 0,9701 | 0,8704 | present on IMGT | -42 to -36 | delCCCAGTC |
| 31272119 | rs2074491 | 5'Upstream | T | C | 0,8732 | 0,8148 | present on IMGT | -48 | A>G |
| 31272135 | rs115120050 | 5'Upstream | G | T | 0,9701 | 0,8704 | present on IMGT | -64 | C>A |
| 31272160 | rs2074490 | 5'Upstream | C | T | 0,9031 | 0,9444 | present on IMGT | -89 | G>A |
| 31272174 | rs778577883 | 5'Upstream | G | A | 0,9988 | 1,0000 | alternative allele is not present on IMGT | -103 | - |
| 31272222 | rs6457358 | 5'Upstream | A | T | 0,9043 | 0,9167 | present on IMGT | -151 | T>A |
| 31272231 | rs962515335 | 5'Upstream | G | A | 0,9976 | 1,0000 | alternative allele is not present on IMGT | -160 | - |
| 31272232 | rs9281301 | 5'Upstream | G | GAA | 0,9031 | 0,9306 | present on IMGT | -162;-165 to -163 | C>T; delTCT |
| 31272233 | rs10657191 | 5'Upstream | G | GAGA,GA | 0,2345 | 0,2037 | present on IMGT | | |
| 31272254 | rs879925566 | 5'Upstream | T | TG | 0,9701 | 0,8704 | present on IMGT | -189;-181 to -180 | T>C;insACTCCC |
| 31272257 | rs201049006 | 5'Upstream | A | AGTGGG | 0,9701 | 0,8704 | present on IMGT | | |
| 31272264 | rs2524094 | 5'Upstream | C | T | 0,4234 | 0,6065 | present on IMGT | -196 | A>G |
| 31272271 | rs281860310 | 5'Upstream | G | T | 0,9701 | 0,8704 | present on IMGT | -203 | C>A |
| 31272283 | rs2844622 | 5'Upstream | A | G | 0,2524 | 0,3333 | present on IMGT | -215 | C>T |
| 31272299 | rs281860309 | 5'Upstream | T | G | 0,9701 | 0,8704 | present on IMGT | -231 | A>C |
| 31272305 | rs35045183 | 5'Upstream | C | T | 0,9701 | 0,8704 | present on IMGT | -237 | G>A |
| 31272314 | rs59057661 | 5'Upstream | T | A | 0,9701 | 0,8704 | present on IMGT | -246 | A>T |
| 31272319 | rs9366775 | 5'Upstream | G | A | 0,9761 | 0,9861 | present on IMGT | -251 | T>C |
| 31272325 | rs281860308 | 5'Upstream | C | T | 0,9845 | 1,0000 | present on IMGT | -257 | G>A |
| 31272351 | rs2074489 | 5'Upstream | C | T | 0,7321 | 0,6713 | present on IMGT | -283 | G>A |
| 31272366 | rs553330871 | 5'Upstream | C | G | 0,9988 | 1,0000 | alternative allele is not present on IMGT | -298 | - |
| 31272369 | rs115573554 | 5'Upstream | A | T | 0,9629 | 0,9954 | present on IMGT | -301 | T>A |
| 31272403 | rs2524093 | 5'Upstream | A | C | 0,2345 | 0,2037 | present on IMGT | -335 | G>T |
| 31272439 | rs2524092 | 5'Upstream | A | C | 0,2345 | 0,2037 | present on IMGT | -371 | G>T |
| 31272469 | rs549359394 | 5'Upstream | TCTC | T | 0,2656 | 0,3333 | present on IMGT | -402 to -401 | insGAG |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 31272481 | rs57800660 | 5'Upstream | A | T | 0,9701 | 0,8704 | present on IMGT | -410 | T>A |
| 31272483 | rs192719521 | 5'Upstream | C | T | 0,9988 | 1,0000 | alternative allele is not present on IMGT | -412 | - |
| 31272516 | rs2524091 | 5'Upstream | C | T | 0,2345 | 0,2037 | present on IMGT | -445 | A>G |
| 31272528 | rs992610800 | 5'Upstream | C | T | 0,9976 | 1,0000 | present on IMGT | variants between -449 and -459 | C>T;insAAGGC; A>G |
| 31272530 | rs554163072 | 5'Upstream | TACCTC | T | 0,2189 | 0,1713 | present on IMGT | | |
| 31272576 | rs28498059 | 5'Upstream | A | C | 0,0837 | 0,0463 | It is not tracked by the IMGT | - | - |
| 31272597 | rs142561742 | 5'Upstream | T | TC | 0,9629 | 0,9954 | It is not tracked by the IMGT | - | - |
| 31272654 | rs2074488 | 5'Upstream | G | T | 0,9031 | 0,9444 | It is not tracked by the IMGT | - | - |
| 31272702 | rs9357121 | 5'Upstream | T | G | 0,9761 | 0,9861 | It is not tracked by the IMGT | - | - |
| 31272721 | rs1025824108 | 5'Upstream | T | C | 0,9988 | 1,0000 | It is not tracked by the IMGT | - | - |
| 31272837 | rs115710087 | 5'Upstream | C | T | 0,9880 | 1,0000 | It is not tracked by the IMGT | - | - |
| 31272858 | rs4361609 | 5'Upstream | G | C | 0,8050 | 0,7130 | It is not tracked by the IMGT | - | - |
| 31272859 | rs5009853 | 5'Upstream | C | T | 0,9629 | 0,9954 | It is not tracked by the IMGT | - | - |
| 31272886 | rs9264679 | 5'Upstream | T | A | 0,0837 | 0,0463 | It is not tracked by the IMGT | - | - |
| 31272915 | rs2395471 | 5'Upstream | G | A | 0,4581 | 0,4583 | It is not tracked by the IMGT | - | - |
| 31272935 | rs2249741 | 5'Upstream | A | C | 0,3457 | 0,3565 | It is not tracked by the IMGT | - | - |
| 31272944 | rs2249742 | 5'Upstream | C | T | 0,4450 | 0,5046 | It is not tracked by the IMGT | - | - |
| 31272994 | rs13203722 | 5'Upstream | T | C | 0,8433 | 0,9722 | It is not tracked by the IMGT | - | - |
| 31272999 | rs7454487 | 5'Upstream | A | G | 0,9605 | 0,9954 | It is not tracked by the IMGT | - | - |
| 31273037 | rs34622008 | 5'Upstream | C | T | 0,9031 | 0,9444 | It is not tracked by the IMGT | - | - |
| 31273041 | rs4298345 | 5'Upstream | G | A | 0,9761 | 0,9861 | It is not tracked by the IMGT | - | - |
| 31273046 | rs2524090 | 5'Upstream | C | A | 0,8756 | 0,8981 | It is not tracked by the IMGT | - | - |
| 31273049 | rs7454567 | 5'Upstream | G | A | 0,9629 | 0,9954 | It is not tracked by the IMGT | - | - |
| 31273059 | rs115345338 | 5'Upstream | T | C | 0,9880 | 1,0000 | It is not tracked by the IMGT | - | - |
| 31273094 | rs9264681 | 5'Upstream | G | C | 0,7787 | 0,8426 | It is not tracked by the IMGT | - | - |
| 31273095 | rs9264682 | 5'Upstream | T | G,C | 0,7787 | 0,8426 | It is not tracked by the IMGT | - | - |
| 31273107 | rs7454526 | 5'Upstream | A | G | 0,9629 | 0,9954 | It is not tracked by the IMGT | - | - |
| 31273155 | rs13218306 | 5'Upstream | C | A | 0,8481 | 0,9722 | It is not tracked by the IMGT | - | - |
| 31273162 | rs66633038 | 5'Upstream | TA | T | 0,8481 | 0,9722 | It is not tracked by the IMGT | - | - |
| 31273163 | rs9264683 | 5'Upstream | A | T,* | 0,7189 | 0,8704 | It is not tracked by the IMGT | - | - |
| 31273182 | rs7773175 | 5'Upstream | C | G | 0,7225 | 0,8704 | It is not tracked by the IMGT | - | - |
| 31273211 | rs7759127 | 5'Upstream | T | G | 0,8876 | 0,8981 | It is not tracked by the IMGT | - | - |
| 31273224 | rs59663747 | 5'Upstream | GA | G | 0,9952 | 1,0000 | It is not tracked by the IMGT | - | - |
| 31273225 | rs13191099 | 5'Upstream | A | G,* | 0,8385 | 0,9722 | It is not tracked by the IMGT | - | - |
| 31273255 | rs5010528 | 5'Upstream | A | G | 0,8289 | 0,7269 | It is not tracked by the IMGT | - | - |
| 31273258 | rs139073773 | 5'Upstream | G | A | 0,9988 | 1,0000 | It is not tracked by the IMGT | - | - |
| 31273282 | rs536373059 | 5'Upstream | CT | C | 0,7943 | 0,8148 | It is not tracked by the IMGT | - | - |
| 31273300 | rs35976302 | 5'Upstream | T | C | 0,9031 | 0,9444 | It is not tracked by the IMGT | - | - |
| 31273315 | rs2523599 | 5'Upstream | C | G | 0,3002 | 0,3148 | It is not tracked by the IMGT | - | - |
| 31273332 | rs13191343 | 5'Upstream | C | T | 0,8505 | 0,9722 | It is not tracked by the IMGT | - | - |
| 31273339 | rs28367582 | 5'Upstream | C | T | 0,9366 | 0,8565 | It is not tracked by the IMGT | - | - |
| 31273350 | rs13207315 | 5'Upstream | T | C | 0,8493 | 0,9722 | It is not tracked by the IMGT | - | - |
| 31273368 | rs200939896 | 5'Upstream | G | T | 0,9880 | 1,0000 | It is not tracked by the IMGT | - | - |
| 31273371 | rs150036304 | 5'Upstream | A | C | 0,9880 | 1,0000 | It is not tracked by the IMGT | - | - |
| 31273373 | rs201913856 | 5'Upstream | T | C | 0,9988 | 1,0000 | It is not tracked by the IMGT | - | - |
| 31273405 | rs6900444 | 5'Upstream | C | T | 0,4043 | 0,4769 | It is not tracked by the IMGT | - | - |
| 31273430 | rs6900458 | 5'Upstream | C | T | 0,8074 | 0,8611 | It is not tracked by the IMGT | - | - |
| 31273438 | rs58019823 | 5'Upstream | A | G | 0,8289 | 0,7269 | It is not tracked by the IMGT | - | - |
| 31273459 | rs6900474 | 5'Upstream | C | G | 0,9880 | 0,9537 | It is not tracked by the IMGT | - | - |
| 31273464 | rs114183633 | 5'Upstream | C | T | 0,9629 | 0,9954 | It is not tracked by the IMGT | - | - |
| 31273483 | rs2524088 | 5'Upstream | T | C | 0,8170 | 0,8194 | It is not tracked by the IMGT | - | - |
| 31273493 | rs6900323 | 5'Upstream | G | A | 0,1722 | 0,2546 | It is not tracked by the IMGT | - | - |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 31273512 | rs181731481 | 5'Upstream | C | T | 0,9964 | 1,0000 | It is not tracked by the IMGT | - | - |
| 31273517 | rs2524087 | 5'Upstream | C | G,T | 0,7165 | 0,6250 | It is not tracked by the IMGT | - | - |
| 31273534 | rs2524086 | 5'Upstream | A | G | 0,8768 | 0,8750 | It is not tracked by the IMGT | - | - |
| 31273576 | rs34090104 | 5'Upstream | G | T | 0,9031 | 0,9444 | It is not tracked by the IMGT | - | - |
| 31273593 | rs6923313 | 5'Upstream | C | T | 0,3409 | 0,2824 | It is not tracked by the IMGT | - | - |
| 31273596 | rs150356282 | 5'Upstream | C | T | 0,9964 | 1,0000 | It is not tracked by the IMGT | - | - |

Notes: The *HLA-C* gene is encoded in the hg38 reverse strand. Therefore, variants in the IPD-IMGT/HLA database is represented as the inverse of the reference and alternative alleles. The *HLA-C* reference sequence in hg38 corresponds to allele *C*07:02:01:03*, while the reference allele in the IMGT/HLA database is *C*01:02:01:01*. Because of that, the reference and alternative alleles do not always match between hg38 and IMGT/HLA. * The *miRNA148a* binding region.

**Table S2. List of the *HLA-C* genomic alleles and their frequencies in Brazil and Benin.**

| *HLA-C* coding alleles | Brazilian frequency (2n=836) | Beninese Frequency (2n=216) |
|---|---|---|
| **C*01:02:01:01**,C*01:02:01:02,C*01:02:01:05 | 0,0239 | 0,0139 |
| **C*02:02:02:01**,C*02:02:02:04,C*02:02:02:08,C*02:02:02:11,C*02:02:02:16,C*02:02:02:18 | 0,0335 | 0,0139 |
| **C*02:02:02:03**,C*02:02:02:08,C*02:02:02:11,C*02:02:02:16 | 0,0024 | - |
| **C*02:02:02:08**,C*02:02:02:11,C*02:02:02:16,C*02:02:02:20 | 0,0012 | - |
| **C*02:10:01:01**,C*02:10:01:06,C*02:10:01:07 | 0,0179 | 0,0602 |
| C*02:10:01:02 | 0,0012 | - |
| C*02:14:02 | 0,0012 | - |
| C*03:02:02:04 | 0,0012 | - |
| C*03:02:02:05 | 0,0072 | 0,0324 |
| **C*03:03:01:01**,C*03:03:01:06,C*03:03:01:09,C*03:03:01:12 | 0,0431 | - |
| C*03:03:04 | - | 0,0046 |
| **C*03:04:01:01**,C*03:04:01:09 | 0,0191 | - |
| **C*03:04:01:02**,C*03:04:01:12 | 0,0120 | - |
| C*03:04:01:13 | 0,0024 | - |
| C*03:04:02 | 0,0096 | 0,0185 |
| C*03:04:58 | 0,0012 | - |
| **C*04:01:01:02,**C*04:01:01:11,C*04:01:01:12,C*04:01:01:14,C*04:01:01:16 | 0,1148 | 0,2639 |
| **C*04:01:01:02**,C*04:01:01:12,C*04:01:01:16 | 0,0036 | - |
| C*04:01:01:05 | 0,0120 | - |
| **C*04:01:01:06**,C*04:01:01:29 | 0,0347 | - |
| C*04:01:01:15 | 0,0012 | - |
| C*04:01:01:32 | - | 0,0046 |
| C*04:07:01 | 0,0012 | - |
| C*04:09N | 0,0024 | - |
| C*05:01:01:01 | 0,0132 | 0,0046 |
| **C*05:01:01:02**,C*05:01:01:08,C*05:01:01:17 | 0,0287 | - |
| **C*06:02:01:01**,C*06:02:01:12 | 0,0562 | 0,0139 |
| **C*06:02:01:02**,C*06:02:01:04,C*06:02:01:16 | 0,0263 | - |
| C*06:02:01:03 | 0,0072 | - |
| C*06:02:01:04 | 0,0012 | - |
| **C*07:01:01:01**,C*07:01:01:12,C*07:01:01:16,C*07:01:01:23,C*07:01:01:26,C*07:01:01:28 | 0,0897 | 0,0602 |
| C*07:01:02 | 0,0156 | - |
| C*07:01:09 | 0,0012 | - |
| **C*07:02:01:01**,C*07:02:01:05,C*07:02:01:15 | 0,0323 | 0,0417 |
| **C*07:02:01:03**,C*07:02:01:13,C*07:02:01:16,C*07:02:01:20,C*07:02:01:24,C*07:02:01:31 | 0,0490 | - |
| C*07:02:01:20 | 0,0024 | - |
| **C*07:04:01:01**,C*07:04:01:03 | 0,0120 | - |
| C*07:06:01:01 | - | 0,0231 |
| C*07:18:01:01 | 0,0167 | 0,0324 |
| C*07:35 | - | 0,0093 |
| **C*08:01:01:01**,C*08:01:01:02 | 0,0012 | - |
| **C*08:02:01:01**,C*08:02:01:03,C*08:02:01:04 | 0,0383 | 0,0370 |
| C*08:02:01:02 | 0,0120 | 0,0046 |
| C*08:03:01 | 0,0012 | - |
| C*08:04:01 | 0,0012 | 0,0231 |
| **C*12:02:02:01**,C*12:02:02:03 | 0,0084 | - |
| **C*12:03:01:01**,C*12:03:01:02,C*12:03:01:05,C*12:03:01:07,C*12:03:01:13 | 0,0538 | 0,0093 |
| C*12:03:01:13 | 0,0024 | - |
| **C*14:02:01:01**,C*14:02:01:04,C*14:02:01:06,C*14:02:01:08,C*14:02:01:09 | 0,0251 | - |
| C*14:02:01:02 | 0,0036 | - |
| C*14:03:01 | 0,0036 | - |
| **C*15:02:01:01**,C*15:02:01:05,C*15:02:01:06,C*15:02:01:07,C*15:02:01:08N | 0,0251 | - |
| C*15:05:02:01 | 0,0024 | 0,0046 |
| C*15:08:01 | 0,0012 | - |
| C*15:09 | 0,0012 | - |
| C*15:13:01:01 | 0,0036 | - |
| C*16:01:01:01 | 0,0407 | 0,1435 |
| C*16:01:01:02 | 0,0012 | - |
| C*16:02:01 | 0,0120 | - |
| **C*16:04:01:01**,C*16:04:01:02 | 0,0072 | - |
| C*17:01:01:02 | 0,0239 | 0,1204 |
| **C*17:03:01:01**,C*17:03:01:04 | 0,0048 | - |
| C*17:38 | 0,0012 | - |
| C*18:02:01 | 0,0072 | 0,0324 |
| Unknown sequences | 0,0191 | 0,0278 |

Obs: Some alleles are grouped together because they differ in nucleotides outside the region we have sequenced. The most likely allele is marked in bold.

**Table S3. Amino acid exchanges and their frequencies in Brazil and Benin.**

| Amino acid residue | Obs | Amino Acid | Brazil Frequency 2n=836 | Benin Frequency 2n=216 | Amino Acid | Brazil Frequency 2n=836 | Benin Frequency 2n=216 | Amino Acid | Brazil Frequency 2n=836 | Benin Frequency 2n=216 | Amino Acid | Brazil Frequency 2n=836 | Benin Frequency 2n=216 | Amino Acid | Brazil Frequency 2n=836 | Benin Frequency 2n=216 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LP2 | Leader peptide | P | 0,000 | 0,009 | R | 1,000 | 0,991 | | | | | | | | | |
| LP7 | Leader peptide | **Q** | **0,030** | **0,130** | **R** | **0,970** | **0,870** | | | | | | | | | |
| LP8 | Leader peptide | A | 0,264 | 0,333 | T | 0,736 | 0,667 | | | | | | | | | |
| LP10 | Leader peptide | I | 0,640 | 0,583 | L | 0,360 | 0,417 | | | | | | | | | |
| LP16 | Leader peptide | A | 0,766 | 0,796 | G | 0,234 | 0,204 | | | | | | | | | |
| LP20 | Leader peptide | **I** | **0,030** | **0,130** | **T** | **0,970** | **0,870** | | | | | | | | | |
| LP24 | Leader peptide | A | 0,995 | 1,000 | T | 0,005 | 0,000 | | | | | | | | | |
| **1*** | Alpha 1 | C | **0,702** | **0,542** | G | **0,298** | **0,458** | | | | | | | | | |
| 6 | Alpha 1 | K | 0,024 | 0,014 | R | 0,976 | 0,986 | | | | | | | | | |
| **9** | Alpha 1 | D | **0,327** | **0,222** | F | 0,024 | 0,014 | S | 0,203 | 0,273 | Y | 0,446 | 0,491 | | | |
| 11 | Alpha 1 | A | 0,773 | 0,713 | S | 0,227 | 0,287 | | | | | | | | | |
| **14** | Alpha 1 | R | **0,829** | **0,727** | W | **0,171** | **0,273** | | | | | | | | | |
| 16 | Alpha 1 | G | 0,941 | 0,921 | S | 0,059 | 0,079 | | | | | | | | | |
| 21 | Alpha 1 | H | 0,196 | 0,139 | R | 0,804 | 0,861 | | | | | | | | | |
| **24** | Alpha 1 | A | **0,651** | **0,764** | S | **0,349** | **0,236** | | | | | | | | | |
| 35 | Alpha 1 | Q | 0,097 | 0,069 | R | 0,903 | 0,931 | | | | | | | | | |
| **49** | Alpha 1 | A | **0,830** | **0,727** | E | **0,170** | **0,273** | | | | | | | | | |
| 66 | Alpha 1 | K | 0,840 | 0,870 | N | 0,160 | 0,130 | | | | | | | | | |
| 73 | Alpha 1 | A | 0,592 | 0,634 | T | 0,408 | 0,366 | | | | | | | | | |
| 77 | Alpha 1 | N | 0,457 | 0,542 | S | 0,543 | 0,458 | | | | | | | | | |
| 80 | Alpha 1 | K | 0,457 | 0,542 | N | 0,543 | 0,458 | | | | | | | | | |
| 90 | Alpha 1 | A | 0,504 | 0,505 | D | 0,496 | 0,495 | | | | | | | | | |
| 91 | Alpha 2 | G | 0,956 | 0,995 | R | 0,044 | 0,005 | | | | | | | | | |
| 94 | Alpha 2 | I | **0,134** | **0,060** | T | **0,866** | **0,940** | | | | | | | | | |
| 95 | Alpha 2 | F | 0,012 | 0,000 | I | 0,156 | 0,157 | L | 0,833 | 0,843 | | | | | | |
| 97 | Alpha 2 | R | 0,734 | 0,815 | W | 0,266 | 0,185 | | | | | | | | | |
| 99 | Alpha 2 | C | 0,024 | 0,014 | F | 0,219 | 0,306 | S | 0,084 | 0,046 | Y | 0,673 | 0,634 | | | |
| 103 | Alpha 2 | L | 0,903 | 0,944 | V | 0,097 | 0,056 | | | | | | | | | |
| 113 | Alpha 2 | C | 0,000 | 0,009 | H | 0,037 | 0,005 | Y | 0,963 | 0,986 | | | | | | |
| **114*** | Alpha 2 | D | **0,687** | **0,495** | N | **0,313** | **0,505** | | | | | | | | | |
| **116*** | Alpha 2 | **F** | **0,328** | **0,509** | L | 0,033 | 0,000 | S | 0,526 | 0,454 | **Y** | **0,112** | **0,037** | | | |
| 138 | Alpha 2 | K | 0,094 | 0,069 | T | 0,906 | 0,931 | | | | | | | | | |
| **143** | Alpha 2 | S | **0,030** | **0,130** | T | **0,970** | **0,870** | | | | | | | | | |
| 147 | Alpha 2 | L | 0,249 | 0,301 | W | 0,751 | 0,699 | | | | | | | | | |
| 152 | Alpha 2 | A | 0,281 | 0,315 | E | 0,717 | 0,685 | T | | | | | | | | |
| 156 | Alpha 2 | D | 0,012 | 0,000 | L | 0,373 | 0,384 | **Q** | **0,055** | **0,144** | R | 0,337 | 0,366 | **W** | **0,222** | **0,106** |
| **163** | Alpha 2 | E | **0,089** | **0,208** | L | 0,097 | 0,056 | T | 0,815 | 0,736 | | | | | | |
| **170** | Alpha 2 | G | **0,030** | **0,130** | R | **0,970** | **0,870** | | | | | | | | | |
| 173 | Alpha 2 | E | 0,903 | 0,944 | K | 0,097 | 0,056 | | | | | | | | | |
| 175 | Alpha 2 | G | 0,998 | 1,000 | R | 0,002 | 0,000 | | | | | | | | | |
| 177 | Alpha 2 | E | 0,891 | 0,931 | K | 0,109 | 0,069 | | | | | | | | | |
| 184 | Alpha 3 | H | 0,751 | 0,699 | P | 0,219 | 0,171 | R | 0,030 | 0,130 | | | | | | |
| 193 | Alpha 3 | L | 0,062 | 0,144 | P | 0,938 | 0,856 | | | | | | | | | |
| 194 | Alpha 3 | L | 0,219 | 0,171 | V | 0,781 | 0,829 | | | | | | | | | |
| 211 | Alpha 3 | A | 0,962 | 0,986 | T | 0,038 | 0,014 | | | | | | | | | |

| Amino acid residue | Obs | Amino Acid | Brazil Frequency 2n=836 | Benin Frequency 2n=216 | Amino Acid | Brazil Frequency 2n=836 | Benin Frequency 2n=216 | Amino Acid | Brazil Frequency 2n=836 | Benin Frequency 2n=216 | Amino Acid | Brazil Frequency 2n=836 | Benin Frequency 2n=216 | Amino Acid | Brazil Frequency 2n=836 | Benin Frequency 2n=216 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 219 | Alpha 3 | R | 0,660 | 0,625 | W | 0,340 | 0,375 | | | | | | | | | |
| 229 | Alpha 3 | E | 0,996 | 1,000 | Q | 0,004 | 0,000 | | | | | | | | | |
| 248 | Alpha 3 | M | 0,024 | 0,014 | V | 0,976 | 0,986 | | | | | | | | | |
| 253 | Alpha 3 | E | 0,751 | 0,699 | Q | 0,249 | 0,301 | | | | | | | | | |
| 261 | Alpha 3 | M | 0,219 | 0,171 | V | 0,781 | 0,829 | | | | | | | | | |
| 267 | Alpha 3 | P | 0,751 | 0,699 | Q | 0,249 | 0,301 | | | | | | | | | |
| **270** | Alpha 3 | C | **0,030** | **0,130** | L | **0,970** | **0,870** | | | | | | | | | |
| 273 | Alpha 3 | R | 0,781 | 0,829 | S | 0,219 | 0,171 | | | | | | | | | |
| **275*** | Transmembran | E | **0,687** | **0,495** | G | 0,097 | 0,069 | K | **0,217** | **0,435** | | | | | | |
| **284** | Transmembran | I | **0,970** | **0,870** | N | **0,030** | **0,130** | | | | | | | | | |
| **285** | Transmembran | L | **0,030** | **0,130** | M | 0,219 | 0,171 | V | 0,751 | 0,699 | | | | | | |
| **289** | Transmembran | A | **0,970** | **0,870** | S | **0,030** | **0,130** | | | | | | | | | |
| **291** | Transmembran | L | **0,970** | **0,870** | P | **0,030** | **0,130** | | | | | | | | | |
| 295 | Transmembran | A | 0,773 | 0,829 | V | 0,227 | 0,171 | | | | | | | | | |
| **301** | Transmembran | - | **0,970** | **0,870** | A | **0,030** | **0,130** | | | | | | | | | |
| **302** | Transmembran | - | **0,970** | **0,870** | V | **0,030** | **0,130** | | | | | | | | | |
| **303** | Transmembran | - | **0,970** | **0,870** | L | **0,030** | **0,130** | | | | | | | | | |
| **304** | Transmembran | - | **0,970** | **0,870** | A | **0,030** | **0,130** | | | | | | | | | |
| **305** | Transmembran | - | **0,970** | **0,870** | V | **0,030** | **0,130** | | | | | | | | | |
| **306** | Transmembran | - | **0,970** | **0,870** | L | **0,030** | **0,130** | | | | | | | | | |
| **309 = 303 at the IMGT** | Transmembran | M | **0,171** | **0,273** | V | 0,829 | 0,727 | | | | | | | | | |
| **310*= 304 at the IMGT** | Transmembran | M | **0,291** | **0,102** | V | **0,709** | **0,898** | | | | | | | | | |
| 311 = 305 at the IMGT | Transmembran | A | 0,781 | 0,829 | T | 0,219 | 0,171 | | | | | | | | | |
| 312 = 306 at the IMGT | Transmembran | A | 0,249 | 0,301 | V | 0,751 | 0,699 | | | | | | | | | |
| 313 = 307 at the IMGT | Transmembran | K | 0,000 | 0,023 | M | 0,219 | 0,148 | V | 0,781 | 0,829 | | | | | | |
| **314 = 308 at the IMGT** | Transmembran | I | **0,030** | **0,130** | M | **0,970** | **0,870** | | | | | | | | | |
| **315 = 309 at the IMGT** | Transmembran | C | **0,970** | **0,870** | H | **0,030** | **0,130** | | | | | | | | | |
| 330 =324 at the IMGT | Cytoplasmatic | A | 0,983 | 0,944 | V | 0,017 | 0,056 | | | | | | | | | |
| 332 = 326 at the IMGT | Cytoplasmatic | C | 0,219 | 0,171 | S | 0,781 | 0,829 | | | | | | | | | |
| 345 = 339 at the IMGT | Cytoplasmatic | A | 0,780 | 0,829 | T | 0,220 | 0,171 | | | | | | | | | |
| 348 = 342 at the IMGT | Cytoplasmatic | A | 0,998 | 1,000 | P | 0,002 | 0,000 | | | | | | | | | |
| 349 | Null | - | 0,998 | 1,000 | X | 0,002 | 0,000 | | | | | | | | | |

Frequency differences greater or equal to 10% are marked in boldface

## Reference


APPS, R. et al. Influence of HLA-C expression level on HIV control. **Science,** v. 340, n. 6128, p. 87-91, Apr 05 2013. ISSN 0036-8075.

AUGUSTO, D. G.; PETZL-ERLER, M. L. KIR and HLA under pressure: evidences of coevolution across worldwide populations. **Hum Genet,** v. 134, n. 9, p. 929-40, Sep 2015. ISSN 0340-6717.

BARRETT, J. C. et al. Haploview: analysis and visualization of LD and haplotype maps. **Bioinformatics (Oxford, England),** v. 21, n. 2, p. 263-5, 2005.

BLAIS, M.-E.; DONG, T.; ROWLAND-JONES, S. HLA-C as a mediator of natural killer and T-cell activation: spectator or key player? **Immunology,** v. 133, n. 1, p. 1-7, 2011.

BRANDT, D. Y. et al. Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. **G3 (Bethesda),** v. 5, n. 5, p. 931-41, Mar 17 2015. ISSN 2160-1836.

BROWNING, S. R.; BROWNING, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. **Am J Hum Genet,** v. 81, n. 5, p. 1084-97, Nov 2007. ISSN 0002-9297.

CASTELLI, E. C. et al. HLA-G variability and haplotypes detected by massively parallel sequencing procedures in the geographicaly distinct population samples of Brazil and Cyprus. **Mol Immunol,** v. 83, p. 115-126, Mar 2017. ISSN 0161-5890.

CASTELLI, E. C. et al. Hla-mapper: An application to optimize the mapping of HLA sequences produced by massively parallel sequencing procedures. **Human Immunology**, 2018/07/03/ 2018. ISSN 0198-8859.

CASTELLI, E. C. et al. Insights into HLA-G Genetics Provided by Worldwide Haplotype Diversity. **Front Immunol,** v. 5, p. 476, 2014. ISSN 1664-3224.

CASTELLI, E. C. et al. Transcriptional and posttranscriptional regulations of the HLA-G gene. **J Immunol Res,** v. 2014, p. 734068, 2014. ISSN 2314-7156.

CELIK, A. A. et al. The diversity of the HLA-E-restricted peptide repertoire explains the immunological impact of the Arg107Gly mismatch. **Immunogenetics,** v. 68, n. 1, p. 29-41, Jan 2016. ISSN 1432-1211.

CHAZARA, O.; XIONG, S.; MOFFETT, A. Maternal KIR and fetal HLA-C: a fine balance. **J Leukoc Biol,** v. 90, n. 4, p. 703-16, Oct 2011. ISSN 0741-5400.

CHEN, H. et al. Psoriasis patients are enriched for genetic variants that protect against HIV-1 disease. **PLoS genetics,** v. 8, n. 2, p. e1002514-e1002514, 2012.

EXCOFFIER, L.; LISCHER, H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. **Molecular ecology resources,** v. 10, n. 3, p. 564-567, 2010.

FELICIO, L. P. et al. Worldwide HLA-E nucleotide and haplotype variability reveals a conserved gene for coding and 3' untranslated regions. **Tissue antigens,** v. 83, n. 2, p. 82-93, 2014.

FELLAY, J. et al. A whole-genome association study of major determinants for host control of HIV-1. **Science,** v. 317, n. 5840, p. 944-7, Aug 17 2007. ISSN 0036-8075.

GAIDATZIS, D. et al. Inference of miRNA targets using evolutionary conservation and pathway analysis. **BMC Bioinformatics,** v. 8, p. 69, Mar 2007. ISSN 1471-2105.

GARRIGAN, D.; HEDRICK, P. W. Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. **Evolution,** v. 57, n. 8, p. 1707-22, Aug 2003. ISSN 0014-3820 (Print)0014-3820.

GONZÁLEZ-GALARZA, F. F. et al. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. **Nucleic Acids Res,** v. 43, n. Database issue, p. D784-8, Jan 2015. ISSN 1362-4962.

GRIMSON, A. et al. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. **Mol Cell,** v. 27, n. 1, p. 91-105, Jul 2007. ISSN 1097-2765.

HACKMON, R. et al. Definitive class I human leukocyte antigen expression in gestational placentation: HLA-F, HLA-E, HLA-C, and HLA-G in extravillous trophoblast invasion on placentation, pregnancy, and parturition. **American Journal of Reproductive Immunology,** v. 77, n. 6, p. e12643-n/a, 2017. ISSN 1600-0897.

HEDRICK, P. W.; THOMSON, G. Evidence for balancing selection at HLA. **Genetics,** v. 104, n. 3, p. 449-56, Jul 1983. ISSN 0016-6731 (Print)0016-6731.

HEDRICK, P. W.; WHITTAM, T. S.; PARHAM, P. Heterozygosity at individual amino acid sites: extremely high levels for HLA-A and -B genes. **Proc Natl Acad Sci U S A,** v. 88, n. 13, p. 5897-901, Jul 01 1991. ISSN 0027-8424 (Print)0027-8424.

HEGDE, R. S. Targeting and Beyond: new roles for old signal sequences. **Molecular Cell,** v. 10, n. 4, p. 697-698, 2002. ISSN 1097-2765.

HIBY, S. E. et al. Combinations of Maternal KIR and Fetal HLA-C Genes Influence the Risk of Preeclampsia and Reproductive Success. **J Exp Med,** v. 200, n. 8, p. 957-65, 2004 Oct 18 2004. ISSN 0022-1007 (Print)1540-9538 (Electronic).

HUNDHAUSEN, C. et al. Allele-specific cytokine responses at the HLA-C locus: implications for psoriasis. **J Invest Dermatol,** v. 132, n. 3 Pt 1, p. 635-41, Mar 2012. ISSN 0022-202x.

KAUR, G. et al. Structural and regulatory diversity shape HLA-C protein expression levels. **Nat Commun,** v. 8, p. 15924, Jun 2017. ISSN 2041-1723.

KULKARNI, S.; MARTIN, M. P.; CARRINGTON, M. The Ying and Yang of HLA and KIR in Human Disease. **Semin Immunol,** v. 20, n. 6, p. 343-52, 2008 Dec 2008. ISSN 1044-5323 (Print)1096-3618 (Electronic).

KULKARNI, S. et al. Genetic interplay between HLA-C and MIR148A in HIV control and Crohn disease. **Proceedings of the National Academy of Sciences of the United States of America,** v. 110, n. 51, p. 20705-20710, 2013.

KULKARNI, S. et al. Differential microRNA regulation of HLA-C expression and its association with HIV control. **Nature,** v. 472, n. 7344, p. 495-498, 2011.

KUMAR, S.; STECHER, G.; TAMURA, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. **Mol Biol Evol,** v. 33, n. 7, p. 1870-4, 07 2016. ISSN 1537-1719.

KUŚNIERCZYK, P. Killer Cell Immunoglobulin-Like Receptor Gene Associations with Autoimmune and Allergic Diseases, Recurrent Spontaneous Abortion, and Neoplasms. **Front Immunol,** v. 4, 2013 Jan 29 2013. ISSN 1664-3224 (Electronic).

LANCASTER, A. K. et al. PyPop update--a software pipeline for large-scale multilocus population genomics. **Tissue Antigens,** v. 69 Suppl 1, p. 192-7, Apr 2007. ISSN 0001-2815.

LEMBERG, M. K. et al. Intramembrane proteolysis of signal peptides: an essential step in the generation of HLA-E epitopes. **J Immunol,** v. 167, n. 11, p. 6441-6, 2001 Dec 1 2001. ISSN 0022-1767 (Print)0022-1767.

LIMA, T. H. A. et al. HLA-A promoter, coding, and 3'UTR sequences in a Brazilian cohort, and their evolutionary aspects. **HLA,** v. 93, n. 2-3, p. 65-79, Feb 2019. ISSN 2059-2310.

MAJOROS, W. H.; OHLER, U. Spatial preferences of microRNA targets in 3' untranslated regions. **BMC Genomics,** v. 8, p. 152, Jun 2007. ISSN 1471-2164.

MCKENNA, A. et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. **Genome Research,** v. 20, n. 9, p. 1297-1303, 2010.

MEYER, D. et al. A genomic perspective on HLA evolution. **Immunogenetics,** v. 70, n. 1, p. 5-27, Jul 2018. ISSN 1432-1211.

MEYER, D. et al. Signatures of demographic history and natural selection in the human major histocompatibility complex Loci. **Genetics,** v. 173, n. 4, p. 2121-42, Aug 2006. ISSN 0016-6731 (Print)0016-6731.

MEYER, D.; THOMSON, G. How selection shapes variation of the human major histocompatibility complex: a review. **Ann Hum Genet,** v. 65, n. Pt 1, p. 1-26, Jan 2001. ISSN 0003-4800 (Print)0003-4800.

MOFFETT-KING, A. Natural killer cells and pregnancy. **Nat Rev Immunol,** v. 2, n. 9, p. 656-63, 2002 Sep 2002. ISSN 1474-1733 (Print)1474-1733.

PARHAM, P. Killer cell immunoglobulin-like receptor diversity: balancing signals in the natural killer cell response. **Immunol Lett,** v. 92, n. 1-2, p. 11-3, Mar 29 2004. ISSN 0165-2478 (Print)0165-2478.

PARHAM, P. MHC class I molecules and kirs in human history, health and survival. **Nat Rev Immunol,** v. 5, n. 3, p. 201-214, 2005.

PARHAM, P.; MOFFETT, A. Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. **Nat Rev Immunol,** v. 13, n. 2, p. 133-44, Feb 2013. ISSN 1474-1733.

PARHAM, P. et al. Human-specific evolution of killer cell immunoglobulin-like receptor recognition of major histocompatibility complex class I molecules. **Philos Trans R Soc Lond B Biol Sci,** v. 367, n. 1590, p. 800-11, Mar 2012. ISSN 1471-2970.

PENMAN, B. S. et al. Reproduction, infection and killer-cell immunoglobulin-like receptor haplotype evolution. **Immunogenetics,** v. 68, n. 10, p. 755-764, Nov 2016. ISSN 0093-7711.

PETERSDORF, E. W. et al. HLA-C expression levels define permissible mismatches in hematopoietic cell transplantation. **Blood,** v. 124, n. 26, p. 3996-4003, Dec 18 2014. ISSN 0006-4971.

RAMALHO, J. et al. HLA-E regulatory and coding region variability and haplotypes in a Brazilian population sample. **Mol Immunol,** v. 91, p. 173-184, 11 2017. ISSN 1872-9142.

RAMSURAN, V. et al. Sequence and Phylogenetic Analysis of the Untranslated Promoter Regions for HLA Class I Genes. **J Immunol,** v. 198, n. 6, p. 2320-2329, Mar 15 2017. ISSN 0022-1767.

RICE, P.; LONGDEN, I.; BLEASBY, A. EMBOSS: the European Molecular Biology Open Software Suite. **Trends Genet,** v. 16, n. 6, p. 276-7, Jun 2000. ISSN 0168-9525.

ROBINSON, J. et al. The IPD and IMGT/HLA database: allele variant databases. **Nucleic Acids Res,** v. 43, n. Database issue, p. D423-31, Jan 2015. ISSN 1362-4962.

ROCK, K. L.; REITS, E.; NEEFJES, J. Present Yourself! By MHC Class I and MHC Class II Molecules. **Trends Immunol,** v. 37, n. 11, p. 724-737, Nov 2016. ISSN 1471-4981.

RODRIGUES, C. et al. Allele and haplotype frequencies of HLA-A, B, C, DRB1 and DQB1 genes in polytransfused patients in ethnically diverse populations from Brazil. **International Journal of Immunogenetics,** v. 42, n. 5, p. 322-328, 2015. ISSN 1744-313X.

SALTER, R. D. et al. A binding site for the T-cell co-receptor CD8 on the alpha 3 domain of HLA-A2. **Nature,** v. 345, n. 6270, p. 41-6, May 1990. ISSN 0028-0836.

SALTER, R. D. et al. Polymorphism in the α3 domain of HLA-A molecules affects binding to CD8. **Nature,** v. 338, p. 345, 03/23/online 1989.

SANCHEZ-MAZAS, A. An apportionment of human HLA diversity. **Tissue Antigens,** v. 69 Suppl 1, p. 198-202, Apr 2007. ISSN 0001-2815 (Print)0001-2815.

SCHMIDT, D.; POOL, J. **The Effect of Population History on the Distribution of the Tajima's D Statistic**. 2002.

SHARKEY, A. M. et al. Killer Ig-like receptor expression in uterine NK cells is biased toward recognition of HLA-C and alters with gestational age. **J Immunol,** v. 181, n. 1, p. 39-46, 2008 Jul 1 2008. ISSN 0022-1767 (Print)0022-1767.

SIBILIO, L. et al. A single bottleneck in HLA-C assembly. **J Biol Chem,** v. 283, n. 3, p. 1267-74, Jan 2008. ISSN 0021-9258.

SONON, P. et al. HLA-G, -E and -F regulatory and coding region variability and haplotypes in the Beninese Toffin population sample. **Mol Immunol,** v. 104, p. 108-127, 12 2018. ISSN 1872-9142.

STAJICH, J. E.; HAHN, M. W. Disentangling the effects of demography and selection in human history. **Mol Biol Evol,** v. 22, n. 1, p. 63-73, Jan 2005. ISSN 0737-4038.

SZOLEK, A. et al. OptiType: precision HLA typing from next-generation sequencing data. **Bioinformatics,** v. 30, n. 23, p. 3310-6, Dec 2014. ISSN 1367-4811.

THOMAS, R. et al. HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. **Nat Genet,** v. 41, n. 12, p. 1290-1294, 2009.

TIERCY, J.-M. HLA-C Incompatibilities in Allogeneic Unrelated Hematopoietic Stem Cell Transplantation. **Frontiers in immunology,** v. 5, p. 216-216, 2014.

VEIGA-CASTELLI, L. C. et al. Non-classical HLA-E gene variability in Brazilians: a nearly invariable locus surrounded by the most variable genes in the human genome. **Tissue antigens,** v. 79, n. 1, p. 15-24, 2012.

VILELLA, A. J. et al. VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. **Bioinformatics,** v. 21, n. 11, p. 2791-3, Jun 2005. ISSN 1367-4803.

VINCE, N. et al. HLA-C Level Is Regulated by a Polymorphic Oct1 Binding Site in the HLA-C Promoter Region. **Am J Hum Genet,** v. 99, n. 6, p. 1353-1358, Dec 01 2016. ISSN 0002-9297.

WESLEY, P. K. et al. The CD8 coreceptor interaction with the α3 domain of HLA class I is critical to the differentiation of human cytotoxic t-lymphocytes specific for HLA-A2 and HLA-Cw4. **Human Immunology,** v. 36, n. 3, p. 149-155, 1993/03/01/ 1993. ISSN 0198-8859.

YAWATA, M. et al. Roles for HLA and KIR polymorphisms in natural killer cell repertoire selection and modulation of effector function. **J Exp Med,** v. 203, n. 3, p. 633-45, Mar 20 2006. ISSN 0022-1007 (Print)0022-1007.

# CONSIDERAÇÕES FINAIS

A medotologia proposta nesse estudo permitiu a avaliação da variabilidade genética de todo gene *HLA-C* (genótipos e haplótipos). Nós observamos uma correlação direta entre os haplótipos promotores, codificadores e de 3'NT, de forma que cada grupo de alelos codificadores (que codificam a mesma proteína ou proteínas semelhantes) está relacionado com sequências regulatórias específicas. Os testes de desvio de neutralidade revelaram uma forte evidência de seleção balanceadora atuando em todo o gene *HLA-C,* que pode ser explicado pelo padrão de elevado LD detectado ao longo do gene *HLA-C*. Adicionalmente, foi detectada seleção positiva no éxon 1, que codifica o 'peptídeo líder' de HLA-C. Além disso, foram observadas regiões altamente conservadas relacionadas principalmente com a regulação da expressão do gene. Na região promotora, observou-se uma região altamente conservada próximo a posição -700, um segmento monomórfico de 115pb que é compartilhado entre diferentes primatas, e um *'core promoter'* conservado comparado a outros segmentos do gene *HLA-C*. Para o segmento da 3'NT, nós detectamos uma conservação no trecho inicial (região de alvos conservados para miRNA). É possível que estes achados estejam relacionados ao distinto perfil de expressão gênica e a função imunomodulatória de *HLA-C*, porém estudos funcionais ainda são necessários para o melhor entendimento sobre a relação entre o padrão de variabilidade de *HLA-C*, expressão e função do gene.

# Anexos

**Anexo 1**. Prêmio relacionado a esta tese

GENÉTICA 2017

BRAZILIAN-INTERNATIONAL CONGRESS OF GENETICS

SEPTEMBER 12-15, 2017
MONTE REAL RESORT - ÁGUAS DE LINDÓIA/SP - BRAZIL

## *Certificado*

A Sociedade Brasileira de Genética concede a *ANDREIA DA SILVA SOUZA* menção honrosa pela participação no Prêmio Painel Pós-Graduação com o trabalho na área de GENÉTICA HUMANA, intitulado "SIGNATURES OF NATURAL SELECTION AT THE HLA-C LOCUS EVALUATED IN A BRAZILIAN POPULATION SAMPLE" autoria de SOUZA, A.S.; PAZ, M.A.; VEIGA-CASTELLI, L.C.; DONADI, E.A.; MENDES-JUNIOR, C..T; CASTELLI, E.C., apresentado durante o *GENÉTICA 2017 - Brazilian-International Congress of Genetics* , em Águas de Lindóia, SP, período de 12 a 15 de Setembro de 2017.

**Patrocinador:** Sociedade Brasileira de Genética GH074

Márcia Maria A N P Margis
Presidente da SBG

SBG SOCIEDADE BRASILEIRA DE GENÉTICA

Nilce Maria Martinez Rossi
Primeira Secretária da SBG